ECMiner™
User's Guide

목차

제 1 장 제품 개요

ECMiner™는 데이터 마이닝 전문 기업인 **㈜이씨마이너**와 포항공과대학교 석/박사 연구진이 공동 개발한 **데이터 마이닝 소프트웨어**로 데이터에서 유용한 정보와 관계를 탐색하고 정보화, 지식화하여 의사결정 활동을 지원합니다. **ECMiner™**는 **우수한 성능**을 기반으로 **다양한 기능**을 제공하며, 데이터 입출력, 전처리, 분석, 모델링, 모델 평가, 차트 등 데이터 마이닝 작업을 위한 기능을 지원하는 **통합 분석 소프트웨어**로써 비전문가도 쉽게 활용할 수 있는 **편리성**으로 사용자를 만족시켜 드릴 것입니다.

ECMiner™는 데이터 분석을 위한 다양한 기능을 제공하며 다음과 같은 특장점을 가지고 있습니다.

- **대용량 데이터 처리**

  상용 데이터베이스, ECL(자체 개발 데이터 구조), TEXT, EXCEL 등의 다양한 데이터 형식을 지원하며 자체 개발된 ECL 데이터 구조를 통해 대용량 데이터의 처리 성능을 향상시켰습니다.

- **사용자 편의성**

  노드 제공을 통해 간단한 마우스 조작만으로도 데이터 마이닝 시 가장 많은 노력이 요구되는 데이터 탐색 및 전처리 과정을 쉽게 수행할 수 있습니다. 또한, 데이터 마이닝 작업을 위한 통합 UI 를 제공함으로써, 여러 프로그램을 사용할 필요가 없습니다.

- **확장 용이성**

  데이터 마이닝을 위한 각 요소는 모듈로 개발되어 확장이 용이하고, 모듈을 조합하여 다양한 데이터 마이닝 프로시저를 구성할 수 있습니다.

- **다양한 기능**

  예측, 분류, 군집, 연관성 분석을 위한 알고리즘, 데이터 처리를 위한 강력한 전처리 노드, 데이터 탐색기, 데이터 가시화를 위한 차트 등의 기능을 통하여 다양한 데이터 마이닝 과제를 수행할 수 있으며, 각종 노드의 조합을 통해 구현할 수 있는 기능은 무궁무진합니다.

- **다양한 분야에 활용**

  다양한 산업분야에서 응용 및 적용 가능한 범용적인 데이터 마이닝 소프트웨어이며 효과적인 의사결정에 도움이 될 것입니다.

1.1 시스템 환경

시스템 개발환경

CPU	인텔® 코어 2 듀오 2.4 GHz
메모리	16 Gbytes
운영체제	Windows 7 Professional
하드 디스크	

최소 운영환경

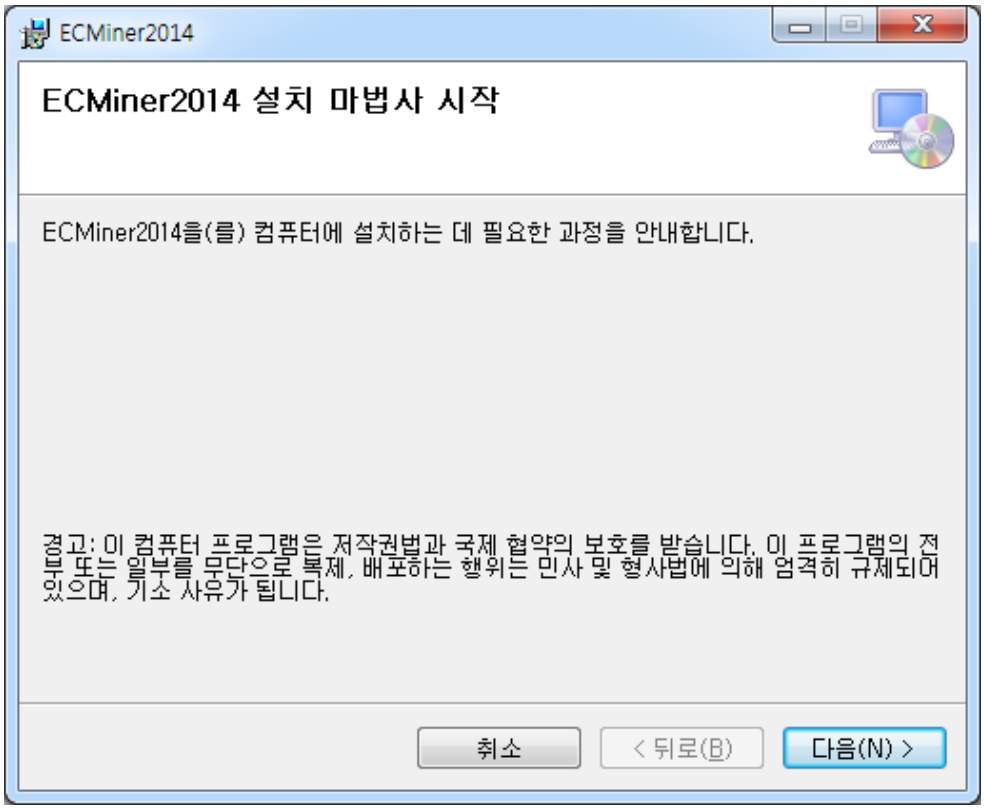
CPU	인텔 펜티엄® D 2.8 GHz 또는 AMD Athlon™ 64 X2 4400+
메모리	4 Gbytes
운영체제	Windows Server 2008 R2, Windows Server 2012, Windows 7, Windows 8
하드 디스크	분석하고자 하는 데이터 양의 2 배 이상

NOTE *분석하고자 하는 데이터의 양에 따라 충분한 하드디스크 용량이 있어야 합니다. 예를 들어 분석하고자 하는 데이터의 크기가 500 MBytes 정도라면 최소 1 GBytes (500 MBytes x 2) 이상의 여유 공간이 하드디스크에 있어야 합니다.*

1.2 ECMiner™ 설치 및 제거

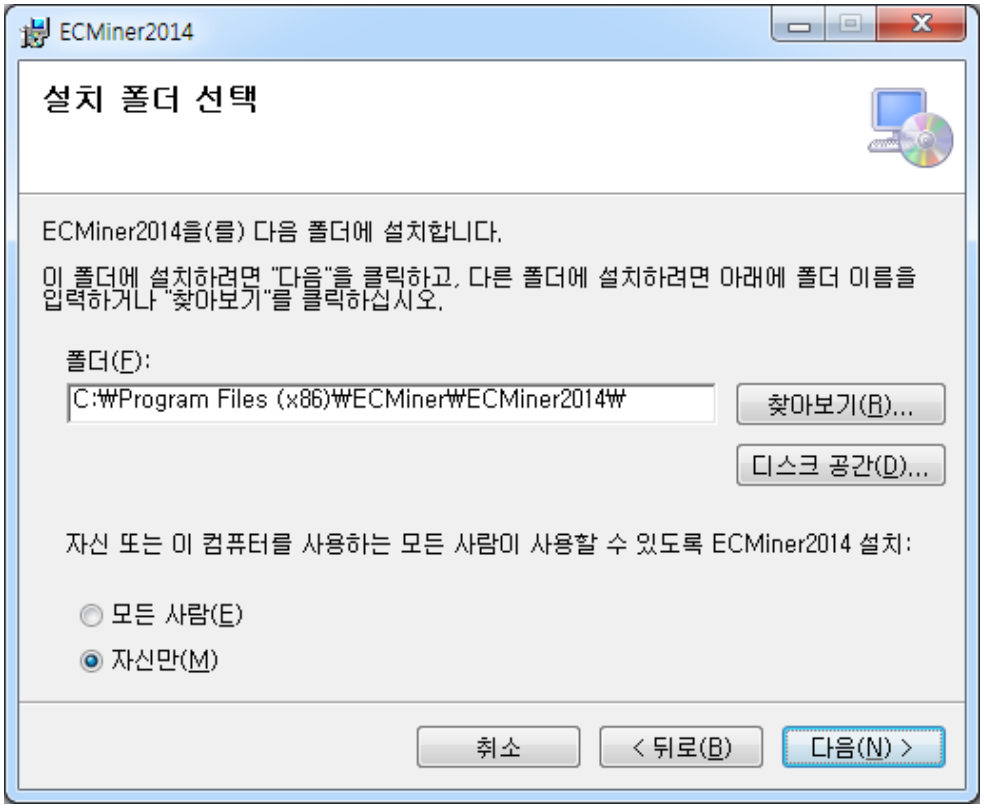
설치

ECMiner™을 설치하려면 다음과 같은 과정을 거칩니다.
제공되는 CD 를 CD-ROM 에 넣으면 자동 실행됩니다. 또는 ㈜이씨마이너 홈페이지인 www.ecminer.com 에서 데모 설치파일을 다운 받아 실행시킵니다.
실행화면과 동시에 다음과 같은 화면이 뜹니다.

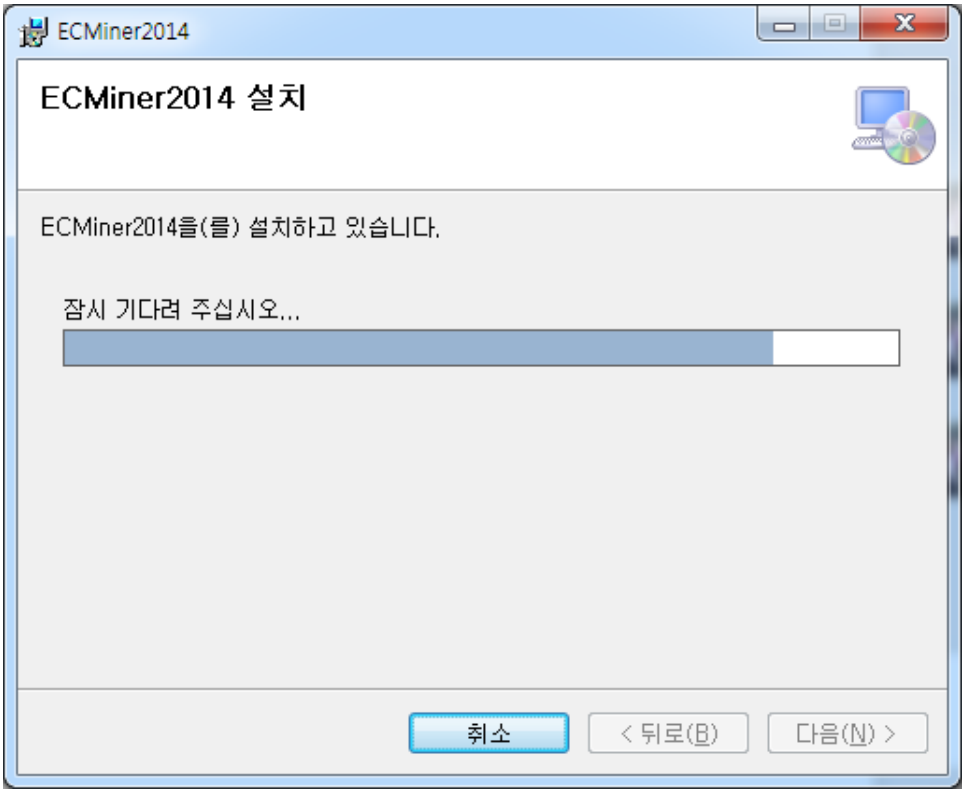


설치 마법사가 실행되면 [다음(N)]을 클릭합니다.

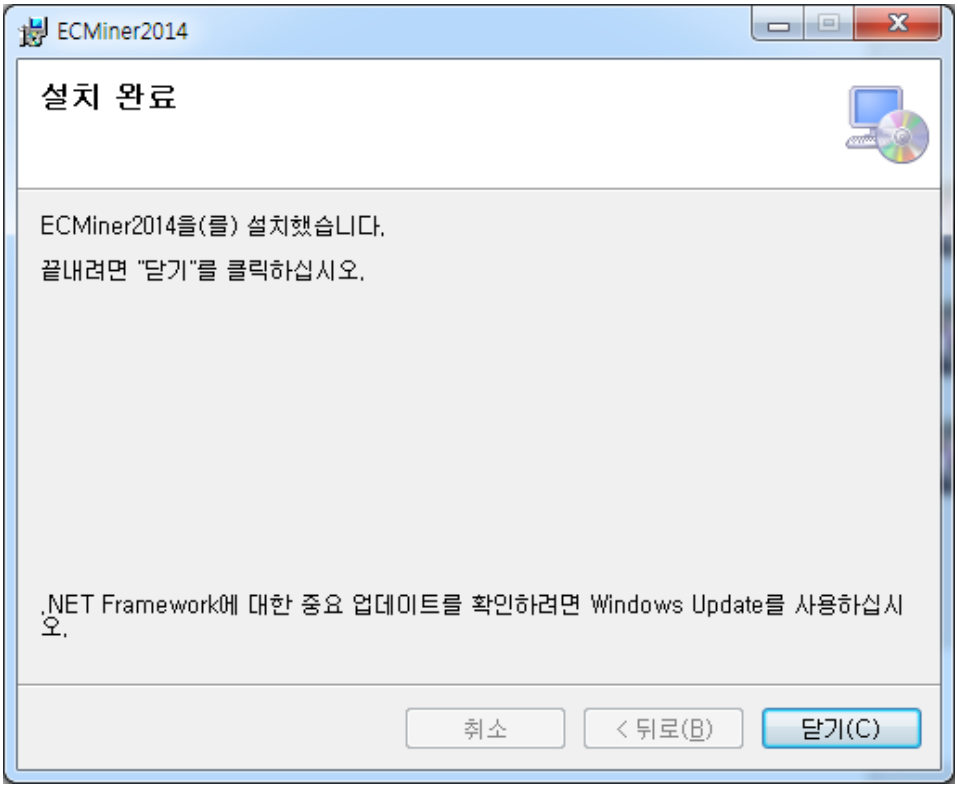
설치할 폴더를 지정하고 [다음(N)]을 클릭합니다. 이때, 사용자 컴퓨터의 남은 디스크 용량을 확인할 수 있으며, 사용자를 정의할 수 있습니다.



최종적으로 설치 진행여부를 묻습니다. 설치를 진행하려면 [다음(N)]을 클릭합니다.



설치가 진행되며, 설치를 중도에 취소하고 싶을 경우 [취소] 버튼을 누릅니다. 설치를 계속 진행하고자 한다면 설치가 종료될 때까지 기다립니다.



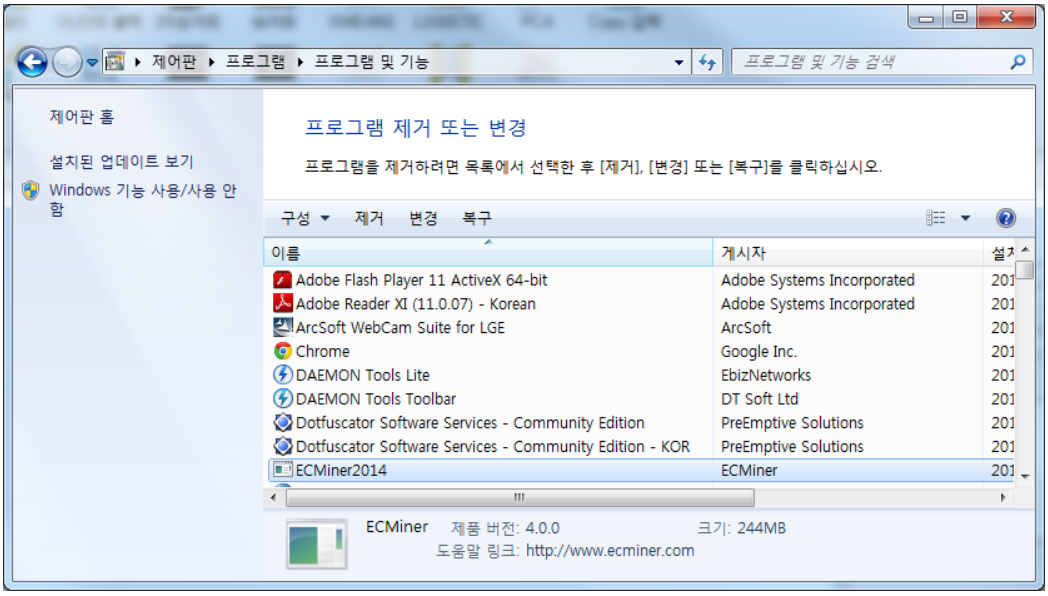
설치가 완료되었습니다. 바탕화면에 생성된 

아이콘을 클릭하면 ECMiner™가 실행됩니다.

제거

ECMiner™를 제거하려면 다음과 같은 과정을 거칩니다.

- 제어판에서 "프로그램 추가/제거"를 선택하여 실행합니다. "프로그램 추가/제거" 대화상자에서 ECMiner™를 선택합니다.



ECMiner™ 제거 확인 대화상자에서 **'예'** 버튼을 클릭합니다. 자동으로 ECMiner™ 제거 과정이 수행되고, 시스템에서 ECMiner™가 제거됩니다.

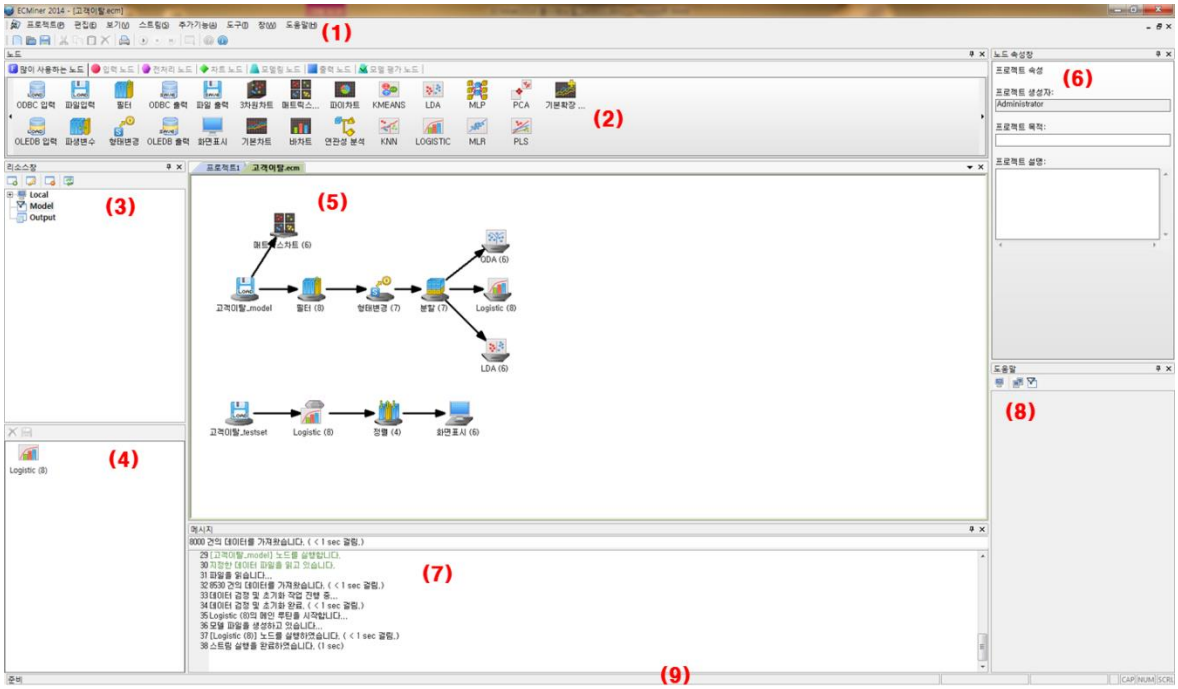
1.3 사용자 인터페이스

1.3.1 화면구성

ECMiner™는 시각적이고 도식적인 인터페이스를 제공하여 사용자가 쉽고 편리하게 사용할 수 있습니다. 사용자는 이 인터페이스를 바탕으로 용도에 맞는 다양한 스트림을 구성할 수 있으며 그에 따른 결과를 확인할 수 있습니다.

주 화면(Main Window)

ECMiner™는 쉽고 효율적인 데이터 마이닝을 수행할 수 있도록 다음과 같은 주 화면을 제공합니다.



번호	창 이름	설명
(1)	메뉴 및 툴바	프로젝트 파일 생성/편집/저장, 스트림 편집 등 ECMiner™ 사용을 위한 기능을 선택할 수 있는 부분입니다. 메뉴의 형태로 정리되어 있으며 자주 사용하는 기능은 툴바에 추가로 나타내어 사용자의 편의를 돕고 있습니다.
(2)	노드창	데이터 마이닝을 수행하기 위한 요소 기능인 노드를 유사한 정도에 따라 구분 지어 놓은 컨트롤 창입니다. ECMiner™ 프로젝트 구성에 있어 기본이 되며 더블-클릭이나 드래그 & 드랍을 이용하여 사용하고자 하는 노드를 추가할 수 있습니다.
(3)	리소스창(메뉴)	리소스창은 리소스 관리의 편의를 위하여 새로 추가된 기능입니다. ECMiner™에서 리소스는 데이터, 수행 결과, 생성된 모델, 프로젝트 파일 등을 의미합니다. 리소스창(메뉴)에는 ECMiner™에서 사용하는 리소스를 트리 형태의 메뉴로 표현되어 있으며 Local, Model, Output 이 해당 메뉴입니다. • Local 은 사용자가 지정한 작업폴더 및 하위 폴더 목록이 나타납니다. • Model 은 프로젝트(스트림) 실행 시 모델링 노드가 있다면 모델링 노드에서 생성된 결과인 모델이 관리되는 곳입니다. • Output 은 프로젝트(스트림) 실행 후 생성된 모델 외의 모든

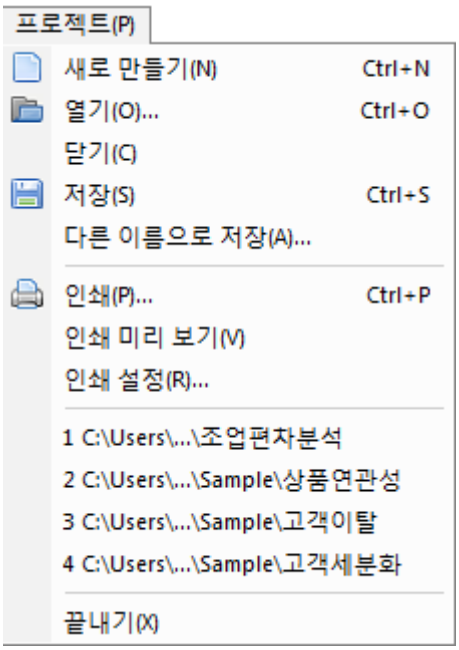
		결과물(화면표시, 차트 등)이 관리되는 곳입니다.
(4)	**리소스창(내용)**	리소스창(메뉴)에서 선택한 메뉴(Local, Model, Output)의 내용물이 나타나는 창입니다. • Local 을 선택하면 선택한 폴더 중 ECMiner™에서 사용하도록 지정된 확장자를 갖는 파일이 나타납니다. • Model 을 선택하면 프로젝트(스트림) 실행 후 생성된 모델 등이 목록화되어 나타납니다. 이 모델들은 임시로 생성된 것이기 때문에 ECMiner™ 종료 시 살아지게 됩니다. 만약 유용한 모델이라면 임시가 아닌 전역 모델화 하여 필요할 때마다 다시 사용할 수 있습니다. • Output 을 선택하면 프로젝트(스트림) 실행 후 결과물이 프로젝트별로 목록화되어 나타나며, 결과물 관리 기능을 제공합니다.
(5)	**작업창**	작업창은 프로젝트를 편집하는 곳입니다. 프로젝트는 스트림으로 이루어지며 스트림은 노드와 이들을 연결 지은 것입니다. 노드창을 이용하여 노드를 작업창에 생성하고, 이들을 연결 짓고, 필요 없는 노드나 연결을 삭제하는 등 프로젝트를 구성하는데 필요한 작업을 수행할 수 있습니다..
(6)	**노드 속성창**	노드는 데이터 마이닝을 위한 요소 기능을 모듈화 한 것이며 각 노드는 노드의 기능에 따라 필요한 옵션(속성)이 있습니다. 예를 들어, 텍스트 파일을 읽어 데이터화 하려고 한다면 "파일입력 노드"를 생성하고 해당 파일을 읽어 데이터화 하라고 "파일입력 노드"에 알려 주어야 합니다. 바로, 파일 경로를 알려 주어야 하는데 이 것이 노드의 속성이며 이를 편집하는 곳이 노드 속성창입니다. 추가로 프로젝트에 대한 간단한 정보도 여기서 편집할 수 있습니다.
(7)	**메시지창**	프로젝트(스트림) 실행 시 발생하는 메시지(로그) 등을 나타내는 곳입니다. 실행 시 오류 등이 발생하였다면 해당 내용을 메시지창에서 확인할 수 있습니다.
(8)	**도움말**	ECMiner™의 도움말이 나타나는 부분입니다. 각 노드의 속성에 대한 도움말과 노드에 대한 도움말 등을 표시하여 ECMiner™ 사용의 편의를 돕습니다.

(9)	**상태 표시줄**	ECMiner™의 상태를 나타냅니다. 실행 중인 프로젝트가 있는지, 키보드에 NUM LOCK 등이 켜졌는지 등이 표시되며, 메뉴 등을 선택할 때 간단한 Tooltip 이 나타납니다.

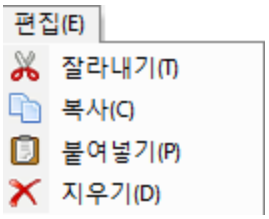
메뉴

메뉴는 화면 최상단에 위치합니다. 프로젝트(P), 편집(E), 보기(V), 스트림(S), 확장기능(A), 도구(T), 창(W), 도움말(H) 등의 메뉴를 제공합니다.

프로젝트(P)

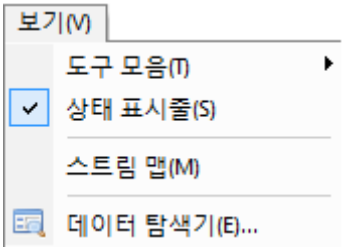


편집(E)

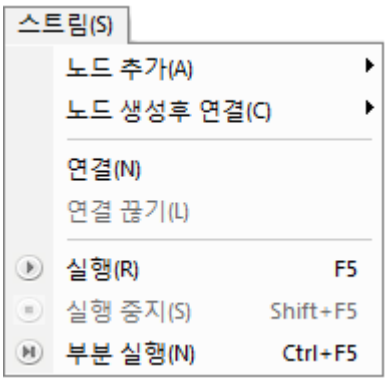


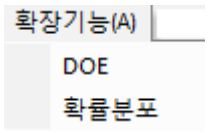
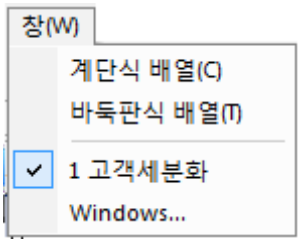
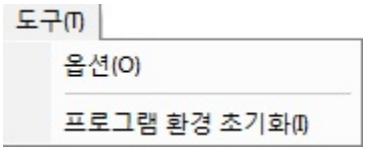
보기(V)

툴바나 각종 창, 그리고 상태바의 표시여부를 결정합니다.



스트림(S)



확장기능(A)	창(W)
실험계획법(DOE)과 확률분포 분석을 수행할 수 있습니다. 	
도움말(H)	**도구(T)**
	

도구의 옵션을 클릭하면 프로그램 관련 각종 옵션을 지정할 수 있습니다.

	**프로그램 설정**	날짜/시간 형식, 데이터 읽기, 리소스창 파일 확장자 지정, 프로그램 룩 등의 프로그램 기본 설정을 합니다.
	**기본 데이터베이스 설정**	ODBC 혹은 OLE DB 의 기본 드라이버와 로그인 정보를 설정합니다.
	**많이 사용하는 노드 편집**	많이 사용하는 노드를 추가하거나 삭제할 수 있습니다.

툴바(Tool Bar)

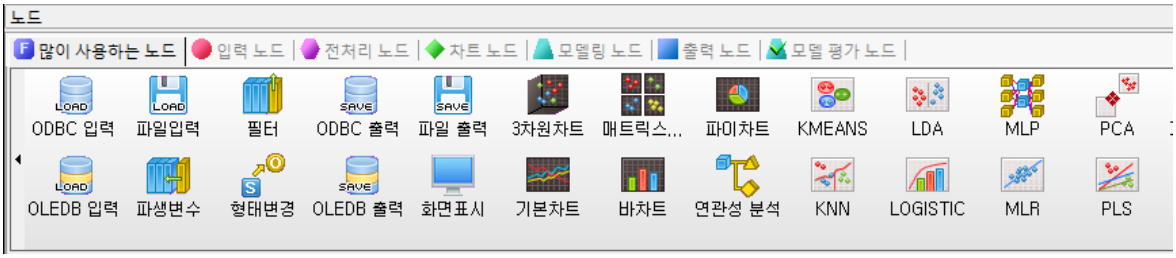
메뉴와 노드창 사이에 유용한 기능을 제공하는 아이콘들의 모음인 툴바가 위치합니다. 아이콘들과 그들의 기능은 다음과 같습니다.

 새로 만들기(Ctrl + N)
 저장하기(Ctrl + S)
 복사
 삭제
 실행(F5)
 부분 실행(Ctrl + F5)
 노드 도움말

 열기(Ctrl + O)
 잘라내기
 붙여넣기
 인쇄(Ctrl + P)
 실행 정지(Shift + F5)
 데이터 탐색기
 정보

1.3.2 노드창

노드창은 데이터 마이닝 작업 절차를 고려하여 특정 기능을 수행하는 노드들을 기능적 유사성에 따라 분류하였습니다. 각 노드는 그의 역할에 해당하는 기능을 완벽하게 수행한 뒤 이후에 연결되어 있는 노드에 수행한 결과를 넘기도록 구성되어 있고, 노드의 조합에 따라 다양한 기능을 구현할 수 있습니다.



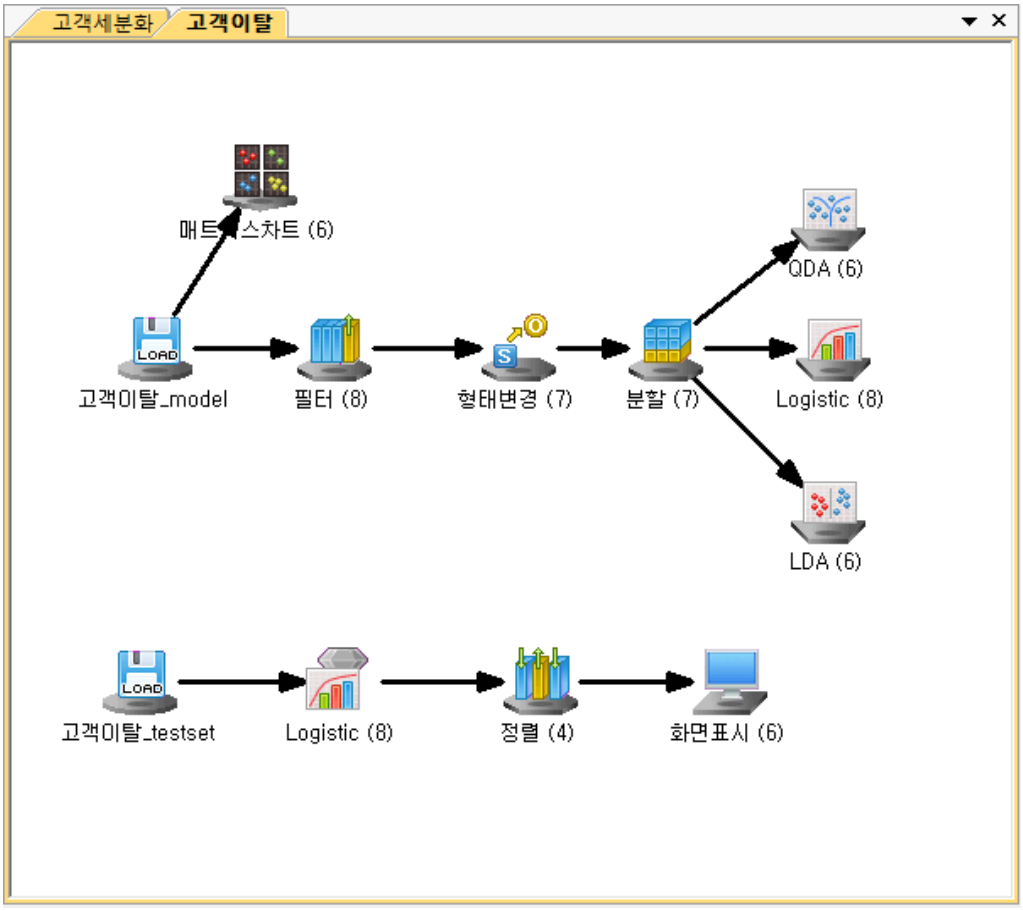
노드창은 총 7 가지로 분류되어 있습니다.

노드 분류	설명
많이 사용하는 노드	자주 사용하는 노드를 모아 놓은 곳입니다. 주 화면의 도구 옵션에서 등록, 삭제가 가능합니다.
입력 노드	데이터 마이닝의 출발점으로써 데이터를 가져오기 위한 노드입니다.
전처리 노드	입력 받은 데이터를 검정하고 적절한 형태로 변형하기 위한 노드입니다.
차트 노드	데이터를 시각적으로 분석하기 위한 차트 기능을 제공하는 노드입니다.
모델링 노드	데이터 마이닝 프로세스의 엔진 역할을 하는 알고리즘을 수행하는 노드입니다.
출력 노드	데이터를 DB 혹은 파일에 저장하거나 화면으로 출력해 볼 수 있는 노드입니다.
모델 평가 노드	생성된 모델의 예측 정확도를 비교, 평가하는 노드입니다.

1.3.3 프로젝트창

프로젝트창에서는 **데이터 마이닝 스트림**을 구성할 수 있습니다. 노드창에 있는 노드를 더블 클릭 혹은 드래그하여 새로 생성한 후 원하는 스트림에 생성된 노드를 연결하여 구성할 수

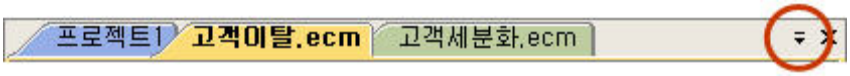
있습니다. 스트림은 여러 개 구성될 수 있으며 **데이터 입력 노드**를 시작 지점으로 하여 각 스트림이 순차적으로 실행됩니다. 스트림의 실행 순서는 특별히 지정하지 않은 경우 만들어진 순서대로 실행 됩니다. 프로젝트도 [새로 만들기]를 이용하여 여러 개 구성될 수 있습니다.



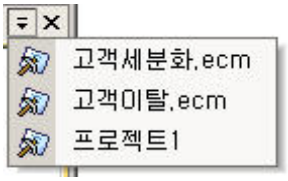
프로젝트창에서 각 노드는 아이콘으로 표시됩니다. 특히 노드는 기능적 유사성에 따라 그룹으로 구분되어 있고 그 분류에 따라 아이콘의 바닥 모양이 바뀌도록 설계되어 있어, 쉽게 스트림의 실행 구조를 이해할 수 있습니다.

프로젝트 선택

- 작업할 프로젝트를 활성화 하려면 위 그림에서 해당 프로젝트의 탭을 누릅니다. 프로젝트가 많아 탭 내에 나타나 있지 않은 경우 다음그림의 버튼을 누릅니다.



해당 버튼을 누르면 현재 열려 있는 프로젝트 목록이 다음 그림과 같이 나타납니다.



이 목록 중 작업하고자 하는 프로젝트를 선택하면 해당 프로젝트가 나타납니다.

프로젝트 닫기

- 프로젝트를 닫으려면 메인 메뉴 중 "프로젝트 > 닫기" 메뉴를 선택하거나 아래 그림에서 표시한 X 버튼을 누릅니다.

프로젝트가 저장되지 않았다면 저장 여부를 묻는 메시지 박스가 나타납니다.

프로젝트창 내의 노드

- 프로젝트창 내에서 각 노드는 아이콘으로 표시됩니다. 노드의 아이콘은 노드에 따라 다르며 해당 아이콘은 노드 설명을 참조하기 바랍니다.
  프로젝트창에서 노드는 노드의 범주별로 다음과 같은 밑받침이 함께 그려집니다. 이 모양을 보고 이 노드는 어떤 작업을 수행하는 것이지 알 수 있습니다.

아이콘	종류	바닥 모양
	입력노드	원형
	전처리노드	육각형
	차트노드	마름모
	모델링노드	사다리꼴
	출력노드, 모델평가노드	사각형
	모델노드	다이아몬드

컨텍스트 메뉴

- 프로젝트창에서 오른쪽 마우스 버튼을 누르면 프로젝트창에서 수행할 수 있는 기능을 모은 컨텍스트 메뉴가 나타납니다.

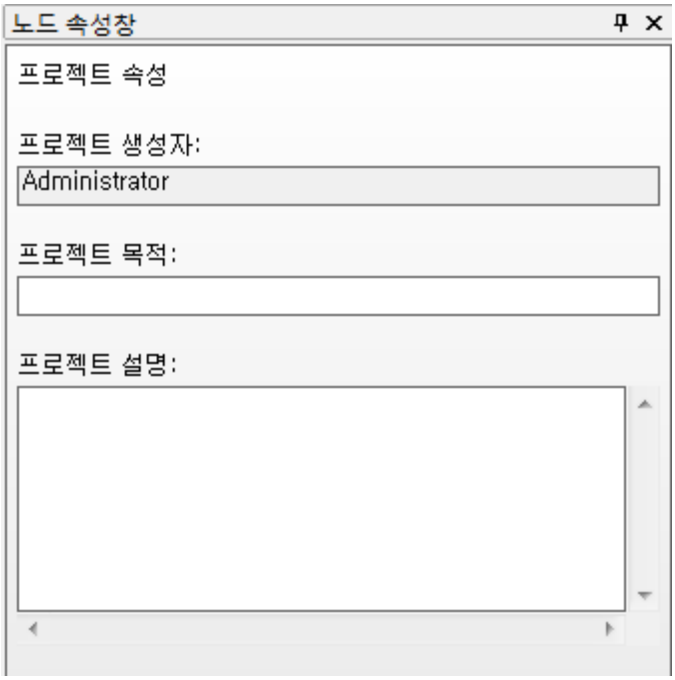
- 이번 최신 버전에서 새로 추가된 기능인 '포스트잇 삽입' 은 사용자가 노드 구성 및 프로젝트를 구성했을 경우, 메모가 가능한 용도로써 추가된 기능입니다.
- 이 컨텍스트 메뉴는 메인 메뉴에 설명되어 있는 메뉴와 같은 메뉴이며 자세한 내용은 메인 메뉴를 참조합니다.

1.3.4 속성창

속성창은 프로젝트의 속성 또는 선택된 노드의 속성을 편집할 수 있는 UI 입니다.

프로젝트 속성창

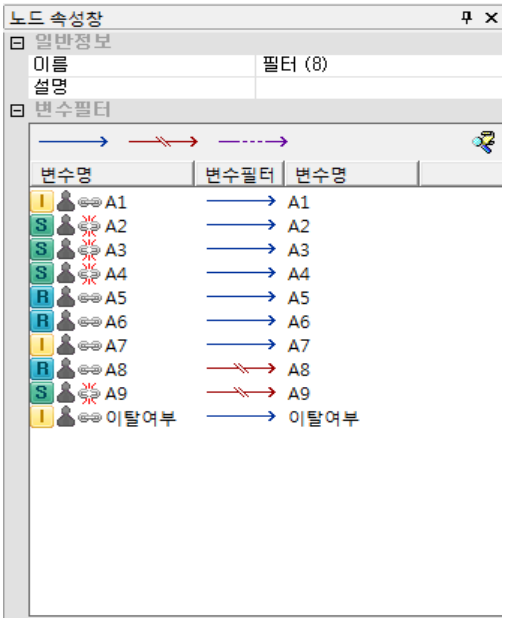
- (노드 연결이 없는)프로젝트창의 빈 영역을 클릭하면, 속성창에 다음과 같은 프로젝트 속성창이 나타납니다.



프로젝트 속성창에서 프로젝트의 목적, 간단한 설명 등을 기록할 수 있습니다.
프로젝트 생성자는 Windows 에 로그인한 사용자명이 나타나며 변경할 수 없습니다.

노드 속성창

- 모든 노드는 그에 해당하는 작업을 수행하기 위하여 입력해야 할 선택사항들이 있습니다. 이 선택사항들을 노드의 속성이라고 하며, 이 속성을 편집하기 위한 것이 노드 속성창입니다. 프로젝트창에서 노드를 선택하면 해당 노드의 속성이 다음 그림과 같이 나타납니다.



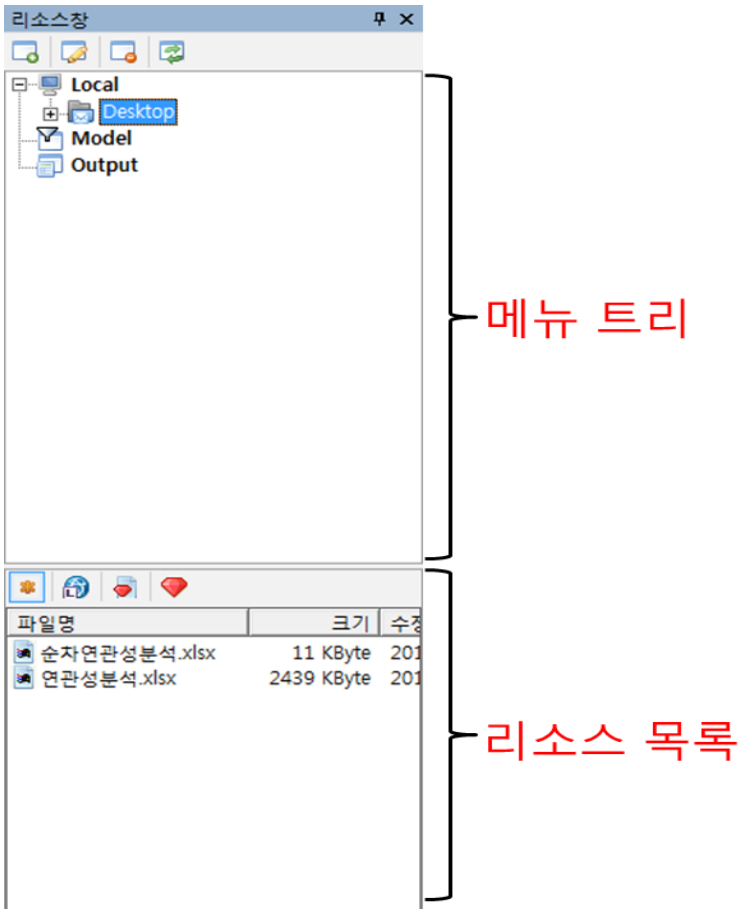
위 그림은 필터 노드의 속성창이며 이 속성창은 노드마다 다릅니다. 노드 속성에 대한 자세한 내용은 노드 도움말을 참조합니다.

노드 속성 편집

- 노드 속성창은 속성 컨트롤들로 구성되며 속성 컨트롤은 문자/숫자를 입력하거나 마우스 버튼을 클릭하는 식으로 편집할 수 있도록 제작되어 있습니다. 마우스로 편집할 속성 컨트롤을 선택/클릭하고 값을 입력하는 방식으로 편집을 진행하시면 됩니다.
  그룹화 노드 등 특별한 경우, 필요한 속성을 편집하기 위한 특화된 속성 컨트롤이 있으며 이들의 사용법은 각 노드를 참조하기 바랍니다.

1.3.5 리소스창

ECMiner™에서 리소스는 데이터, 수행 결과, 생성된 모델, 프로젝트 파일 등 ECMiner™가 사용 또는 생성하는 모든 파일, 결과를 의미합니다. 이들의 관리 및 스트림 구성의 편의를 위하여 제공된 것이 바로 리소스창입니다.



리소스창은 리소스의 종류를 선택할 수 있는 메뉴 트리 부분과 선택된 리소스의 내용을 나타내는 리소스 목록 부분으로 나누어져 구성되어 있습니다.

메뉴 트리

- 메뉴 트리는 세가지 큰 메뉴가 존재합니다. Local, Model, Output 이 그 세 가지이며 각 메뉴의 의미는 다음과 같습니다.
    - Local

      사용자가 지정한 작업폴더 및 작업폴더의 하위 폴더 목록이 나타나는 곳입니다. Local 메뉴 내의 작업폴더 및 하위 폴더를 선택하면 그 폴더 내의 파일 목록이 리소스 목록에 나타납니다. 단, ECMiner™에서 사용하도록 지정된 확장자를 갖는 파일만 나타납니다.

    - Model

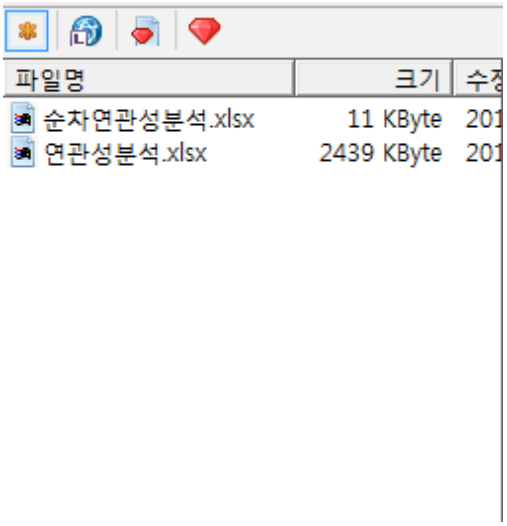
      스트림 실행 시 실행 대상 노드 중 모델링 노드가 있다면 그 모델링 노드의 결과인 모델 파일이 생성됩니다. 이 메뉴를 선택하면 생성된 모델 파일 목록이 리소스 목록에 나타나며 스트림 구성 시 재사용할 수 있도록 합니다.

    - Output

      스트림 실행 시 실행 대상 노드 중 출력과 관련된 노드(출력 노드, 모델 평가 노드, 차트 노드)가 있다면 각 노드에 해당하는 결과물이 생성되며 이들을 관리하기 위한 메뉴가 Output 메뉴입니다. 이 메뉴를 선택하면 리소스 목록에 프로젝트 별로 생성된 출력물이 나타납니다.

리소스 목록

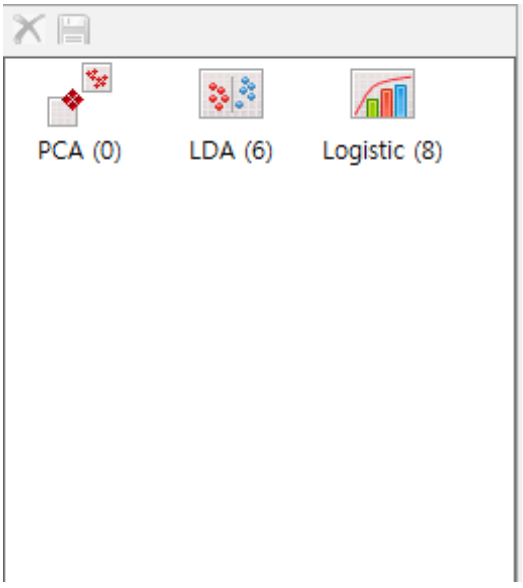
- 메뉴 트리에서 선택된 메뉴의 내용이 나타나는 부분입니다. 각 메뉴마다 그 메뉴에 맞는 형태의 UI 가 나타나도록 디자인되어 있습니다.
    - Local 메뉴 선택 시
      Local 메뉴 선택 시 다음과 같은 화면이 리소스 목록에 나타납니다.



선택된 폴더의 내용이 위 그림과 같이 나타나며 위쪽의 툴바 버튼을 이용하여 목록화 될 파일을 필터링 할 수 있습니다.

- Model 메뉴 선택 시

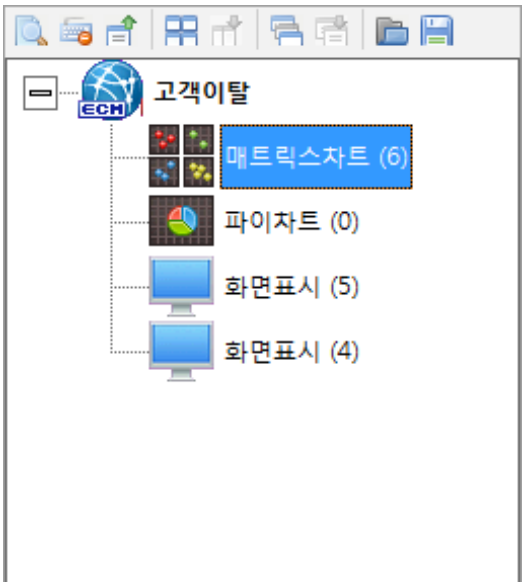
  Model 메뉴 선택 시 다음과 같은 화면이 리소스 목록에 나타납니다.



  스트림 실행 후 생성된 모델이 목록화 되어 위 그림과 같이 나타납니다. 필요 없는 모델이라 판단되면 삭제할 수 있는 기능을 제공하며, 임시 모델을 전역 모델로 변환하는 기능을 제공합니다.

- Output 메뉴 선택 시

  Output 메뉴 선택 시 다음과 같은 화면이 리소스 목록에 나타납니다.

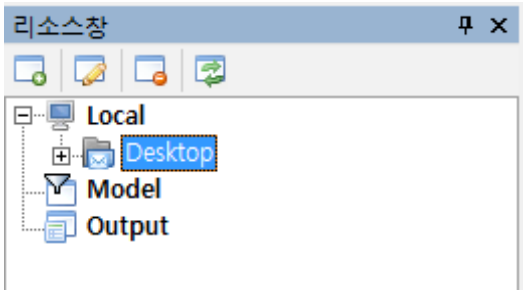


스트림 실행 후 생성된 결과물이 스트림이 속한 프로젝트 별로 트리화 되어 위 그림처럼 나타납니다. 보고자 하는 결과물을 선택하여 나타낼 수 있으며, 현재 보여지고 있는 결과물을 서로 비교하여 보기 위한 결과물 정리 기능을 제공합니다. 또한, 결과물을 저장하거나 기 저장된 결과물을 다시 볼 수 있는 기능도 제공하고 있습니다.

1.3.6 리소스창-Local

작업폴더

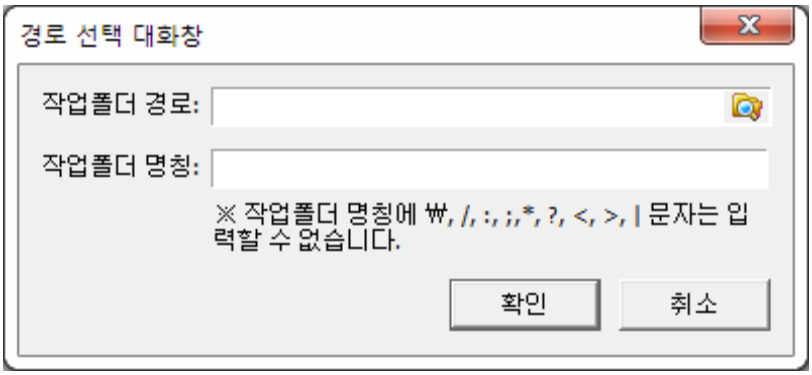
작업폴더란 ECMiner™을 사용하면서 자주 사용되는 폴더를 다른 이름으로 대변시킨(aliasing) 폴더입니다. 사용자가 추가하여야 하며 추가된 작업폴더는 리소스창 내에 Local 메뉴에 나타납니다.



리소스 트리 메뉴를 나타내는 위 그림에서 보이는 툴바를 이용하여 작업폴더를 추가/삭제/편집할 수 있습니다.

- 작업폴더 추가

  툴바 중  버튼을 누릅니다. 버튼을 누르면 다음과 같은 대화상자가 나타납니다.



위 대화상자에서 작업폴더로 지정할 폴더의 경로를 "작업폴더 경로"에 직접 입력하거나 버튼을 눌러 선택합니다. "작업폴더 명칭"에 지정된 폴더를 대변할 이름을 입력합니다. 작업폴더 명칭은 위 그림에 명시된 것처럼 일부 특수 문자는 사용할 수 없습니다. 모든 사항을 입력한 뒤 확인 버튼을 누르면 작업폴더가 추가됩니다.

- 작업폴더 변경

  기 설정된 작업폴더의 내용(실제 폴더 경로, 작업폴더 명칭)을 변경하려면, 툴바 중 버튼을 누릅니다. 작업폴더 추가 시 나타나는 대화상자가 나타납니다. 대화상자에는 현재 선택한 작업폴더의 폴더 경로와 명칭이 입력되어 있습니다.
  작업폴더를 추가할 때와 같이 변경하고자 하는 것을 수정한 뒤 확인 버튼을 누릅니다.

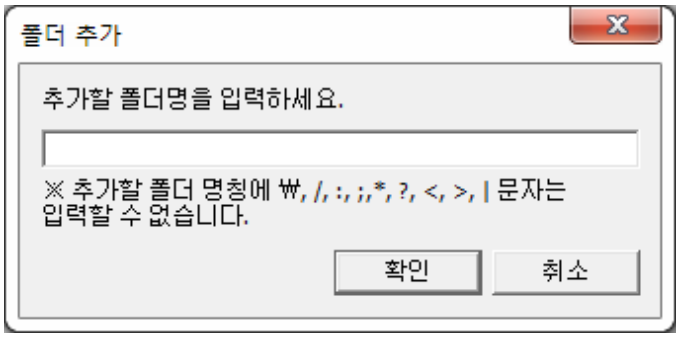
- 작업폴더 삭제

  기 설정된 작업폴더를 삭제하려면, 툴바 중 버튼을 누릅니다. "삭제하시겠습니까?"를 묻는 메시지 박스가 나타나고 **'예'** 버튼을 누르면 삭제됩니다.

- 작업폴더 내용 다시 읽기

  설정된 작업폴더 및 하위 폴더의 파일 목록을 다시 읽어야 할 경우, 툴바 중 버튼을 누릅니다.

- 작업폴더 새로 만들기
  리소스 트리 메뉴의 작업폴더에서 마우스 우클릭을 하면 작업폴더 새로 만들기를 할 수 있습니다. 작업폴더 새로 만들기를 실행하면 아래의 대화상자가 나타납니다.



추가할 폴더 명칭은 위 그림에 명시된 것처럼 일부 특수 문자는 사용할 수 없습니다.
추가할 폴더명을 입력한 뒤 확인 버튼을 누르면 작업폴더가 추가됩니다

작업폴더 활용

작업폴더를 이용하면, 프로젝트의 관리 효율을 높이고 사용자 간의 데이터 공유가 가능해집니다.

- 프로젝트 관리

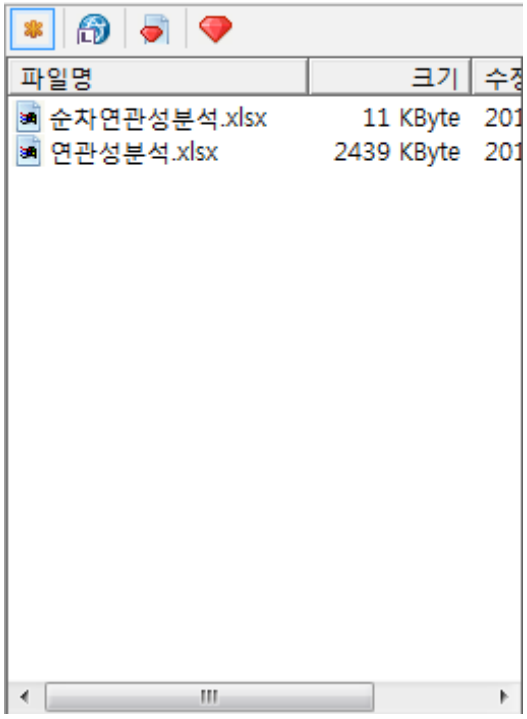
  여기서 언급한 프로젝트는 ECMiner™의 프로젝트 (파일)이 아니라 한 과제에 대한 수행 내용 전반을 의미합니다.
  한 프로젝트 또는 과제에 대한 작업폴더를 하나 생성하고 분석 작업 수행 중 생성된 데이터 파일, ECMiner™의 프로젝트 파일, 모델 파일 등을 생성한 작업 폴더에 모아 관리합니다. 이렇게 하면 해당 프로젝트의 내용을 알기 위하여 여러 폴더를 뒤져 볼 필요 없이 생성한 작업폴더의 내용만 살펴 보면 됩니다.

- 데이터 공유

  사용자간의 사용 환경이 조금씩 다르므로 한 사용자가 지정한 입력 노드의 선택사항이 다른 사용자에게는 적용되지 않을 수 있습니다. 이런 경우 같은 이름의 작업폴더를 서로 생성해 놓고 생성한 작업폴더 내의 파일만 사용하게 되면 사용자 간의 데이터 파일 공유가 추가 작업 없이 가능해 집니다.

리소스 목록

메뉴 트리에서 선택된 폴더의 파일 목록이 나타납니다.



파일 목록은 선택된 폴더의 파일 목록이 나타나는 부분이며, 파일 필터링 툴바는 표시할 파일의 형태를 결정합니다.

- 파일 필터링 툴바

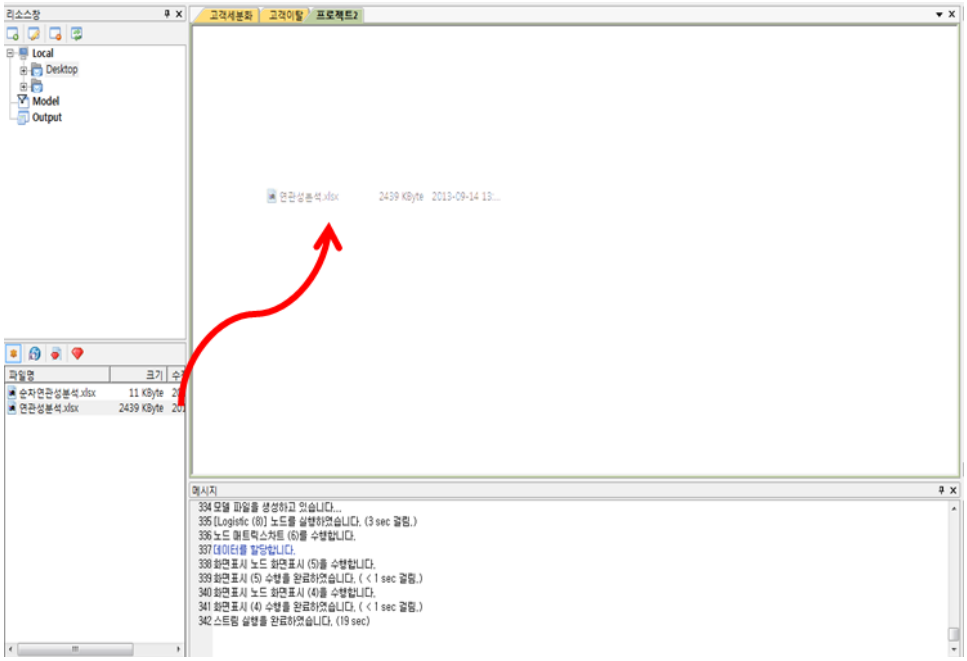
ICON	설명
	ECMiner™ 옵션에서 지정된 모든 형태의 파일을 나타냅니다.
	ECMiner™의 데이터 파일인 *.ecl 파일만 나타냅니다.
	ECMiner™에서 생성되어 전역화된 모델 파일(*.gms)만 나타냅니다.
	ECMiner™의 프로젝트 파일(*.ecm)만 나타납니다.

- 이 툴바에서 선택된 파일 형태만 파일 목록에 나타납니다.

리소스 목록 활용

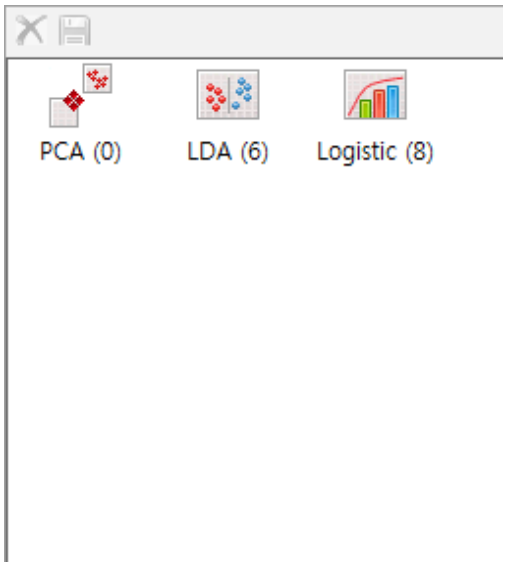
리소스 목록을 이용하여 ECMiner™ 프로젝트 파일을 열거나 입력 노드 추가를 보다 쉽게 수행할 수 있습니다.

텍스트 파일을 데이터 소스로 하는 파일입력 노드를 추가할 경우 노드창에서 파일입력 노드를 생성한 후 데이터 파일을 선택하는 식으로 입력 노드를 추가할 수 있습니다. 만약, 지정된 작업폴더 내에 해당 파일이 있다면 다음과 같은 방법으로 보다 쉽게 입력 노드를 생성할 수 있습니다.



1.3.7 리소스창-Model

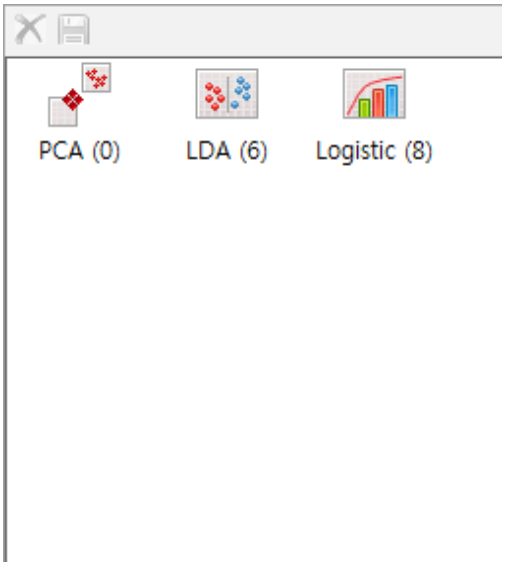
스트림 실행 후 생성된 임시 모델 파일이 리소스 목록에 나타납니다.



임시 모델 생성

- 위 목록에 나타나는 임시 모델은 스트림 실행 후 모델링 노드의 결과물로 모델링 노드를 포함한 스트림 실행 시 생성되어 목록에 나타나게 됩니다. 생성될 때 스트림

구성 시 모델링 노드에 사용한 명칭을 그대로 사용하며 이 목록은 프로젝트 등으로 구분되지 않기 때문에 실행한 모델링 노드의 명칭이 같은 경우 아래 그림처럼 같은 이름의 모델이 생성되게 됩니다.



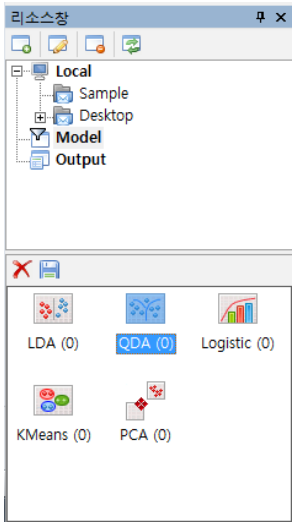
프로그램 내부에서 이들을 달리 처리하기 때문에 문제가 되지 않지만 사용자가 혼동할 수 있는 부분입니다. 따라서 노드를 생성한 다음에 이름이 중복되지 않도록 그에 걸맞은 명칭을 부여하는 것이 중요합니다.

임시 모델 삭제

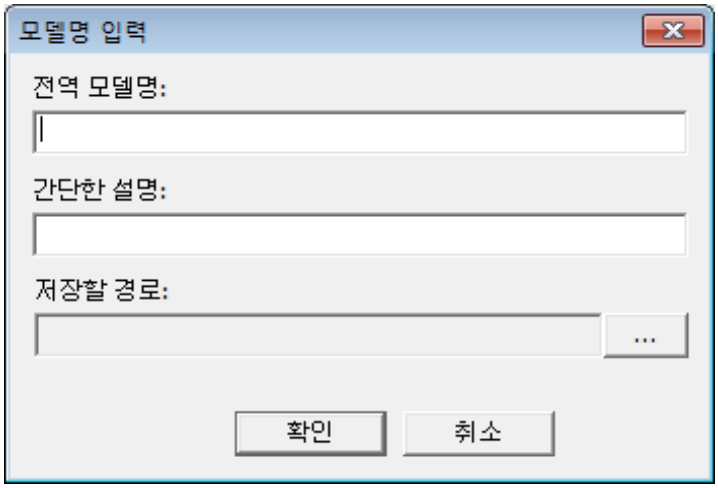
- 목록에 나타나 있는 임시 모델 중 필요 없다고 판단되는 것은 삭제할 수 있습니다. 삭제할 임시 모델을 선택한 뒤 툴바에 있는  버튼을 누릅니다.

임시 모델 전역화

- 생성된 임시 모델 중 재사용하는 것이 좋겠다는 판단이 들면 전역 모델로 변경할 수 있습니다. 임시 모델을 전역 모델로 만들려면 아래 그림과 같이 전역화 할 임시 모델을 선택하여 툴바에 있는  버튼을 누릅니다.

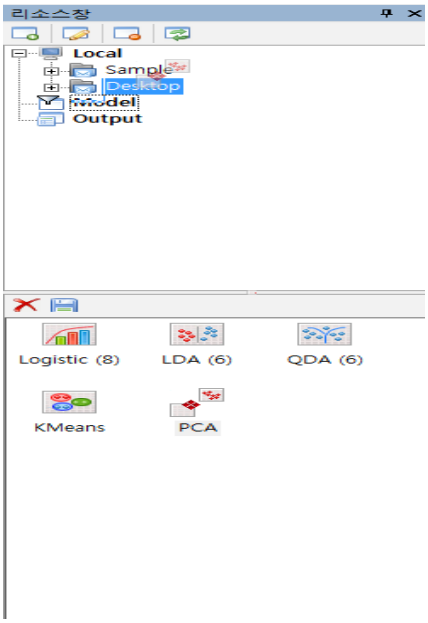


모델을 전역화 하기 위한 옵션을 묻는 대화상자가 나타납니다.

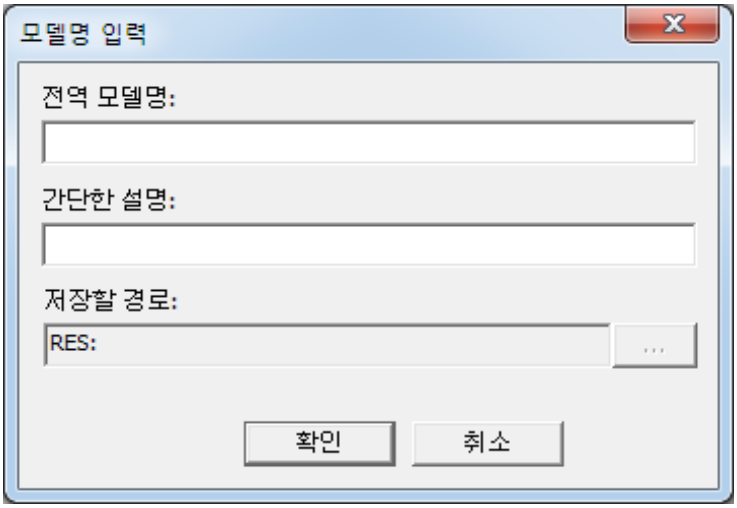


나타난 대화상자에 전역화 하였을 경우 가질 모델명과 간단한 설명을 입력한 후 저장할 경로를 지정하여 확인 버튼을 누르면 선택한 작업폴더에 입력한 "전역 모델명.gms"라는 전역 모델 파일이 생성됩니다.

- 또 다른 방법으로는 아래 그림과 같이 전역화 할 임시 모델을 선택하여 저장할 작업폴더(하위 폴더 포함)로 Drag & Drop 을 합니다.



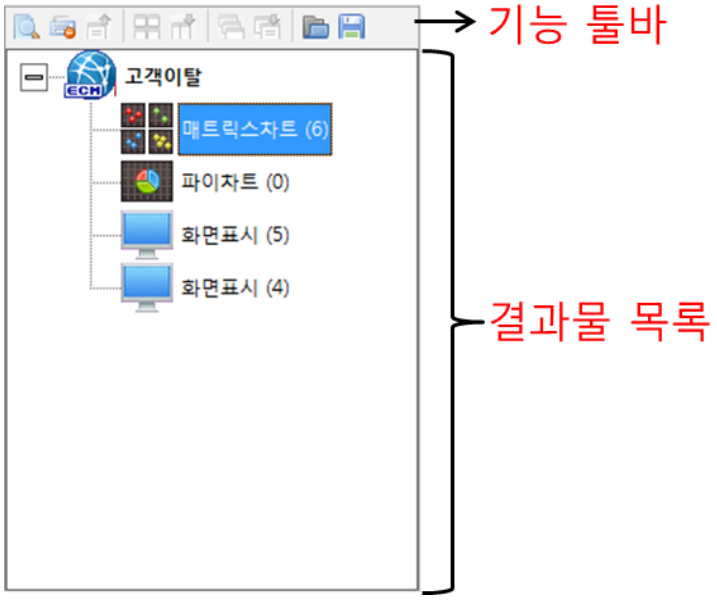
모델을 전역화 하기 위한 옵션을 묻는 대화상자가 나타납니다.



나타난 대화상자에 전역화 하였을 경우 가질 모델명과 간단한 설명을 입력한 후 저장할 경로를 지정하여 확인 버튼을 누르면 선택한 작업폴더에 입력한 "전역 모델명.gms"라는 전역 모델 파일이 생성됩니다. 저장할 경로는 Drop 한 작업폴더(혹은 하위 폴더)를 나타냅니다.

1.3.8 리소스창-Output

스트림 실행 후 생성된 결과물은 Output 메뉴 선택 시 목록에 나타납니다.



결과물 목록은 프로젝트 별로 생성된 결과물이 관리되어 나타나는 목록이며, 기능 툴바는 출력물 관리 및 보기의 편의를 위한 기능을 제공하는 툴바입니다.

결과물 다시 보기

- 생성된 결과물을 다시 보려면 결과물 목록에서 보고자 하는 결과물을 선택한 뒤 더블 클릭합니다. 또는 결과물을 선택한 다음 툴바의 제일 앞에 있는 버튼을 누릅니다. 목록화 되어 있는 결과물에 대하여 다중 선택이 가능하며 선택된 결과물들은 버튼을 누르면 모두 나타나게 됩니다. 다중 선택이 가능한 이유는 다중 선택을 필요로 하는 기능이 있기 때문입니다. 자세한 내용은 다음을 참조합니다.

Output 툴바

- 기능 툴바는 결과물을 효율적으로 보기 위한 기능과 관리 기능을 제공합니다.
  - 결과물 열기

    목록에서 선택한 모든 결과물을 화면에 표시합니다.

  - 결과물 삭제

    목록에서 선택한 모든 결과물을 삭제 합니다. 삭제하면 다시 열어 볼 수가 없습니다.

- 열린 결과물 닫기

  열린 모든 결과물 창을 닫습니다. 삭제되는 것이 아니며 화면 상에 나타나지만 않게 되는 것입니다.

- 열린 결과물 바둑판 정렬

  현재 열려 있는 결과물이 두 개 이상일 경우 바둑판 모양으로 정렬하여 화면에 표시합니다.

  위 그림에서 현재 열린 결과물이 상단 그림처럼 겹쳐 있을 경우 이 버튼을 누르면 하단 그림처럼 겹치지 않게 바둑판 모양으로 정렬되어 나타납니다. 결과물들을 한번에 보면서 비교해야 할 작업이 필요할 경우 유용합니다.

- 선택된 결과물들 바둑판으로 정렬하여 열기

  결과물을 바둑판 형식으로 정렬하여 보여주는 것은 위 기능 버튼과 비슷한 역할입니다. 차이점은, 위 버튼은 열려 있는 결과물에 대한 바둑판 정렬이고, 이 버튼은 선택된 결과물들을 (닫혀있다면) 열면서 바둑판 모양으로 정렬하는 것입니다. 두 개 이상의 결과물을 선택하였을 경우 활성 됩니다.

- 열린 결과물 계단식으로 정렬

  열린 결과물을 계단 형식으로 정렬합니다.

- 선택된 결과물들 계단식으로 정렬하여 열기

  위 기능 버튼과 마찬가지고 결과물을 계단 형태로 정렬합니다. 차이점은 정렬하는 대상이 열린 결과물이 아니라 목록에서 선택한 결과물이라는 것입니다. 두 개 이상의 결과물을 선택하였을 경우 활성 됩니다.

- 기 저장된 결과물 열기

  결과물을 저장한 ECMiner™용 결과 파일(*.ept)을 열어 다시 볼 수 있습니다. 이 버튼을 누르면 기 저장한 결과 파일을 선택할 수 있는 대화상자가 나타나며 보고자 하는 파일을 선택한 뒤 열기 버튼을 누르면 해당 결과물을 볼 수 있습니다.

  이렇게 연 결과물은 위 그림과 같이 "기타"라는 항목으로 목록에 추가되며 다른 결과물들과 마찬가지로 작동합니다.

- 선택된 결과물 결과 파일로 저장

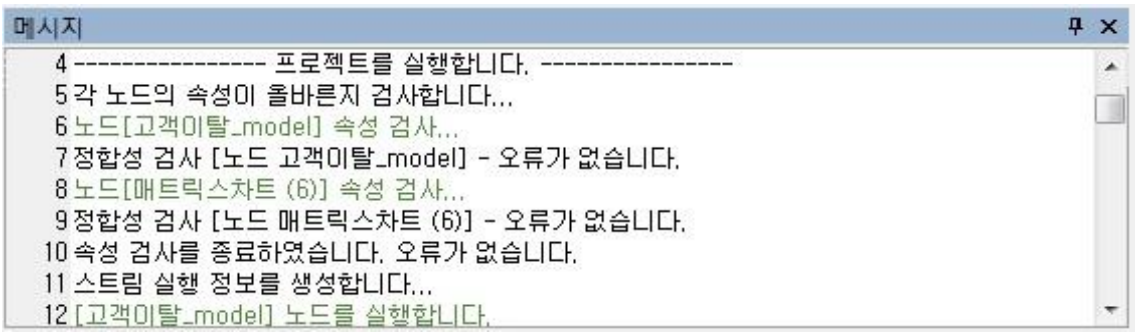
  ECMiner™은 생성된 결과물을 파일로 저장할 수 있는 기능을 가지고 있습니다. 다른 사람에게 보여줘야 한다든지, 유용한 정보라서 보관해 놓고 싶은 경우

사용할 수 있습니다.
저장할 결과물을 선택하고 이 기능 버튼을 누릅니다. 그러면 어떤 폴더에 무슨 이름으로 저장할 지를 입력하는 대화상자가 나타나며, 여기에 저장할 폴더 위치와 이름을 입력한 뒤 저장 버튼을 누르면 결과 저장 파일(*.ept)가 생성됩니다. 이렇게 저장된 결과 파일은 위 기능 버튼을 이용하여 다시 열람할 수 있습니다.

NOTE
화면 표시 노드의 결과물은 생성된 데이터를 그대로 대변하고 있기 때문에 따로 저장할 필요가 없습니다. 화면 표시 노드의 결과를 데이터 파일 형태로 저장합니다.

1.3.9 메시지창

프로젝트창 내의 스트림 실행 시 발생하는 노드의 실행 정보 혹은 에러 등을 표시하는 Log 창입니다. ECMiner™는 스트림을 실행하기 전 스트림을 구성하고 있는 노드의 속성값이 올바르게 지정되어 있는지 검사합니다. 만약, 잘못 입력되거나 필수 항목을 입력하지 않았다면 에러가 발생하게 되며 그 정보는 메시지창에 기록되게 됩니다. 에러가 발생하여 스트림이 실행되지 않는다면 메시지창의 내용을 참조하여 에러에 대한 내용을 수정합니다. 에러에 대한 정보뿐 아니라 노드를 실행할 때 발생하는 Log 도 메시지창에 표시되며 수행 시간도 같이 표시됩니다.



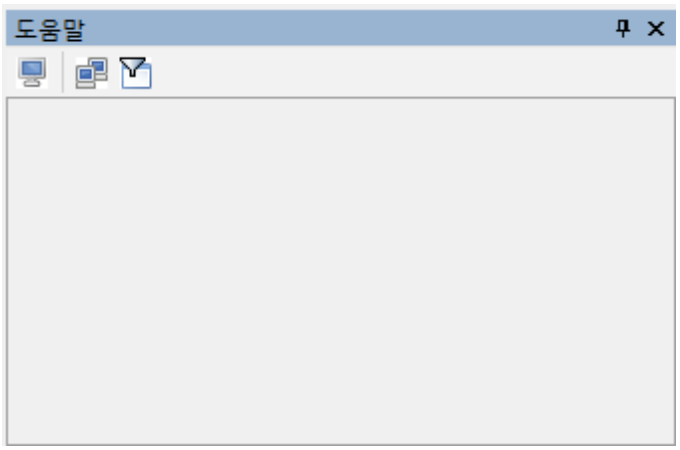
메시지 종류

메시지창에 나타나는 메시지 종류는 크게 5 가지입니다. 이들은 색상으로 구별되며 다음과 같습니다.

순번	색상	내용
1		일반 메시지
2		일반 메시지 중 강조가 필요한 경우에 사용됩니다.
3		작은 오류를 표현하기 위하여 사용됩니다.
4		오류의 내용이 치명적일 경우 사용됩니다.
5		메시지를 구분 지을 때 사용됩니다. 예를 들어 파일입력 노드의 속성을 검사하다가 형태변경 노드의 속성을 검사할 때 이들을 구분 짓기 위하여 사용되는 것입니다.

1.3.10 동적 도움말

노드의 속성 편집 및 ECMiner™ 사용을 위한 동적 도움말이 나타나는 창입니다.



노드 속성 편집 시 속성을 클릭하면 위 그림과 같이 선택한 속성에 대한 간단한 설명이 나타납니다.

 버튼을 누르면 현재 선택된 노드에 해당하는 도움말 부분이 자동 선택되어 나타나며 더블 클릭하면 해당 도움말을 보실 수 있습니다. 이 버튼은 Toggle 형태도 다시 한번 누르면 속성 도움말로 변경됩니다.

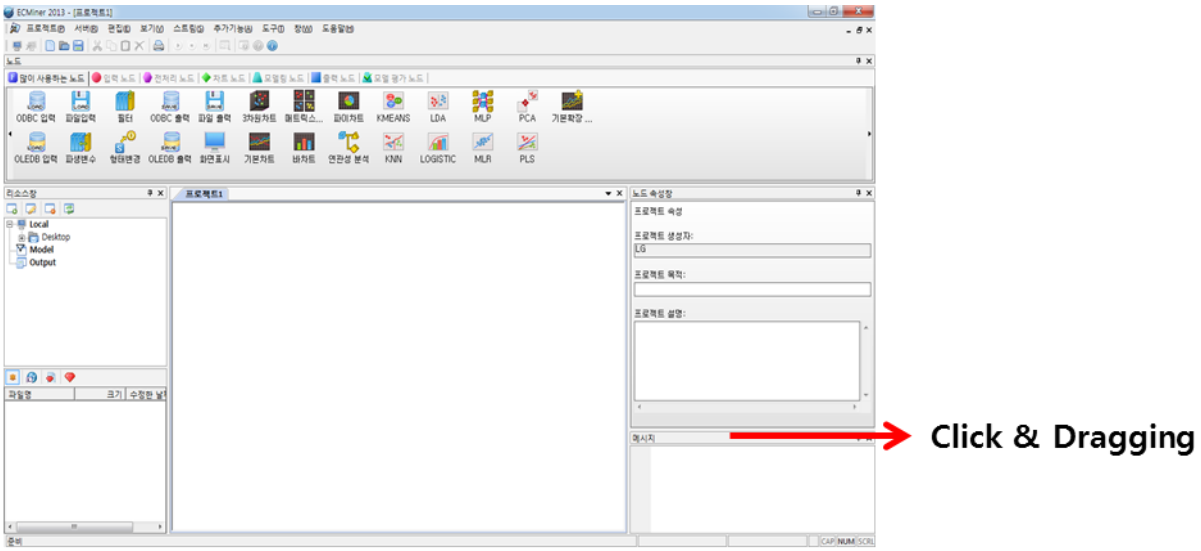
나머지 버튼은 각각 색인 검색, 도움말 검색 기능입니다.

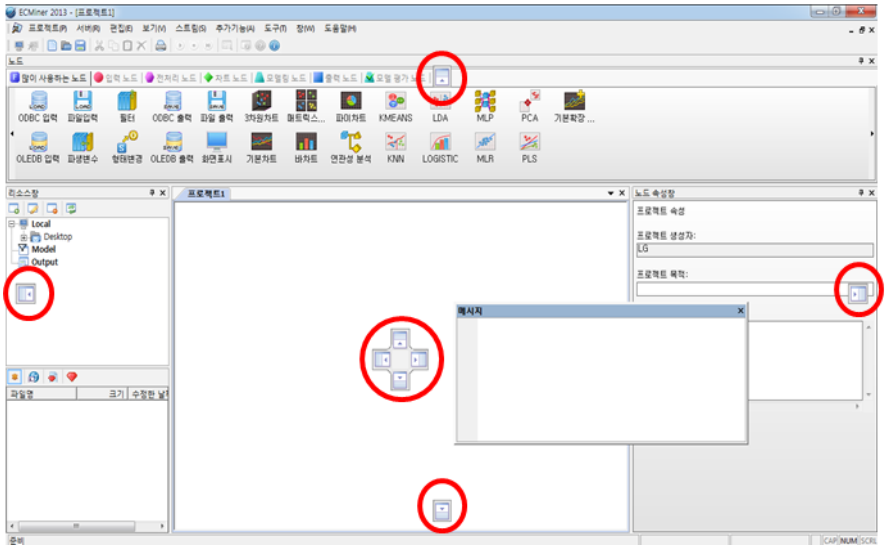
1.3.11 UI 변경

ECMiner™ UI 상의 모든 컨트롤창은 사용자가 나름대로 위치 및 크기를 변경할 수 있습니다.

위치 변경

- 컨트롤창의 위치를 변경하고자 할 경우 각 컨트롤창의 제목 부분을 마우스로 클릭한 상태에서 드래깅을 합니다. 만약 도움말 컨트롤창의 위치를 변경하고자 할 경우 도움말 컨트롤창의 제목 부분을 클릭한 후 드래깅을 합니다. 그러면 다음 그림과 같이 화살표 모양의 아이콘이 나타납니다.





화살표 모양의 아이콘은 현재 드래깅하고 있는 컨트롤창(위 그림에서는 도움말 컨트롤창)을 위치시킬 곳을 가리킵니다.

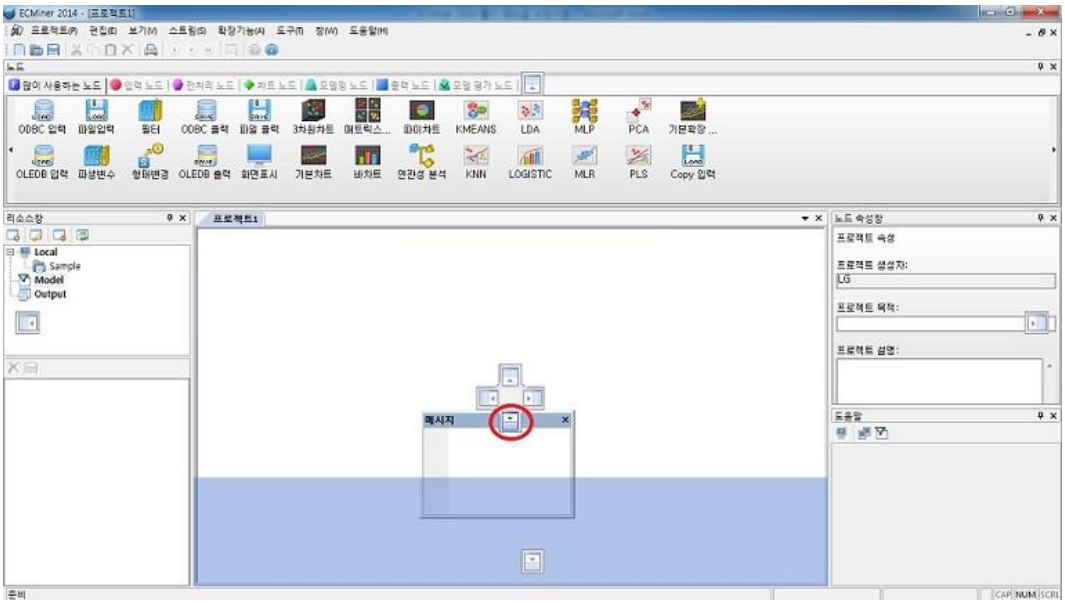
- 가운데 4 방향 아이콘

  이 아이콘들은 드래깅하는 과정 중 마우스가 속한 컨트롤창 내의 상, 하, 좌, 우를 의미합니다.

- 주변의 4 방향 아이콘

  이 아이콘들은 프로그램 전체 윈도우에 대한 상, 하, 좌, 우를 의미합니다.

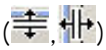
- 드래깅 중인 상태에서 컨트롤창을 위치시키고 싶은 곳의 화살표 위에 마우스 포인터를 이동합니다. 위 그림에서 도움말 컨트롤창을 프로젝트창 밑으로 위치시키고자 한다면 가운데 4 방향 아이콘 중 아래 있는 화살표에 마우스 포인터를 이동시킵니다.



그러면 드래깅 중인 컨트롤창이 위치할 부분이 위 그림처럼 파란색 사각형과 같이 나타납니다. 위치를 확인 후 마우스 버튼을 놓으면 해당 컨트롤창의 위치가 변경됩니다.
화살표 이외의 위치에 놓으면 프로그램의 전체 프레임과 독립적으로 떠다니는(floating) 컨트롤창이 됩니다.

NOTE
프로젝트창의 위치는 다른 컨트롤창들의 기준이 되므로 변경될 수 없습니다.

크기 조정

- 각 컨트롤창의 경계선 위에 마우스 포인트를 위치시키면 마우스 포인트 모양이 크기를 조정할 때의 모양()으로 변경됩니다. 이 때 마우스 왼쪽 버튼을 클릭하고 드래깅하여 크기를 조정할 수 있습니다.

컨트롤창 닫기

- 각 컨트롤창의 제목 부분을 보면 다음 그림과 같은 닫기 버튼이 있습니다.

  이 버튼을 누르면 컨트롤창이 닫혀 보이지 않게 됩니다. 메인 메뉴 중 "보기 > 도구모음"을 이용하여 컨트롤창을 닫을 수도 있습니다.

컨트롤창 열기

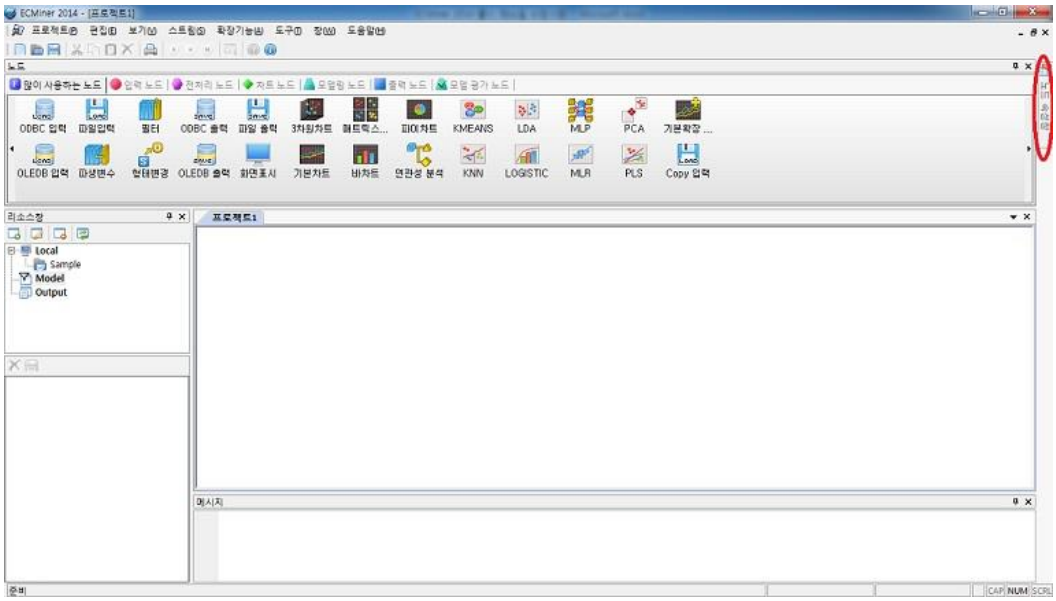
- 닫힌(안 보이는) 컨트롤창을 다시 보이게 하려면 메인 메뉴 중 "보기 > 도구모음"을 이용합니다.

컨트롤창 자동 숨기기

- 각 컨트롤창의 제목 부분을 보면 다음 그림과 같은 자동 숨기기 버튼이 있습니다.

- 버튼을 누르면 다음 그림과 같이 컨트롤창이 자동으로 숨겨지며(아래 그림은 노드 속성창과 도움말창을 자동 숨기기 한 경우임), 마킹된 탭 위로 마우스 포인트를 올리면 다시 나타나게 됩니다.



화면 크기가 작을 경우 프로젝트창을 크게 보면서 작업하고자 할 때 유용합니다.

1.4 마우스 인터페이스

ECMiner™의 많은 기능들은 마우스 조작으로 이루어집니다. 왼쪽 버튼 클릭은 결정을, 오른쪽 버튼 클릭은 지원 메뉴 나열을, 왼쪽 버튼 더블 클릭은 정보 표시 같은 기능을 주로 지원합니다.

드래그 & 드롭

노드창에서 프로젝트창으로 노드를 생성시킬 때 해당되는 노드를 마우스로 끌어서 놓습니다.

리소스창의 Model 에서 프로젝트창으로 노드를 생성시킬 때 해당되는 노드를 마우스로 끌어서 놓습니다.

리소스창의 Model 에 있는 모델을 Local 이나 Server 의 폴더로 드래그하면 파일이 이동하며 광역모델로 저장됩니다.

왼쪽 버튼 클릭

프로젝트창에서 노드를 클릭하면 속성창에 해당 노드의 속성이 나타나게 됩니다.

창, 노드 그리고 속성에 관계없이 공통적으로 왼쪽버튼을 클릭하면 해당되는 부분을 선택한다는 의미입니다.

프로젝트창의 빈 공간을 클릭한 경우, 노드 속성창에서 프로젝트 관련 설정을 할 수 있습니다.

오른쪽 버튼 클릭

프로젝트창에서 마우스를 노드 위에 위치 시킨 후 오른쪽 버튼을 클릭하면 노드 연결 및 해제, 노드 삭제 여부 그리고 데이터 편집기 실행 등의 메뉴가 나타나게 됩니다. 물론 선택은 왼쪽 버튼으로 합니다.

왼쪽 버튼 더블 클릭

리소스창의 Output 에서 차트, 화면보기 등의 결과를 더블 클릭하면 결과 정보가 화면에 나타납니다.

리소스창의 Model 에서 모델을 더블 클릭하면 프로젝트창에 해당 노드가 생성됩니다.

노드창에서 노드를 더블 클릭하면 프로젝트창에 해당 노드가 생성됩니다.

프로젝트창에서 모델 노드를 더블 클릭하면 그 모델의 정보가 화면에 나타납니다.

파일 읽기 노드를 더블 클릭하면 데이터 탐색기가 화면에 나타납니다.

1.5 단축키 인터페이스

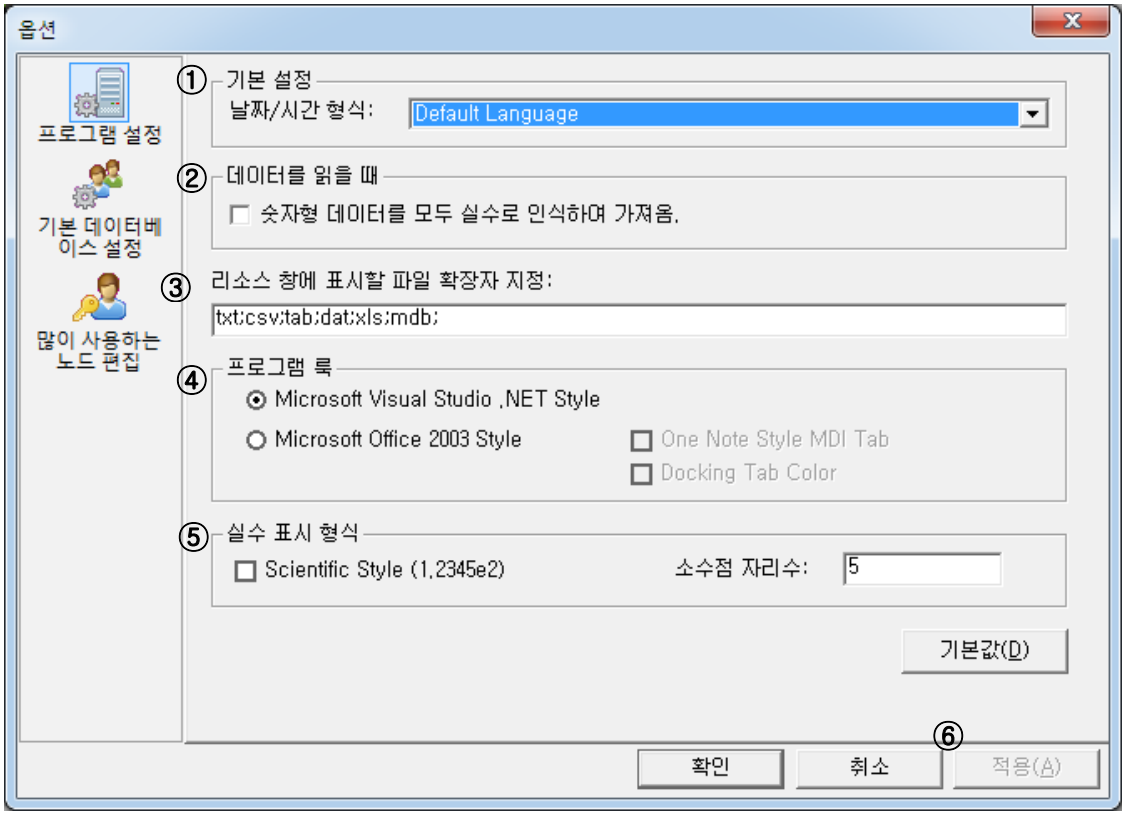
ECMiner™는 기능과 관련된 많은 단축키들을 지원합니다. 예를 들면 프로젝트창에서 노드를 삭제할 때 키보드상의 삭제(Delete)키를 사용함으로써 마우스 버튼과 같은 기능이 가능합니다. 마찬가지로 Ctrl 키를 누른 상태에서 S 를 누르면 프로젝트의 저장이 가능하게 됩니다.
아래 테이블은 단축키와 그 키가 지원하는 기능을 보여줍니다.

단축키	기능	단축키	기능
Ctrl + N	새로 만들기	Alt + V, T	도구 모음
Ctrl + O	열기	Alt + V, T, S	표준 툴바
Ctrl + S	저장하기	Alt + V, T, N	노드창
Ctrl + P	인쇄	Alt + V, T, P	노드 속성창
F5	실행	Alt + V, T, E	리소스 창
Shift + F5	실행 정지	Alt + V, T, H	동적 도움말
Ctrl + F5	부분 실행	Alt + V, S	상태 표시줄
Alt + P	프로젝트	Alt + V, E	데이터 탐색기

Alt + P, N	새로 만들기	Alt + S	스트림
Alt + P, O	열기	Alt + S, A	노드 추가
Alt + P, C	닫기	Alt + S, C	노드 생성 후 연결
Alt + P, S	저장	Alt + S, N	연결
Alt + P, A	다른 이름으로 저장	Alt + S, L	연결 끊기
Alt + P, P	인쇄	Alt + S, R	실행
Alt + P, V	인쇄 미리 보기	Alt + S, S	실행 중지
Alt + P, R	인쇄 설정	Alt + S, N	부분 실행
Alt + P, X	끝내기	Alt + T	도구
Alt + E	편집	Alt + T, O	옵션
Alt + E, T	잘라내기	Alt + T, I	프로그램 환경 초기화
Alt + E, C	복사	Alt + W	창
Alt + E, P	붙여넣기	Alt + W, C	계단식 배열
Alt + E, D	지우기	Alt + W, T	바둑판식 배열
Alt + V	보기	Alt + H	정보
Alt + V, T, M	메시지 창	Alt + H, A	ECMiner™ 정보

1.6 프로그램 설정 개요

프로그램 설정의 대화창은 ECMiner™ 최상단에 위치한 메뉴 중 **도구(T)- 옵션(O)**을 선택하면 나타납니다. 이 대화창을 통해 사용자가 프로그램 관련 각종 옵션을 지정할 수 있습니다.



옵션을 구분 지어 놓은 메뉴 탭과 해당 메뉴 탭을 선택하였을 때 나타나는 편집창으로 구성되어 있습니다. 설정할 수 있는 옵션은 프로그램의 일반적인 내용, 기본적으로 사용할 데이터베이스 설정, 많이 사용하는 노드 편집 등 세가지가 있습니다.

대화창에는 다음과 같은 메뉴들이 탭으로 나눠져 있어 해당되는 옵션에 따라 선택적으로 설정이 가능합니다.

	**프로그램 설정**	날짜/시간 형식 같은 프로그램 기본 설정을 합니다.
	**기본 데이터베이스 설정**	ODBC 혹은 OLE DB 의 기본 드라이버와 로그인 정보를 설정합니다.
	**많이 사용하는 노드 편집**	많이 사용하는 노드를 추가하거나 삭제할 수 있습니다.

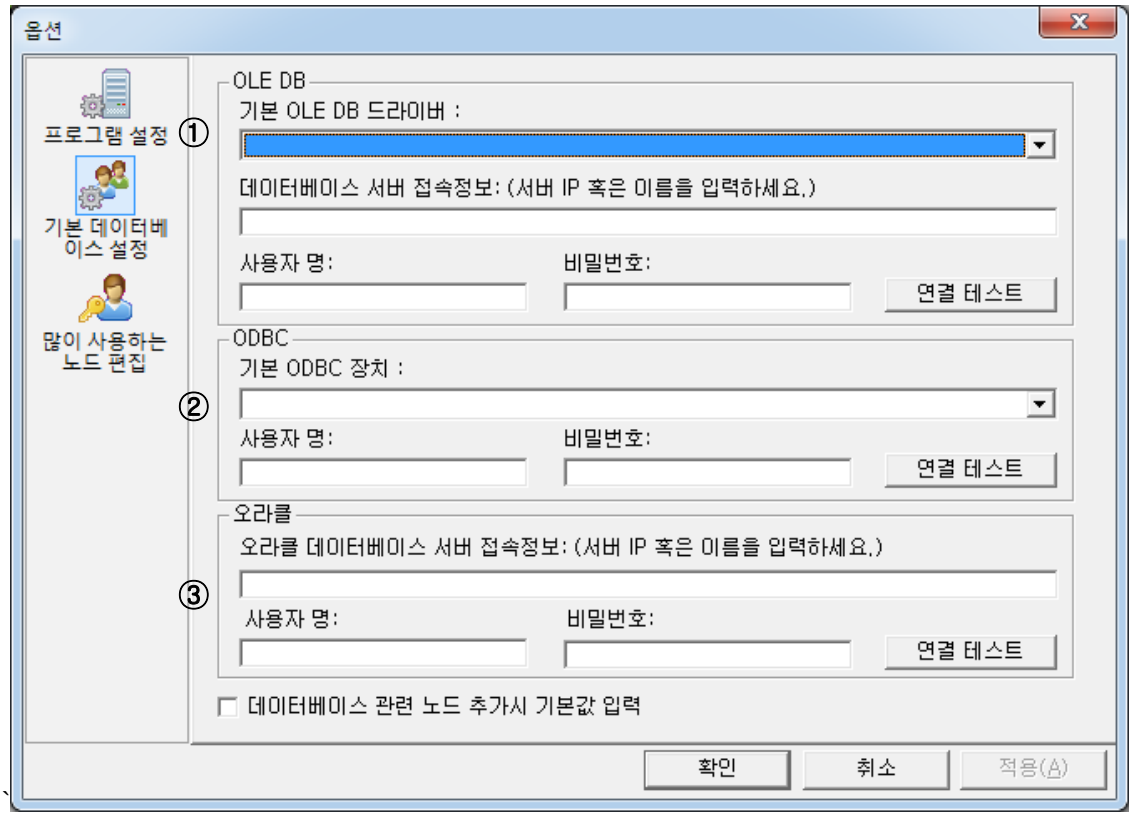
1.6.1 프로그램 설정

도구(T) - 옵션(O)의 **프로그램 설정**을 선택하면 다음과 같은 옵션창이 뜨게 됩니다. 이 옵션창에서는 국가별로 **날짜/시간** 형식, **숫자형 데이터 인식 방식**, **리소스창 파일 확장자 지정** 그리고 **프로그램 룩 방식**을 설정합니다.

번호	설명
①	날짜/시간 표시 형식을 언어를 기반으로 하여 선택합니다.
②	데이터를 읽을 때 숫자형식의 데이터를 모두 실수형으로 간주하여 처리하고자 할 경우 선택합니다.
③	리소스창에 표시될 파일 형태를 지정합니다. 표시할 파일 확장자를 입력하면 되며 ";"으로 구분하여 여러 개의 확장자를 지정할 수 있습니다. 지정된 것 외에 ecm, ecl, ept, gms 등은 지정하지 않아도 기본적으로 선택되어 있는 것들 입니다.
④	프로그램의 전체적인 look 을 지정합니다. 사용자가 보기 좋은 형태로 설정합니다.
⑤	데이터를 화면이나 보고서에 나타낼 경우 실수에 대한 표시 형식을 지정합니다.
⑥	상기 옵션을 기본값으로 할당합니다.

1.6.2 기본 데이터베이스 설정

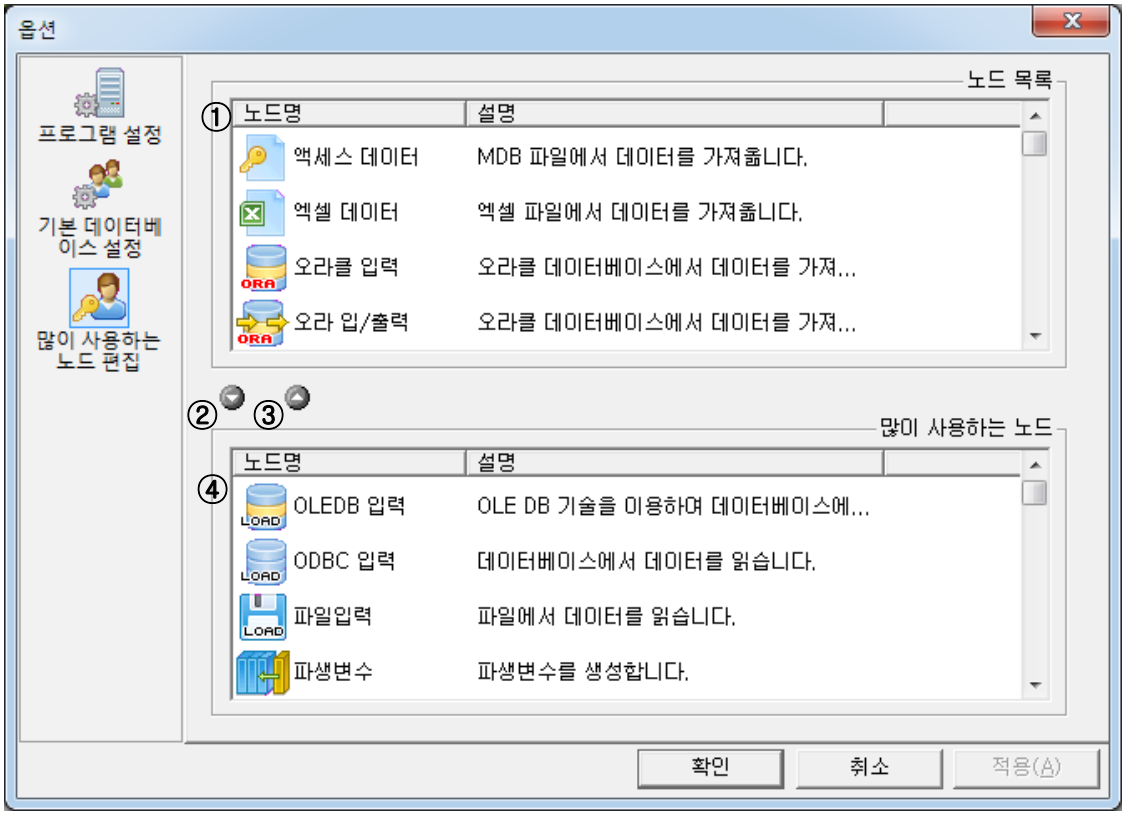
도구(T) - 옵션(O)의 기본 데이터베이스 설정을 선택하면 다음과 같은 옵션창이 뜨게 됩니다. ECMiner™가 지원하는 데이터베이스는 ODBC 와 OLE DB 입니다. 아래 옵션 창에서 드라이버 및 로그인 정보를 입력하면 프로젝트창에서 데이터베이스 작업 시 기본 데이터베이스 정보로 설정되어 편리하게 이용할 수 있습니다.



번호	설명
①	OLE DB 관련 노드에 입력할 데이터베이스 접속 정보를 입력합니다.
②	ODBC 관련 노드에 입력할 데이터베이스 접속 정보를 입력합니다.
③	이 항목을 선택해 놓아야 데이터베이스 관련 노드를 추가할 때 데이터베이스 접속 정보가 자동 입력됩니다.

1.6.3 많이 사용하는 노드 편집

도구(T) - 옵션(O)의 **많이 사용하는 노드 편집**(Favorites)을 선택하면 다음과 같은 옵션창이 뜨게 됩니다. 기능 버튼을 사용하여 **많이 사용하는 노드**에 노드를 추가하거나 삭제합니다. 추가되거나 삭제된 노드는 노드창의 **많이 사용하는 노드** 탭에서 확인할 수 있습니다.



번호	설명
①	ECMiner™에서 지원하는 모든 노드가 나열되어 있습니다.
②	①에 선택된 노드를 "많이 사용하는 노드"에 추가할 때 사용합니다.
③	"많이 사용하는 노드"에 있는 노드를 제거할 때 ④에서 제거할 노드를 선택하고 이 버튼을 누릅니다.
④	현재 "많이 사용하는 노드"에 추가된 노드 목록입니다.

기능 버튼

아이콘	기능
	선택된 노드를 많이 사용하는 노드 목록에 추가합니다.
	선택된 노드를 많이 사용하는 노드 목록에서 삭제합니다.

제 2 장 스트림

데이터 마이닝을 수행하기 위해서는 단위 기능으로 구현된 각각의 노드들을 연결하여 스트림으로 구성하여야 합니다. ECMiner™는 데이터입력, 전처리, 모델링, 데이터출력 등의 데이터 마이닝 작업 절차에 따라, 필요한 노드들을 스트림으로 연결함으로써 쉽고, 편리하게 다양한 기능을 구현할 수 있습니다. 노드 연결 조합에 따라 구현될 수 있는 기능은 무궁무진합니다. 한 개의 ECMiner™ 프로젝트 파일(ECMiner™의 프로젝트창에 사용자가 기능을 구현한 작업 파일로 ecm 확장자로 저장됨)에는 여러 개의 스트림이 존재할 수 있습니다.

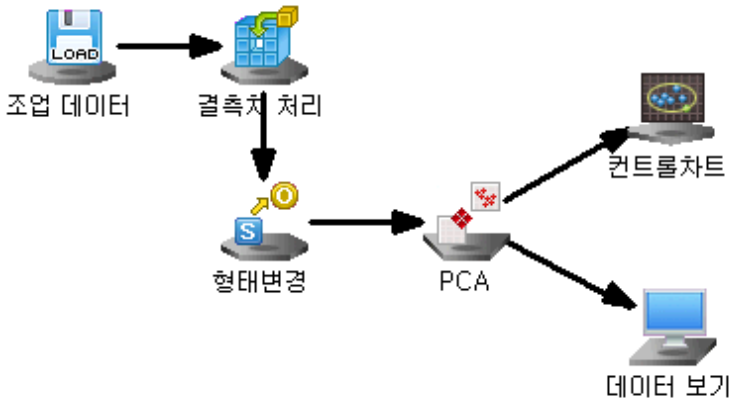
2.1 스트림 개요

ECMiner™를 사용하여 데이터 마이닝 작업을 수행한다는 것은 다양한 종류의 데이터를 분석, 가공 처리하여 예측, 분류, 군집, 연관성 분석 등의 모델을 구현하고, 업무에 적용하는 것입니다. 이 작업을 위해 여러 노드를 연결하여 기능을 구현하게 되고, 이렇게 기능이 구현된 연결 노드 묶음이 스트림입니다. 여러분은 앞으로 스트림을 구성하고 변경하면서 데이터 마이닝 작업을 수행하게 됩니다.

간략한 예:

1. 조업 데이터 생성
2. 생성된 데이터 보정 (결측치 처리, 형태 변경 등)
3. 조업 편차 분석을 위한 PCA 모델 수행
4. 컨트롤 차트, 생성된 데이터 등 결과물 보기

위와 같은 과정을 다음과 같이 표현해 보겠습니다.



위 그림은 앞쪽에서 예를 든 과정을 그대로 대변하고 있으며 ECMiner™에서는 이런 그림을 스트림이라고 명명하였습니다. 즉, 스트림은 데이터 마이닝을 위한 일련의 과정을 ECMiner™가 수행할 수 있도록 하는 기본 단위입니다. 스트림이라고 명명한 이유는 강에서 물이 흘러 가듯이 데이터가 연결된 노드를 따라 흘려가기 때문입니다.

2.2 스트림 구성

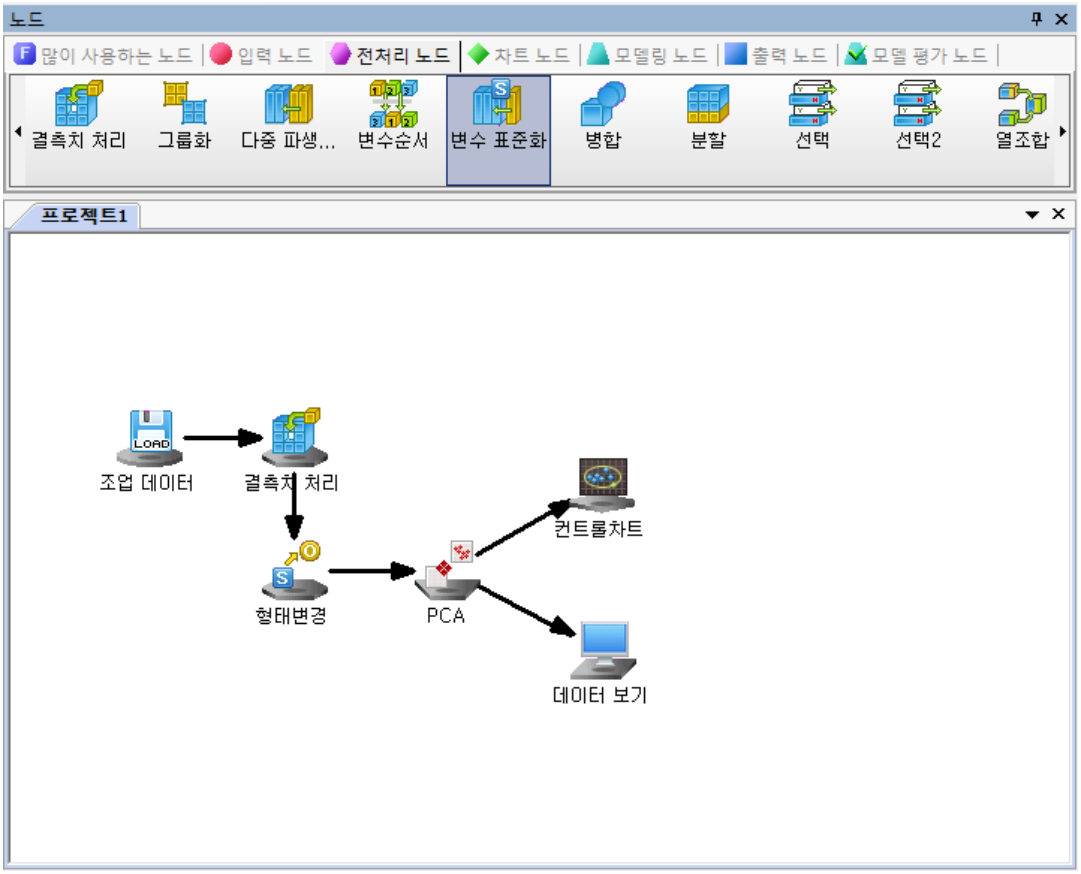
스트림은 실행 단위인 노드와 그들을 연결함으로써 생성할 수 있습니다. 따라서 스트림을 구성하려면 다음과 같은 과정을 일반적으로 거칩니다.

1. 노드 생성
2. 노드의 속성 편집
3. 노드 연결
4. 상기 과정을 반복

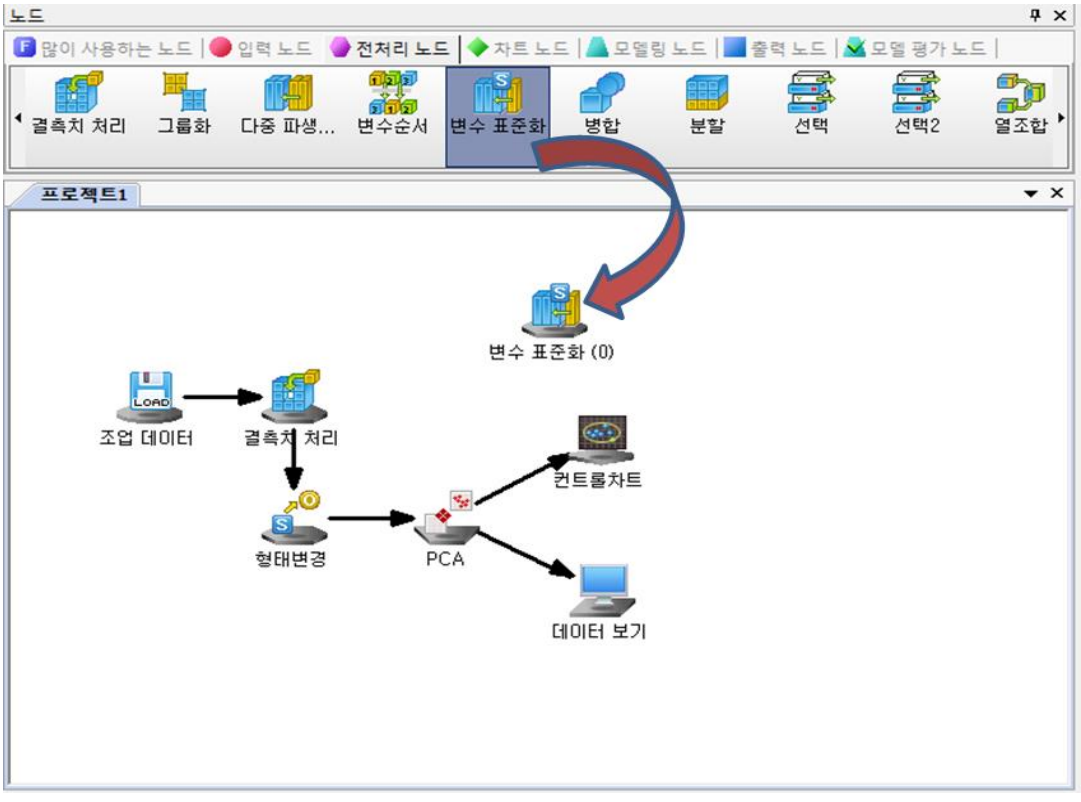
위에서 언급한 과정은 일반적인 순서이며 노드를 생성하고 연결한 뒤 속성을 편집하는 식으로 하여도 가능합니다.

노드 생성

- 노드를 프로젝트창에 생성하려면 노드창에서 생성하고자 하는 노드를 선택한 뒤 더블 클릭하거나,



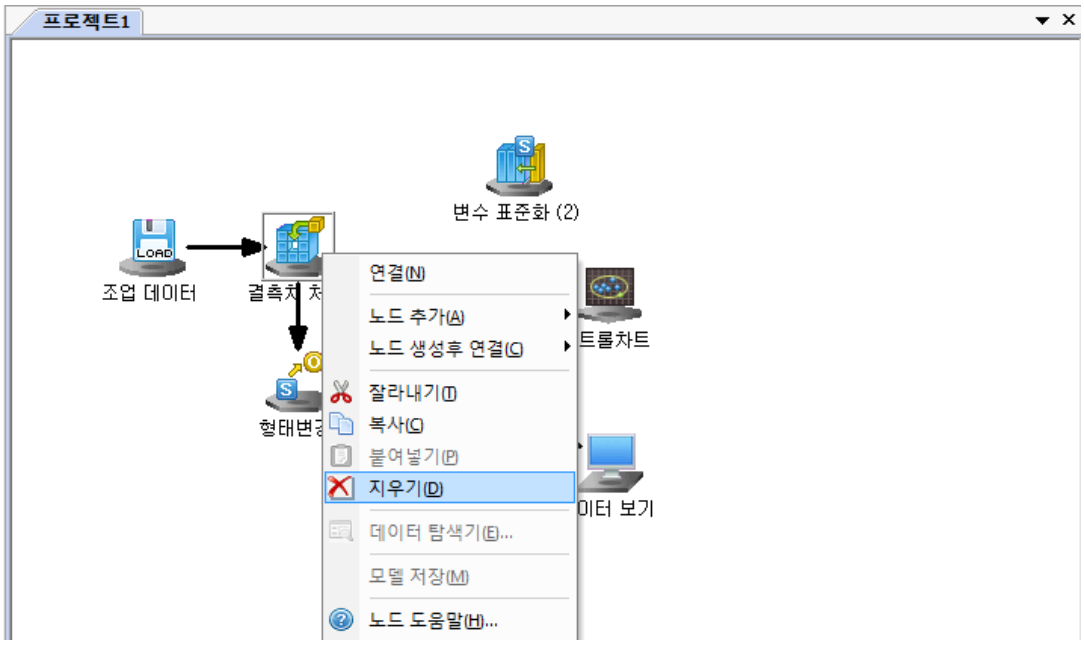
드래그하여 프로젝트창에 드랍합니다. (드래그 & 드랍)



다른 방법으로, 메인 메뉴 혹은 프로젝트창의 컨텍스트 메뉴를 이용하여 노드를 추가할 수 있습니다.

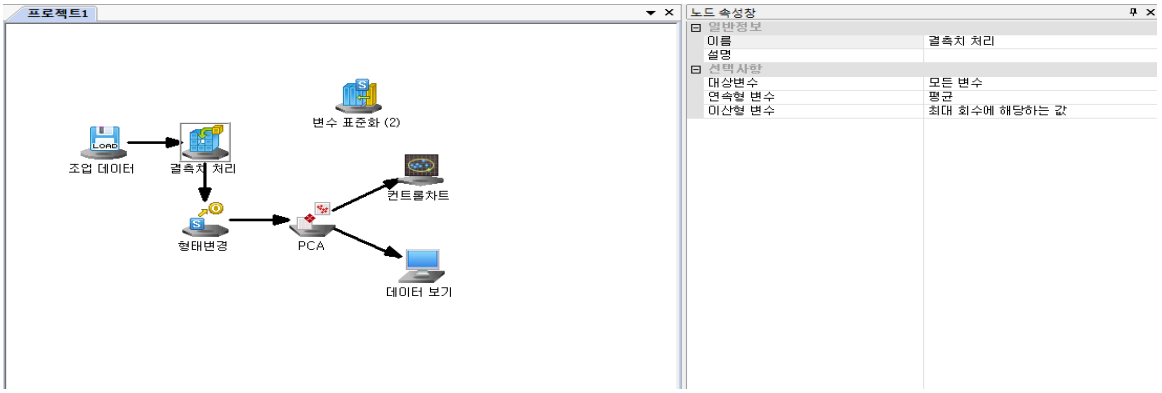
노드 제거

- 생성된 노드를 제거하려면 제거할 노드를 선택한 뒤 DELETE 키를 누르거나 마우스 오른쪽 버튼을 누르면 나타나는 다음 메뉴 중 지우기를 선택합니다.



노드 속성 편집

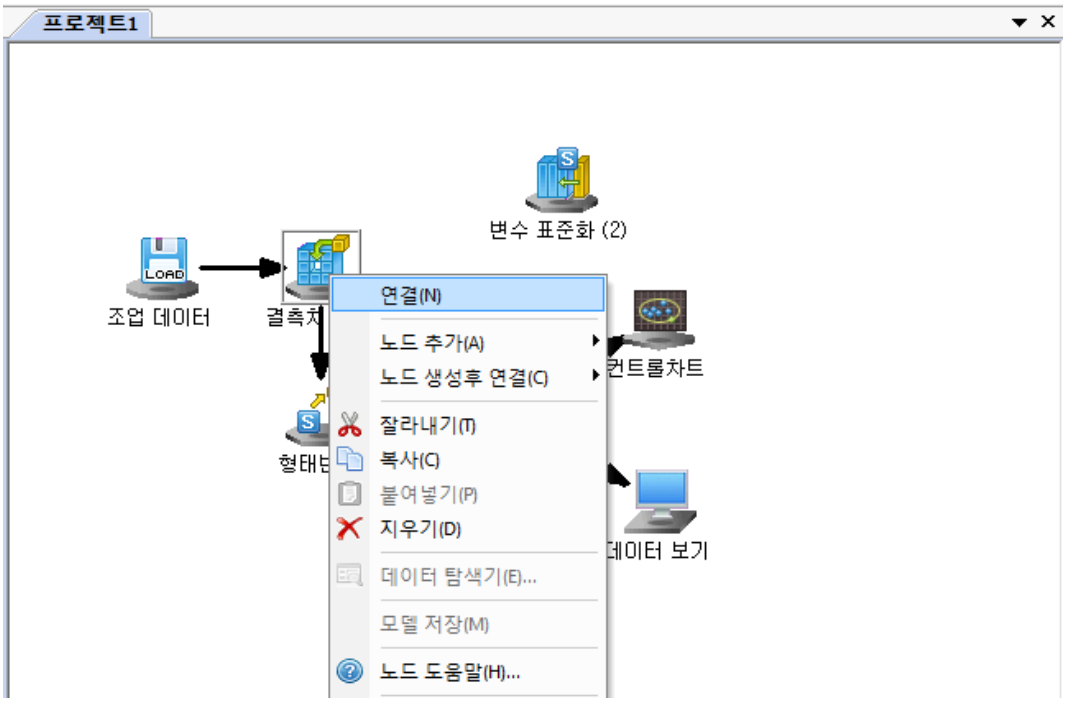
- 속성을 편집하고자 하는 노드를 프로젝트창에서 선택하여 클릭하면, 속성창에 선택한 노드의 속성을 편집할 수 있는 UI 가 나타납니다.



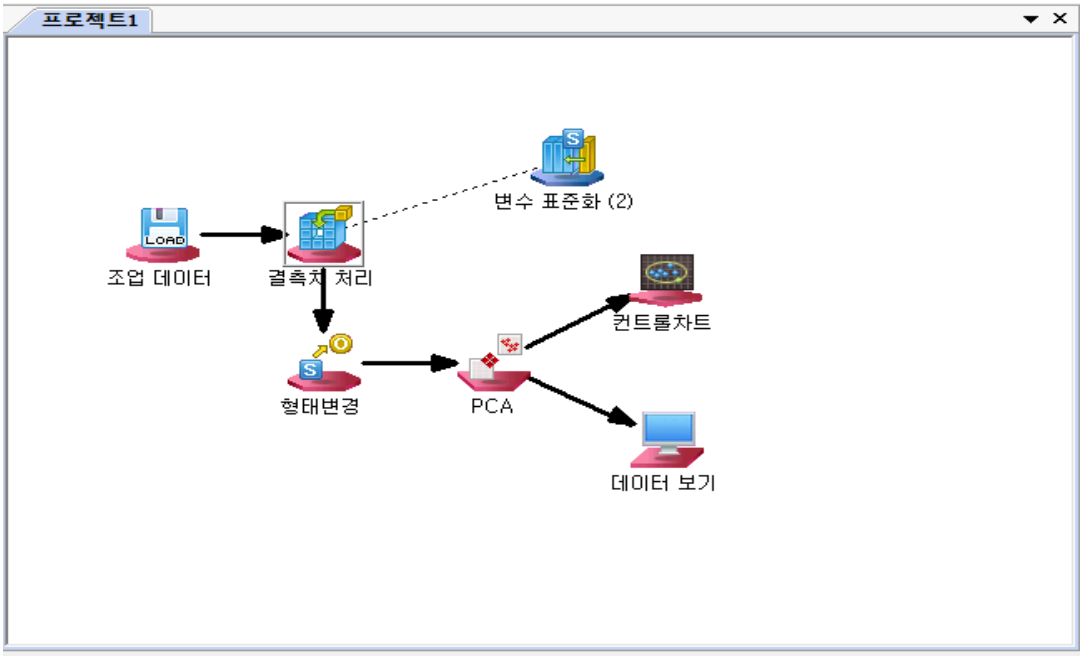
나타난 속성창을 이용하여 선택한 노드의 속성을 변경할 수 있습니다.

노드 연결

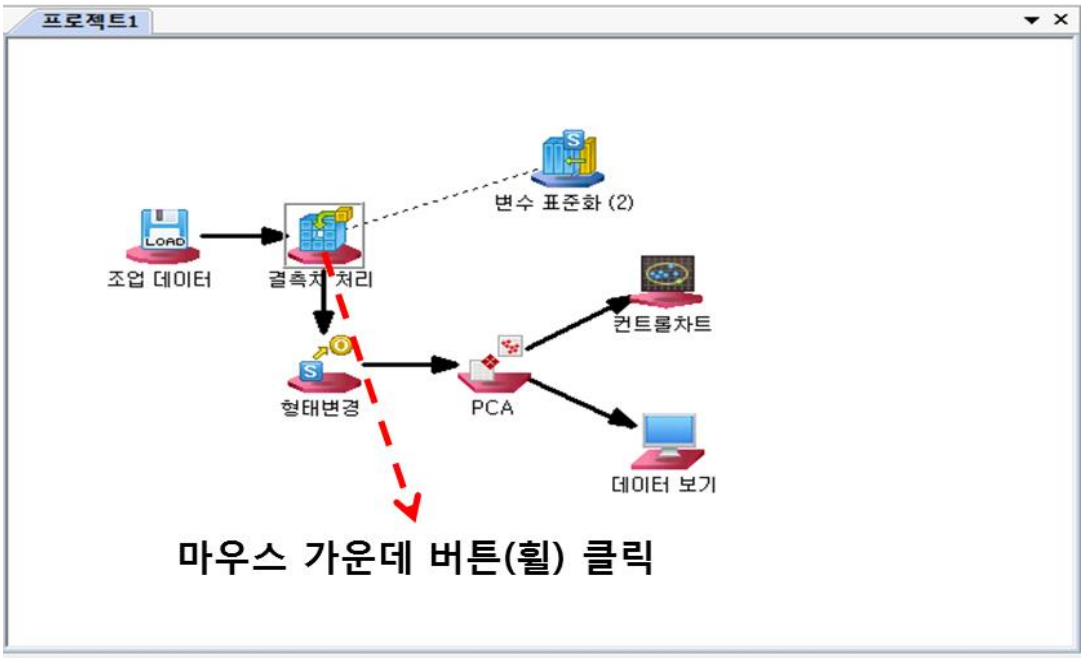
- 노드를 연결하려면 출발점에 해당하는 노드를 선택한 후 마우스 오른쪽 버튼을 누릅니다. 나타나는 컨텍스트 메뉴 중 연결을 선택합니다.



메뉴를 선택하면 프로젝트창이 다음 그림과 같이 변경되며



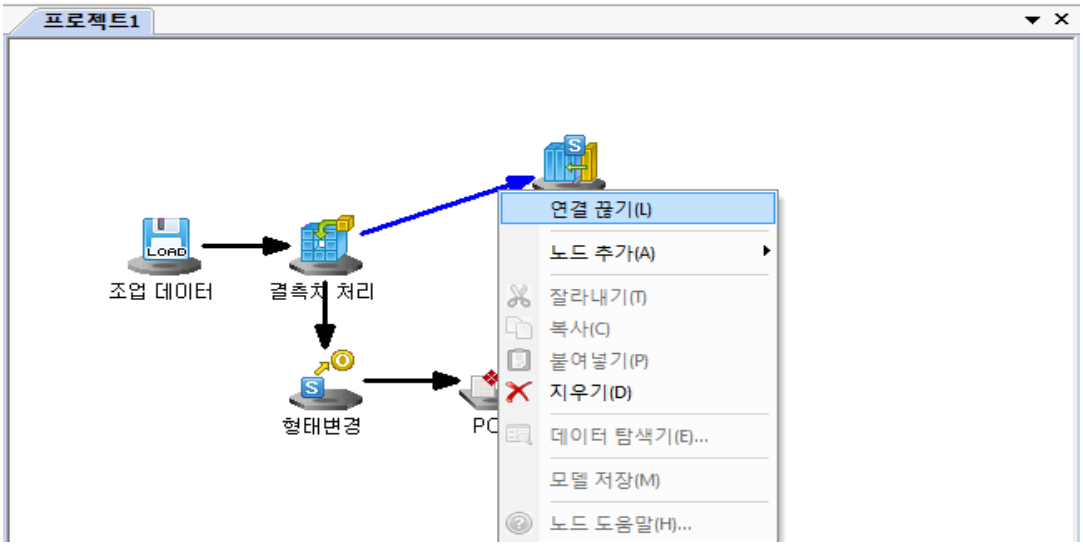
연결하고자 하는 노드로 마우스 포인트를 이동한 후 클릭합니다.
보다 편한 방법으로 출발점에 해당하는 노드를 선택한 후 마우스의 가운데 버튼(휠)을 누릅니다.



가운데 버튼을 누른 상태에서 연결하고자 하는 노드로 마우스 포인트를 이동한 후 버튼을 놓습니다.

노드 연결 끊기

- 노드 연결을 끊으려면 끊고자 하는 연결을 클릭하여 선택한 뒤 마우스 오른쪽 버튼을 누릅니다. 다음과 같은 메뉴가 나타나며 나타난 메뉴 중 연결 끊기를 선택합니다.



노드의 연결 가능 여부

- 연결은 다음 그림과 같은 요소를 갖습니다.
  노드는 기본적으로 여러 출발점이 될 수 있지만 종착점으로는 하나만 가능합니다(병합, 추가, 열조합 등은 여러 종착점이 가능). 또한, 출력 노드나 차트 노드 등은 결과물을 산출하는 종단의 성격이 있기 때문에 출발점이 될 수 없습니다. 스트림 구성에 있어 이와 같은 규칙이 존재하며 이를 고려하여 스트림을 구성하여야 합니다. 하지만 사용자가 신경 쓸 필요는 없습니다. 왜냐하면 ECMiner™에서는 노드의 연결 가능 여부를 시각적으로 표현하기 때문에 이를 참조하여 스트림을 구성할 수 있습니다.

	입력 노드	전처리 노드	차트 노드	모델링 노드	출력 노드	등록된 모델
기본						

	입력 노드	전처리 노드	차트 노드	모델링 노드	출력 노드	등록된 모델
연결 가능						
연결 불가능						

- 위 테이블은 노드의 연결 가능 여부를 나타내기 위하여 변화되는 노드의 모습입니다.

2.3 스트림 구성 규칙

스트림은 다음과 같은 규칙을 준수하여야 합니다.

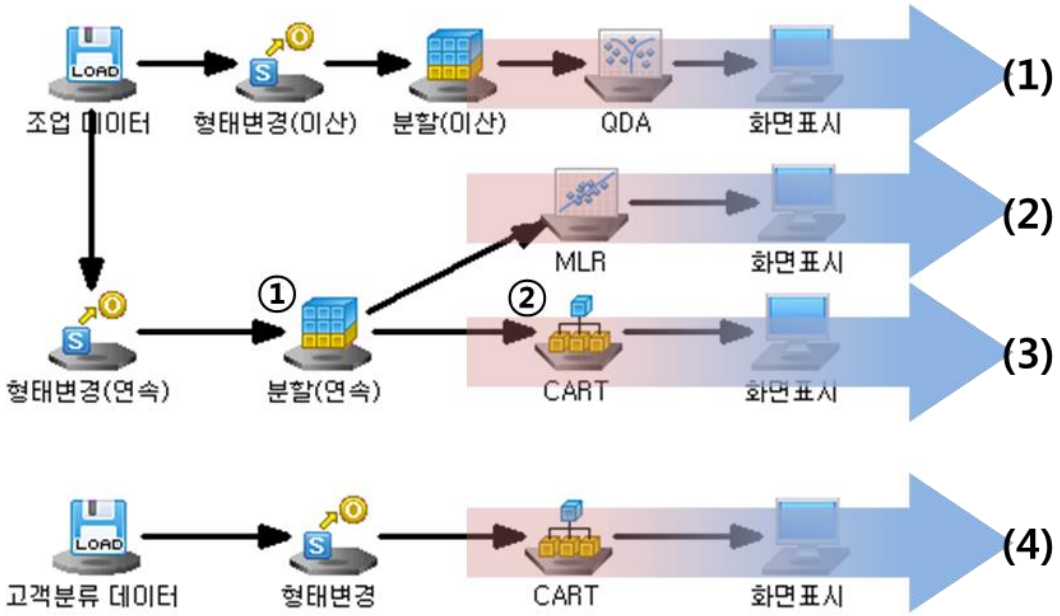
- 기본적인 스트림의 형식은 입력, 전처리, 모델링, 출력입니다. 따라서 입력 노드, 전처리 노드, 모델링 노드, 출력 노드 탭에서 필요한 노드들을 가져온 후, 이들을 스트림으로 연결하면 기본적인 스트림이 구성됩니다. 이런 기본 스트림에서 노드를 추가하거나 삭제하면서 스트림에 변형을 주어 다양한 스트림을 구성할 수 있습니다.
- 기본적인 스트림 구조를 굳이 따를 필요는 없습니다. 각 노드는 그의 기능을 완벽하게 실행 한 후 다음 단계로 데이터만 넘겨주기 때문에 데이터 마이닝의 모든 단위 요소가 있을 필요는 없습니다. 사용자가 단순히 파일에서 데이터를 가져와 약간의 변형을 한 후 파일에 저장하고 싶다면, "파일 입력 -> 파생 변수 -> 파일 출력"과 같이 스트림을 구성하여 실행할 수 있습니다.
- 데이터 입력 노드를 출발점으로 스트림은 실행이 됩니다. 데이터 마이닝 시 제일 중요한 부분이 바로 데이터입니다. 따라서 출발점이 데 Gabage in 이터 입력 노드가 아닌 혹은 없는 스트림은 실행되지 않습니다.
- 스트림에 연결되지 않고 독립적으로 존재하는 노드들은 실행이 되지 않습니다.
- 프로젝트창에 여러 개의 스트림이 있을 경우, 스트림의 실행 순서는 ECMiner™가 나름대로 결정합니다. 순차적으로 실행해야 할 필요가 있는 스트림들이 있을 경우 먼저 실행하여야 하는 것부터 순차적으로 부분 실행하여야 합니다.
- 노드의 속성을 잘못 지정하거나 입력하지 않은 경우 실행되지 않습니다. 이런 경우, 메시지창에 나타나는 에러 메시지를 참조하여 속성을 수정한 후 실행할 수 있습니다.
- 스트림에서 노드는 일대다 연결이 가능합니다. 즉, 데이터 소스를 받을 노드는 하나이어야 하지만 데이터를 보내줄 노드는 여러 개 일 수 있습니다. 그러나 추가, 병합, 열조합 등 여러 데이터를 필요로 하는 경우 여러 개의 데이터 소스를 가질 수

있기 때문에 다대일 연결이 가능합니다. 결과를 출력하는 노드(출력 노드, 차트 노드, 모델 평가 노드)는 다른 노드의 데이터 소스로 연결될 수 없습니다.

- 스트림을 순환 루프가 되도록 구성할 수 없으며 ECMiner™에서 자동으로 연결이 불가능 하도록 처리합니다.
- 스트림 구성을 위하여 노드 연결 / 연결 끊기 등의 기능을 수행할 때 연결 가능한 노드 및 불가능한 노드에 대한 정보가 프로젝트창에 나타납니다. 이를 참조하면 보다 쉽게 스트림을 구성할 수 있습니다.

2.4 스트림 실행하기

스트림을 실행하려면 메인 메뉴 중 "스트림 > 실행" 혹은 "스트림 > 부분 실행"을 선택합니다. 다음과 같이 프로젝트를 구성하였다고 가정합니다.



위 프로젝트에는 (1)부터 (4)까지 총 4 개의 스트림이 있습니다.

- 메인 메뉴 중 "스트림 > 실행"을 누르거나 F5 키를 누르면 (1)부터 (4)까지의 스트림이 모두 실행되어 결과를 산출합니다. 이 때 어떤 스트림이 먼저 실행되는 지는 ECMiner™가 나름대로 결정합니다.
- 노드 ①을 선택한 후 "부분 실행"을 선택하였다면, 이 노드를 포함하고 있는 스트림 (2)와 (3)만 실행됩니다.

- ②의 노드를 선택한 후 "부분 실행"을 선택하였다면, 이 노드를 포함하고 있는 스트림 (3)만 실행됩니다.

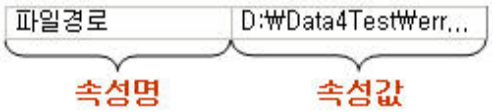
제 3 장 노드 설명

ECMiner™에서 노드란 데이터 마이닝을 위한 단위 프로세스입니다. 각 노드는 데이터 가져오기, 변경하기, 모델링 등 특정 기능을 수행하며, 이러한 노드들이 모여 스트림을 구성하고 데이터 마이닝을 위한 최적의 프로세스를 산출할 수 있습니다.
ECMiner™에서는 크게 입력 노드, 전처리 노드, 차트 노드, 모델링 노드, 출력 노드, 모델 평가 노드, 생성된 모델 노드 등 7 종류의 노드를 지원합니다.

노드 분류	설명	지원하는 노드
입력	데이터 마이닝의 출발점인 데이터를 가져오기 위한 노드입니다.	ODBC 입력, OLEDB 입력, 액세스 데이터, 엑셀 데이터, 오라클 입력, 오라클 입/출력 파일 입력, 파일입력 2, 파일 입/출력, 추가 입력, Copy 입력
전처리	입력 받은 데이터를 정제하고 적절한 형태로 변형하기 위한 노드입니다.	결측치 처리, 그룹화, 다중 파생변수, 변수순서, 병합, 변수표준화, 분할, 선택, 열조합, 정렬, 채우기, 추가, 피봇, 파생변수, 표본추출, 필터, 형태변경, COUNTER, RANKING, 그룹통계량, 선택 2, 구간화
차트	데이터를 시각적으로 분석하기 위한 차트 기능을 제공하는 노드입니다.	3 차원차트, 바차트, 컨튜어차트, 컨트롤차트, 히스토그램, 매트릭스차트, 파레토차트, 파이차트, 기본차트, 통계차트, 관리도, 기본 확장 차트, 산포 차트, 다변량 관리 차트
모델링	데이터 마이닝 프로세스의 엔진 역할을 하는 알고리즘을 수행하는 노드입니다. **ECMiner™**에서는 총 24 가지 알고리즘을 지원합니다.	연관성 분석, CART, HIERARCHICAL, KMEANS, KNN, LDA, LOGISTIC, MANUAL CART, MLP, MLR, PCA, PCR, PLS, QDA, RBF, 순차연관성, Score card, SOM, RBF-DDA, Factor Analysis, SVM, SVR, LOF, One class SVM
출력	데이터를 DB 혹은 파일에 저장하거나 화면으로 출력해 볼 수 있는 노드입니다.	ODBC 출력, 파일 출력, 통계 분석, OLEDB 출력, 화면표시, 피봇, 원인/결과 연관, 분리저장, 오라클 출력
모델 평가	생성된 모델의 예측정확도를 비교/ 평가하는 노드입니다.	모델평가, 이익도표, ROC 차트

생성된 모델 노드	생성된 모델은 모델링 노드의 결과물로 생성되는 노드입니다.	생성된 모델 노드를 이용하여 새로운 데이터에 대한 데이터 마이닝적인 관점에서의 새로운 결과를 산출할 수 있습니다.

ECMiner™의 노드는 모두 해당 작업을 수행하기 위한 속성을 갖습니다. 속성은 기본적으로 속성명과 속성값으로 구성됩니다. 예를 들어 파일입력 노드라면 어떤 파일을 읽어야 할 지에 대한 속성이 필요하고 이는 속성명(파일경로) 그리고 속성값으로 대변됩니다.



공통속성

ECMiner™의 노드는 ECMiner™에서 사용될 이름과 간단한 설명을 공통적으로 갖습니다.

속성그룹	속성명	설명
일반정보	이름	프로젝트창에 표시될 노드의 이름이며 이 이름을 기본으로 메시지창 등에 노드 관련 메시지가 나타납니다. 노드에 기능에 따라 사용자가 알아 보기 쉽도록 이름을 짓는 것이 중요합니다.
	설명	노드에 대한 간단한 설명을 입력하는 부분으로 선택 입력사항입니다. 설명이 입력된 경우 프로젝트창에서 노드 위에 마우스를 올리면 입력한 설명이 나타납니다.

노드 관련 기타 정보

노드에 공통적으로 사용되는 내용을 정리한 부분으로 다음의 내용을 포함하고 있습니다.

- **변수 형태 아이콘**
  변수가 목록 되는 컨트롤의 경우 직관성을 높이기 위하여 변수 형태를 아이콘으로 표시하였습니다. 이 아이콘에 대한 설명입니다.
- **변수 정보 편집**
  파일입력 노드, 엑셀 데이터 노드 등에서 변수정보를 편집하는 방법을 설명하고 있습니다.

- **변수 조건부 선택 대화상자**
  변수 목록 UI 에서 변수를 선택하여야 할 경우 보다 쉽게 변수를 선택하기 위한 기능으로 변수 선택 조건을 입력할 수 있는 대화상자입니다.
- **변수 선택 속성 컨트롤**
  노드 중 변수를 다중으로 선택하여야 하는 노드들이 있습니다. 이런 노드에 공통적으로 포함된 속성 컨트롤로 변수를 다중으로 선택할 수 있도록 구현되어 있습니다.

변수 형태 아이콘

변수가 목록 되는 컨트롤의 경우 직관성을 높이기 위하여 변수 형태를 아이콘화하여 표시하였습니다.

ICON	형태 종류	설명
D	데이터 형태	날짜형 데이터를 의미합니다.
S		문자형 데이터를 의미합니다.
I		정수형 데이터를 의미합니다.
R		실수형 데이터를 의미합니다.
	입/출력 형태	모델링 시 독립변수로 사용됨을 의미합니다.
		모델링 시 종속변수로 사용됨을 의미합니다.
	통계학적 형태	이산형 데이터를 의미합니다.
		연속형 데이터를 의미합니다.

변수 정보 편집

변수정보는 다음과 같은 사용자 인터페이스를 갖습니다.

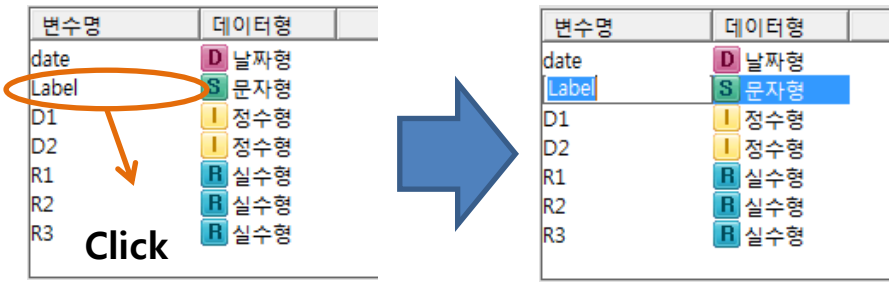
변수 목록이 리스트 컨트롤에 목록화 되어 나타나며 변수명과 변수의 데이터형을 컬럼으로 갖습니다.
변수정보는 거의 모든 노드에 속성으로 존재하며 값을 변경할 수 있는 것과 변경할 수 없는 것 두 가지 형태가 있습니다. 입력노드 중 일부 노드(파일입력, 엑셀 데이터)만 값을 변경할 수 있는 변수정보 UI 를 사용합니다.

변수정보 변경

- 변수정보 UI 에서는 변수에 대한 두 가지 속성을 변경할 수 있습니다. 보이는 그대로 변수명과 데이터형이 그 두 가지입니다.

  - 변수명 변경

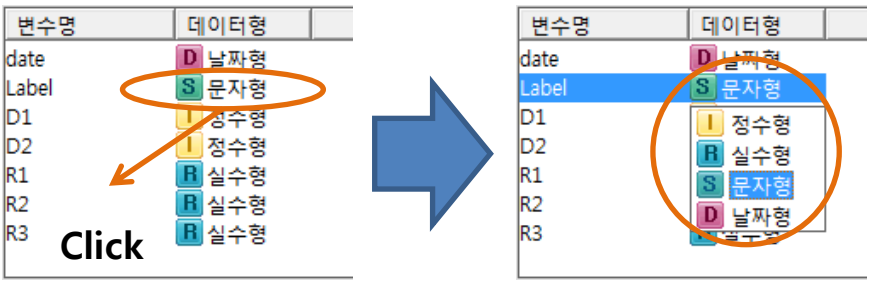
    변수명을 변경하려면 변경하고자 하는 변수의 변수명을 클릭합니다. 위 그림에서 LABEL 이라는 변수명을 변경하고자 한다면 다음 그림의 왼쪽과 같이 LABEL 을 클릭합니다. 그러면 오른쪽 그림과 같이 UI 가 변경되며, 바꾸고자 하는 이름을 입력합니다.



  - 데이터형 변경

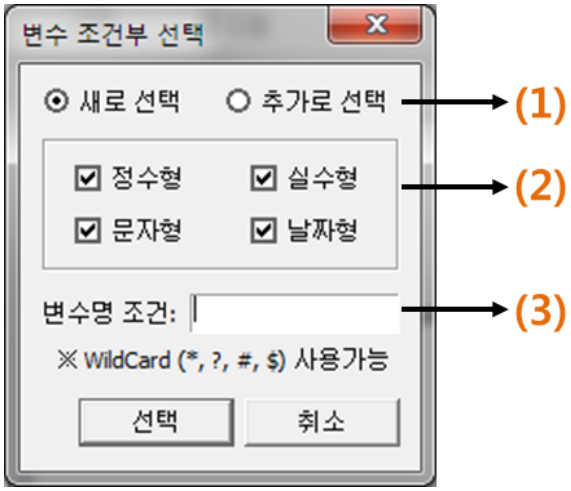
    데이터형을 변경하려면 변경하고자 하는 변수의 데이터형 컬럼을 클릭합니다. 위 그림에서 LABEL 의 데이터형을 변경하고자 한다면 다음 그림의 왼쪽과 같이 LABEL 변수에 해당하는 데이터형 컬럼을 클릭합니다. 그러면 오른쪽 그림과 같이 선택할 수 있는 항목이 나타나며 이 항목 중 바꾸고자 하는

데이터형을 선택합니다.



변수 조건부 선택 대화상자

변수 조건부 선택 대화상자를 이용하여 선택하고자 하는 변수의 조건을 입력함으로써 변수 목록이 많을 경우보다 편리하게 변수를 선택할 수 있습니다.



(1)은 변수가 기 선택된 변수를 유지할 것인지 여부를 지정합니다. "새로 선택"을 클릭하면 기존의 선택을 무시하고 조건에 맞는 변수만 새로 선택합니다. "추가로 선택"은 기존에 선택된 변수는 유지하면서 조건에 맞는 변수를 추가로 선택하는 것입니다.
(2)에서는 선택할 변수의 데이터형 조건을 지정합니다.
(3)에서는 선택할 변수의 명칭 조건을 지정하며, Wild card 를 사용할 수 있습니다.

Wild Card

- 변수명 조건입력 시 사용할 수 있습니다. 4 가지 Wild card 를 지원하며 의미는 다음과 같습니다.
  - *

임의의 문자열을 의미합니다. "A*"와 같이 하면 A 로 시작되는 모든 변수명, "*B"와 같이 하면 B 로 끝나는 모든 변수명을 의미합니다.

- ?

  임의의 문자 한자를 의미합니다. "A??B"와 같이 하면 A 로 시작되고 B 로 끝나는 4 자리수의 변수명을 의미합니다.

- #

  임의의 숫자 한자를 의미합니다. "A##B"와 같이 하면 A 로 시작되고 B 로 끝나며 가운데 2 자리는 숫자(0 ~ 9)를 갖는 변수명을 의미합니다.

- $

  숫자를 제외한 임의의 문자 한자를 의미합니다. "A$$B"와 같이 하면 A 로 시작되고 B 로 끝나며 가운데 2 자리는 숫자가 아닌 문자를 갖는 변수명을 의미합니다.

NOTE: 다음을 유의 하시기 바랍니다.

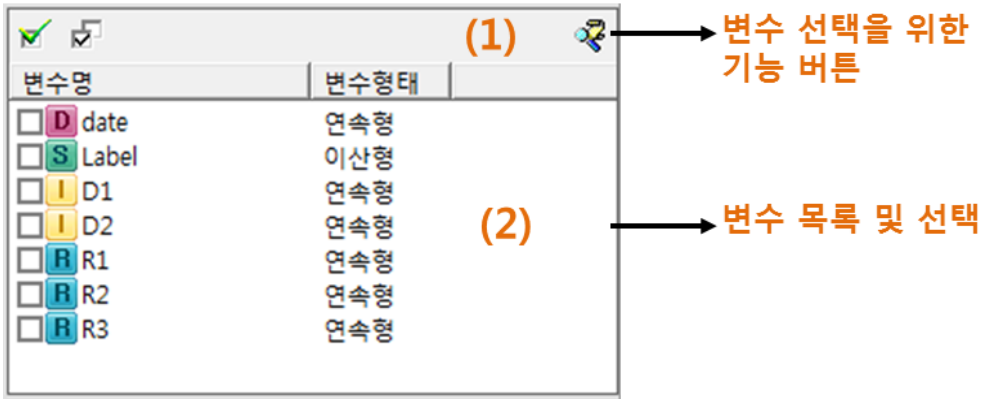
- 변수명 조건에 "*"을 입력하면 모든 변수명을 의미합니다.
- ABC 를 변수명에 포함하는 변수를 선택하고자 할 때 "*ABC*"와 같이 입력하여야 합니다. "*"를 넣지 않고 "ABC"와 같이 입력하면 변수명이 ABC 인 변수를 의미합니다.
- 공백도 하나의 문자로 취급됩니다. "* ABC *"와 같이 하면 "ABC"가 아니라 " ABC "를 포함하는 변수명을 찾습니다.

변수 선택 속성 컨트롤

일부 노드에서 변수를 다중으로 선택하여야 할 경우 포함된 속성 컨트롤입니다. 말 그대로 변수를 다중으로 지정할 때 유용한 기능들을 제공하며 노드의 속성으로 사용됩니다.

UI 설명

- 변수 선택 속성 컨트롤은 다음과 같은 UI 를 갖습니다.



- (1)은 변수 선택을 용이하게 하기 위한 기능 버튼입니다.

ICON	기능 설명
	(2)의 목록 중 현재 선택되어 있는 변수를 노드가 각 노드의 형태에 맞게 사용하도록 지정(체크)합니다.
	(2)의 목록 중 변수의 지정(체크)된 상태를 반전합니다.
	변수 조건부 선택 대화상자를 나타냅니다. 이 대화상자에서 지정한 조건에 맞는 변수만 (2)의 목록에서 선택됩니다. 이렇게 변수를 선택한 후  버튼을 눌러 변수를 노드가 사용하도록 지정할 수 있습니다. 자세한 내용은 변수 조건부 선택 대화상자을 참조하시기 바랍니다.

- (2)은 선택할 수 있는 변수 목록이 나타나고 실제 변수를 선택하거나 선택된 변수를 확인할 수 있는 부분입니다.
  앞쪽의 체크 박스를 체크하여 노드가 변수를 사용하도록 지정할 수 있으며 체크된 것을 다시 누르면 지정 해제됩니다.

변수 목록

- (2)에 나열되는 변수는 노드에서 사용할 수 있는 모든 변수가 아니라 노드의 기능에 따라 지정될 수 있는 변수들만 나타납니다.
  예를 들어 "매트릭스 차트" 노드의 경우 숫자형/연속형인 변수만 취급하므로 해당 형태의 변수만 나타나게 됩니다.

이번 장에서는 위에서 언급한 노드의 속성 및 사용법에 대하여 알아 보겠습니다.

- **입력 노드**
- **전처리 노드**
- **차트 노드**
- **모델링 노드**
- **모델 노드**
- **출력 노드**
- **모델 평가 노드**

3.1 입력 노드

입력 노드는 여러 데이터 소스를 ECMiner™에서 사용할 수 있는 형태의 데이터 포맷으로 변형하여 데이터 마이닝을 위한 초기 데이터를 구축하는 기능을 수행합니다. 데이터 소스의 종류에 따라 노드가 구분되어 있으며 이들은 다음과 같습니다.

- ODBC 입력 노드

  데이터베이스의 데이터를 가져와 ECMiner™에서 사용할 수 있는 데이터를 생성합니다. 데이터베이스 핸들링을 위하여 ODBC 기술을 이용합니다. 따라서 ODBC 를 지원하는 데이터베이스에만 사용할 수 있습니다.

- OLEDB 입력 노드

  데이터베이스의 데이터를 가져와 ECMiner™에서 사용할 수 있는 데이터를 생성합니다. 데이터베이스 핸들링을 위하여 OLE DB 기술을 이용합니다. 따라서 OLE DB 를 지원하는 데이터베이스에만 사용할 수 있습니다.

- 액세스 데이터 노드

Microsoft(R) Access 데이터베이스의 데이터를 ECMiner™에서 사용할 수 있는 데이터로 변형하고자 할 때 사용합니다.

- 엑셀 데이터 노드

  Microsoft(R) Excel 파일의 데이터를 ECMiner™에서 사용할 수 있는 데이터로 변형하고자 할 때 사용합니다.

- 오라클 입력 노드

  데이터베이스의 데이터를 가져와 ECMiner™에서 사용할 수 있는 데이터를 생성합니다. 데이터베이스 핸들링을 위하여 오라클 DB 기술을 이용합니다.

- 오라클 입/출력 노드

  데이터베이스의 데이터를 가져와 ECMiner™에서 사용할 수 있는 데이터를 생성합니다. 데이터베이스 핸들링을 위하여 오라클 DB 기술을 이용합니다. 같은 형태의 여러 파일을 모아서 세로로 붙여 분석용 데이터를 생성하고 나중에 다시 읽지 않고 재사용할 수 있도록 ECL 형태로 저장하는 기능을 수행합니다.

- 파일입력 노드

  텍스트 형태의 파일을 데이터 소스로 하여 ECMiner™에서 사용할 수 있는 데이터를 생성합니다.

- 파일입력 2 노드

  변수정보와 데이터가 분리된 파일을 읽습니다.

- 파일 입/출력 노드

  같은 형태의 여러 파일을 모아서 세로로 붙여 분석용 데이터를 생성하고 나중에 다시 읽지 않고 재사용할 수 있도록 ECL 형태로 저장하는 기능을 수행합니다.

- 추가 입력 노드

  같은 형태의 여러 ecl 파일을 모아서 세로로 붙여 새로운 ecl 파일을 만드는 기능을 하는 노드입니다.

- Copy 입력 노드

  정해진 데이터 파일, DB 의 형태가 아닌 사용자가 Web 또는 원하는 데이터를 Copy&Paste 를 통해서 데이터 입력을 할 수 있는 노드입니다.

효과적인 데이터 마이닝을 수행하려면 데이터의 유효성이 중요합니다. 적절치 못한 데이터를 사용하면 적절치 못한 결과 밖에 산출되지 않습니다(Garbage in, garbage out). 따라서 입력되는 데이터의 유효성을 판단하는 것이 매우 중요합니다. ECMiner™에서는 이를 위하여 데이터 탐색기를 제공합니다. 입력 노드를 더블 클릭함으로써 데이터 탐색기를 띄울 수 있으며, 데이터 탐색기를 이용하여 데이터 도식화, 통계 정보 등 데이터를 탐색해 볼 수 있습니다.

입력 노드는 분석할 대상인 초기 데이터를 생성하는 것으로 필요 이상으로 많은 데이터를 생성하면 뒤이어 연결된 모든 노드에 영향을 줍니다. 즉, 각 노드의 성능 저하의 원인이 될 수 있습니다. 만약 입력 노드에서 분석에 영향을 주지 않는 범위로 데이터의 양을 줄일 수 있다면 보다 효율적인 데이터 마이닝을 수행할 수 있을 것입니다. 이를 위하여 ECMiner™에서는 초기 데이터 샘플링 기능을 제공하며, 이를 위하여 모든 입력 노드는 다음과 같은 공통 속성을 갖습니다.

입력 노드의 공통 속성

속성그룹	속성명	설명
부분읽기	부분읽기 방법	부분읽기 방법을 지정합니다. "전체 읽기", "처음부터", "임의 추출" 등 세가지 선택사항이 있습니다. • 전체 읽기 지정된 데이터 소스의 모든 데이터를 읽어 초기 데이터를 구축합니다. 이 경우 "부분읽기 인수" 속성은 무시됩니다. • 처음부터 처음부터 "부분읽기 인수" 속성에서 지정된 만큼의 데이터만 지정된 데이터 소스에서 읽어 초기 데이터를 구축합니다. • 임의 추출

		지정된 데이터 소스에서 임의의 레코드를 선택하여 "부분읽기 인수" 속성에서 지정된 만큼의 초기 데이터를 생성합니다.
	부분읽기 인수	부분읽기 방법에서 지정된 방법으로 얼마만큼의 데이터를 읽을 것인지를 지정합니다. 부분읽기 방법이 "전체 읽기"인 경우 이 속성은 무시됩니다.

NOTE: 부분읽기 방법을 "전체 읽기" 이외의 값으로 지정한 경우 부분읽기 인수 속성을 사용하며 의미는 다음과 같습니다.

- 부분읽기 방법이 "처음부터"인 경우

  "부분읽기 인수" 속성의 값은 읽을 레코드 개수를 의미합니다. 만약 지정된 값이 실제 데이터의 레코드 수보다 클 경우는 "전체 읽기"와 같은 동작을 합니다.

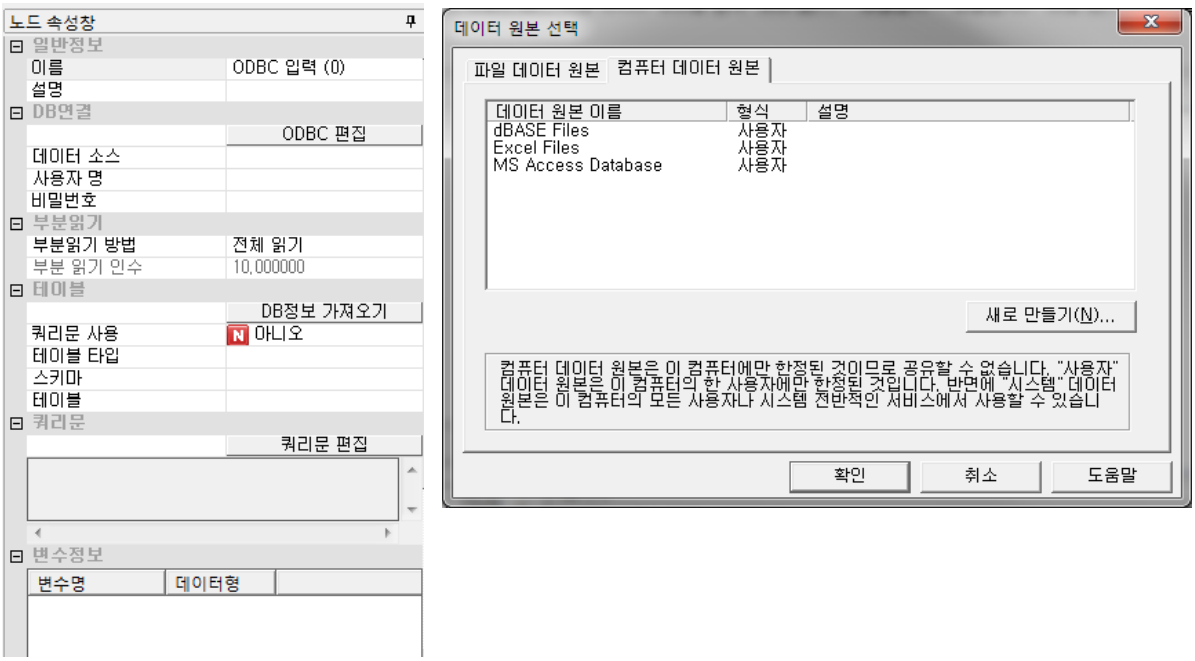
- 부분읽기 방법이 "임의 추출"인 경우

  "부분읽기 인수" 속성의 값은 백분율을 의미합니다. 만약 지정된 값이 100 이상이라면 100%로 간주하여 "전체 읽기"와 같은 동작을 합니다. 또 한가지 주의할 사항은 전체가 100 건인 데이터에 대하여 10%라고 지정했다고 해서 10 건의 데이터만 생성되는 것이 아닙니다. 지정한 백분율의 의미는 10% 정도의 확률로 한 레코드를 선택하라는 의미이기 때문입니다. 따라서 앞에 언급한 예의 경우 10 건 전후로 데이터가 생성되게 됩니다.

3.1.1 ODBC 입력 노드

ODBC 입력 노드는 ODBC(Open Database Connectivity) 기술을 이용하여 데이터베이스에서 데이터를 가져오는 노드입니다. 테이블명 혹은 View 명을 직접 선택하여 해당하는 데이터를 가져오거나 쿼리문을 입력하여 데이터를 가져올 수 있습니다.

사용법



- 데이터베이스에 접속하기 위하여 데이터 소스, 사용자명, 비밀번호를 입력합니다.
- 연결하고자 하는 데이터베이스가 없다면 **ODBC 편집** 버튼을 눌러 새로운 DSN 을 추가합니다.
- DB 정보 가져오기 버튼을 눌러 지정된 데이터베이스의 정보를 가져옵니다.
- 데이터를 가져올 테이블을 지정하거나 쿼리문을 입력합니다.
- 쿼리문을 사용하려면 **쿼리문 사용**을 `예'로 지정합니다.
- 입력된 속성이 제대로 되었다면 **변수** 부분에 테이블 혹은 쿼리문이 가지고 있는 변수들이 목록으로 나열됩니다.
- 만약 **변수정보**에 목록으로 나열되지 않는다면 입력된 속성을 다시 확인하거나 데이터베이스를 확인합니다.
- DB **부분 읽기** 방식을 지정합니다.

속성

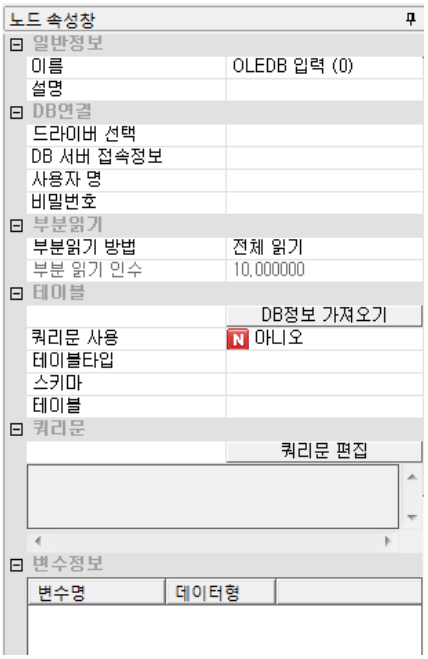
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
DB 연결	ODBC 편집	ODBC 연결을 위한 DSN(Data Source Name)을 새로 만들기 위하여 사용합니다.	버튼	

	데이터 소스	현재 PC 내에 등록되어 있는 ODBC DSN 의 목록을 나타내며 이 중 사용하고자 하는 DSN 을 선택하면 됩니다. 만약 존재하지 않는다면 **[ODBC 편집]**을 이용하여 추가할 수 있습니다.	필수	
	사용자명	데이터베이스에 접속하기 위한 **사용자명**을 입력합니다.	필수	
	비밀번호	데이터베이스에 접속하기 위한 사용자의 **비밀번호**를 입력합니다.	필수	
부분 읽기	부분 읽기 방법	DB 를 부분적으로 읽는 방법을 지정합니다.	필수	전체 읽기, 처음부터, 임의 추출
	부분 읽기 인수	부분 읽기 방법이 '처음부터'라면 인수가 관측치 수가 되고 '임의 추출'이라면 관측치 퍼센티지가 됩니다.		
테이블	DB 정보 가져오기	입력된 데이터베이스 접속정보를 이용하여 데이터베이스의 스키마, 테이블, 뷰 등의 정보를 읽어옵니다. 가져온 정보는 아래쪽에 목록으로 표시됩니다.	버튼	
	쿼리문 사용	쿼리문을 사용하여 데이터를 가져올지 테이블에서 직접 가져올지 여부를 지정합니다.		예, 아니오
	테이블 타입	테이블 속성에 목록으로 나열될 데이터베이스 스키마를 선택합니다. 테이블과 뷰를 지원합니다.		All, Table, View
	스키마	테이블스페이스 혹은 소유자 목록이 표시됩니다.		
	테이블	데이터를 가져올 테이블을 지정합니다.		
쿼리문	쿼리문 편집	**쿼리문 사용** 속성이 **'예'**일 경우 활성화 됩니다. 쿼리문을 편집할 수 있는 대화상자가 나타 나며 이를 이용하여 쿼리문을 편집 하거나 불러오기/저장하기를 할 수 있습니다.	버튼	기능이 강화된 편집 툴로 확장 예정임.
	쿼리문 보기	현재 입력된 쿼리문을 확인할 수 있습니다. 편집하려면 "쿼리문 편집" 버튼을 누릅니다.		
변수정보	변수정보	읽을 데이터의 변수명 및 형태를 나타냅니다.		

3.1.2 OLEDB 입력 노드

OLEDB 입력 노드는 OLEDB 기술을 이용하여 데이터베이스에서 데이터를 가져오는 노드입니다. 테이블명 혹은 View 명을 직접 선택하여 해당하는 데이터를 가져오거나 쿼리문을 입력하여 데이터를 가져올 수 있습니다. ODBC 및 OLEDB 를 모두 지원하는 데이터베이스라면 OLEDB 를 사용할 것을 권장합니다.

사용법



- 접속하려는 데이터베이스에 맞는 **드라이버**를 지정합니다.
- 데이터베이스에 접속하기 위하여 **DB 서버 접속정보, 사용자명, 비밀번호**를 입력합니다.
- [DB 정보 가져오기] 버튼을 눌러 지정된 데이터베이스의 정보를 가져옵니다.
- 데이터를 가져올 테이블을 지정하거나 쿼리문을 입력합니다.
- 쿼리문을 사용하려면 **쿼리문 사용**을 **'예'**로 지정합니다.
- 입력된 속성이 제대로 되었다면 **변수정보** 부분에 테이블 혹은 쿼리문이 가지고 있는 변수들이 목록으로 나열됩니다.
- 만약 **변수정보**에 목록으로 나열되지 않는다면 입력된 속성을 다시 확인하거나 데이터베이스를 확인합니다.
- DB 부분 읽기 방식을 지정합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	

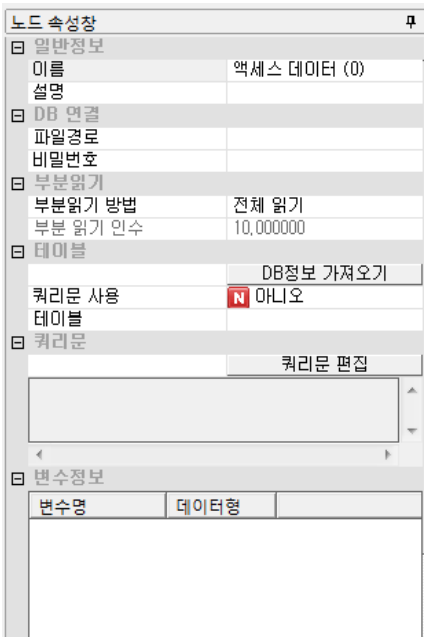
DB 연결	드라이버 선택	접속할 데이터베이스 종류에 맞는 **OLE DB 드라이버**를 선택합니다. ECMiner™은 선택된 드라이버를 통하여 데이터베이스에 접근합니다.	필수	
	DB 서버 접속정보	**데이터베이스를 인식할 수 있는 정보**를 입력합니다. SQL 서버인 경우 데이터베이스 서버의 IP 어드레스를 입력하며, Oracle 인 경우 TNS 명을 입력합니다. 기타 데이터베이스인 경우 데이터베이스를 인식할 수 있는 특정 정보를 입력합니다.	필수	
	사용자명	데이터베이스에 접속하기 위한 **사용자명**을 입력합니다.	필수	
	비밀번호	데이터베이스에 접속하기 위한 사용자의 **비밀번호**를 입력합니다.	필수	
부분 읽기	부분 읽기 방법	DB 를 부분적으로 읽는 방법을 지정합니다.	필수	전체 읽기, 처음부터, 임의 추출
	부분 읽기 인수	부분 읽기 방법이 '처음부터'라면 인수가 관측치 수가 되고 '임의 추출'이라면 관측치 퍼센티지가 됩니다.		
테이블	DB 정보 가져오기	입력된 데이터베이스 접속정보를 이용하여 데이터베이스의 스키마, 테이블, 뷰 등의 정보를 읽어옵니다. 가져온 정보는 아래쪽에 목록으로 나열됩니다.	버튼	
	쿼리문 사용	쿼리문을 사용하여 데이터를 가져올지 테이블에서 직접 가져올지 여부를 지정합니다.		예, 아니오
	테이블 타입	테이블 속성에 목록으로 나열될 데이터베이스 스키마를 선택합니다. 테이블과 뷰를 지원 합니다.		All, Table, View
	스키마	테이블스페이스 혹은 소유자 목록이 표시됩니다.		
	테이블	데이터를 가져올 테이블을 지정합니다.		
쿼리문	쿼리문 편집	**쿼리문 사용** 속성이 **'예'**일 경우 활성 됩니다. 쿼리문을 편집할 수 있는 대화상자가 나타나며 이를 이용하여 쿼리문을 편집하거나 불러오기 / 저장하기를 할 수 있습니다.	버튼	기능이 강화된 편집 툴로 확장 예정임.
	쿼리문	현재 입력된 쿼리문을 확인할 수 있습니다.		

	보기	편집하려면 "쿼리문 편집" 버튼을 누릅니다.		
변수정보	변수정보	읽을 데이터의 변수명 및 형태를 나타냅니다.		

3.1.3 액세스 데이터 노드

액세스 데이터 노드는 MDB 형태로 저장된 데이터를 가져오는 노드입니다. 테이블을 직접 선택하여 해당하는 데이터를 가져오거나 쿼리문을 입력하여 데이터를 가져옵니다.

사용법



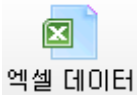
- MDB 파일 경로와 비밀번호를 입력합니다.
- [DB 정보 가져오기] 버튼을 눌러 테이블 정보를 가져옵니다.
- 데이터를 가져올 테이블을 지정하거나 쿼리문을 입력합니다. 쿼리문을 사용하려면 쿼리문 사용을 '예'로 지정합니다.
- 입력된 속성이 제대로 되었다면 변수정보 부분에 테이블 혹은 쿼리문이 가지고 있는 변수 들이 목록으로 나열됩니다. 만약 변수정보에 목록으로 나열되지 않는다면 입력된 속성을 다시 확인하거나 MDB 파일을 확인합니다.
- DB 부분 읽기 방식을 지정합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	

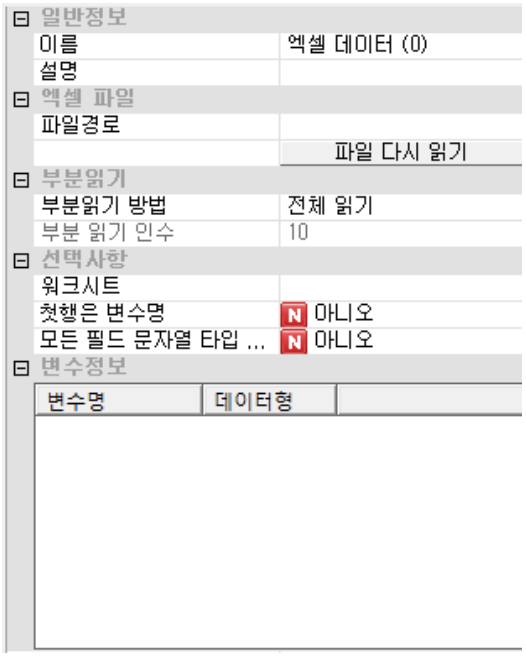
DB 연결	파일 경로	MDB 파일의 경로를 지정합니다. 대화창이 뜨면 선택합니다.	버튼	대화창 인터페이스.
	비밀번호	MDB 에 접속하기 위한 사용자의 비밀번호를 입력합니다.	필수	
부분읽기	부분 읽기 방법	DB 를 부분적으로 읽는 방법을 지정합니다.	필수	전체 읽기, 처음부터, 임의 추출
	부분 읽기 인수	부분 읽기 방법이 '처음부터'라면 인수가 관측치 수가 되고 '임의 추출'이라면 관측치 퍼센티지가 됩니다.		
테이블	DB 정보 가져오기	입력된 접속정보를 이용하여 데이터베이스의 테이블 정보를 읽어옵니다. 가져온 정보는 아래쪽에 목록으로 표시됩니다.	버튼	
	쿼리문 사용	쿼리문을 사용하여 데이터를 가져올지 테이블에서 직접 가져올지 여부를 지정합니다.		예, 아니오
	테이블	데이터를 가져올 테이블을 지정합니다.		
쿼리문	쿼리문 편집	**쿼리문 사용** 속성이 **'예'**일 경우 활성 됩니다. 쿼리문을 편집할 수 있는 대화상자가 나타나며 이를 이용하여 쿼리문을 편집하거나 불러오기 / 저장하기를 할 수 있습니다.	버튼	기능이 강화된 편집툴로 확장 예정임.
	쿼리문 보기	현재 입력된 쿼리문을 확인할 수 있습니다. 편집하려면 "쿼리문 편집" 버튼을 누릅니다.		
변수정보	변수정보	읽을 데이터의 변수명 및 형태를 나타냅니다.		

3.1.4 엑셀 데이터 노드



엑셀 데이터 노드는 excel 파일 형태로 저장된 데이터를 가져오는 노드입니다. 파일을 직접 선택하여 해당하는 데이터를 가져옵니다.

사용법



- 엑셀 파일 경로를 지정합니다.
- 워크시트를 선택합니다.
- 해당되는 경로의 데이터를 읽어옵니다.
- 변수정보를 확인합니다.
- 파일 부분 읽기 방식을 지정합니다.

속성

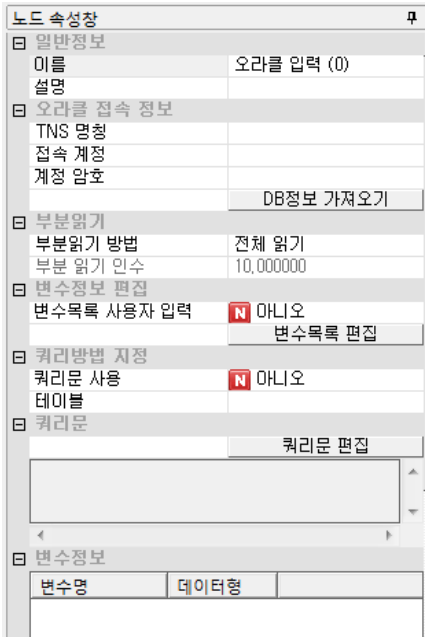
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
엑셀파일	파일경로	excel 파일의 경로를 지정합니다. 대화창이 뜨면 선택합니다.	버튼	대화창 인터페이스.
	파일 다시 읽기	파일이 제대로 읽혀지지 않은 경우 또는 파일에 수정이 있을 경우 [파일 다시 읽기]를 통해 다시 데이터를 가져옵니다.	버튼	
부분읽기	부분 읽기 방법	파일을 부분적으로 읽는 방법을 지정합니다.	필수	전체읽기, 처음부터, 임의 추출
	부분 읽기 인수	부분 읽기 방법이 '처음부터'라면 인수가 관측치 수가 되고 '임의 추출'이라면 관측치 퍼센티지가 됩니다.		
선택사항	워크시트	해당 데이터가 있는 워크 시트를 선택합니다.	필수	
	첫행은	파일의 첫 행에 변수명이 있다면 '예'로 설정	자동	예, 아니오

	변수명	합니다. **'아니오'**로 설정되면 변수명을 자동으로 생성합니다. 첫 행에 변수명이 있는데 **'아니오'**로 설정할 경우 변수의 형태가 모두 문자형으로 설정될 수 있습니다.	분석	
	모든 필드 문자열 타입 처리	입력되는 모든 데이터의 형태를 문자형으로 입력하도록 하는 옵션입니다.		예, 아니오
변수정보	변수정보	읽을 데이터의 변수명 및 형태를 나타냅니다.		

3.1.5 오라클 입력 노드

오라클 입력 노드는 오라클 데이터베이스로부터 데이터를 불러오는 노드입니다.

사용법



- 데이터베이스에 접속하기 위하여 TNS 명칭, 접속 계정, 계정 암호를 입력합니다
- DB 정보 가져오기 버튼을 눌러 지정된 데이터베이스의 정보를 가져옵니다.
- 데이터를 가져올 테이블을 지정하거나 쿼리문을 입력합니다.
- 쿼리문을 사용하려면 쿼리문 사용을 `예' 로 지정합니다.
- 입력된 속성이 제대로 되었다면 변수 부분에 테이블 혹은 쿼리문이 가지고 있는 변수들이 목록으로 나열됩니다.
- 만약 변수정보에 목록으로 나열되지 않는다면 입력된 속성을 다시 확인하거나 데이터베이스를 확인합니다.

속성

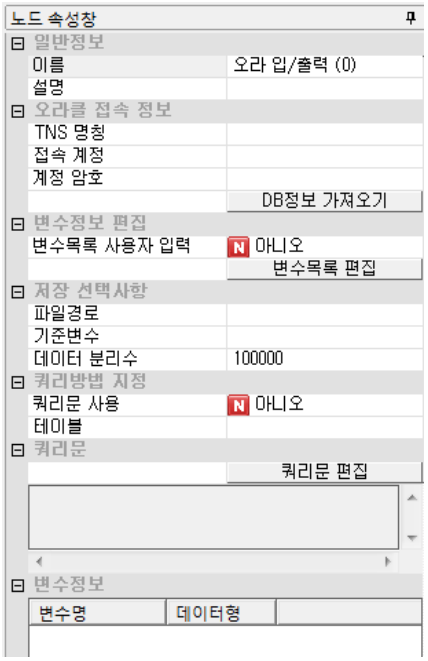
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
오라클 접속 정보	TNS 명칭	접속할 오라클 데이터베이스의 TNS 명을 입력합니다..	버튼	
	접속 계정	데이터베이스 접속 계정명을 입력합니다.	필수	
	계정 암호	데이터베이스 접속 계정의 암호를 입력합니다.	필수	
부분읽기	부분 읽기 방법	부분 읽기 방법을 지정합니다.		
	부분 읽기 인수	부분 읽기에 필요한 인수를 입력합니다. 부분 읽기 방법이 '처음부터'일 경우는 개수, '임의 추출'일 경우는 확률을 입력합니다.		
변수정보 편집	변수 목록 사용자 입력	변수목록을 사용자가 직접 입력하고자 할 경우 '예'를 선택합니다. **'아니오'** 일 경우 DB 에서 반환된 컬럼 정보를 사용합니다.		
	변수목록 편집	변수 목록 사용자 입력 속성이 **'예'** 일 경우 활성 됩니다. 변수목록을 편집할 수 있는 대화상자가 나타나며 이를 이용하여 변수목록을 편집할 수 있습니다.	버튼	
쿼리방법 지정	쿼리문 사용	쿼리문을 사용하여 데이터를 가져올지 테이블에서 직접 가져올지 여부를 지정합니다.		
	테이블	테이블을 선택합니다.		
쿼리문	쿼리문 편집	쿼리문 사용 속성이 **'예'**일 경우 활성 됩니다. 쿼리문을 편집할 수 있는 대화상자가 나타나며 이를 이용하여 쿼리문을 편집하거나 불러오기 / 저장하기를 할 수 있습니다.	버튼	
변수정보	변수정보	읽을 데이터의 변수명 및 형태를 나타냅니다		

3.1.6 오라클 입/출력 노드

오라클 입/출력

오라클 입/출력 노드는 같은 형태의 여러 파일을 모아서 세로로 붙여 분석용 데이터를 생성하고 나중에 다시 읽지 않고 재사용할 수 있도록 저장하는 기능을 수행합니다.

사용법



- 데이터베이스에 접속하기 위하여 TNS 명칭, 접속 계정, 계정암호를 입력합니다
- DB 정보 가져오기 버튼을 눌러 지정된 데이터베이스의 정보를 가져옵니다.
- 데이터를 가져올 테이블을 지정하거나 쿼리문을 입력합니다.
- 쿼리문을 사용하려면 쿼리문 사용을 '예' 로 지정합니다.
- 저장할 파일 경로 및 기준변수를 지정 합니다.
- 입력된 속성이 제대로 되었다면 변수 부분에 테이블 혹은 쿼리문이 가지고 있는 변수들이 목록으로 나열됩니다.
- 만약 변수정보에 목록으로 나열되지 않는다면 입력된 속성을 다시 확인하거나 데이터베이스를 확인합니다.

속성

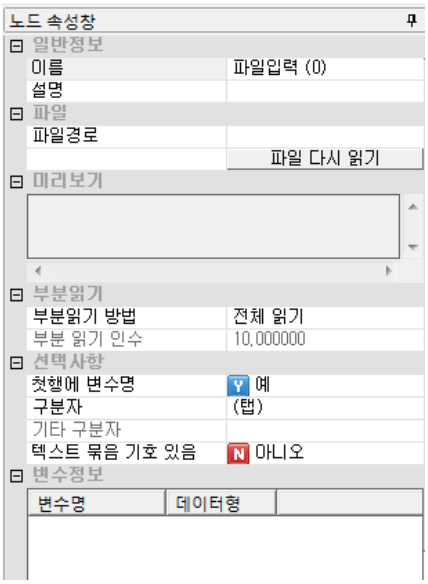
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
오라클 접속 정보	TNS 명칭	접속할 오라클 데이터베이스의 TNS 명을 입력합니다.	버튼	
	접속 계정	데이터베이스 접속 계정명을 입력합니다.	필수	

	계정 암호	데이터베이스 접속 계정의 암호를 입력합니다.	필수	
변수정보 편집	변수 목록 사용자 입력	변수목록을 사용자가 직접 입력하고자 할 경우 **'예'**를 선택합니다.**'아니오'** 일 경우 DB 에서 반환된 컬럼 정보를 사용합니다.	버튼	
	변수목록 편집	변수목록사용자 입력 속성이 '예'일 경우 활성 됩니다. 변수목록을 편집할 수 있는 대화상자가 나타나며 이를 이용하여 변수목록을 편집할 수 있습니다.	버튼	
저장 선택사항	파일경로	저장할 파일의 경로를 지정합니다.		
	기준변수	기준변수를 지정합니다.		
	데이터 분리수	데이터 분리수를 지정합니다.		
쿼리방법 지정	쿼리문 사용	쿼리문을 사용하여 데이터를 가져오도록 설정 합니다.	버튼	
	테이블	테이블을 선택합니다.		
쿼리문	쿼리문 편집	쿼리문 사용 속성이 예일 경우 활성 됩니다. 쿼리문을 편집할 수 있는 대화상자가 나타나며 이를 이용하여 쿼리문을 편집하거나 불러오기 / 저장하기를 할 수 있습니다.	버튼	
변수정보	변수정보	데이터의 변수명 및 형태를 나타냅니다		

3.1.7 파일 입력 노드

파일 입력 노드는 txt, csv, dat, tab, ecl 등의 확장자를 갖는 파일의 데이터를 가져오는 노드입니다. 자동으로 구분자 및 변수명, 변수형태 등을 탐색하는 기능을 가지고 있습니다. 만약 데이터가 잘못 읽혀졌을 경우 직접 읽기 옵션을 변경하여 데이터를 다시 읽을 수 있습니다.

사용법



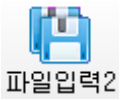
- 데이터를 가지고 있는 파일의 파일경로를 **파일경로** 속성에 지정합니다. 직접 입력할 수도 있으며 파일 대화상자를 통하여 지정할 수도 있습니다.
- 제대로 읽혀지지 않은 경우 **[파일 다시 읽기]** 버튼을 눌러 파일을 다시 분석 합니다.
- 제대로 분석이 되었다면 **선택사항** 항목들이 자동으로 분석되어 나타나며 **변수정보** 속성에 변수정보들이 목록으로 나열됩니다.
- 만약 제대로 분석되지 않았다면 **선택사항** 항목을 직접 지정합니다.
- **선택사항** 항목들이 잘못 입력되었을 경우 올바른 데이터마이닝을 수행할 수 없을 뿐 아니라 경우에 따라 **프로그램에 치명적인 영향**을 미칠 수도 있으므로 주의하시기 바랍니다.
- **변수정보** 목록 내용 중 잘못된 것이 있다면 수정합니다. 변수명, 변수형태 등을 수정할 수 있습니다. 변수형태는 정수, 실수, 문자, 날짜 등 총 4 가지입니다.
- 변수형태를 잘못 지정하면 원하는 결과를 얻을 수 없습니다.
- 파일 **부분 읽기** 방식을 지정합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
파일	파일경로	데이터 파일의 경로와 파일명을 입력합니다. 직접 입력할 수도 있으며, [...] 버튼을 누르면 대화상자를 통하여 파일경로를 지정할 수 있습니다.	필수	
	파일 다시 읽기	파일이 제대로 읽혀지지 않은 경우 또는 파일에 수정이 있을 경우 [파일 다시 읽기]를 통해 다시 데이터를 가져옵니다.	버튼	

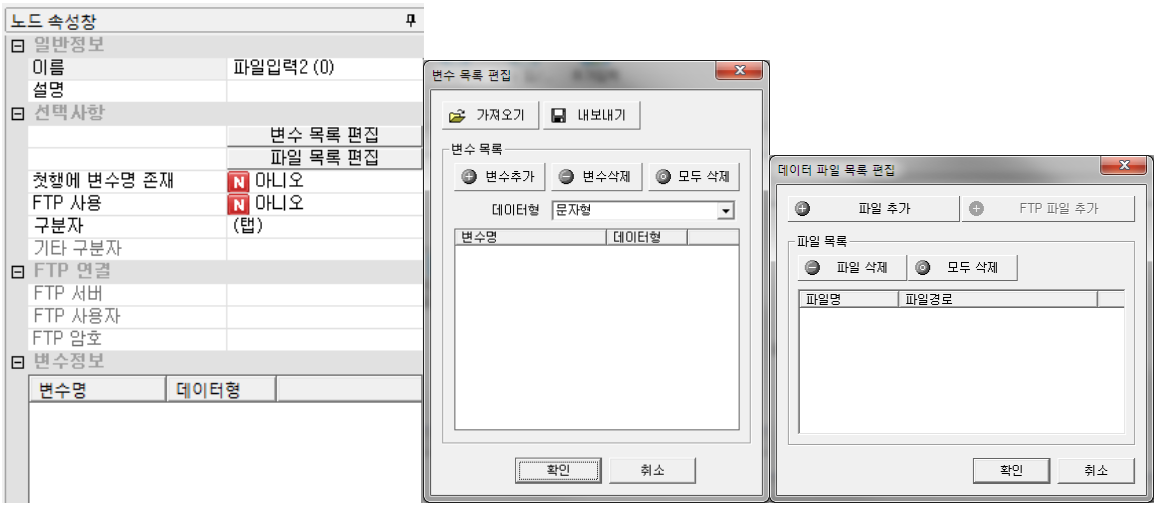
	미리보기	지정된 파일의 일부를 볼 수 있습니다. 이를 통하여 다른 파일 편집기를 열지 않고 파일의 구조를 파악할 수 있습니다. 자동으로 인식된 선택사항 중 잘못된 것이 있다면 미리보기의 내용을 참조하여 수동으로 수정할 수 있습니다.		
부분읽기	부분읽기 방법	파일을 부분적으로 읽는 방법을 지정합니다.	필수	전체읽기, 처음부터, 임의추출
	부분읽기 인수	부분 읽기 방법이 '처음부터'라면 인수가 관측치 수가 되고 '임의 추출'이라면 관측치 퍼센티지가 됩니다.		
선택사항	첫 행에 변수명	파일의 첫 행에 변수명이 있다면 **'예'**로 설정합니다. **'아니오'**로 설정되면 변수명을 자동으로 생성합니다. 첫 행에 변수명이 있는데 **'아니오'**로 설정할 경우 변수의 형태가 모두 문자형으로 설정될 수 있습니다.	자동 분석	예, 아니오
	구분자	컬럼을 구분하는 구분자를 입력합니다. 기본적으로 **(탭), (공백), |(바), ,(콤마), :(콜론), ;(세미콜론)**을 지원하며 이들 이외의 구분자를 사용할 경우 **(기타)**를 선택합니다.	필수	(탭) '|' (공백) ',' ';' ':' (기타)
	기타 구분자	**구분자** 속성이 **(기타)**일 경우 활성화되며 기타 구분자를 입력합니다.		
	텍스트 묶음 기호 있음	따옴표를 가지고 있는 문자형 데이터에서 따옴표를 제거하고 읽을 것인지 따옴표를 포함하여 데이터로 읽을 것인지 여부를 지정합니다.	예 일 경우 "ABC"➔ ABC	예, 아니오
변수정보	변수정보	파일을 분석하여 얻는 변수의 정보가 나타납니다. 총 변수의 개수, 변수명, 변수형태 등을 알 수 있으며 변수명, 변수형태의 변경이 가능합니다.		

3.1.8 파일 입력 2 노드



파일 입력 2 노드는 같은 형태의 여러 파일을 모아서 세로로 붙여 새로운 파일을 만드는 기능을 하는 노드입니다.

사용법



- **변수 목록 편집**을 선택하여 데이터 상의 변수 목록을 적어줍니다. (이 때 변수명은 데이터 상의 변수명과 일치할 필요는 없지만 데이터형은 일치해야 합니다.)
- **변수 목록 편집**의 **가져오기**를 통해서 변수 목록에 대한 정보가 저장되어 있는 파일을 불러올 수 있습니다.(혹은 수동으로 편집할 수도 있습니다.)
- **변수 목록 편집**의 **내보내기**를 통해서 변수 목록에 대한 정보를 저장할 수 있습니다.
- **파일 목록 편집**의 **파일 추가**를 통해서 세로로 붙일 파일들을 선택합니다. 이 때 **FTP 파일 추가**를 선택하면 FTP 에 서버에 있는 파일을 추가할 수도 있습니다.
- 선택사항을 통해 세부 사항을 선택하고 FTP 연결을 통해서 FTP 연결 시 FTP 서버, FTP 사용자, FTP 암호를 설정할 수 있습니다.

속성

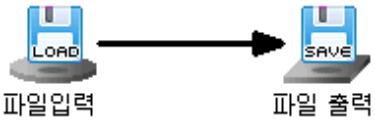
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
선택사항	변수	불러올 데이터의 변수가 어떠한지를 설정합니다.	필수	

	목록 편집	수동으로 설정할 수도 있고 변수 정보가 저장되어 있는 파일을 불러올 수도 있습니다. 변수 목록을 저장할 수 있는 기능 또한 제공합니다.		
	파일 목록 편집	불러올 데이터의 목록을 설정해 주도록 합니다.	필수	
	첫행에 변수명 존재	파일의 첫 행에 변수명이 있다면 **'예'**로 설정합니다. **'아니오'**로 설정되면 변수명을 자동으로 생성합니다. 첫 행에 변수명이 있는데 **'아니오'**로 설정할 경우 변수의 형태가 모두 문자형으로 설정될 수 있습니다.	필수	예 아니오
	FTP 사용	FTP 의 사용여부를 설정합니다.	필수	예 아니오
	구분자	컬럼을 구분하는 구분자를 입력합니다. 기본적으로 **(탭), (공백), |(바), ,(콤마), :(콜론)**, **;(세미콜론)**을 지원하며 이들 이외의 구분자를 사용할 경우 **(기타)**를 선택합니다.	필수	(탭) '|' (공백) ',' ';' ':' (기타)
	기타 구분자	**구분자** 속성이 **(기타)**일 경우 활성화되며 기타 구분자를 입력합니다.		
FTP 연결	FTP 서버	파일이 있는 FTP 서버가 무엇인지를 설정합니다.	FTP 사용시 필수	
	FTP 사용자	FTP 의 사용자 ID 를 입력합니다.	FTP 사용시 필수	
	FTP 암호	FTP 의 암호를 입력합니다	FTP 사용시 필수	
변수정보	변수정보	파일을 분석하여 얻는 변수의 정보가 나타납니다. 총 변수의 개수, 변수명, 변수형태 등을 알 수 있습니다.		

3.1.9 파일 입/출력 노드

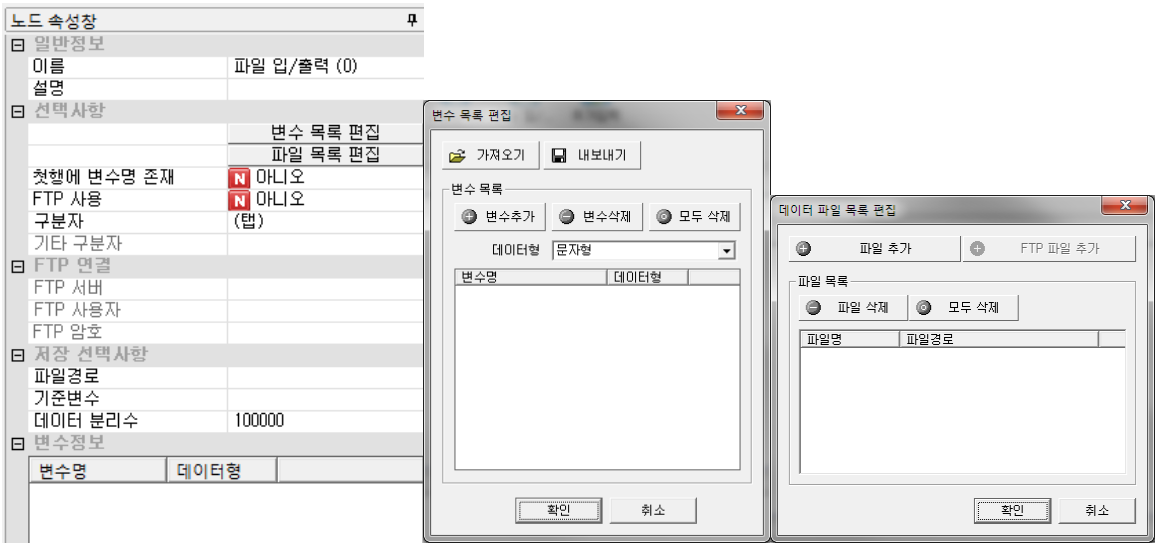
파일 입/출력 노드는 같은 형태의 여러 파일을 모아서 세로로 붙여 분석용 데이터를 생성하고 나중에 다시 읽지 않고 재사용할 수 있도록 ECL 형태로 저장하는 기능을 수행합니다.

분석을 위한 원천 데이터를 생성하고자 할 때 다음과 같이 단순한 스트림을 구성하여 실행하여야 할 경우가 있습니다.



파일 입/출력 노드는 위와 같은 스트림을 동시에 수행하여 보다 빠르게 원천데이터를 구성할 수 있도록 하는 노드입니다.

사용법



- **변수 목록 편집**을 선택하여 데이터 상의 변수 목록을 적어줍니다. (이 때 변수명은 데이터 상의 변수명과 일치할 필요는 없지만 데이터형은 일치해야 합니다.)
- **변수 목록 편집**의 **가져오기**를 통해서 변수 목록에 대한 정보가 저장되어 있는 파일을 불러 올 수 있습니다. (혹은 수동으로 편집할 수도 있습니다.)
- **변수 목록 편집**의 **내보내기**를 통해서 변수 목록에 대한 정보를 저장할 수 있습니다.
- **파일 목록 편집**의 **파일 추가**를 통해서 세로로 붙일 파일들을 선택합니다. 이 때 **FTP 파일 추가**를 선택하면 FTP 에 서버에 있는 파일을 추가할 수도 있습니다.
- 선택사항을 통해 세부 사항을 선택하고 FTP 연결을 통해서 FTP 연결 시 FTP 서버, FTP 사용자, FTP 암호를 설정할 수 있습니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
선택사항	변수 목록 편집	불러올 데이터의 변수가 어떠한지를 설정합니다. 수동으로 설정할 수도 있고 변수 정보가 저장되어 있는 파일을 불러올 수도 있습니다. 변수 목록을 저장할 수 있는 기능 또한 제공합니다.	필수	
	파일 목록 편집	불러올 데이터의 목록을 설정해 주도록 합니다.	필수	
	첫행에 변수명 존재	파일의 첫 행에 변수명이 있다면 **'예'**로 설정합니다. **'아니오'**로 설정되면 변수명을 자동으로 생성합니다. 첫 행에 변수명이 있는데 **'아니오'**로 설정할 경우 변수의 형태가 모두 문자형으로 설정될 수 있습니다.	필수	예 아니오
	FTP 사용	FTP 의 사용여부를 설정합니다.	필수	예 아니오
	구분자	컬럼을 구분하는 구분자를 입력합니다. 기본적으로 **(탭), (공백), |(바), ,(콤마), :(콜론)**, **;(세미콜론)**을 지원하며 이들 이외의 구분자를 사용할 경우 **(기타)**를 선택합니다.	필수	(탭) '|' (공백) ',' ';' ':' (기타)
	기타 구분자	**구분자** 속성이 **(기타)**일 경우 활성화되며 기타 구분자를 입력합니다.		
FTP 연결	FTP 서버	파일이 있는 FTP 서버가 무엇인지를 설정합니다.	FTP 사용시 필수	
	FTP 사용자	FTP 의 사용자 ID 를 입력합니다.	FTP 사용시 필수	
	FTP 암호	FTP 의 암호를 입력합니다	FTP 사용시 필수	

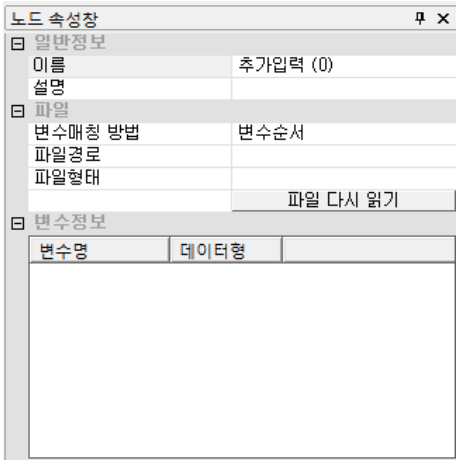
저장 선택 사항	파일 경로	저장할 파일의 경로 및 이름을 설정합니다.	필수	
	기준 변수	데이터를 분리하여 저장할 때 기준이 되는 변수를 선택합니다.	필수	None 을 선택할 수도 있음
	데이터 분리수	데이터를 분리하여 저장할 때 몇 개의 데이터를 기준으로 분리할지를 설정합니다.	필수	
변수정보	변수정보	파일을 분석하여 얻는 변수의 정보가 나타납니다. 총 변수의 개수, 변수명, 변수형태 등을 알 수 있습니다.		

NOTE: 기준변수 옵션이 NONE 이 아닌 임의의 변수로 지정되어 있을 경우, 데이터 분리수 옵션이 우선순위가 되어 데이터 분리수 기준으로 먼저 분리한 후 기준 변수 옵션에 따라 분리 저장이 진행됩니다.
예) 데이터 분리수가 5 이며, A 라는 변수를 기준변수로 할 경우, 첫 5 개의 데이터를 분리한 후, 기준 변수가 5 번째 행 이후로 변화하는지 확인합니다. 즉 A 변수 5 번째 데이터와 6 번째 데이터가 같을 경우 하나로 합치며, 다를 경우는 분리하여 저장하게 됩니다. 이와 같은 과정을 반복하여 데이터가 분리 저장이 끝날 때까지 진행하며, 그에 따라 파일수가 달라집니다.

3.1.10 추가 입력 노드

추가 입력 노드는 같은 형태의 여러 ecl 파일을 모아서 세로로 붙여 새로운 ecl 파일을 만드는 기능을 하는 노드입니다.

사용법



- **변수매칭 방법**을 선택합니다.
- 같은 형식의 파일이 있는 파일 경로를 선택합니다..
- 파일 이름을 구분하는 접두어 혹은 접미어를 *로 하여 **파일 형태**를 써 줍니다.
- **파일 다시 읽기**를 통해 설정 변경 후 새롭게 파일을 읽을 수 있습니다.

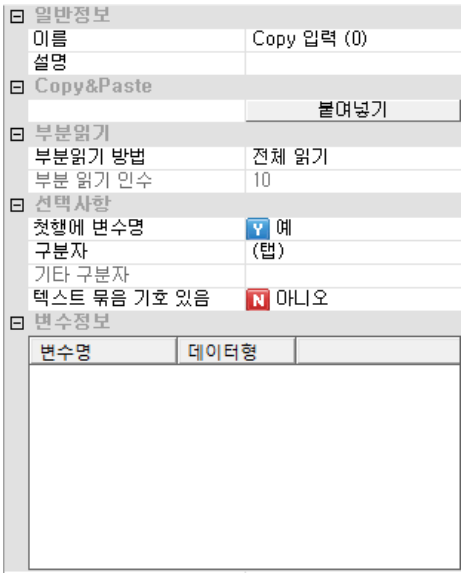
속성

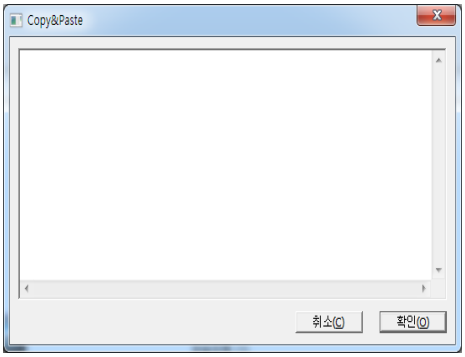
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
파일	변수 매칭 방법	여러 개의 ecl 파일들의 변수 매칭을 어떻게 시킬 것인지를 선택합니다. 변수 순서를 선택할 경우, 변수 이름에 무관하게 순서가 같은 변수를 같은 변수로 봅니다. 또한, 변수 이름을 선택할 경우, 변수의 순서와 무관하게 변수의 이름이 같으면 같은 변수로 봅니다.	필수	
	파일 경로	같은 형태의 ecl 파일들이 모여있는 파일의 경로를 선택합니다.	필수	
	파일 형태	만약 데이터가 test1, test2, …이런 식으로 되어 있으면 test*을 입력하고 1test, 2test, …이런 식으로 되어 있으면 *test 를 입력하는 식으로 하여 파일을 구별해 주는 접두어 혹은 접미어를 *로 하여 파일 형태를 써 줍니다.	필수	
변수정보	변수정보	파일을 분석하여 얻는 변수의 정보가 나타납니다. 총 변수의 개수, 변수명, 변수형태 등을 알 수 있습니다.		

NOTE: '변수 매칭 방법'에서 변수이름 옵션에 따른 특이사항
'변수 매칭 방법'에서 변수이름 옵션을 설정한 후 입력된 데이터 셋의 변수가 매칭이 되지 않은 경우가 발생하면 해당 데이터 셋을 추가할 때 공백으로 추가하게 됩니다. 즉 여러 개 파일을 합칠 때, 변수가 매칭되지 않으면 최종 데이터 셋에서 빈 공백 부분이 발생할 수 있으며, 이는 사용자가 결측치처리 방식을 이용하여 처리 후 분석을 진행하셔야 됩니다.

3.1.11 Copy 입력 노드

Copy 입력





정해진 데이터 파일, DB 의 형태가 아닌 사용자가 Web 또는 원하는 데이터를 Copy&Paste 를 통해서 데이터 입력을 할 수 있는 노드입니다.

- **Copy&Paste** 의 **붙여넣기** 버튼을 누르면 아래의 창이 뜹니다. 아래 창에서 원하는 데이터를 입력하거나 붙여 넣으면 됩니다.
- 제대로 분석이 되었다면 **선택사항** 항목들이 자동으로 분석되어 나타나며 **변수정보** 속성에 변수정보들이 목록으로 나열됩니다.
- 만약 제대로 분석되지 않았다면 **선택사항** 항목을 직접 지정합니다.
- **선택사항** 항목들이 잘못 입력되었을 경우 올바른 데이터마이닝을 수행할 수 없을 뿐 아니라 경우에 따라 **프로그램에 치명적인 영향**을 미칠 수도 있으므로 주의하시기 바랍니다.
- **변수정보** 목록 내용 중 잘못된 것이 있다면 수정합니다. 변수명, 변수형태 등을 수정할 수 있습니다. 변수형태는 정수, 실수, 문자, 날짜 등 총 4 가지입니다.
- 변수형태를 잘못 지정하면 원하는 결과를 얻을 수 없습니다.
- 파일 **부분 읽기** 방식을 지정합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
Copy&Paste	붙여넣기	데이터를 입력하거나 붙여 넣을 수 있는 다이얼로그가 활성화 됩니다.	필수	버튼

부분읽기	부분읽기 방법	파일을 부분적으로 읽는 방법을 지정합니다.	필수	전체읽기, 처음부터, 임의추출
	부분읽기 인수	부분 읽기 방법이 '처음부터'라면 인수가 관측치 수가 되고 '임의 추출'이라면 관측치 퍼센티지가 됩니다.		
선택사항	첫 행에 변수명	파일의 첫 행에 변수명이 있다면 **'예'**로 설정합니다. **'아니오'**로 설정되면 변수명을 자동으로 생성합니다. 첫 행에 변수명이 있는데 **'아니오'**로 설정할 경우 변수의 형태가 모두 문자형으로 설정될 수 있습니다.	자동 분석	예, 아니오
	구분자	컬럼을 구분하는 구분자를 입력합니다. 기본적으로 **(탭), (공백), |(바), ,(콤마), :(콜론), ;(세미콜론)**을 지원하며 이들 이외의 구분자를 사용할 경우 **(기타)**를 선택합니다.	필수	(탭) '|' (공백) ',' ';' ':' (기타)
	기타 구분자	**구분자** 속성이 **(기타)**일 경우 활성화되며 기타 구분자를 입력합니다.		
	텍스트 묶음 기호 있음	따옴표를 가지고 있는 문자형 데이터에서 따옴표를 제거하고 읽을 것인지 따옴표를 포함하여 데이터로 읽을 것인지 여부를 지정합니다.	예 일 경우 "ABC"➔ ABC	예, 아니오
변수정보	변수정보	파일을 분석하여 얻는 변수의 정보가 나타납니다. 총 변수의 개수, 변수명, 변수형태 등을 알 수 있으며 변수명, 변수형태의 변경이 가능합니다.		

3.2 전처리 노드

데이터 마이닝 프로세스에 있어 데이터는 매우 중요합니다. 올바르지 못한 데이터를 사용하여 분석할 경우 제대로 된 분석결과를 산출하기 어렵습니다. 따라서 데이터를 정제 / 변경하는 과정이 필수적이며 이러한 과정을 통틀어 **데이터 전처리**라고 합니다.

전처리 과정은 경우에 따라 매우 복잡하거나 난해해질 수 있으나, **ECMiner™**에서는 간단한 조작만으로 쉽게 수행할 수 있도록 다음과 같은 전처리 노드를 제공합니다.

결측치 처리

결측치가 있을 경우 지정된 값으로 채워 넣습니다.

그룹화

지정된 변수들의 값 별로 데이터를 그룹핑하여 새로운 데이터 셋을 만듭니다.

다중 파생변수

지정된 변수들에 같은 룰을 적용하여 새로운 변수들을 생성합니다.

변수순서

변수의 순서를 변경합니다.

변수 표준화

선택한 변수를 표준화한 변수로 새롭게 생성합니다.

병합

입력된 데이터 소스를 키값을 기준으로 병합합니다.

분할

데이터를 모델링용과 평가용으로 나눕니다.

선택

지정된 조건이 참인 행만 선택합니다.

선택 2

지정된 키변수를 기준으로 그룹핑 한 후 조건에 맞는 값을 선택하는 노드입니다.

열조합

모델/모델링 결과를 하나의 데이터 소스로 묶습니다.

정렬

지정된 변수들을 정렬합니다.

채우기

특정 변수에 대하여 지정된 조건이 참이라면 지정된 새로운 값으로 채워 넣습니다.

추가

입력된 데이터 소스를 하나로 합칩니다.

파생변수

파생변수를 생성합니다.

표본추출

전체 데이터 중 일부만 표본 추출합니다.

전처리 피봇

피봇팅을 수행합니다.

필터

변수의 사용여부 및 변수명을 변경합니다.

형태변경

변수의 형태를 변경합니다.

COUNTER

지정된 그룹핑 조건에 따라 그룹별로 counting 한 결과를 파생변수로 추가합니다.

그룹 통계량

지정된 그룹핑 조건에 따라 지정된 변수의 통계량을 파생변수로 붙입니다.

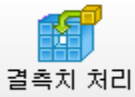
RANKING

RANKING 을 계산하여 변수로 추가합니다.

구간화 노드

종속변수에 미치는 영향도에 따라 독립변수를 여러 개의 구간으로 나눕니다.

3.2.1 결측치 처리 노드



입력된 변수에 결측치가 존재할 경우, 지정값으로 채워 넣는 노드입니다.

사용법

- **연속형 변수**에 대한 채우기 값을 지정합니다.
- **이산형 변수**에 대한 채우기 값을 지정합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 주석을 입력합니다.	선택	
선택사항	대상변수	처리 대상이 되는 변수를 선택합니다.		모든 변수, 이산형 변수만, 연속형 변수만
	연속형 변수	변수가 **연속형**일 경우, 채워 넣을 값을 지정합니다.		평균, 중앙값, 이전값, 이후값, 구간 평균, 0 할당.
	이산형 변수	변수가 **이산형**일 경우 채워 넣을 값을 지정합니다.		최대회수에 해당하는 값, 최소회수에 해당하는 값, 이전값, 이후값

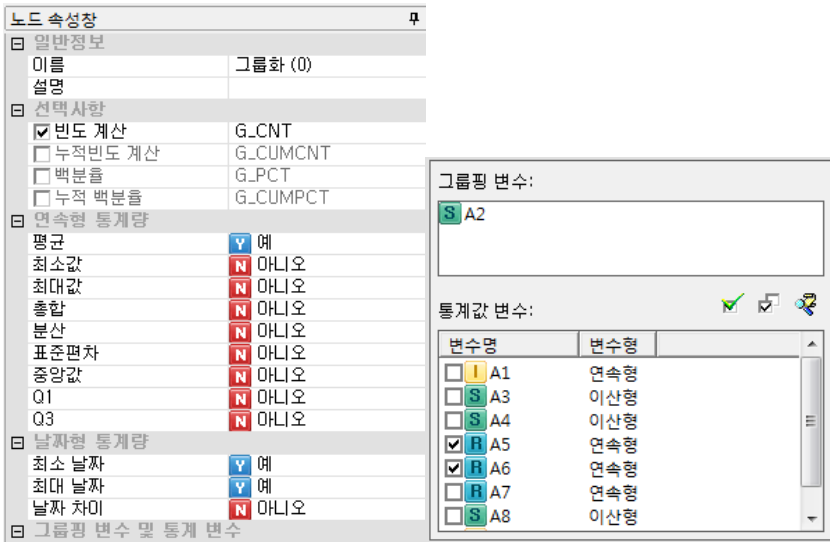
채우기 값 정의

변수형태	값	설명
연속형	평균	결측치를 해당 변수의 평균값으로 대체합니다.
	중앙 값	결측치를 해당 변수의 중앙값으로 대체합니다.
	이전 값	결측치를 해당 변수의 이전값으로 대체합니다.
	이후 값	결측치를 해당 변수의 이후값으로 대체합니다.
	구간평균	결측치를 (이전값 + 이후값)/2 의 값으로 대체합니다.
	0 할당	결측치를 0 으로 대체합니다.
이산형	최대 횟수에 해당하는 값	결측치를 최대 회수 범주값으로 대체합니다.
	최소 횟수에 해당하는 값	결측치를 최소 회수 범주값으로 대체합니다.
	이전 값	결측치를 해당 변수의 이전값으로 대체합니다.
	이후 값	결측치를 해당 변수의 이후값으로 대체합니다.

3.2.2 그룹화 노드

그룹화 노드를 이용하여 기존 데이터를 그룹화 할 수 있습니다. 이 때 이산형 필드를 기준으로 연속형 필드의 값들을 그룹화합니다.

사용법



- 형태변경 노드에서 연속형, 이산형으로 데이터 형을 지정합니다.
- 노드 속성창에서 선택사항과 데이터 처리 부분의 각종 옵션을 선택합니다.
- 스트림을 실행합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
선택사항	선택사항	선택사항에 선택된 값에 대해 파생변수를 생성합니다.	빈도계산, 누적빈도계산, 백분율, 누적백분율,	
통계량	연속형통계량	평균, 최소값, 최대값, 총합, 분산, 표준편차를 구하여 파생변수로 생성합니다.	필수(최소 1 개 이상 선택)	
	날짜형통계량	최소 날짜, 최대 날짜, 날짜 차이를 구하여 파생변수로 생성합니다.	선택	
그룹핑 변수 및	그룹핑 변수	그룹핑 변수를 지정합니다. 그룹핑 변수는 이산형만 가능합니다.	필수	

통계 변수	통계값 변수	그룹핑변수에 의한 통계값을 확인할 변수를 선택합니다. 통계값 변수는 연속형만 가능합니다.	필수	

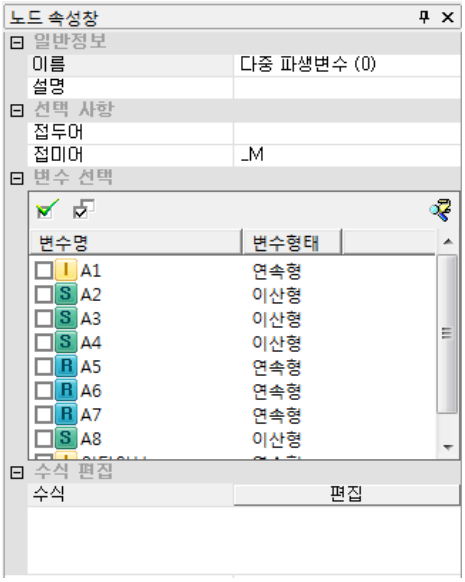
3.2.3 다중 파생변수 노드


다중 파생변수

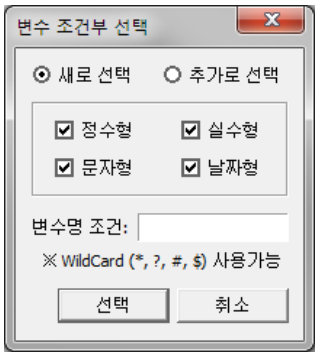
다중 파생변수 노드를 이용하여 여러 개의 기존 필드에 공통적인 변환규칙을 적용하여 새로운 필드들을 생성할 수 있습니다. 이때 공통된 수식에 각각의 필드가 대입되게 되는데, 따라서 새로운 필드들을 생성할 기존 필드들의 데이터 형은 서로 일치해야 합니다.

사용법

다중 파생필드 노드는 다음과 같은 속성변경 인터페이스를 갖습니다.



	선택된 변수를 사용하도록 설정합니다.
	선택된 변수와 선택되지 않은 변수들을 반전시킵니다.



버튼 클릭 시 왼쪽과 같은 창이 생성되어 사용자가 선택한 변수형 또는 변수명 조건에 일치하는 변수들을 선택합니다

- 새롭게 생성될 파생변수들의 접두어와 접미어를 설정합니다.
- 공통의 변환규칙을 적용할 변수들을 선택합니다.
- 스트림을 실행합니다.

속성

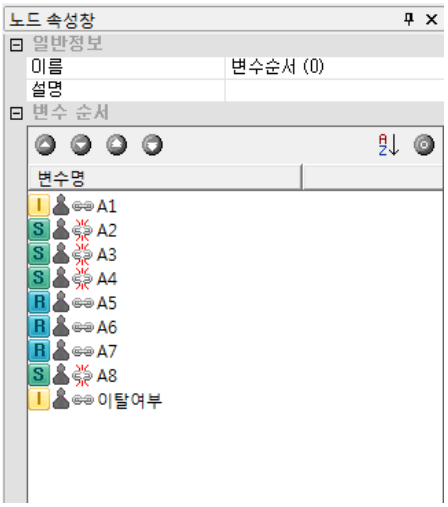
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 입력합니다.	선택	
선택사항	접두어	파생변수들의 접두어를 입력합니다.	선택	
	접미어	파생변수들의 접미어를 입력합니다.	선택	
변수선택	변수선택	변수들의 이름과 형태를 나열합니다. 파생변수 생성시 바탕이 되는 변수를 선택하는 인터페이스를 제공합니다.	기타 UI	사용자 인터페이스 참조.
수식편집	수식편집	각 파생변수들에 공통으로 적용할 변수 변환 규칙을 정의합니다.	기타 UI	사용자 인터페이스 참조.

3.2.4 변수순서 노드

입력 데이터 변수의 순서를 조정하기 위한 노드입니다.

사용법

변수순서 노드는 다음과 같은 속성변경 인터페이스를 갖습니다.



	변수순서를 앞으로 당기는 버튼입니다.
	변수순서를 뒤로 미는 버튼입니다.
	변수순서를 가장 앞으로 당기는 버튼입니다.
	변수순서를 가장 뒤로 미는 버튼입니다.
	변수명으로 오름차순 정렬합니다.
	변수순서를 초기화하는 버튼입니다.

변수명 리스트에서 기능버튼을 이용하여 변수들의 순서를 변경합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 입력합니다.	선택	
변수 순서	변수명	입력된 변수명 리스트로, 순서 결정을 할 수 있고 그 결과를 볼 수 있습니다.	기타 UI	사용자 인터페이스 참조.

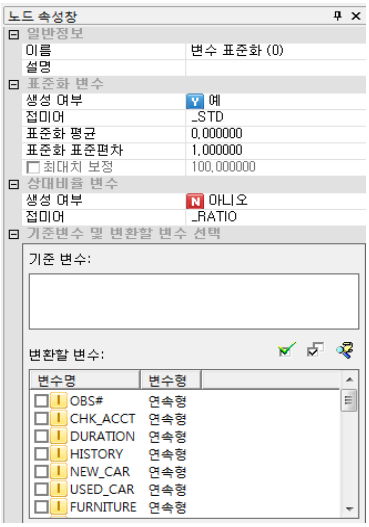
3.2.5 변수 표준화 노드

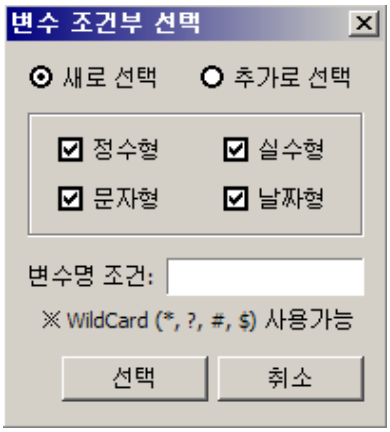
변수 표준화

선택한 변수를 표준화한 변수로 새롭게 생성하는 노드입니다.

사용법

변수표준화 노드는 다음과 같은 속성변경 인터페이스를 갖습니다.



	선택된 변수를 사용하도록 설정합니다.
	선택된 변수와 선택되지 않은 변수들을 반전시킵니다.
	 버튼 클릭 시 왼쪽과 같은 창이 생성되어 변수형 또는 변수명과 같은 조건에 맞는 변수를 선택합니다.

변환할 변수 리스트에서 기능버튼을 이용하여 변수들을 선택, 반전 시킵니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 입력합니다.	선택	
표준화 변수	생성 여부	표준화 변수를 생성할지 여부를 지정합니다. 상대비율 변수 혹은 표준화 변수 생성 여부 중 하나는 '예'로 선택되어야 합니다.	조건부 필수	
	접미어	생성된 표준화 변수에 덧붙일	선택	

		문자열을 지정합니다.		
	표준화 평균	변수 표준화 후 표준화된 변수가 가질 평균을 입력합니다.	선택	
	표준화 표준편차	변수 표준화 후 표준화된 변수가 가질 표준편차를 입력합니다.	선택	
	최대치 보정	생성된 표준화 변수의 최대값이 지정된 값이 되도록 하고자 할 때 선택합니다.	선택	
상대비율 변수	생성여부	상대비율 변수를 생성할 지 여부를 지정합니다.	조건부 필수	
	접미어	생성한 상대비율 변수에 덧붙일 문자열을 지정합니다.		
변수선택	기준변수	변수를 지정하지 않으면 각각의 레코드를 대상으로, 변수를 지정하면 지정된 기준변수의 범주별로 값을 계산합니다.	선택	사용자 인터페이스 참조.
수식편집	변환할 변수	변환할 변수를 선택합니다. 숫자형태의 변수만 가능합니다.	필수	사용자 인터페이스 참조.

참고 : 상대비율 변수는 지정된 변수의 총합을 계산한 후 각 행의 값을 계산된 총합으로 나눈 값이 저장됩니다.

3.2.6 병합 노드

둘 이상의 데이터를 하나로 합쳐 새로운 데이터 소스를 만들고자 한다면 **병합 노드**를 이용합니다. 단, 데이터를 합칠 때 기준이 되는 필드는 각각의 데이터에 모두 있어야 합니다. 예를 들어 고객 정보에 대한 데이터가 데이터 소스 A 와 소스 B 에 나눠져 있을 경우 이를 고객 번호(키필드)를 기준으로 하나로 합쳐 새로운 데이터 소스 C 로 만들 때 사용합니다.

내부 조인, 외부 조인, 부분 외부 조인, 안티 조인 등 다양한 병합 방법을 제공하며 이를 이용하여 얻고자 하는 새로운 데이터 소스를 얻을 수 있습니다.

사용법

병합 노드는 다음과 같은 속성변경 인터페이스를 갖습니다.

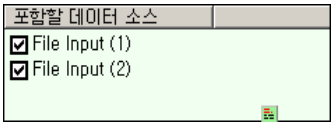
데이터 소스 순위 인터페이스는 추가할 데이터 소스가 표시되고, 이들의 우선 순위를 결정하는 인터페이스입니다. 가장 높은 우선순위의 데이터 소스(맨 위에 위치하는 데이터 소스)는 주 소스가 되어서 병합의 기준이 됩니다.

아이콘	기능
	데이터 소스의 순위를 높이는 버튼입니다.
	데이터 소스의 순위를 내리는 버튼입니다.

키필드는 나눠져 있는 데이터를 하나로 합칠 때 기준이 되는 값입니다. 키필드 인터페이스는 키필드로 사용 가능한 변수명이 좌측에 표시되고, 사용자가 키필드로 선택한 변수들이 우측에 표시됩니다. 가운데 버튼을 이용하여 사용하고자 하는 키필드를 선택 혹은 해제할 수 있습니다.

아이콘	기능
	선택된 키필드를 추가 합니다.
	선택된 키필드를 삭제 합니다.

포함할 데이터 소스 인터페이스는 **병합방법**이 **부분 외부 조인**인 경우 활성화되며 포함될 데이터 소스를 선택하게 합니다. **부분 외부 조인**은 **외부 조인**과 마찬가지로 데이터 병합을 하지만 전체 데이터 소스가 아니라 선택된 데이터 소스의 데이터만 반환한다는 차이가 있습니다.



- **데이터 소스 순위** 리스트에서, 소스들의 순서를 변경하고, **주 소스**를 지정합니다.
- **키필드 지정 방법**을 선택합니다. **수동**을 선택하면 인터페이스가 활성화되어 키필드로 사용될 필드를 선택 혹은 삭제할 수 있습니다.
- **병합 방법**을 결정합니다. **변수명 중복 시** 처리법을 선택하고 **병합 방법**을 선택합니다.

- **병합 방법**이 **부분 외부 조인**일 때 **포함할 데이터 소스** 인터페이스가 활성화되어 데이터 소스를 선택할 수 있습니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 입력합니다.	선택	
데이터 소스 순위	데이터 소스	추가할 데이터 소스의 리스트로, 순위 조정을 통해 소스의 우선순위 조정과 **주 소스** 결정을 할 수 있습니다.	기타 UI	사용자 인터페이스 참조.
키필드	키필드 지정 방법	변수병합 시 키가 되는 필드를 지정해줍니다. **수동** 선택 시 키필드 선택 인터페이스가 활성화되어 사용자가 직접 키필드를 선택할 수 있습니다.	필수	자동, 수동
	키필드 지정 인터페이스	**키필드 지정 방법**이 **수동**의 경우 활성화됩니다. 사용자가 직접 키필드를 선택 혹은 삭제할 수 있습니다.	기타 UI	사용자 인터페이스 참조.
병합방법	변수명 중복 시	**변수명 중복 시** 처리법을 선택합니다.	필수	필터링, 인덱스부여, 오류발생
	병합방법	**병합방법**을 선택합니다. **부분 외부 조인**의 경우 포함할 **데이터 소스** 인퍼테이스를 활성화시켜 데이터 소스 선택이 가능하게 합니다.	필수	내부조인, 외부조인, 부분외부조인, Anti-Join
	포함할 데이터 소스	**병합 방법**이 **부분 외부 조인**의 경우 활성화됩니다. 사용자가 직접 체크박스를 통해 데이터 소스를 선택 혹은 해제할 수 있습니다.	기타 UI	사용자 인터페이스 참조.

참고 사항 – 병합 방법

ECMiner™에서는 다음과 같은 병합 방법을 제공합니다.

병합 방법	설명	비고
내부 조인	입력된 데이터 소스 중 지정된 키필드가 모두 일치하는 행을 검색하여 새로운 데이터 소스에 추가합니다. 예를 들어, **고객정보_A** 와 **고객정보_B** 데이터 소스에서 고객번호가 동일한 행만 새로운 데이터 소스에 추가합니다.	
외부 조인	입력된 데이터 소스 중 지정된 키필드의 일치 여부에 상관없이 모든 행을 새로운 데이터 소스에 추가합니다.	
부분 외부 조인	입력된 데이터 소스 중 지정된 키필드의 일치 여부에 상관없이 모든 행을 새로운 데이터 소스에 추가합니다. 단, 지정된 데이터 소스의 데이터만 추가합니다.	포함할 데이터 소스를 지정해야 합니다.
안티 조인	입력된 데이터 소스에서 지정된 키필드 중 일치하지 않는 키값을 주 데이터 소스에서 제거하여 새로운 데이터 소스를 구성합니다.	

예제

다음과 같은 데이터 소스 A, B 가 있다고 가정합니다.

	1	2
	고객번호	구매회수
1	2	1
2	3	3
3	4	5
4	6	7

<데이터 소스 A>

	1	2
	고객번호	구매금액
1	2	8000
2	3	5000
3	9	6000
4	10	8700

<데이터 소스 B>

위 두 데이터 소스 중 중복되는 필드인 "고객번호"를 키필드로 하여 각 병합법을 수행하면 다음과 같은 결과를 얻을 수 있습니다.

(1) 내부 조인

두 데이터 소스 중 중복되는(일치하는) 키값 즉, 고객번호 2, 3 에 대해서만 병합을 수행합니다. 결과는 다음과 같습니다.

	1	2	3
	고객번호	구매회수	구매금액
1	2	1	8000
2	3	3	5000

(2) 외부 조인

두 데이터 소스 중 모든 키값에 대하여 병합을 수행합니다. 키필드로 지정된 고객번호 필드의 경우 데이터 소스 A 는 2, 3, 4, 6, 데이터 소스 B 는 2, 3, 9, 10 의 키값을 가지고 있으며 이들 모두에 대하여 병합을 수행합니다. 따라서, 2, 3, 4, 6, 9, 10 의 고객번호에 대하여 다음과 같은 결과를 얻습니다. 이 경우 값이 없는 변수에 대하여 <NULL> 값을 할당합니다.

	1	2	3
	고객번호	구매회수	구매금액
1	2	1	8000
2	3	3	5000
3	4	5	<NULL>
4	6	7	<NULL>
5	9	<NULL>	6000
6	10	<NULL>	8700

(3) 부분 외부 조인

두 데이터 소스 중 모든 키값에 대하여 외부 조인과 같이 병합을 수행한 후 선택된 데이터 소스에 있는 값만 반환합니다. 아래의 경우 데이터 소스 A 만 선택하여 부분 외부 조인을 한 경우입니다. 결과에서 보듯이 데이터 소스 A 에만 있는 2, 3, 4, 6 의 고객번호에 대해서만 값을 반환합니다.

소스 A 만 선택하여
부분 외부 조인 수행

	1	2	3
	고객번호	구매회수	구매금액
1	2	1	8000
2	3	3	5000
3	4	5	<NULL>
4	6	7	<NULL>

소스 B 만 선택하여
부분 외부 조인 수행

	1	2	3
	고객번호	구매회수	구매금액
1	2	1	8000
2	3	3	5000
3	9	<NULL>	6000
4	10	<NULL>	8700

(4) Anti Join

주 데이터 소스가 A 일 때 주 데이터 소스가 아닌 데이터 소스 B 에 있는 키값 중 중복되는 행을 제거합니다. 즉, A 에 있는 키값 2, 3, 4, 6 중 B 와 중복되는 2, 3 이 제거되어 다음과 같은 결과를 얻을 수 있습니다.

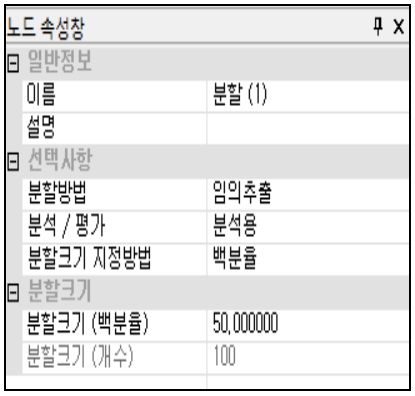
	1	2
	고객번호	구매회수
1	4	5
2	6	7

3.2.7 분할 노드


분할

보다 좋은 모델을 산출하기 위하여 입력된 전체 데이터 중 일부를 평가용으로 사용할 수 있습니다. 즉, 분석용으로 분할된 데이터는 모델을 만드는데 사용되고 평가용으로 분할된 데이터를 이용하여 생성된 모델을 평가함으로써 더 좋은 모델을 얻을 수 있는 것입니다.

사용법



- 분할방법을 선택합니다. (현재는 **임의추출**만 가능합니다.)
- **분석 / 평가** 속성에서 추출된 표본을 분석용으로 사용할지 평가용으로 사용할 지 지정합니다.
- 분할 크기를 결정할 방법을 선택합니다.
- **분할크기 지정방법**에 따라 백분율 혹은 분할 개수를 입력합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
선택사항	분할방법	데이터를 분석용/평가용으로 분할하는 방법을 지정합니다. 현재는 임의추출만 지원합니다.	필수	임의추출

	분석/평가	추출된 일부 데이터를 분석용으로 사용할 지 평가용으로 사용할 지 지정합니다.	필수	분석용, 평가용
	분할크기 지정방법	분할크기를 결정할 방법을 지정합니다.	필수	백분율, 개수
표본크기	분할크기 (백분율)	**분할크기 지정방법**이 **백분율**일 경우 활성화됩니다.		0 ~ 100 사이의 실수
	분할크기 (개수)	**분할크기 지정방법**이 **개수**일 경우 활성화됩니다. 분할하고자 하는 데이터 크기를 정수로 입력합니다. 만약, 전체 데이터 건수보다 클 경우 전체 데이터 수로 변경되어 처리됩니다.		정수

3.2.8 선택 노드


선택

주어진 조건에 맞는 행만 선택하는 노드입니다. 데이터 중 일부만 추출한다는 점에서 **표본추출 노드**와 비슷하지만 **표본추출 노드**는 정해진 방법에 의해서 추출하고 **선택 노드**는 사용자에 의해 지정된 조건으로 추출한다는 차이점이 있습니다.

사용법

조건형식을 지정합니다. 선택 조건을 임의로 지정하려면 **사용자 입력**을 선택하고 결측치만 제거하고자 하며 **결측치 포함 행 제거**를 선택합니다.

조건형식이 **사용자 입력** 일 경우 **선택조건**을 지정하여야 합니다. 편집 버튼을 누르면 나타나는 **수식 편집기**를 이용하여 조건에 해당하는 수식을 입력합니다.

Note *사용자가 정의한 선택 조건일 경우, 입력된 수식의 형태에 관계 없이 결과가 참이기만 하면 그 행을 선택합니다. 즉, 간단한 조건문 (예, {A} < 1) 혹은 IF ~ ENDIF 구문이 입력되었을 지라도 수행된 결과가 참이기만 하면 행을 선택합니다.*

속성

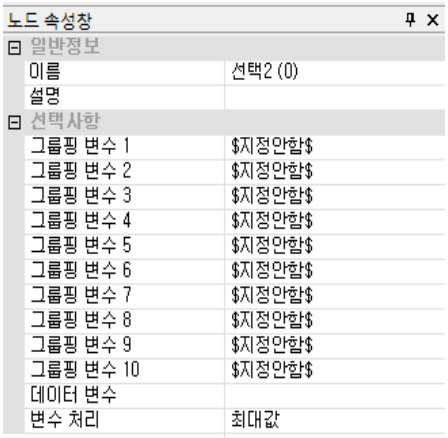
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	

선택사항	조건형식	특정 조건을 지정하려면 **사용자 입력**으로 선택하고 결측치만 제거하고 싶다면 **결측치 포함 행 제거**를 선택합니다.	필수	사용자 입력, 결측치 포함 행 제거.
	선택조건	행을 선택할 조건을 입력합니다. 편집 버튼을 누르면 조건 수식을 편집할 수 있는 **수식 편집기**가 나타납니다. **조건형식**이 **사용자 입력**일 경우 활성화됩니다.	버튼	**수식편집기** 참조.
	조건 열람창	현재 입력된 조건 수식이 나타납니다. 열람만 할 수 있으며 직접 편집할 수는 없습니다.		

3.2.9 선택 2 노드

지정된 키변수를 기준으로 그룹핑 한 후 조건에 맞는 값을 선택하는 노드입니다.

사용법



그룹핑 변수를 지정한 후 그룹핑 된 변수에 의해 분석할 데이터 변수를 정합니다. 그리고 변수처리(최대값, 최소값, 최대/최소 이외의 값, 최대/최소값, 중복제거, 중복제거 & 모두 0 제거)값을 선택하여 결과를 확인합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
선택사항	그룹핑변수	특정 변수를 이용하여 그룹핑 합니다.	필수	
	데이터변수	그룹핑 된 변수에 의해 분석할 데이터 변수를 지정합니다.	필수	
	변수 처리	변수 처리 방식을 지정합니다. • 최대값: 데이터 변수에서 그룹핑 된 그룹별 최대값을 반환합니다. • 최소값: 데이터 변수에서 그룹핑 된 그룹별 최소값을 반환합니다. • 최대 / 최소 이외값: 데이터 변수에서 그룹핑 된 그룹별 최대 / 최소 이외의 값들을 반환합니다. • 최대 / 최소값: 데이터 변수에서 그룹핑 된 그룹별 최대 / 최소값들을 반환합니다. • 중복 제거: 데이터 변수에서 그룹핑 된 그룹별로 중복을 제거한 값들을 반환합니다. • 중복 제거 & 모두 0 제거: 데이터 변수에서 그룹핑 된 그룹별로 중복 및 0 인 값을 제거한 값들을 반환합니다.	필수	

3.2.10 열조합 노드

한 데이터 소스에 대하여 여러 가지 모델링 혹은 모델을 적용한 뒤 각 결과를 하나로 합쳐 보고자 할 때 열조합 노드를 사용합니다. 모델링 / 모델의 결과를 동시에 보고자 할 때만 사용하는 노드이므로 다음과 같은 제한사항이 있습니다.

제한사항

- 바로 앞단의 노드는 모델링 혹은 모델 노드이어야 합니다. 단, 모델링 혹은 모델노드 두 종류가 동시에 연결될 수 없습니다.
- 연결하고자 하는 모델 / 모델링 노드는 같은 원천 데이터 소스를 가지고 있어야 합니다.
- 모델 / 모델링 노드를 제외한 다른 노드는 연결할 수 없습니다.

사용법

- 수행하고자 하는 모델 / 모델링을 위한 스트림을 구성합니다.
- 스트림 내에 포함되어 있는 모델 / 모델링 노드 중 결과를 한데 묶어서 보고자 하는 노드를 열조합 노드에 연결합니다.
- 연결된 노드들의 변수를 분석하여 자동으로 열조합 노드에서 사용될 변수 목록을 나열합니다.
- 만약 중복되는 변수명이 있을 경우 1, 2, 3... 과 같은 인덱스가 추가로 붙습니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 입력합니다.	선택	
변수정보	변수정보	현재 조합된 변수에 대한 정보를 나타냅니다.	열람	

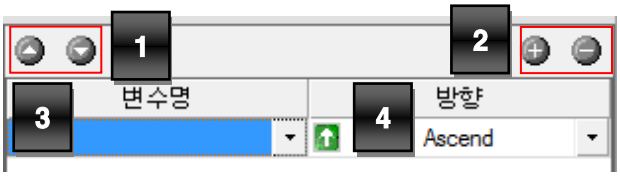
3.2.11 정렬 노드


정렬

정렬 노드를 이용하여 데이터를 원하는 변수를 기준으로 정렬할 수 있습니다. **정렬 노드** 에서는 단순히 한 개의 변수에 대한 정렬뿐 아니라, 여러 변수에 대해서 다중정렬도 가능합니다. 또한 정렬순서도 변경이 가능합니다.

사용자 인터페이스

정렬 노드는 다음과 같은 속성변경 인터페이스를 갖습니다.



번호	기능
1	선택된 변수(필드)의 정렬 순서를 변경하는 버튼들입니다. 왼쪽 버튼을 누르면 정렬 순위가 올라가고, 오른쪽 버튼을 누르면 정렬 순위가 내려갑니다.
2	정렬할 변수(필드)의 수를 추가/삭제하는 버튼들입니다. 왼쪽 버튼을 누르면 정렬할 변수(필드) 개수가 늘어나고, 오른쪽 버튼을 누르면 줄어듭니다.
3	정렬할 변수(필드)를 지정합니다. 콤보 박스를 누르면 변수 리스트가 표시되고, 리스트 내에서 정렬할 변수를 지정하면 됩니다.
4	변수(필드)의 정렬 방향을 지원합니다. 정렬 방향을 지정하는 방법은, 콤보 박스 리스트에서 선택하는 방법과, 방향 표시 영역을 마우스로 더블 클릭하는 방법이 있습니다.

사용법

- 정렬하고자 하는 변수만큼, 정렬 변수 리스트를 추가합니다.(위 그림의 2 번의 왼쪽 버튼을 이용합니다.)
- 각 리스트마다 정렬할 변수(필드)를 지정합니다.(위 그림의 3 번의 콤보박스를 이용합니다.)
- 각 리스트마다 정렬할 방향을 지정합니다.(위 그림의 4 번의 콤보박스를 이용합니다.)
- 만약, 정렬 순서를 변경할 경우, 위 그림의 1 번 버튼들을 이용하여 순서를 변경합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
정렬	정렬 변수	변수 정렬을 위한 사용자 인터페이스를 제공합니다. 이 속성을 이용하여 정렬한 변수와 그 방향을 쉽게 지정할 수 있습니다.	기타 UI	사용자 인터페이스 참조.

3.2.12 채우기 노드

채우기 노드는 특정 변수에 값을 채워 넣을 때 사용합니다. 즉, 특정 변수에 대하여 주어진 조건이 만족할 때 지정된 값으로 변경합니다.

사용법

필러추가 속성을 클릭하면 현재 이 노드에서 사용 가능한 변수들이 나타납니다. 이 변수들 중 채워 넣기를 수행할 변수를 지정합니다. 그러면 다음과 같은 창이 추가됩니다.

(1)에 선택한 변수명이 나타납니다. 단, (1)에서 선택된 변수는 '채우기' 기능이 실행될 대상의 변수이며, (4)에서 설정될 조건의 대상이 되는 변수는 아닙니다.

(2)에서 채워 넣을 조건을 지정합니다. **결측치** 혹은 **사용자 정의**를 선택할 수 있으며 **결측치**로 선택했을 경우 조건을 따로 지정하지 않아도 됩니다.

(2)에서 **사용자 정의**로 선택했을 경우, 버튼 (3)을 누르면 **수식 편집기**가 나타납니다. 이를 이용하여 조건에 해당하는 수식을 편집할 수 있습니다. 이 때, 조건의 대상이 되는 변수를 입력해야 정확한 결과가 나오게 됩니다. 편집된 수식은 (4)에 나타납니다.

NOTE 사용자가 정의한 조건일 경우, 입력된 수식의 형태에 관계 없이 결과가 참이기만 하면 채워 넣기를 수행합니다. 즉, 간단한 조건문 (예, {A} < 1) 혹은 IF ~ ENDIF 구문이 입력되었을 지라도 수행된 결과가 참이기만 하면 채워 넣기를 수행합니다.

버튼 (5)을 눌러 조건이 만족할 경우 입력할 값 또는 수식을 입력합니다. 이 경우도 **수식 편집기**를 이용합니다. 편집된 내용은 (6)에 나타납니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 입력합니다.	선택	
필러 추가/삭제	필러 추가	필러를 추가하고자 하는 변수명을 선택합니다. 지정한 변수에 맞는 필러가 추가됩니다.		가용 변수 목록.
	마지막 필러 삭제	현재 추가된 필러 중 마지막 필러를 삭제합니다.	버튼	
	필러 편집	필러를 추가하고자 하는 변수를 한번에 선택하여 추가합니다.	버튼	
필러 정보	지정한 변수명	지정한 변수명이 나타나며, 채우기 조건을 지정합니다. 조건을 **사용자 정의**로 선택하였을 경우, 조건문을 지정하여야 합니다.	위 그림에서 (1), (2)	사용자 정의, 결측치.
	조건 편집	채우기 조건을 편집합니다. 이 때, 조건의 대상이 되는 변수와 조건식을 정확히 입력하여야 합니다.	버튼	**수식 편집기** 참조.
	값 편집	해당 변수에 대하여 지정한 채우기 조건이 만족할 경우, 채워 넣을 값을 지정합니다.	버튼	

3.2.13 추가 노드

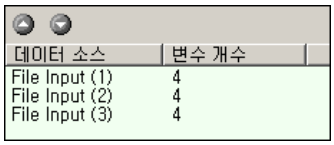

추가

비슷한 종류의 데이터가 흩어져 있는 경우 **추가 노드**를 이용하여 하나로 묶을 수 있습니다. 즉, 데이터 소스 A 와 B 가 같은 종류의 데이터 (고객 데이터, 공정 데이터 등) 이고 이를 하나로 묶어 처리하고자 할 경우 사용합니다. 데이터 소스의 개수에 제한이 없으며 메모리가 허락하는 한 입력된 데이터 소스를 하나로 묶을 수 있습니다.

사용법

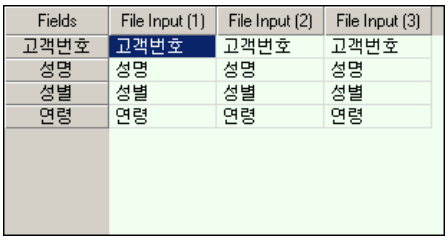
추가 노드는 다음과 같은 속성변경 인터페이스를 갖습니다.

데이터 소스 순위 인터페이스는 추가할 데이터 소스가 표시되고, 이들의 우선 순위를 결정하는 인터페이스입니다. 가장 높은 우선순위의 데이터 소스(맨 위에 위치하는 데이터 소스)는 **주 소스**가 되어서 추가의 기준이 됩니다.



아이콘	기능
	데이터 소스의 순위를 높이는 버튼입니다.
	데이터 소스의 순위를 내리는 버튼입니다.

입력된 데이터 소스의 순위에 따라 다음 인터페이스에서 가용한 변수를 확인할 수 있습니다.



제일 앞에 있는 "Fields"는 현재의 설정으로 추가를 수행한 후 가지게 되는 변수 목록을 나타냅니다. 즉, 이 노드 이후의 노드는 이 값을 변수 목록으로 가지게 됩니다. "Fields" 이후에 있는 "File Input (1)" ~ "File Input (3)"은 데이터 소스를 나타내며 각 데이터 소스에 있는 변수 목록을 나타냅니다. 만약, 각 데이터 소스의 변수 목록이 일치하지 않을 경우 변수 매칭 방법 및 포함할 변수에 따라 "Fields"가 결정됩니다.

- **데이터 소스 순위** 리스트에서, 소스들의 순서를 변경하고, **주 소스**를 지정합니다.
- **변수 매칭 방법**을 선택합니다. 일반적으로 변수명이 같은 소스일 경우 **변수명**을, 그렇지 않은 소스일 경우 **변수 위치**를 선택하면 됩니다.
- **포함할 변수**를 결정합니다.
- **소스 인덱스**를 추가할 지 여부를 결정합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	

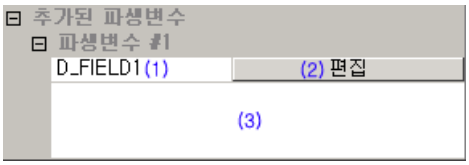
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
데이터 소스 순위	데이터 소스	추가할 데이터 소스의 리스트로, 순위 조정을 통해 소스의 우선순위 조정과 **주 소스** 결정을 할 수 있습니다.	기타 UI	사용자 인터페이스 참조.
선택사항	변수 매칭 방법	변수들을 추가시키는 방법입니다. **변수명** 선택 시에는 데이터를 변수명을 기준으로 추가시킵니다. 따라서 변수명이 같으면 같은 열에, 다르면 새로운 열에 데이터를 추가시킵니다. **변수위치** 선택 시에는 데이터를 열 인덱스에 따라 추가시킵니다.		변수명, 변수위치
	포함할 변수	결과에 포함할 변수를 지정합니다. **주 소스**만을 선택할 경우, 주 소스에 있는 변수명을 가진 **변수**의 데이터만을 결과에 추가시킵니다. **모든 소스**를 선택할 경우, 모든 소스에 있는 변수들과 그 데이터를 결과에 추가시킵니다.	필수	주 소스만, 모든 소스
	소스 인덱스	결과 데이터가 어느 소스에서 추가되었는지를 나타내는 **인덱스 열**을 추가시킬지 유무를 결정합니다. 선택될 경우, **소스 인덱스** 속성에 부여된 변수명을 가진 데이터 소스 변수가 추가됩니다. 이 변수명은 **소스 인덱스** 속성창에서 수정 가능합니다.	체크박스	

3.2.14 파생변수 노드

데이터를 정제하는 과정에서 경우에 따라 새로운 변수를 생성해야 할 경우가 있습니다. 예를 들어 주민등록번호를 이용하여 나이를 추출하고 이를 새로운 변수로 생성하는 것입니다. 이러한 작업을 수행하기 위하여 **파생변수 노드**를 사용할 수 있습니다.

사용법

필요한 파생변수 개수만큼 파생변수 추가 버튼을 눌러 추가합니다. 다음과 같은 속성이 노드의 속성창에 추가됩니다.



- 파생 변수명을 변경하려면 (1)을 클릭하고 변경할 이름을 입력합니다.
- (2)번 버튼을 누르면 **수식 편집기**가 실행됩니다. **수식 편집기**를 이용하여 파생변수에 대한 수식을 입력합니다.
- (3)은 수식열림창으로 이를 통하여 입력된 수식을 확인할 수 있습니다. 그러나, (3)을 이용하여 직접 수식을 편집할 수 없습니다. 꼭 (2)의 편집 버튼을 사용하여야 합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
파생변수 추가 / 삭제	파생변수 추가	추가로 파생변수를 생성하려면 이 버튼을 누릅니다.	버튼	
	마지막 파생변수 삭제	현재 추가된 파생변수 중 마지막 파생변수를 삭제합니다. 파생변수가 하나만 있을 경우 삭제되지 않습니다.	버튼	
	파생변수 편집	선택되지 않은 파생변수를 삭제합니다.	버튼	
추가된 파생변수	파생변수 #00	현재 추가하고자 하는 파생변수의 목록이 나타납니다. 기본적으로 새로 추가된 파생변수는 D_FIELD00 (00 은 인덱스)의 명칭을 가지며 변경도 가능합니다.		
	편집	이 버튼을 누르면 파생변수를 편집할 수 있는 에디터 창이 띄워집니다. 이를 이용하여 쉽게 파생변수를 편집할 수 있습니다.	버튼	수식 편집기 참조.
	수식 열람창	해당 파생변수에 정의된 수식을 열람할 수 있습니다.		

3.2.15 표본추출 노드

대용량의 데이터를 이용한 모델링의 경우 계산 시간 및 비용이 증가하기 때문에 데이터의 일부를 이용하여 초기분석을 함으로써 시간과 비용을 줄일 수 있습니다. 이 때 **표본추출 노드**를 사용합니다.

사용법

- 표본추출방법을 선택합니다.
- **선택여부** 속성에서 추출된 표본을 사용할 것인지 아니면 사용하지 않을 것인지를 지정합니다.
- 표본크기를 결정할 방법을 선택합니다.
- **표본크기 지정방법**에 따라 백분율 혹은 표본 개수를 입력합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
선택사항	표본 추출방법	전체 데이터 중 일부(표본)를 추출할 방법을 지정합니다.	필수	임의추출, 계통추출, 층화추출, 처음부터, 층화추출(등분)
	선택여부	추출된 표본 데이터를 사용할 지 아니면 추출되지 않은 데이터를 사용할 지를 지정합니다.	필수	선택, 배제.
	이산형 변수	표본 추출 방법으로 층화추출(등분)을 선택할 시 선택하는 옵션. 선택한 이산형 변수에 대해서 이산형 변수가 갖는 각 값에 대해서 동일한 개수의 sample 이 추출됩니다.		
표본크기	표본크기 지정방법	표본크기를 결정할 방법을 지정합니다.	필수	백분율, 개수.
	표본크기 (백분율)	**표본크기 지정방법**이 **백분율**일 경우 활성화 됩니다.		0 ~ 100 사이의 실수
	표본크기	**표본크기 지정방법**이 **개수**일 경우 활성화됩니다.		정수.

	(개수)	추출하고자 하는 표본의 크기를 정수로 입력합니다. 만약, 전체 데이터 건수보다 클 경우 전체 데이터 수로 변경되어 처리됩니다.		

NOTE: 표본추출 방법에 대한 설명입니다.

- 임의추출 : 난수표 등을 이용하여 무작위로 데이터를 추출하는 방법.
- 계통추출 : 모든 모집단에 순서를 매기고, 처음 하나의 표본을 임의추출하고, 그 후 일정한 간격으로 추출하는 방법.(단, 전체 데이터의 절반(50%) 이하의 표본수만 추출 가능합니다.)
- 층화추출 : 모집단을 동질적인 소집단들로 층화시키고 그 집단의 크기에 따라 임의 추출하는 방법

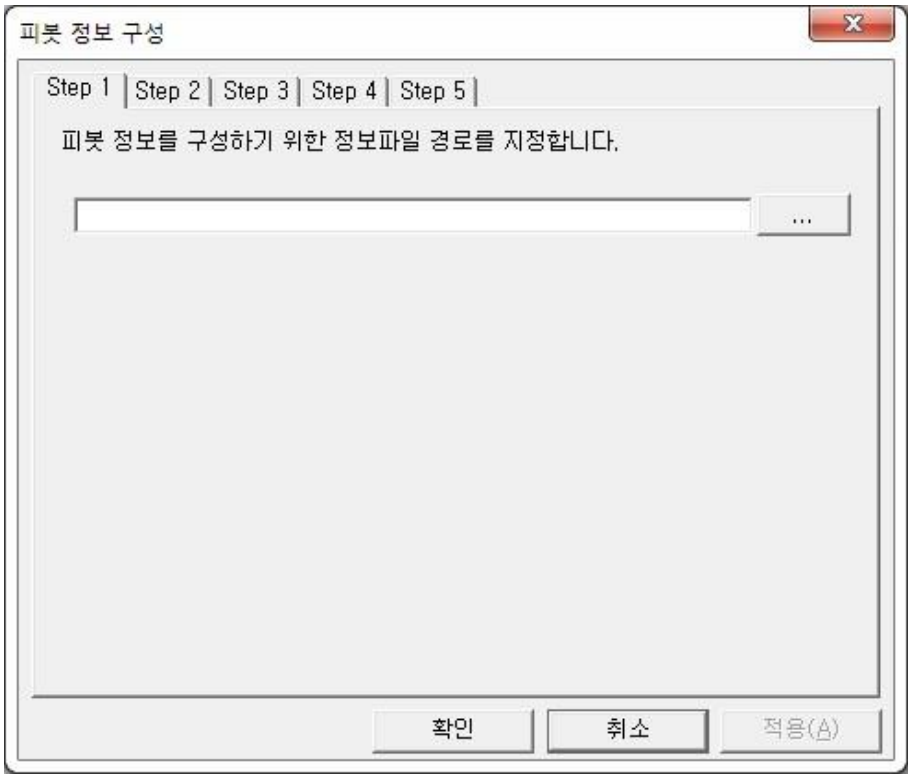
3.2.14 전처리 피봇 노드

피봇 노드는 데이터를 열(column)별, 행(row)별로 **그룹화하여 분석**하는 기능을 하는 노드입니다.

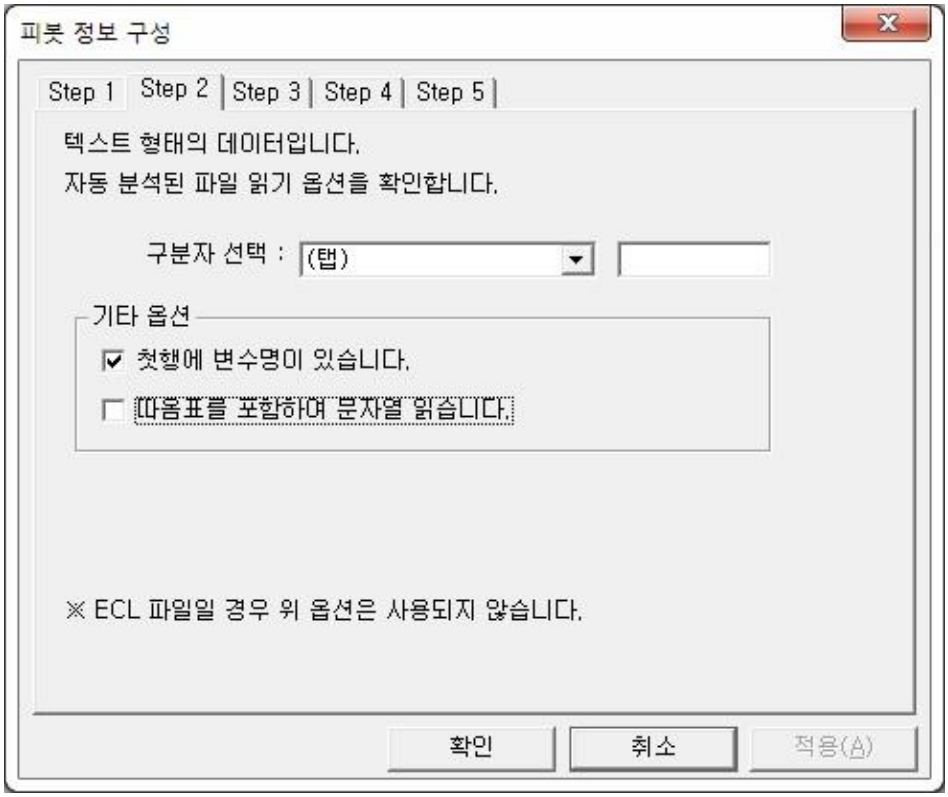
사용법

속성창에 있는 피봇팅 정보 구성 버튼을 클릭하면 다음과 같은 속성 변경창이 뜨게 됩니다.

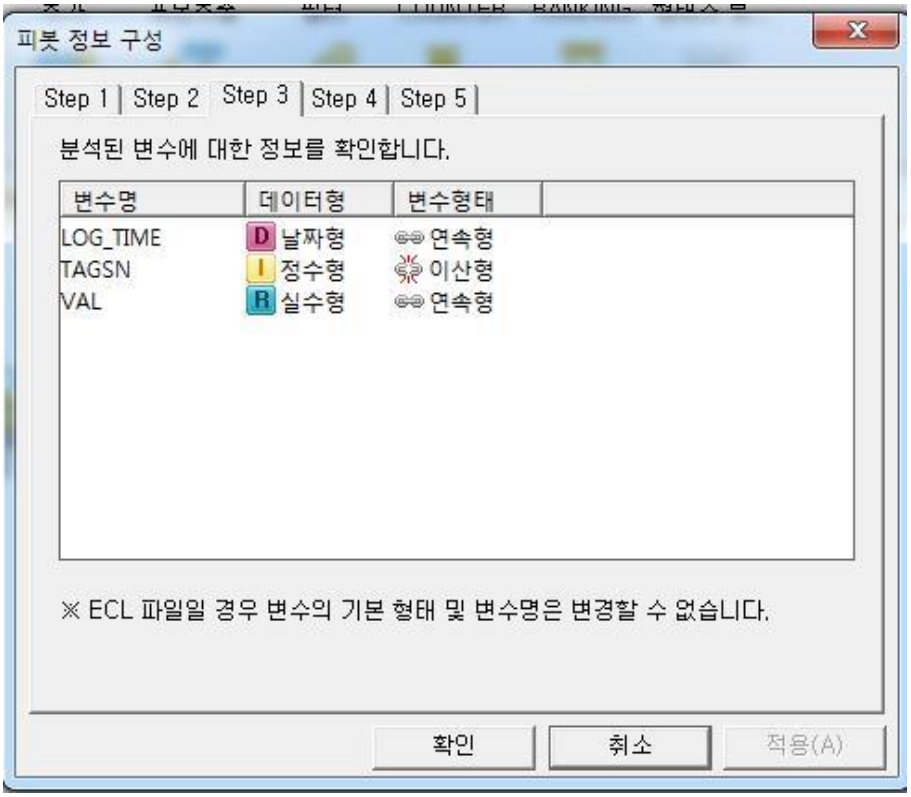
데이터 소스가 되는 파일을 지정합니다.



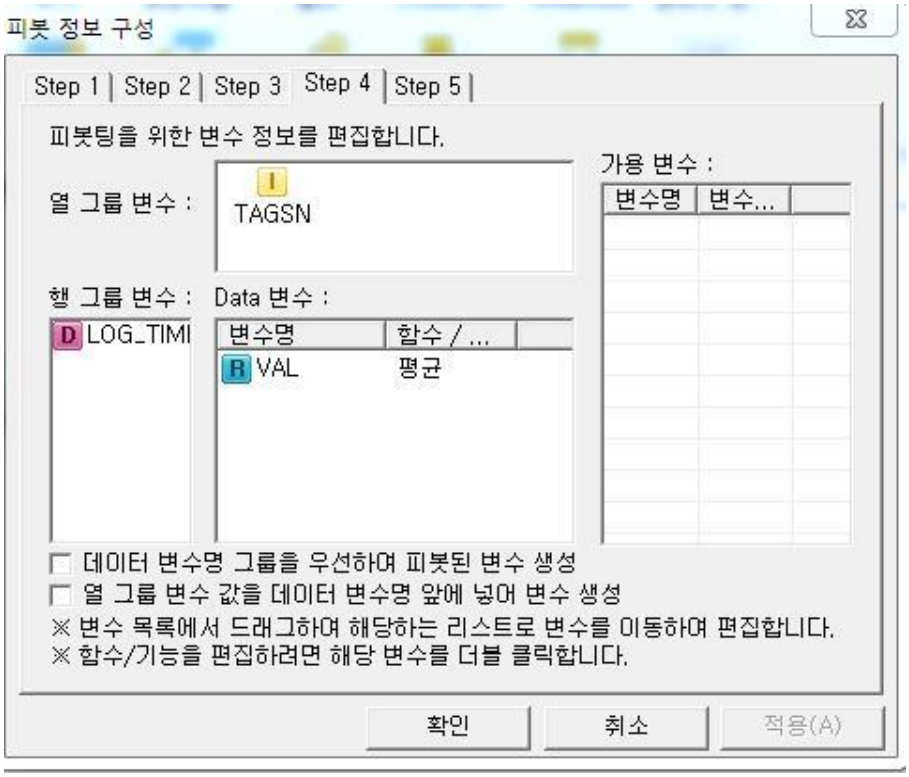
데이터 저장 형식을 파악하고 데이터 읽기 옵션을 선택합니다.



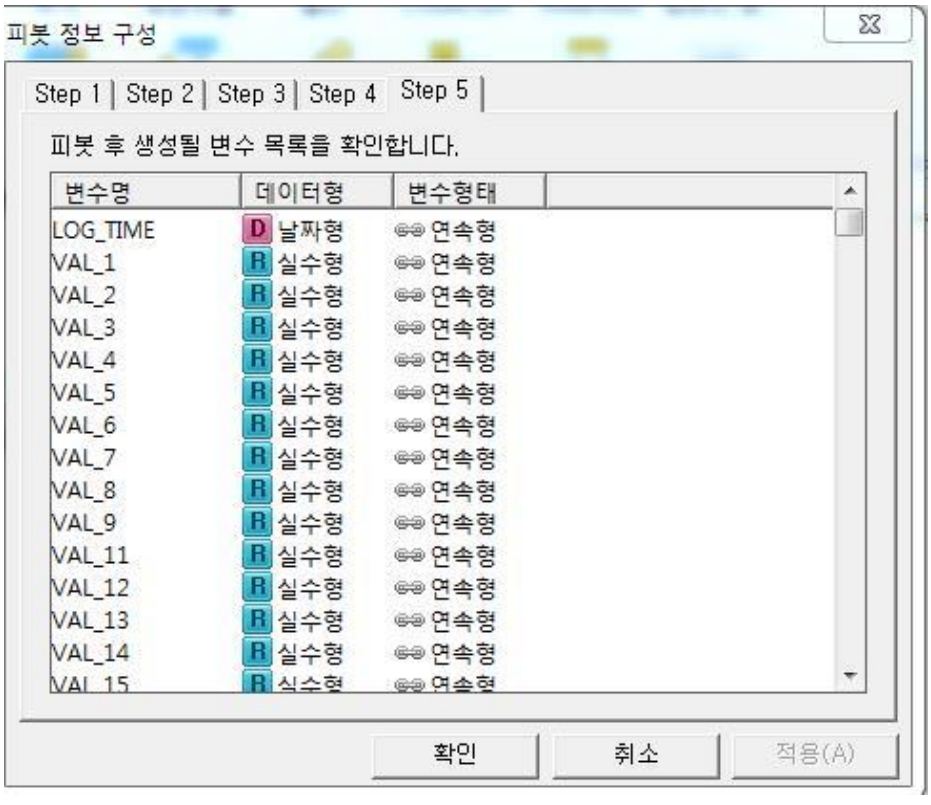
변수형태를 지정합니다.



데이터 변수와 피봇팅에 기준이 될 행, 열 변수(이산형 변수)를 지정합니다. 데이터 변수, 행, 열 변수의 조합만큼 데이터 변수가 분석되게 됩니다. 분석하고자 하는 값을 선택합니다.



피봇 후 생성된 변수에 대한 목록을 확인합니다.



속성창에 있는 피봇 정보 인터페이스는 다음과 같습니다.

- **피봇팅 정보 구성** 버튼을 클릭하여 속성을 지정합니다.
- 스트림을 실행합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
선택사항	피봇팅 정보 구성	피봇팅을 위한 정보를 대화상자를 통해 구성합니다.	버튼	

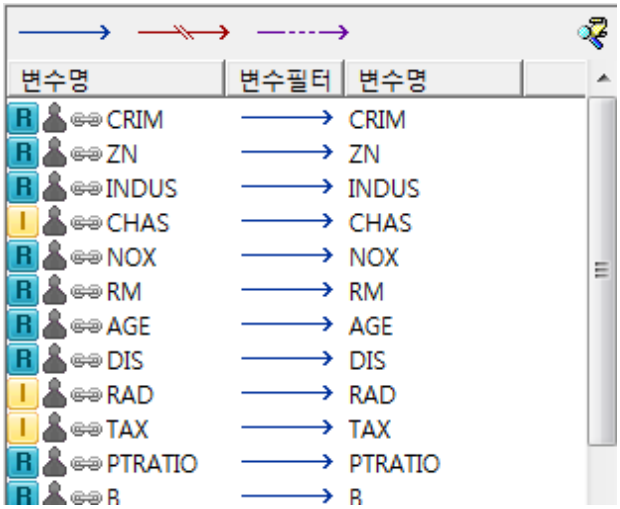
3.2.17 필터 노드


필터

필터 노드를 이용하여 사용하지 않을 변수를 걸러내거나 변수명을 변경할 수 있습니다. 변수를 걸러내면 이 노드 이후의 노드에서 걸러진 변수가 나타나지 않으며 걸러진 변수를 사용할 수 없도록 설정됩니다. 또한 모델링에는 사용하지 않고 결과에만 출력하기 위해 모델링에서만 해당 변수를 건너뛰도록 설정할 수 있습니다. 변수를 다중으로 선택하여 많은 변수를 한꺼번에 필터링 할 수 있습니다. Shift 키, Ctrl 키를 이용하여 다중 선택이 가능합니다.

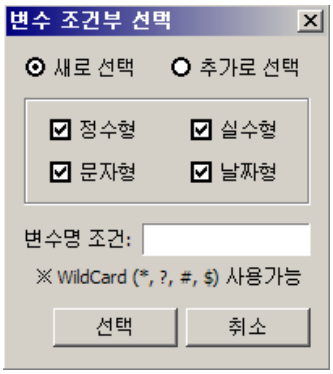
사용법

필터 노드는 다음과 같은 속성변경 인터페이스를 갖습니다.



필터 노드의 속성변경 인터페이스는 속성을 변경할 수 있는 기능 버튼 부분과 현재 변수의 상황을 볼 수 있는 변수정보 부분으로 나눠져 있습니다. 변수정보 부분을 통하여 변수명 변경 및 속성을 변경할 수 있습니다.

(1) 기능 버튼

버튼	기능
→	선택된 변수를 사용하도록 설정합니다.
→	선택된 변수를 사용하지 않도록 설정합니다.
→	선택된 변수를 모델링시에만 사용하지 않도록 설정합니다.
	 버튼 클릭 시 다음과 같은 창이 생성되어 변수형 또는 변수명과 같은 조건에 맞는 변수를 선택하게 합니다.

(2) 변수 정보

컬럼	설명
변수명	현재 변수의 상태를 나타냅니다. 즉, 변수의 수학적 형태, 독립/종속 여부, 통계학적 형태를 아이콘으로 나타내고, 현재 변수명을 나타냅니다.
변수필터	변수를 사용함 / 사용안함 / 건너뜀 등 속성을 나타내며, 더블 클릭함으로써 변경할 수도 있습니다.
변수명 (변경)	변수명을 변경하고자 할 경우 이 행을 클릭하면 에디트 상자가 나타나며, 변경하고자 하는 변수명을 입력하면 변수명이 변경됩니다.

- 변경하고자 하는 변수를 변수 정보창에서 선택합니다.
- 기능 버튼을 이용하여 변수의 형태를 변경하거나 더블 클릭하여 변경합니다.

속성

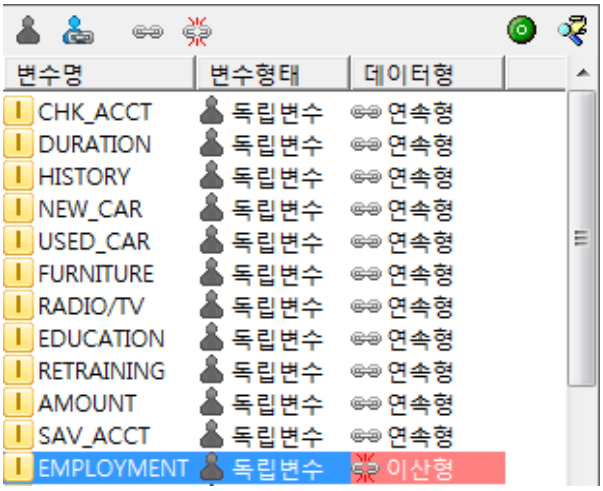
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 입력합니다.	선택	
변수필터	변수필터	변수필터를 변경할 수 있는 사용자 인터페이스를 제공합니다. 이 속성을 이용하여 변수의 사용 여부를 쉽게 변경할 수 있습니다.	기타 UI	사용자 인터페이스 참조.

3.2.18 형태변경 노드

형태변경 노드를 이용하여 변수의 형태를 변경할 수 있습니다. **독립변수 / 종속변수, 연속형 / 이산형** 등의 변수형태를 변경할 수 있으며, 모델링 시 **종속변수**를 필요로 하는 경우 **형태변경 노드**를 이용하여 변수형태를 변경하여야 합니다. 변수를 다중으로 선택하여 한번에 변경할 수 있습니다. Shift 키, Ctrl 키를 이용하여 다중 선택이 가능합니다.

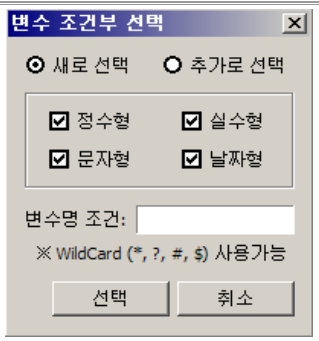
사용법

형태변경 노드는 다음과 같은 속성변경 인터페이스를 갖습니다.



형태변경 노드의 속성변경 인터페이스는 속성을 변경할 수 있는 기능 버튼 부분과 현재 변수의 상황을 볼 수 있는 변수정보 부분으로 나눠져 있습니다. 변수정보 부분을 통하여 변수명, 변수의 수학적 형태, 독립/종속 여부, 통계학적 형태를 볼 수 있으며, 기능 버튼을 이용하여 이들의 속성을 변경할 수 있습니다.

(1) 기능 버튼

버튼	기능
	선택된 변수의 형태를 **독립변수**로 변경합니다.
	선택된 변수의 형태를 **종속변수**로 변경합니다.
	선택된 변수의 통계학적 형태를 **연속형**으로 변경합니다.
	선택된 변수의 통계학적 형태를 **이산형**으로 변경합니다.
	변수 형태를 초기화하는 버튼입니다.
	 버튼 클릭 시 다음과 같은 창이 생성되어 변수형 또는 변수명과 같은 조건에 맞는 변수를 선택하게 합니다.

(2) 변수 정보

컬럼	설명
변수명	변수명 및 수학적 형태를 표시합니다. 변수의 수학적 형태는 변경할 수 없습니다.
변수형태	모델링 시 사용하는 독립 / 종속변수의 형태를 변경합니다. 마우스 왼쪽 버튼을 더블 클릭하여 변경할 수도 있습니다.
데이터형	변수의 통계학적 형태를 표시합니다. 붉은색으로 표시되어 있는 부분은 변경할 수 없음을 나타내며, 변수의 수학적 형태가 문자형 혹은 날짜형과 같은 이산형일 경우 변경할 수 없기 때문에 붉은색으로 표시됩니다. 마우스 왼쪽 버튼을 더블 클릭하여 변경할 수도 있습니다.

- 변경하고자 하는 변수를 변수 정보창에서 선택합니다.
- 기능 버튼을 이용하여 변수의 형태를 변경하거나 또는 컬럼을 더블 클릭하여 변경합니다.

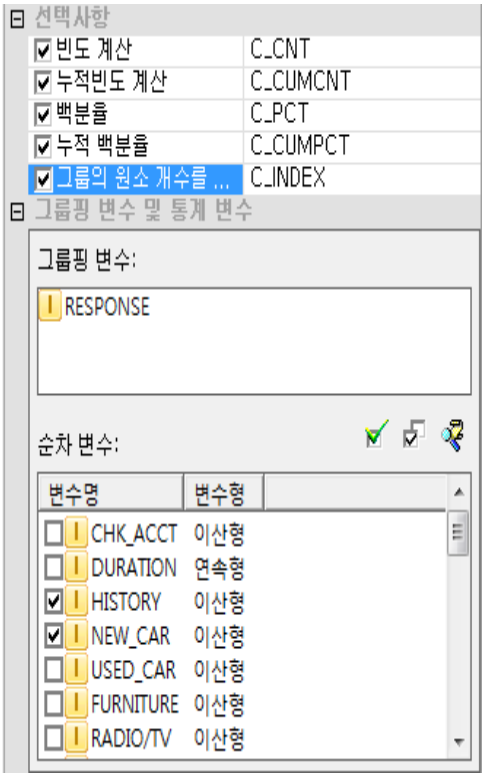
속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
변수형태	변수형태	변수형태를 변경할 수 있는 사용자 인터페이스를 제공합니다. 이 속성을 이용하여 변수의 형태를 쉽게 변경할 수 있습니다.	기타 UI	사용자 인터페이스 참조.

3.2.19 COUNTER 노드

지정된 그룹핑 변수에 따라 그룹별로 Counting 한 결과를 파생변수로 추가합니다.

사용법



Counting 할 그룹핑 변수를 선택합니다. 그리고 선택사항 선택 시 선택한 옵션에 따라 파생변수가 생성됩니다.

11	12	13	14	15
C_CNT	C_CUMCNT	C_PCT	C_CUMPCT	C_INDEX
3846	3846	45,08792	45,08792	1841
4684	8530	54,91208	100	4683
4684	8530	54,91208	100	2004
3846	3846	45,08792	45,08792	3725
4684	8530	54,91208	100	4227

속성

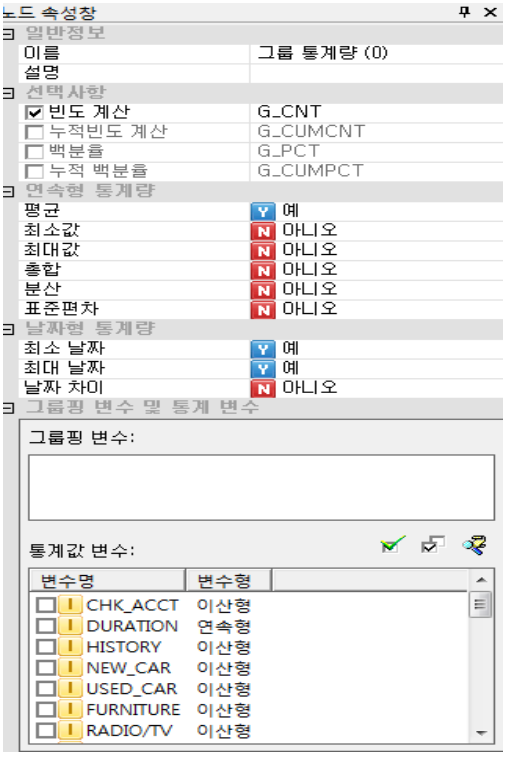
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
선택사항	선택사항	선택사항에 선택된 값의 파생변수를 생성합니다. (빈도계산, 누적빈도계산, 백분율, 누적백분율, 그룹 내 원소번호를 계산하여 변수로 추가합니다.)		
그룹핑 변수 및 순차 변수	그룹핑 변수	Counting 할 그룹핑 변수를 지정합니다. 그룹핑 변수는 이산형만 가능합니다.	필수	

	순차변수	선택사항에서 '그룹 내 원소번호' 선택 시 순차변수의 값 순서대로 원소번호를 부여합니다.	선택	

3.2.20 그룹 통계량 노드

지정된 그룹핑 조건에 따라 선택된 변수의 통계량을 파생변수로 새롭게 생성합니다.

사용법



그룹핑 변수를 지정하여 연속형 통계량과 날짜 통계량을 구할 수 있습니다. 그리고 선택사항 선택 시 선택된 사항에 따라 파생변수가 추가로 생성됩니다.

속성

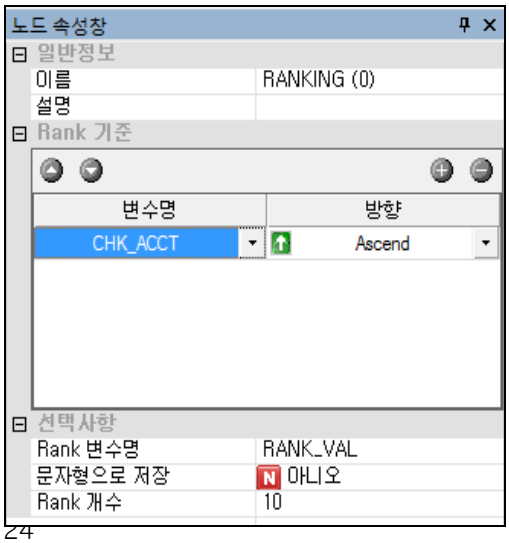
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수	선택	

		있습니다.		
선택사항	선택사항	선택사항에 선택된 값에 대해 파생변수를 생성합니다.	빈도계산, 누적빈도계산, 백분율, 누적백분율,	
통계량	연속형통계량	평균, 최소값, 최대값, 총합, 분산, 표준편차를 구하여 파생변수로 생성합니다.	필수(최소 1 개 이상 선택)	
	날짜형통계량	최소 날짜, 최대 날짜, 날짜 차이를 구하여 파생변수로 생성합니다.	선택	
그룹핑 변수 및 통계 변수	그룹핑 변수	그룹핑 변수를 Drag & Drop 을 통해 지정합니다. 그룹핑 변수는 이산형만 가능합니다.	필수	
	통계값 변수	그룹핑변수에 의한 통계값을 확인할 변수를 선택합니다. 통계값 변수는 연속형만 가능합니다.	필수	

3.2.21 RANKING 노드

RANKING 을 계산하여 파생변수로 추가하는 노드입니다.

사용법



변수를 지정하고 방향 지정 후 RANKING 작업을 합니다. RANKING 노드 실행 후에는 새로운 RANKING 된 변수가 추가됩니다.

	1	2	3	4	5	6
	A1	A6	A7	A8	이탈여부	RANK_VAL
1	24	42	213	921	0	5
2	12	191,4	554	1082,1	1	10
3	23	37,8	33	909	1	5
4	22	128,4	215	990	0	10
5	56	102	421	768,6	1	10
6	30	36,5	515	947	1	5
7	43	93,5	355	1067,6	1	9
8	19	70,8	332	881,4	0	8

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
선택사항 그룹핑 변수 및 통계 변수	Rank 기준	Ranking 할 변수와 방향을 지정합니다	필수	
	Rank 변수명	Ranking 후 새로 생성되는 변수명입니다.	필수	
	문자형으로 저장	문자형으로 저장 여부를 묻습니다.	선택	
	Rank 개수	Rank 개수를 지정합니다. 2 미만의 값이 입력되면 Rank 개수는 전체 데이터 개수로 간주됩니다.	필수	

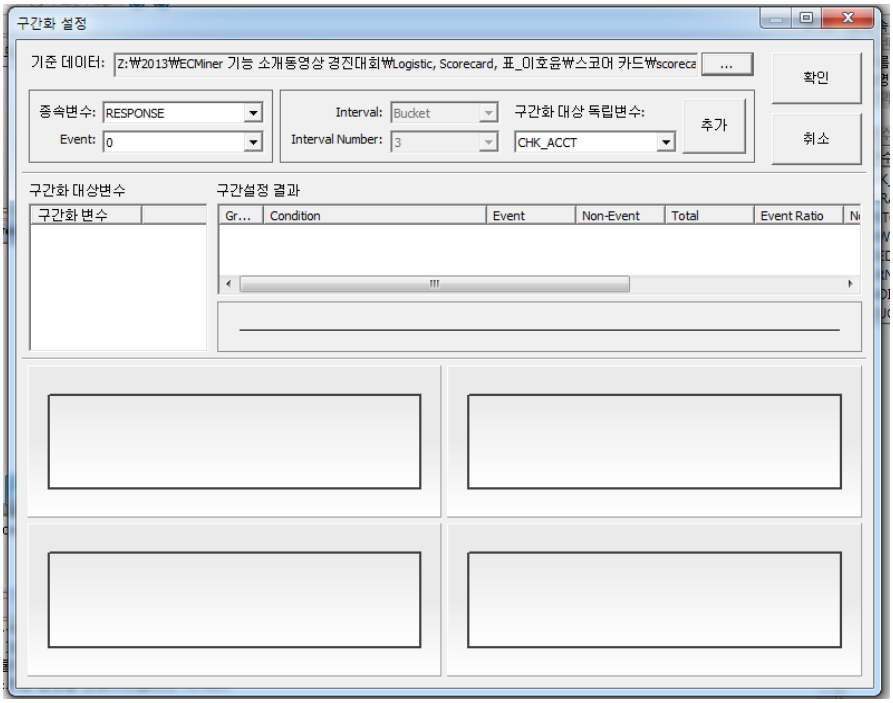
3.2.22 구간화 노드

구간화 노드는 종속변수에 미치는 영향도에 따라 독립변수를 여러 개의 구간으로 나누는 노드입니다.

사용법

- 구간화 할 데이터를 형태변경을 통해 종속변수를 지정한 후 ECL 파일로 저장합니다. 구간화를 수행하기 위해서 종속변수는 Binary 형태만 가능합니다.
- 구간화 편집 버튼을 누르면 아래와 같은 창이 나타납니다.

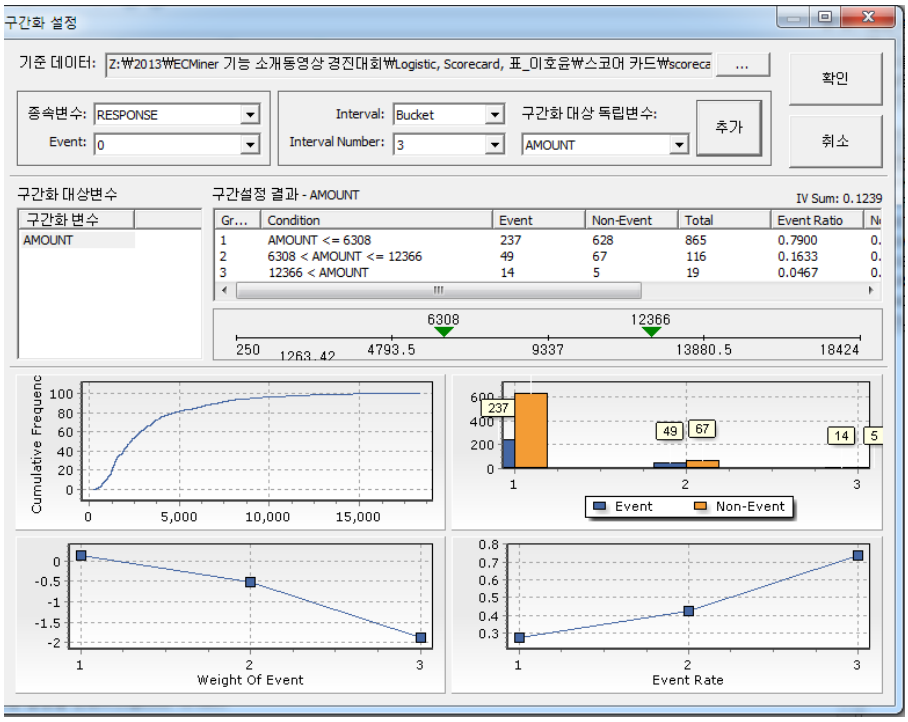
▪ 기준데이터를 앞서 저장한 ECL 파일을 불러옵니다.



▪ 추가 버튼을 누르면 아래와 같은 차트 및 통계량이 나타납니다.

▪ <Note> Interval 과 Interval Number 는 연속형 변수에만 해당합니다.

연속형 변수일 경우 화면



Event : 2 가지 종속변수의 값 중에서 어떤 것을 Event 로 인식할 지를 선택합니다.

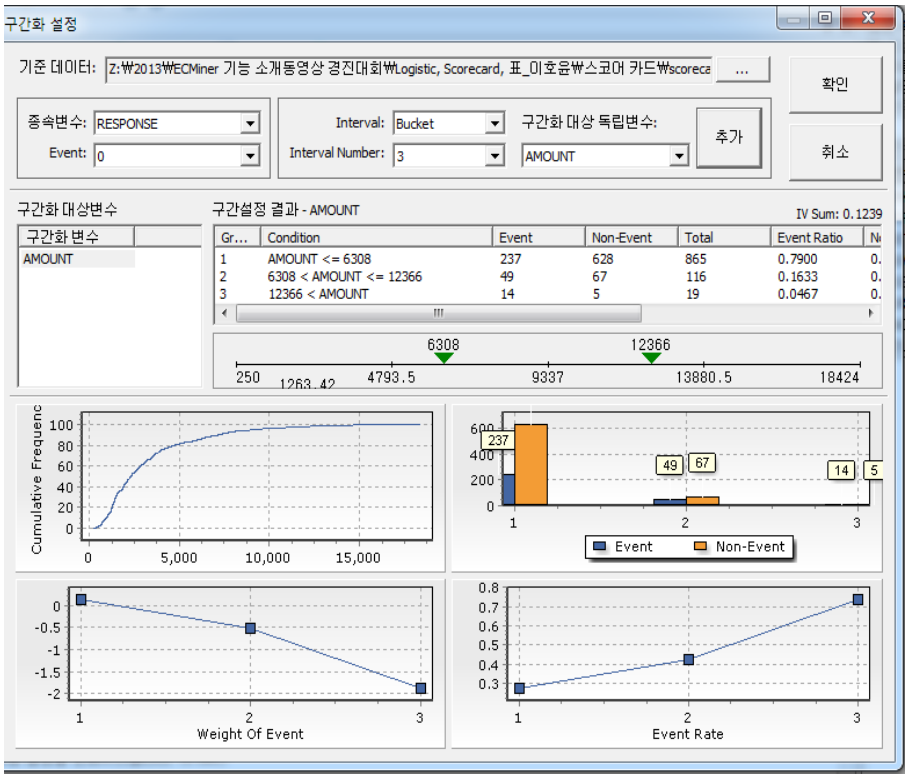
구간화 대상변수: 여러 개의 독립변수에 대해 각각 구간화를 실행합니다.

Interval : 구간의 간격이 동일한 Bucket 과 구간의 수가 동일한 Quantile 방식 중 선택합니다.

Interval Number : 구간 수를 설정합니다.

이산형 변수는 Interval, Interval Number 설정 없이 바로 Run 버튼을 누르면 모든 속성을 각각의 그룹으로 생성합니다.

이산형 변수일 경우 화면



구간수정결과에서 2 개 이상의 구간을 선택한 후 마우스 오른쪽 버튼을 클릭하면 팝업메뉴가 나타납니다.

선택된 구간 합치기: 2 개 이상의 구간을 병합합니다.

연속형 데이터는 인접 구간끼리만 병합할 수 있습니다.
이산형 데이터는 인접 구간이 아니더라도 병합할 수 있으며 조건은 or 로 변경됩니다.
(예: A2=a 와 A2=b 를 병합하면 A2=a or b)

구간 나누기: 1 개의 구간을 2 개 이상으로 분할합니다.

연속형 데이터는 직선상에서 마우스 오른쪽 버튼을 클릭하면 분할할 수 있는 팝업메뉴가 나타납니다.
이산형 데이터는 모든 속성을 각각의 그룹으로 분할합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
선택사항	구간화 편집	구간화를 위한 창을 불러옵니다.	필수	버튼

3.3 차트 노드

차트 노드는 데이터를 차트로 표현하기 위한 노드입니다. 효과적인 데이터 마이닝을 수행하려면 데이터의 시각화가 중요합니다. **ECMiner™**에서는 효과적인 데이터 표현을 위해 다양한 차트들을 제공해 줄 뿐만 아니라, 차트 자체에서도 세부적인 옵션사항을 두어 더욱 다양하고 효과적인 분석이 가능하도록 지원하고 있습니다.
ECMiner™에서는 일부차트에 데이터 슬라이드바 기능을 추가하였습니다. 데이터 슬라이드바 기능은 데이터와 차트를 매칭시킬 수 있는 역할을 하며, 이를 이용하여 차트 내에서 특정 데이터의 분포, 영역 등을 파악할 수 있습니다.

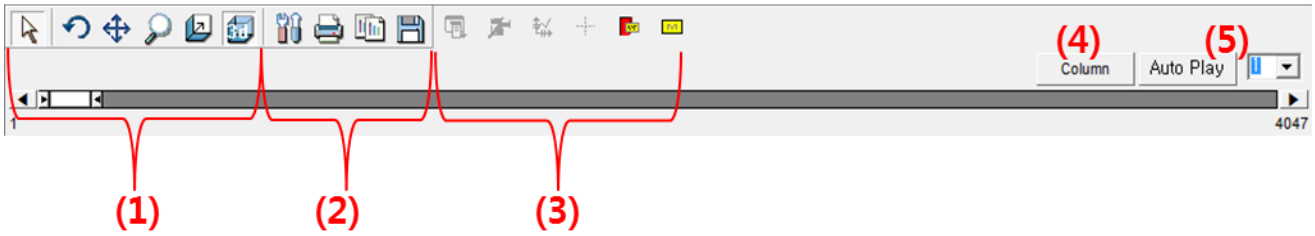
ECMiner™에서는 다음과 같은 차트들을 지원합니다. 괄호 안에 표시된 차트는 데이터 슬라이드바 기능이 지원되는 차트입니다.

- **3 차원차트(√)**
- **관리도(√)**
- **기본차트(√)**
- **매트릭스차트(√)**
- **바차트**
- **컨튜어차트**
- **컨트롤차트(√)**
- **통계차트**
- **파레토차트**
- **파이차트**
- **히스토그램**

- **산포차트**
- **확장 기본 차트(√)**
- **다변량 관리도(√)**

NOTE *차트 노드는 출력 형식이므로, 더 이상 다른 노드로 연결이 불가능 합니다.*

차트 기본 기능



(1) 3 차원차트 노드에서만 활성화 됩니다. T-Chart 의 기본 기능으로 첨부한 TeeChartPro.pdf 문서를 참조하시기 바랍니다.

(2) T-Chart 의 기본 기능으로 첨부한 TeeChartPro.pdf 문서를 참조하시기 바랍니다.

(3)  그래프의 표시된 점의 데이터를 추출하는 기능입니다. 실행할 경우 표시된 부분의 데이터를 가지고 오는 전처리 선택 노드가 실행됩니다. *그래프의 표시된 점의 양이 너무 많거나, 없으면 실행되지 않습니다.

 x 축과 y 축의 스케일을 똑 같이 맞추는 기능입니다.

 마우스를 따라 다니는 지시선을 활성/비활성 하는 버튼 입니다.

 그래프 상의 점을 선택 했을 때 마크를 표시하는 기능을 활성/비활성 하는 기능입니다.

 그래프 상의 전체 점의 마크를 표시하는 기능을 활성/비활성 하는 기능입니다.

(4) 

다른 변수 선택을 통해 차트상에서 표현된 변수 이외의 변수값을 확인할 수 있습니다.

(5) 

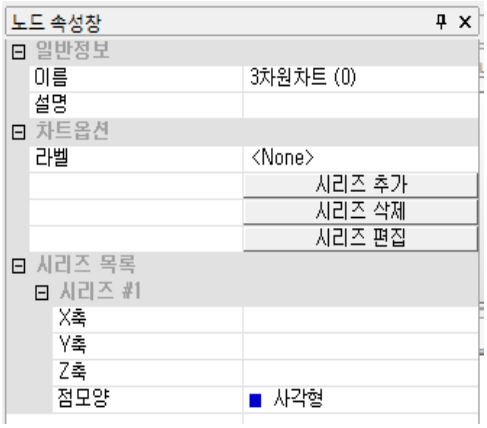
슬라이드 바를 사용하여 원하는 데이터 수를 선택한 후, 해당 버튼 실행을 통해 자동으로 그래프상에서 이동하면서 보여 줍니다. 이동 시간의 간격은 1, 5, 10 초 입니다.

(6) 슬라이드 바 그래프 상의 특정 점을 강조합니다. 특정 점의 개수를 지정합니다. 강조한 점의 위치를 변경합니다.

3.3.1 3 차원차트 노드

3 차원차트 노드는 데이터를 이용하여 일반적인 3 차원차트를 그릴 수 있는 노드입니다.

사용법



- 두 개 이상의 시리즈를 그릴 경우, 필요한 시리즈 개수만큼 **시리즈 추가** 버튼을 눌러 추가합니다.
- **시리즈 목록** 속성창에서 **X 축** 및 **Y 축**, **Z 축**으로 사용할 변수를 선택합니다.
- 선택 사항

  라벨 속성 변경

  점모양 속성 변경

속성

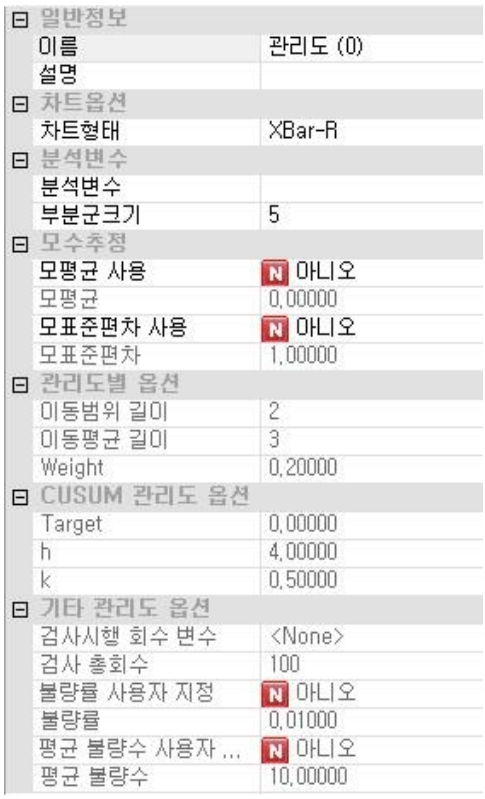
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
차트옵션	라벨	차트 상의 포인터에 표시될 라벨을 지정합니다.		

	시리즈 추가		여러 시리즈를 한 차트에 그리기 위해 사용됩니다. 버튼을 누르면 시리즈 선택 리스트가 추가됩니다.	버튼	
	시리즈 삭제		마지막에 추가된 시리즈가 삭제됩니다.	버튼	
	시리즈 편집		선택되지 않은 시리즈를 삭제합니다.	버튼	
시리즈 목록	시리즈 #00	X 축	X 축으로 사용할 변수를 지정합니다	.	
		Y 축	Y 축으로 사용할 변수를 지정합니다.		
		Z 축	Z 축으로 사용할 변수를 지정합니다.		
		점모양	차트에 그려질 점들의 모양을 지정합니다.	사각형, 원형, 삼각형, 역삼각형, 십자형, X 형,별모양, 다이아몬드, 작은점	

3.3.2 관리도 노드

관리도 노드는 산포에 대한 원인을 찾는 관리도를 그립니다.

사용법



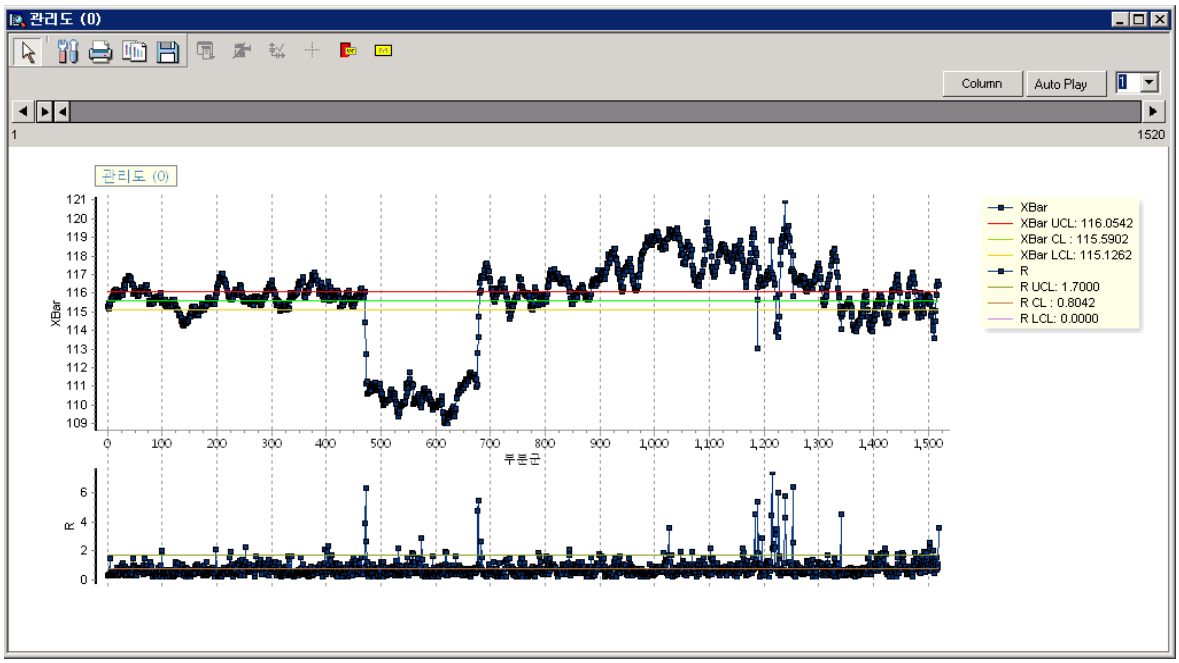
⊟ 일반정보	
이름	관리도 (0)
설명	
⊟ 차트옵션	
차트형태	XBar-R
⊟ 분석변수	
분석변수	
부분군크기	5
⊟ 모수추정	
모평균 사용	N 아니오
모평균	0,00000
모표준편차 사용	N 아니오
모표준편차	1,00000
⊟ 관리도별 옵션	
이동범위 길이	2
이동평균 길이	3
Weight	0,20000
⊟ CUSUM 관리도 옵션	
Target	0,00000
h	4,00000
k	0,50000
⊟ 기타 관리도 옵션	
검사시행 회수 변수	<None>
검사 총회수	100
불량률 사용자 지정	N 아니오
불량률	0,01000
평균 불량수 사용자 ...	N 아니오
평균 불량수	10,00000

관리도 노드는 변수를 선택하여 XBar-R, XBar-S, I-MR, MA, EWMA, CUSUM, P, NP, C, U chart 를 그릴 수 있습니다.

(1) XBar-R chart

사용법

- 차트 형태에서 그리고자 하는 XBar-R 차트를 선택합니다.
- X-Bar-R 차트를 그릴 변수를 지정합니다.
- 분석에 알맞은 부분군의 크기를 지정합니다.
- 모평균 사용과 모표준편차 사용을 예로 바꾸면 활성화된 창에 모평균과 모표준편차를 입력할 수 있습니다.
- 마지막으로 알맞은 부분군의 크기를 지정하여 실행버튼은 누르면 아래와 같은 그림을 볼 수 있습니다.



속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
차트옵션	차트형태	그리고자 하는 차트의 형태를 선택할 수 있습니다.	XBar-R, XBar-S, I-MR, MA, EWMA CUSUM, P, NP, C, U	
분석변수	분석변수	XBar-R 차트를 그릴 변수를 지정합니다.	필수	
	부분군크기	사용자가 정한 부분군의 크기를 입력합니다.	필수	
모수추정	모평균 사용	모평균을 사용하여 차트를 그릴지 결정합니다	예, 아니오	
	모평균	모평균 사용이 **'예'**일 경우 활성화됩니다. 알고 있는 모평균을 입력합니다.	실수	

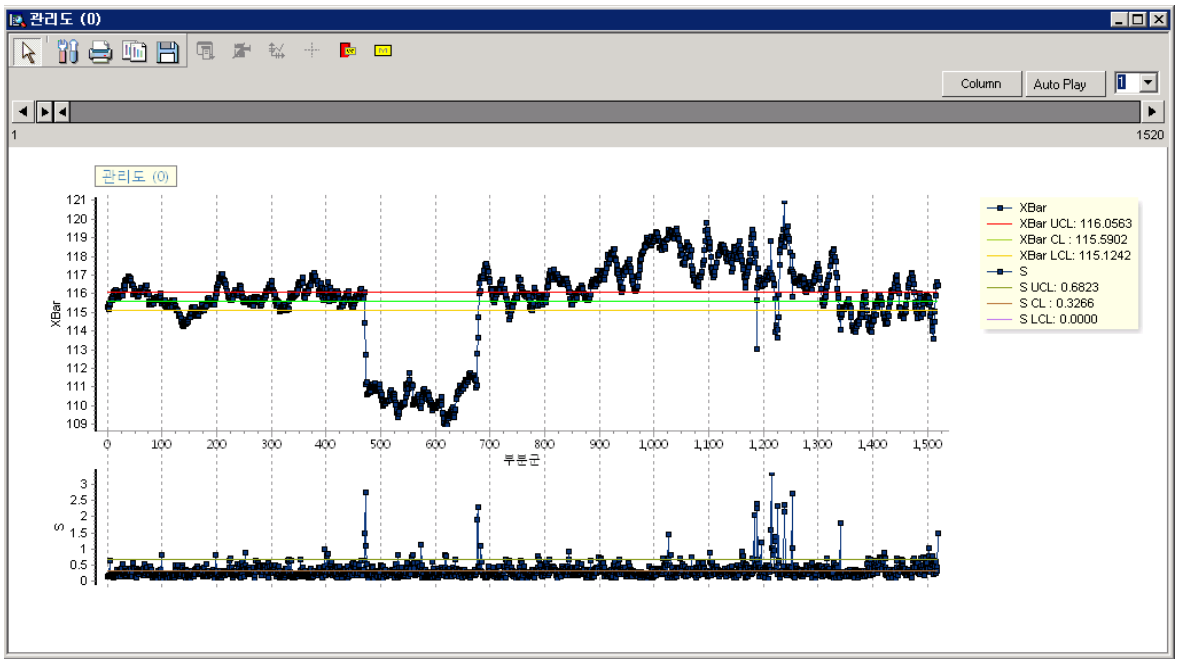
	모표준편차 사용	모표준편차를 사용하여 차트를 그릴지 결정합니다.	예, 아니오	
	모표준편차	모표준편차 사용이 **'예'**일 경우 활성화됩니다. 알고 있는 모표준편차를 입력합니다.	실수	

NOTE: 모평균과 모표준편차를 사용하는 것에 따라서 차트가 다른 모형으로 나옵니다. 모평균을 사용하면 중심값이 모평균으로 바뀌어서 파트가 그려지고, 모표준편차를 사용하면 차트의 형태가 달라지니 직접 확인해 보시는 것이 좋을 것 같습니다.

(2) XBar-S chart

사용법

- 차트형태에서 그리고자 하는 XBar-S 차트를 선택합니다.
- XBar-S 차트를 그릴 변수를 지정합니다.
- 분석에 알맞은 부분군의 크기를 지정합니다.
- 모평균 사용과 모표준편차 사용을 예로 바꾸면 활성화된 창에 모평균과 모표준편차를 입력할 수 있습니다.



속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
	차트형태	그리고자 하는 차트의 형태를 선택할 수 있습니다.	XBar-R, XBar-S, I-MR, MA, EWMA CUSUM, P, NP, C, U	
분석변수	분석변수	XBar-S 차트를 그릴 변수를 지정합니다.	필수	
	부분군크기	사용자가 정한 부분군의 크기를 입력합니다.	필수	
모수추정	모평균 사용	모평균을 사용하여 차트를 그릴지 결정합니다	예, 아니오	
	모평균	모평균 사용이 '예'일 경우 활성화됩니다. 알고 있는 모평균을 입력합니다.	실수	
	모표준편차 사용	모표준편차를 사용하여 차트를 그릴지 결정합니다.	예, 아니오	
	모표준편차	모표준편차 사용이 '예'일 경우 활성화됩니다. 알고 있는 모표준편차를 입력합니다.	실수	

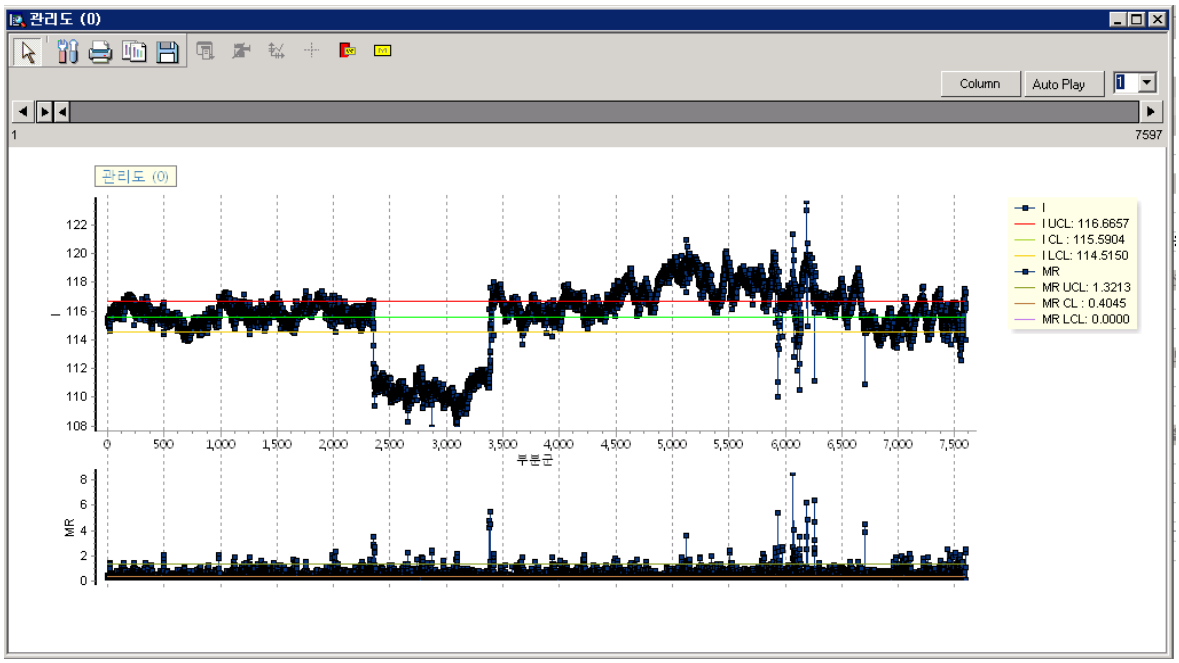
NOTE: XBar-S 차트는 XBar-R 차트와 비슷합니다. XBar-R 차트와의 차이점은 XBar-R 차트는 부분군의 크기가 2~5 사이일 때 사용하고, XBar-S 차트는 부분군의 크기가 6 이상일 때 사용합니다.

(3) I-MR

사용법

- 차트형태에서 그리고자 하는 I-MR 차트를 선택합니다.
- I-MR 차트를 그릴 변수를 지정합니다.
- 모평균 사용과 모표준편차 사용을 예로 바꾸면 활성화된 창에 모평균과 모표준편차를 입력할 수 있습니다.

▪ 이동범위 길이를 정합니다.



속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
	차트형태	그리고자 하는 차트의 형태를 선택할 수 있습니다.	XBar-R, XBar-S, I-MR, MA, EWMA CUSUM, P, NP, C, U	
분석변수	분석변수	I-MR 차트를 그릴 변수를 지정합니다.	필수	
모수추정	모평균 사용	모평균을 사용하여 차트를 그릴지 결정합니다	예, 아니오	
	모평균	모평균 사용이 **'예'**일 경우 활성화됩니다. 알고 있는 모평균을 입력합니다.	실수	

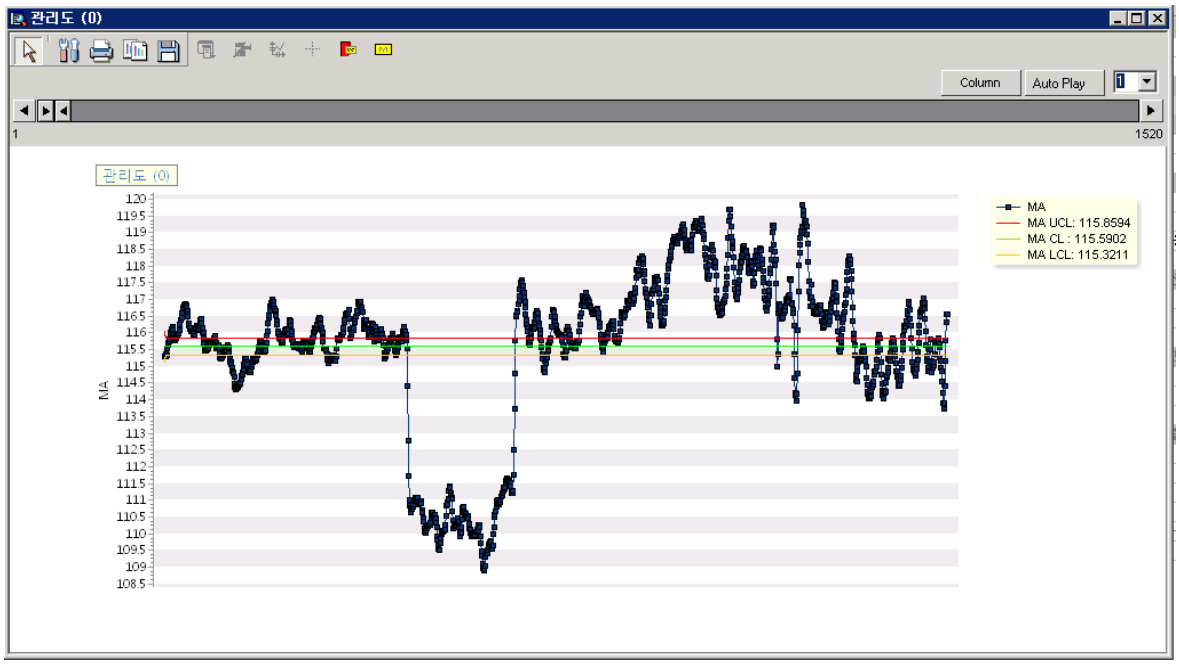
	모표준편차 사용	모표준편차를 사용하여 차트를 그릴지 결정합니다.	예, 아니오	
	모표준편차	모표준편차 사용이 '예'일 경우 활성화됩니다. 알고 있는 모표준편차를 입력합니다.	실수	
관리도별 옵션	이동범위 길이	이동범위의 길이를 입력합니다.	실수	

NOTE: 이동범위의 길이에 따라 관리한계선이 변하는 것을 알 수 있습니다.

(4) MA

사용법

- 차트형태에서 그리고자 하는 MA 차트를 선택합니다.
- MA 차트를 그릴 변수를 지정합니다.
- 분석에 알맞은 부분군의 크기를 지정합니다.
- 모평균 사용과 모표준편차 사용을 예로 바꾸면 활성화된 창에 모평균과 모표준편차를 입력할 수 있습니다.
- 이동평균 길이를 정합니다.



속성

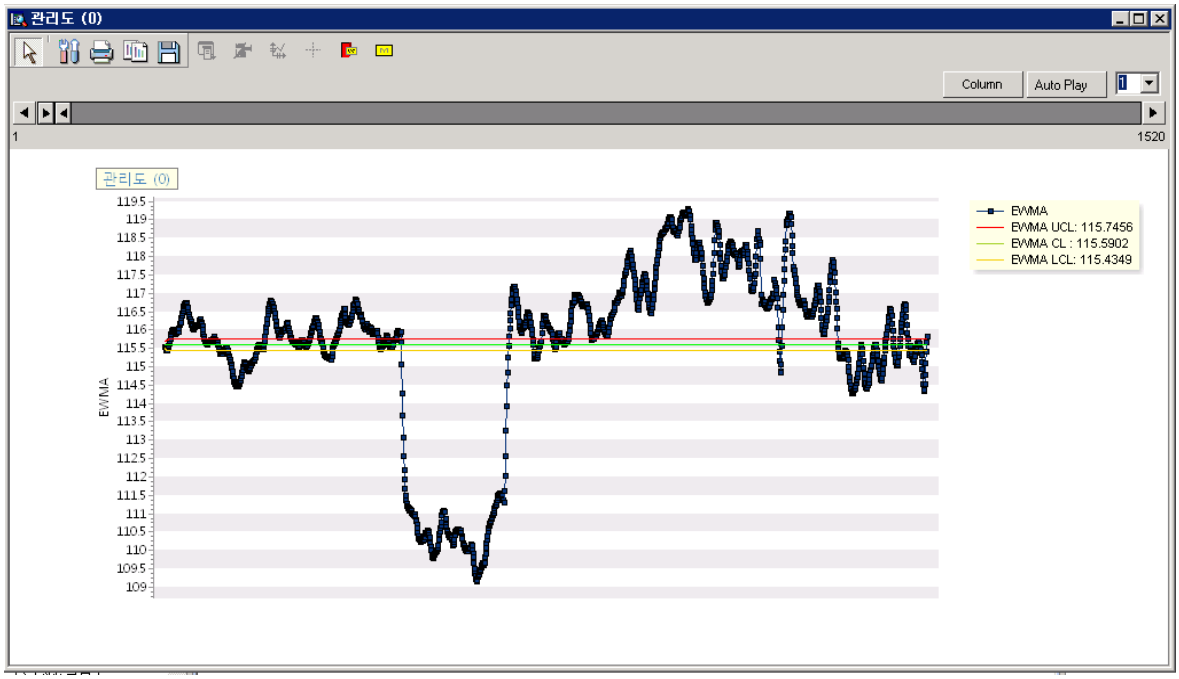
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
	차트형태	그리고자 하는 차트의 형태를 선택할 수 있습니다.	XBar-R, XBar-S, I-MR, MA, EWMA CUSUM, P, NP, C, U	
분석변수	분석변수	MA 차트를 그릴 변수를 지정합니다.	필수	
	부분군크기	사용자가 정한 부분군의 크기를 입력합니다.	필수	
모수추정	모평균 사용	모평균을 사용하여 차트를 그릴지 결정합니다	예, 아니오	
	모평균	모평균 사용이 **'예'**일 경우 활성화됩니다. 알고 있는 모평균을 입력합니다.	실수	
	모표준편차 사용	모표준편차를 사용하여 차트를 그릴지 결정합니다.	예, 아니오	
	모표준편차	모표준편차 사용이 **'예'**일 경우 활성화됩니다. 알고 있는 모표준편차를 입력합니다.	실수	
관리도별 옵션	이동평균 길이	이동평균의 길이를 입력합니다.	실수	

(5) EWMA(Exponentially Weighted Moving Average:지수 가중 이동평균 관리도)

사용법

- 차트형태에서 그리고자 하는 EWMA 차트를 선택합니다.

- EWMA 차트를 그릴 변수를 지정합니다.
- 분석에 알맞은 부분군의 크기를 지정합니다.
- 모평균 사용과 모표준편차 사용을 예로 바꾸면 활성화된 창에 모평균과 모표준편차를 입력할 수 있습니다.
- 가중치 값을 입력합니다.



속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
	차트형태	그리고자 하는 차트의 형태를 선택할 수 있습니다.	XBar-R, XBar-S, I-MR, MA, EWMA CUSUM, P, NP, C, U	
분석변수	분석변수	EWMA 차트를 그릴 변수를 지정합니다.	필수	

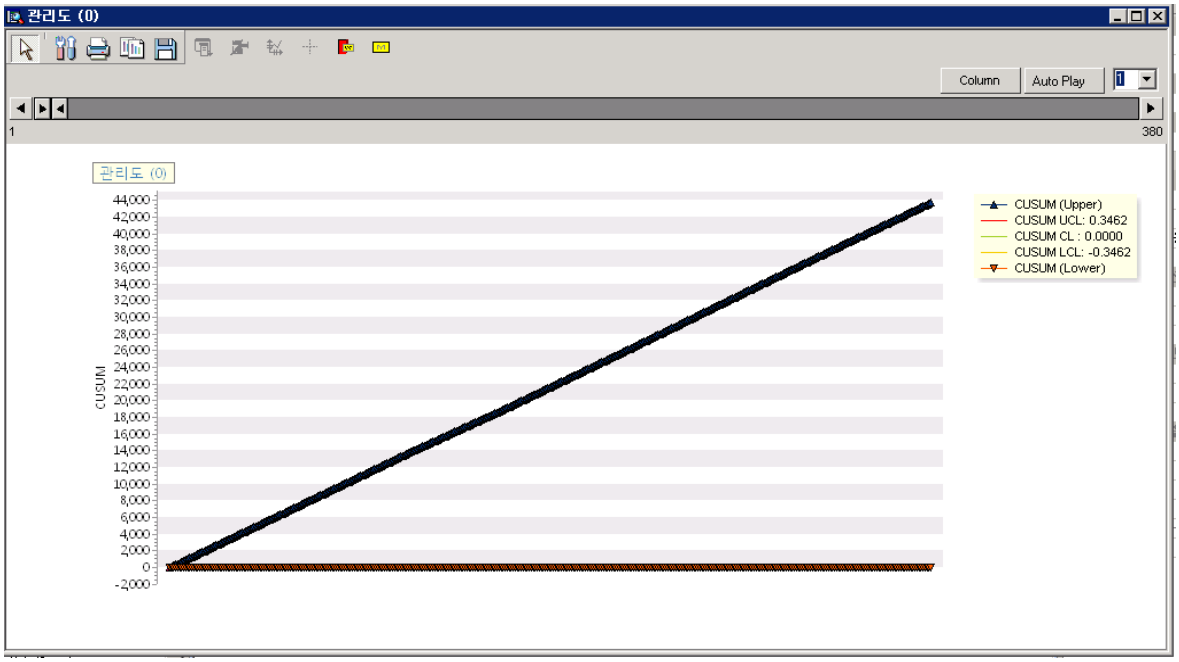
	부분군크기	사용자가 정한 부분군의 크기를 입력합니다.	필수	
모수추정	모평균 사용	모평균을 사용하여 차트를 그릴지 결정합니다	예, 아니오	
	모평균	모평균 사용이 **'예'**일 경우 활성화됩니다. 알고 있는 모평균을 입력합니다.	실수	
	모표준편차 사용	모표준편차를 사용하여 차트를 그릴지 결정합니다.	예, 아니오	
	모표준편차	모표준편차 사용이 **'예'**일 경우 활성화됩니다. 알고 있는 모표준편차를 입력합니다.	실수	
관리도별 옵션	Weight	가중치 값을 입력합니다.	실수	

NOTE: 가중치는 얻고자 하는 데이터 값을 계산할 경우, 과거 데이터와 최근 데이터의 비중을 달리 함으로써 데이터가 결과값에 미치는 영향을 비교하기 위해 어느 정도의 값으로 영향을 주어야 하는지를 결정하는 것입니다.

(6) CUSUM (Cumulative Summation:누적합관리도)

사용법

- 차트형태에서 그리고자 하는 CUSUM 차트를 선택합니다.
- CUSUM 차트를 그릴 변수를 지정합니다.
- 분석에 알맞은 부분군의 크기를 지정합니다.
- 모표준편차 사용을 예로 바꾸면 활성화된 창에 모표준편차를 입력할 수 있습니다.
- 차트를 그리기 적당한 target 과 h, k 값을 입력합니다.



속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
	차트형태	그리고자 하는 차트의 형태를 선택할 수 있습니다.	XBar-R, XBar-S, I-MR, MA, EWMA CUSUM, P, NP, C, U	
분석변수	분석변수	CUSUM 차트를 그릴 변수를 지정합니다.	필수	
	부분군크기	사용자가 정한 부분군의 크기를 입력합니다.	필수	
모수추정	모평균 사용	모평균을 사용하여 차트를 그릴지 결정합니다	예, 아니오	

CUSUM 관리도별 옵션	모평균	모평균 사용이 **'예'**일 경우 활성화됩니다. 알고 있는 모평균을 입력합니다.	실수	
	모표준편차 사용	모표준편차를 사용하여 차트를 그릴지 결정합니다.	예, 아니오	
	모표준편차	모표준편차 사용이 **'예'**일 경우 활성화됩니다. 알고 있는 모표준편차를 입력합니다.	실수	
	target	목표변수를 지정합니다.	실수	
	h	center line 과 Control Limit line 사이에 있어야 하는 표준편차의 수, 이상상태 신호가 발생하는 값입니다.	실수	
	k	부분군의 크기를 설정합니다.	실수	

(7) U chart

사용법

- 차트형태에서 그리고자 하는 U 차트를 선택합니다.
- U 차트를 그릴 변수를 지정합니다.
- 검사시행 횟수 변수를 선택합니다.
- 검사 총 횟수를 입력합니다.
- 평균불량수를 사용자가 지정하려면 예로 선택하고, 평균불량수 칸이 활성화되면 그 수를 입력합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
	차트형태	그리고자 하는 차트의 형태를 선택할 수 있습니다.	XBar-R, XBar-S, I-MR, MA, EWMA	

			CUSUM, P, NP, C, U	
분석변수	분석변수	U 차트를 그릴 변수를 지정합니다.	필수	
기타 관리도 옵션	검사시행 회수 변수	검사시행 회수를 나타내는 변수를 지정합니다.	필수	
	검사 총회수	검사를 시행한 총 회수를 입력합니다.	실수	
	평균 불량수 사용	평균 불량수를 사용자가 지정할지 결정합니다.	예, 아니오	
	평균 불량수	평균 불량수 사용자 지정을 **'예'**라고 했을 때 활성화됩니다. 활성화되면 평균 불량수를 입력합니다.	실수	

(8) P chart

사용법

- 차트형태에서 그리고자 하는 P 차트를 선택합니다.
- P 차트를 그릴 변수를 지정합니다.
- 검사시행 회수 변수를 선택합니다.
- 검사 총 회수를 입력합니다.
- 불량률을 사용자가 지정하려면 **'예'**로 선택하고, 불량률 칸이 활성화되면 그 비율을 입력합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
	차트형태	그리고자 하는 차트의 형태를 선택할 수 있습니다.	XBar-R, XBar-S, I-MR, MA,	

			EWMA CUSUM, P, NP, C, U	
분석변수	분석변수	P 차트를 그릴 변수를 지정합니다.	필수	
기타 관리도 옵션	검사시행 회수 변수	검사시행 회수를 나타내는 변수를 지정합니다.	필수	
	검사 총회수	검사를 시행한 총 회수를 입력합니다.	실수	
	불량률 사용자 지정	불량률을 사용자가 지정할 지 결정합니다.	예, 아니오	
	불량률	불량률 사용자 지정을 '예'라고 했을 때 활성화됩니다. 활성화되면 불량률을 입력합니다.	실수	

(9) NP chart

사용법

- 차트형태에서 그리고자 하는 NP 차트를 선택합니다
- NP 차트를 그릴 변수를 지정합니다.
- 검사시행 회수 변수를 선택합니다.
- 검사 총 회수를 입력합니다.
- 불량률을 사용자가 지정하려면 예로 선택하고, 불량률 칸이 활성화되면 그 비율을 입력합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
	차트형태	그리고자 하는 차트의 형태를 선택할 수 있습니다.	XBar-R, XBar-S, I-MR, MA,	

			EWMA CUSUM, P, NP, C, U	
분석변수	분석변수	NP 차트를 그릴 변수를 지정합니다.	필수	
기타 관리도 옵션	검사시행 회수 변수	검사시행 회수를 나타내는 변수를 지정합니다.	필수	
	검사총회수	검사를 시행한 총 회수를 입력합니다.	실수	
	불량률 사용자 지정	불량률을 사용자가 지정할 지 결정합니다.	예, 아니오	
	불량률	불량률 사용자 지정을 '예'라고 했을 때 활성화됩니다. 활성화되면 불량률을 입력합니다.	실수	

(10) C chart

사용법

- 차트형태에서 그리고자 하는 C 차트를 선택합니다.
- C 차트를 그릴 변수를 지정합니다.
- 평균 불량수를 사용자가 지정하려면 예로 선택하고, 평균 불량수 칸이 활성화되면 그 수를 입력합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
	차트형태	그리고자 하는 차트의 형태를 선택할 수 있습니다.	XBar-R, XBar-S, I-MR, MA, EWMA CUSUM, P,	

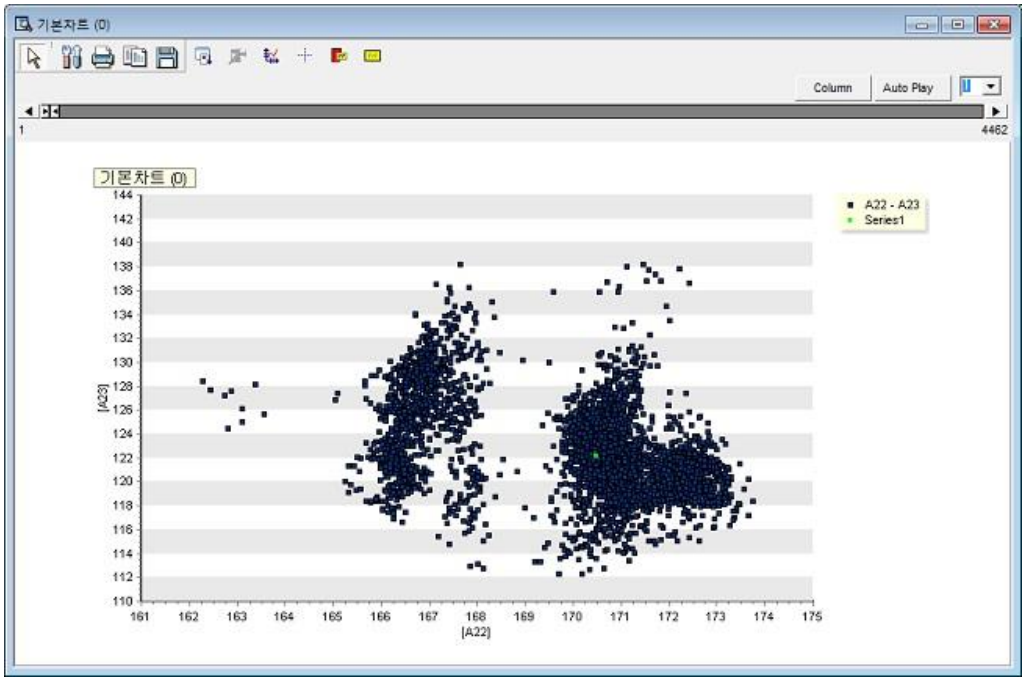
			NP, C, U	
분석변수	분석변수	C 차트를 그릴 변수를 지정합니다.	필수	
기타 관리도 옵션	평균 불량수 사용	평균 불량수를 사용자가 지정할지 결정합니다.	예, 아니오	
	평균 불량수	평균 불량수 사용자 지정을 **'예'**라고 했을 때 활성화됩니다. 활성화되면 평균 불량수를 입력합니다.	실수	

3.3.3 기본차트 노드

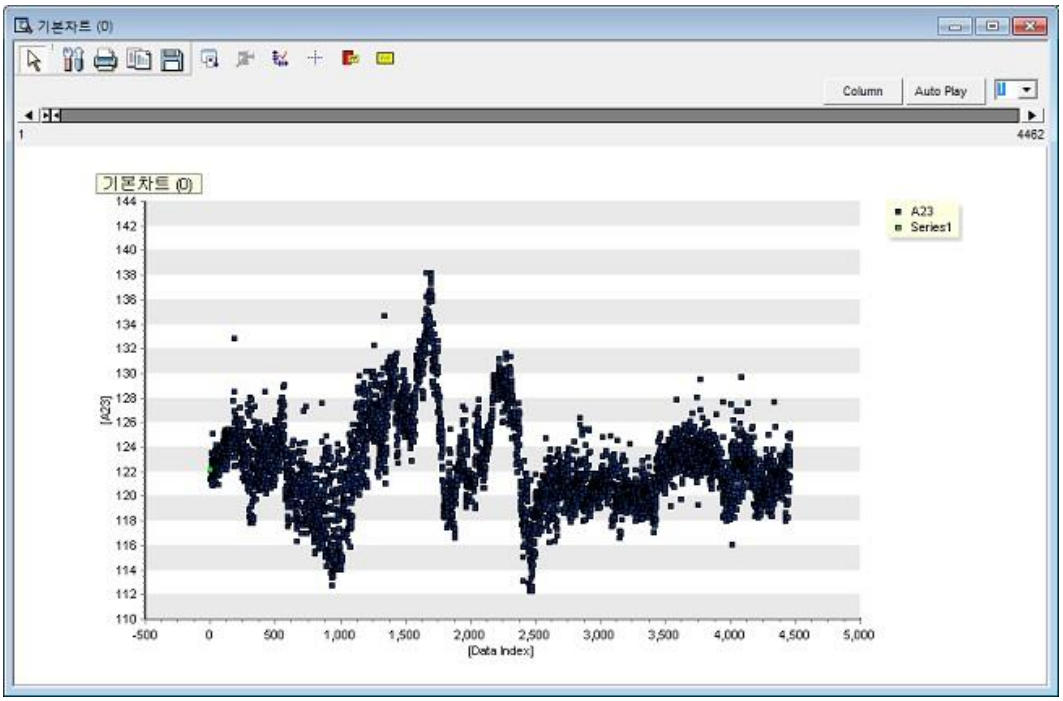
기본차트 노드는 데이터를 이용하여 일반적인 이차원 차트를 그릴 수 있는 노드입니다. **기본차트 노드**를 사용하면 일반적인 이차원 차트뿐만 아니라 아래에서 설명되는 데이터/시리즈 인덱스 차트도 표현이 가능합니다.

지원 플롯

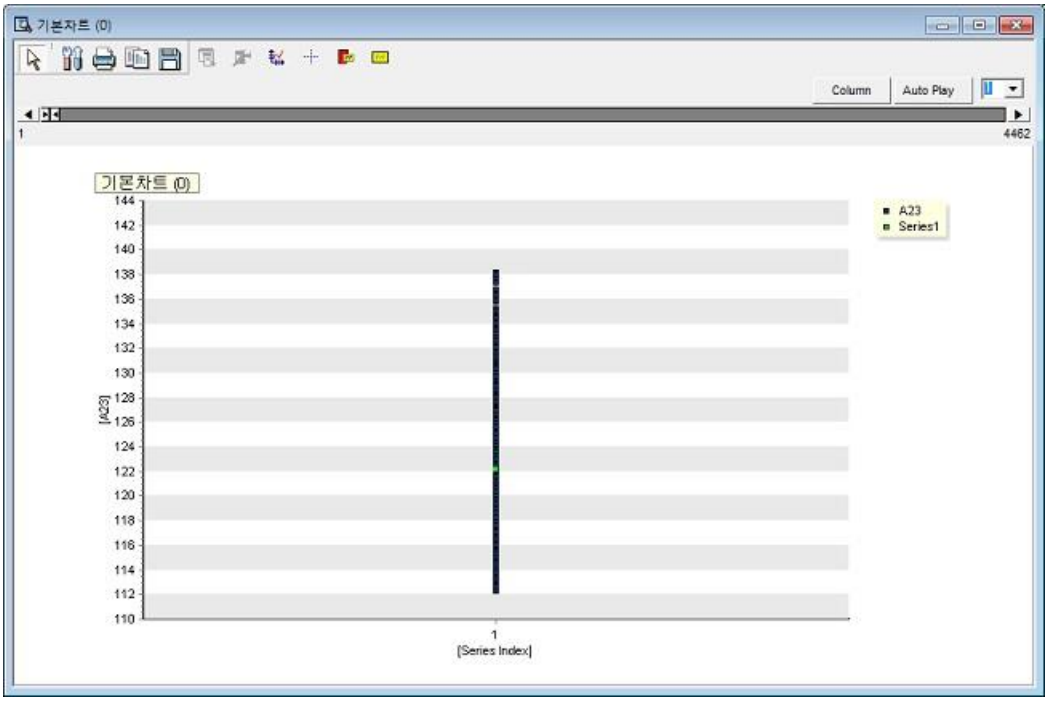
일반 이차원 차트: X, Y 축으로 지정된 두 변수 값을 사용하여 이들에 대한 이차원 차트를 그립니다.



데이터 인덱스 차트: 변수의 인덱스를 X 축으로, 변수의 값을 Y 축으로 사용하여 이들에 대한 이차원 차트를 그립니다.



시리즈 인덱스 차트: 변수의 시리즈 인덱스를 X 축으로, 변수의 값을 Y 축으로 사용하여 이들에 대한 이차원 차트를 그립니다.



사용법

- 두 개 이상의 시리즈를 그릴 경우, 필요한 시리즈 개수만큼 **시리즈 추가** 버튼을 눌러 추가합니다.
- **Value** 외 다른 인덱스를 X 축으로 사용하고자 한다면, **X 축 형태** 속성을 변경합니다.
- **시리즈 목록** 속성창에서 **X 축** 및 **Y 축**으로 사용할 변수를 선택합니다. 만약 **Data Index** 또는 **Series Index** 를 사용하면, **Y 축**으로 사용할 변수만 지정하면 됩니다.
- 선택 사항

  라벨 속성 변경

  차트 형태 속성 변경

  점모양 속성 변경

  X 축, Y 축 지정 속성 변경

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
차트옵션	라벨	차트 상의 포인터에 표시될 라벨을 지정합니다.		
	차트 형태	차트의 형태를 지정합니다. **Scatter Plot** 일 경우 흩뿌려진 점들로만 그려지고, **Line Plot** 일 경우 점들이 선으로 이어져서 그려집니다.	Scatter Plot, Line Plot	
	X 축 형태	X 축 타입을 결정합니다. **Value** 가 선택되면 X 축으로 지정된 변수가 X 축이 되고, **Data**	Value, Data Index,	

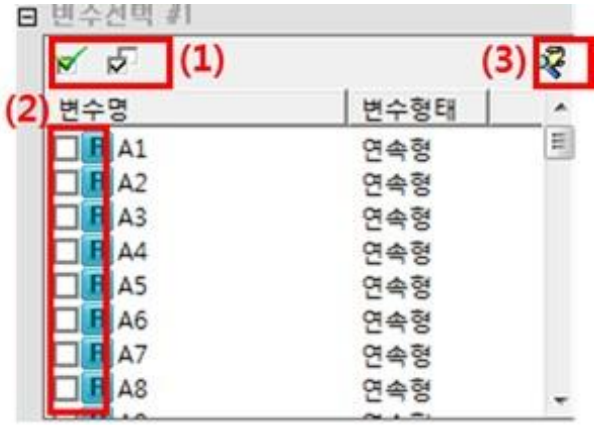
			Index 가 선택되면 Y 축 변수의 데이터 인덱스 값이 X 축이 되면, **Series Index** 가 선택되면, Y 축 변수의 시리즈 인덱스값이 X 축이 됩니다.	Series Index	
	시리즈 추가		여러 시리즈를 한 차트에 그리기 위해 사용됩니다. 버튼을 누르면 시리즈 선택 리스트가 추가됩니다.	버튼	
	시리즈 삭제		마지막에 추가된 시리즈가 삭제됩니다.	버튼	
	시리즈 편집		선택되지 않은 시리즈를 삭제합니다.	버튼	
사용자 지정 최대/최소	Y 축 지정		Y 축의 사용자 지정을 결정합니다.	예, 아니오	
	Y 축 최소값		Y 축 지정 시에만 활성화됩니다. Y 축의 최소값을 결정합니다.	실수	
	Y 축 최대값		Y 축 지정 시에만 활성화됩니다. Y 축의 최대값을 결정합니다.	실수	
	X 축 지정		X 축의 사용자 지정을 결정합니다.	예, 아니오	
	X 축 최소값		X 축 지정 시에만 활성화됩니다. Y 축의 최소값을 결정합니다.	실수	
	X 축 최대값		X 축 지정 시에만 활성화됩니다. Y 축의 최대값을 결정합니다.	실수	
시리즈 목록	시리즈 #00	X 축	X 축으로 사용할 변수를 지정합니다. **차트 옵션**에서 **Data Index** 또는 **Series Index** 를 선택하면, 자동으로 이들 인덱스로 X 축이 지정됩니다		
		Y 축	Y 축으로 사용할 변수를 지정합니다.		
		점모양	차트에 그려질 점들의 모양을 지정합니다.	사각형, 원형, 삼각형, 역삼각형, 십자형, X 형, 별모양, 작은점, 다이아몬드	

3.3.4 매트릭스차트 노드


매트릭스차트

매트릭스차트 노드는 데이터를 이용하여, 매트릭스차트를 그릴 수 있는 노드입니다. 매트릭스차트를 사용하면, 여러 변수들의 상관 관계를 하나의 그래프로 표현할 수 있습니다.

사용법



- 차트로 그릴 변수를 선택합니다
- 변수를 선택하기 위해서는 변수 앞의 체크박스(이미지의 2 번)를 선택합니다.
- 모든 변수를 선택 또는 해제할 때는 아래 이미지 1 번 그림의 버튼을 이용합니다.
- 조건부 변수(이미지의 3 번)을 이용하여 조건에 맞는 변수를 선택할 수 있다.
- 선택사항

  그룹변수 옵션 변경

  X 축, Y 축 지정 속성 변경

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
차트옵션	차팅 방식	차트 표현 방식을 설정합니다.	단순 행렬도, X, Y 지정 행렬도	
	히스토그램 그리기	차팅 방식이 X, Y 지정 행렬도일 경우 활성화됩니다.	예, 아니오	

		히스토그램을 그릴지 결정합니다.		
	Bin 개수	차팅 방식이 X, Y 지정 행렬도일 경우 활성화됩니다. Bin 개수를 입력합니다.	입력	
그룹 변수 옵션	그룹표시	데이터 내에 그룹 관련 변수가 있을 때 그룹 표시 방법을 선택합니다.	그룹 사용 안함, 색상으로 구분 중 택일	
	그룹변수 #1	첫번째 그룹변수를 선택합니다	선택	
	그룹변수 #2	두번째 그룹변수를 선택합니다	선택	
	그룹변수 #3	세번째 그룹변수를 선택합니다.	선택	
	그룹변수 #4	네번째 그룹변수를 선택합니다.	선택	
	그룹변수 #5	다섯번째 그룹변수를 선택합니다.	선택	
사용자 지정 최대/최소	Y 축 지정	Y 축 최대/최소값을 사용자가 수동으로 지정할 것인지 결정합니다.	예, 아니오	
	Y 축 최소값	Y 축 지정이 **'예'**일 경우 활성화됩니다. Y 축 최소값을 입력합니다.	실수	
	Y 축 최대값	Y 축 지정이 **'예'**일 경우 활성화됩니다. Y 축 최대값을 입력합니다.	실수	
	X 축 지정	X 축 최대/최소값을 사용자가 수동으로 지정할 것인지 결정합니다.	예, 아니오	
	X 축 최소값	X 축 지정이 **'예'**일 경우 활성화됩니다. X 축 최소값을 입력합니다.	실수	
	X 축 최대값	X 축 지정이 **'예'**일 경우 활성화됩니다. X 축 최대값을 입력합니다.	실수	
변수선택	전체 변수 선택/해제	전체 변수를 선택하거나, 선택 해제하는 버튼으로 위의 이미지	버튼	

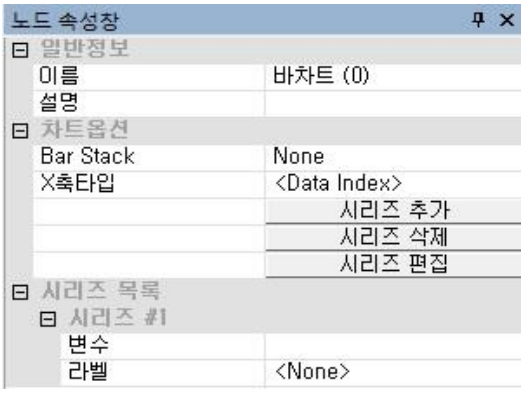
	버튼	1 번 그림의 두 버튼 입니다.		
	변수 목록	매트릭스차트로 그릴 변수를 지정합니다. 위의 이미지 2 번 그림에 표현된 체크박스를 이용하여, 변수를 선택할 수 있습니다.		
	조건부 변수 선택	3 번 이미지를 이용하여 변수 목록 중 제시하는 조건에 맞는 변수를 선택할 수 있습니다.		

3.3.5 바차트 노드


바차트

바차트 노드는 데이터를 이용하여, 바차트를 그릴 수 있는 노드입니다.

사용법



- 두 개 이상의 시리즈를 그릴 경우, 필요한 시리즈 개수만큼 **시리즈 추가** 버튼을 눌러 추가합니다.
- **시리즈 목록** 속성창의 변수 속성에 차트로 표현할 변수를 지정합니다.
- 선택 사항

  라벨 속성 변경

  Bar Stack 속성 변경

속성

속성 그룹	속성명	설명	기타	비고

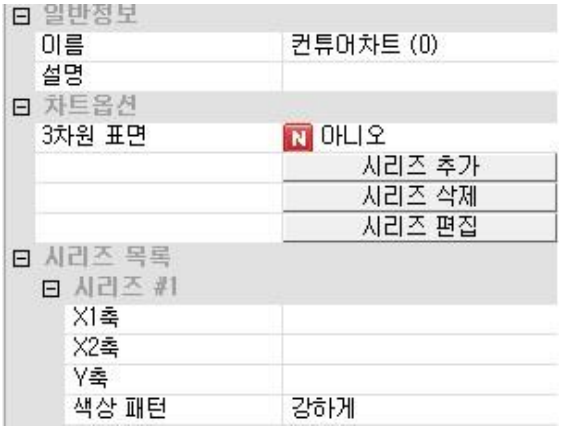
일반정보	이름		노드의 이름을 입력합니다.	선택	
	설명		노드에 대한 간단한 주석을 달 수 있습니다.	선택	
차트 옵션	Bar Stack		바차트의 유형을 선택합니다.(단, X 축 형태가 series index 로 선택될 경우 Bar chart 특성상 이 기능은 비활성화 됩니다.)	None Stacked Stacked 100%	
	X 축 형태		X 축 형태를 결정합니다. **Data Index** 가 선택되면, 지정된 변수의 데이터 인덱스 값이 X 축이 되고, **Series Index** 가 선택되면, series 별로 차트가 구분되며, 이는 범례를 통해 파악할 수 있습니다.	Data Index, Series Index	
	시리즈 추가		여러 시리즈를 한 차트에 그리기 위해 사용됩니다. 버튼을 누르면 시리즈 선택 리스트가 추가됩니다.	버튼	
	시리즈 삭제		마지막에 추가된 시리즈가 삭제됩니다.	버튼	
	시리즈 편집		선택되지 않은 시리즈를 제거합니다.	버튼	
시리즈 목록	시리즈 #00	변수	바차트로 표현될 변수를 지정합니다.		
		라벨	라벨로 사용할 변수를 지정합니다. 변수 속성에 지정된 변수의 형태가 **이산형**이라면, 이 속성은 무시됩니다.		

- Note: 변수 형태에 따른 Bar chart 의미
  변수 형태가 연속형인 경우, x 값이 데이터 인덱스가 되며 y 값이 해당 변수값이 표현됩니다. 변수 형태가 이산형(범주형)인 경우, x 값은 범주 이름이 되며 y 값은 해당 범주에 속하는 데이터 빈도수가 됩니다.

3.3.6 컨튜어차트 노드

컨튜어차트 노드는 데이터를 이용하여 등고선 차트를 그릴 수 있는 노드입니다.

사용법



- 두 개 이상의 시리즈를 그릴 경우, 필요한 시리즈 개수만큼 **시리즈 추가** 버튼을 눌러 추가합니다.
- **시리즈 목록** 속성창에서 **X1 축** 및 **X2 축**, **Y 축**으로 사용할 변수를 선택합니다.
- **색상 패턴** 속성 변경
- **3 차원 표면** 속성 변경

NOTE X1 축 및 X2 축은 2 차원의 두 축으로 등고선의 좌표에 해당됩니다. Y 축은 각 좌표에서의 높이를 나타냅니다.

속성

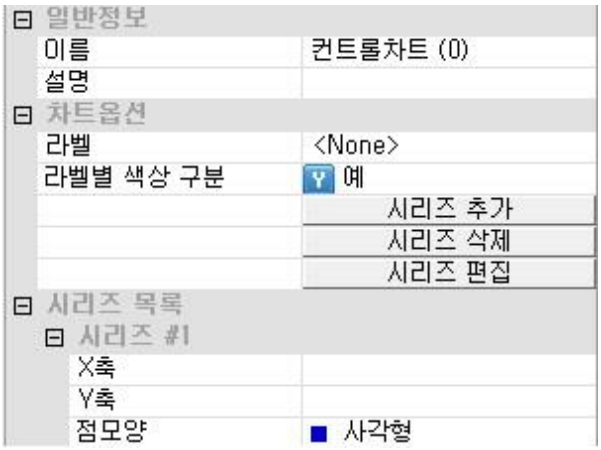
속성그룹	속성명		설명	기타	비고
일반정보	이름		노드의 이름을 입력합니다.	선택	
	설명		노드에 대한 간단한 주석을 달 수 있습니다.	선택	
차트옵션	3 차원 표면		3 차원 입체로 표현할지 2 차원 평면으로 표현할지를 지정합니다.	예, 아니오	
	시리즈 추가		여러 시리즈를 한 차트에 그리기 위해 사용됩니다. 버튼을 누르면 시리즈 선택 리스트가 추가됩니다.	버튼	
	시리즈 삭제		마지막에 추가된 시리즈가 삭제됩니다.	버튼	
	시리즈 편집		선택되지 않은 시리즈를 제거합니다.	버튼	
시리즈 목록	시리즈 #00	X1 축	X1 축(첫번째 위치 좌표) 변수를 지정합니다.		
		X2 축	X2 축(두번째 위치 좌표) 변수를 지정합니다.		
		Y 축	Y 축, 즉 높이 변수를 지정합니다.		
		색상 패턴	차트에 그려질 등고선의 색상 패턴을 지정합니다.	강하게, 약하게, 세가지 색상, 두가지 색상, 단색	

3.3.7 컨트롤차트 노드


컨트롤차트

컨트롤차트 노드는 데이터를 이용하여 컨트롤 차트를 그릴 수 있는 노드입니다. 컨트롤 차트를 사용하면, 데이터의 분포 상태를 알 수 있습니다.

사용법



- 두 개 이상의 시리즈를 그릴 경우, 필요한 시리즈 개수만큼 **시리즈 추가** 버튼을 눌러 추가합니다.
- **시리즈 목록** 속성 창에서 **X 축** 및 **Y 축**으로 사용할 변수를 선택합니다.
- 선택 사항

  라벨 속성 변경

  라벨별 색상 구분 속성 변경

  점모양 속성 변경

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
차트옵션	라벨	차트 상의 포인터에 표시될 라벨을 지정합니다.		
	라벨별 색상 구분	각 라벨에 대한 색상을 다르게 할 것인지를 결정합니다.	예, 아니오	
	시리즈 추가	여러 시리즈를 한 차트에 그리기 위해 사용됩니다. 버튼을 누르면 시리즈 선택	버튼	

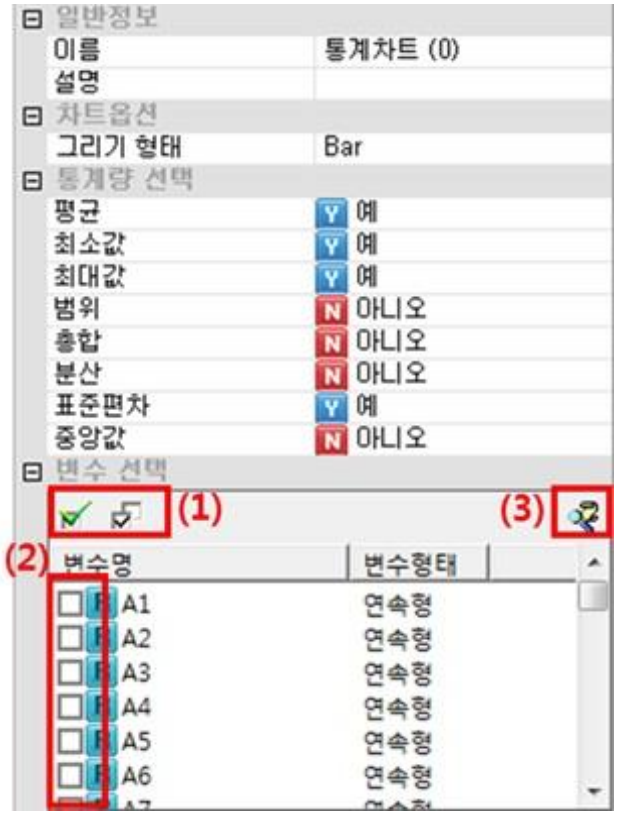
			리스트가 추가됩니다.		
	시리즈 삭제		마지막에 추가된 시리즈가 삭제됩니다.	버튼	
	시리즈 편집		선택되지 않은 시리즈를 제거합니다.	버튼	
시리즈 목록	시리즈 #00	X 축	X 축으로 사용할 변수를 지정합니다.		
		Y 축	Y 축으로 사용할 변수를 지정합니다.		
		점모양	차트에 그려질 점들의 모양을 지정합니다.	사각형, 원형, 삼각형, 역삼각형, 십자형, X 형, 별모양, 다이아몬드, 작은점	

3.3.8 통계차트 노드


통계차트

통계차트 노드는 데이터를 이용하여, 변수들의 통계량을 비교 차트를 그릴 수 있는 노드입니다. **통계차트**에서는 X 축에는 변수가, Y 축에는 통계량이 표시되어 변수들간의 통계량 변화를 한 눈에 볼 수 있습니다.

사용법



- 차트로 그릴 변수를 선택합니다. 변수를 선택하기 위해서는 변수 앞의 체크박스(아래 이미지 2 번 그림)를 선택하면 됩니다. 모든 변수를 선택 또는 해제할 때는 아래 이미지 1 번 그림의 버튼을 이용하면 됩니다.
- 조건부 변수(이미지의 3 번)을 이용하여 조건에 맞는 변수를 선택할 수 있다.
- 선택 사항

  통계량 선택 속성 변경

  차트 옵션 속성 변경

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
차트옵션	그리기 형태	차트의 출력 형태를 결정합니다.	Line Point Bar	

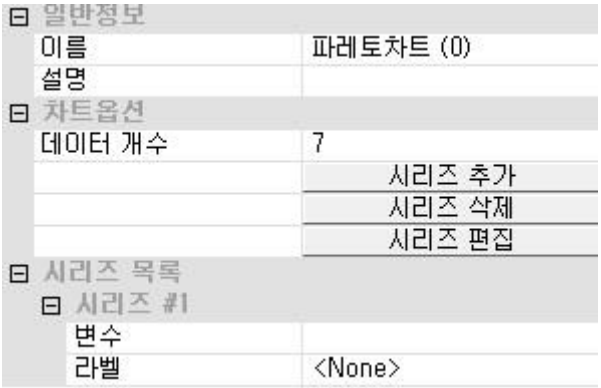
변수선택	전체 변수 선택/해제 버튼	전체 변수를 선택하거나, 선택 해제하는 버튼으로 위의 **이미지 1 번** 그림의 두 버튼입니다.	버튼	
	변수 목록	차트로 그릴 변수를 지정합니다. 위의 **이미지 2 번** 그림에 표현된 체크박스를 이용하여, 변수를 선택할 수 있습니다.		
통계량 선택	통계량 목록	차트로 표현할 통계량을 선택합니다.	평균, 최소값, 최대값, 범위, 총합, 분산, 표준편차, 중앙값	
변수선택	변수 목록	차트를 그릴 변수를 선택합니다. 다중선택도 가능합니다.		

3.3.9 파레토차트 노드


파레토차트

파레토차트 노드는 데이터를 이용하여, 파레토차트를 그릴 수 있는 노드입니다. 파레토차트는 주로 이산형 변수에 대해 데이터 분포 상태를 보기 위해 사용됩니다.

사용법



- 두 개 이상의 시리즈를 그릴 경우, 필요한 시리즈 개수만큼 **시리즈 추가** 버튼을 눌러 추가합니다.
- **시리즈 목록** 속성창의 **변수** 속성에 플롯으로 표현할 변수를 지정합니다.
- 선택 사항

  라벨 속성 변경

속성

속성그룹	속성명		설명	기타	비고
일반정보	이름		노드의 이름을 입력합니다.	선택	
	설명		노드에 대한 간단한 주석을 달 수 있습니다.	선택	
차트옵션	데이터 개수		플롯으로 표현할 데이터 종류의 최대 개수입니다. 데이터 종류를 그 수가 많은 것부터 내림차순으로 정렬한 후 여기서 지정된 개수만큼 플롯으로 표현됩니다. 데이터 개수를 초과하는 소량 데이터들은 합쳐져서 기타 종류로 표현됩니다.		
	시리즈 추가		여러 시리즈를 한 차트에 그리기 위해 사용됩니다. 버튼을 누르면 시리즈 선택 리스트가 추가됩니다.	버튼	
	시리즈 삭제		마지막에 추가된 시리즈가 삭제됩니다.	버튼	
	시리즈 편집		선택되지 않은 시리즈를 삭제합니다.	버튼	
시리즈 목록	시리즈 #00	변수	차트로 표현될 변수를 지정합니다.		
		라벨	라벨로 사용할 변수를 지정합니다. 변수 옵션에 지정된 변수의 형태가 **이산형**이라면, 이 속성은 무시됩니다.		

3.3.10 파이차트 노드

파이차트 노드는 데이터를 이용하여, 파이차트를 그리는 노드입니다. 파이차트는 주로 이산형 변수에 대해 데이터 상대적인 비율을 보기 위해 사용됩니다.

사용법

- 두 개 이상의 시리즈를 그릴 경우, 필요한 시리즈 개수만큼 **시리즈 추가** 버튼을 눌러 추가합니다.
- **시리즈 목록** 속성창의 **변수** 속성에 차트로 표현할 변수를 지정합니다.

속성

속성그룹	속성명		설명	기타	비고
일반정보	이름		노드의 이름을 입력합니다.	선택	
	설명		노드에 대한 간단한 주석을 달 수 있습니다.	선택	
차트옵션	시리즈 추가		여러 시리즈를 한 차트에 그리기 위해 사용됩니다. 버튼을 누르면 시리즈 선택 리스트가 추가됩니다.	버튼	
	시리즈 삭제		마지막에 추가된 시리즈가 삭제됩니다.	버튼	
	시리즈 편집		선택되지 않은 시리즈를 삭제합니다.	버튼	
시리즈 목록	시리즈 #00	변수	차트로 표현될 변수를 지정합니다.		
		라벨	라벨로 사용할 변수를 지정합니다. 변수 옵션에 지정된 변수의 형태가 **이산형**이라면, 이 속성은 무시됩니다.		

3.3.11 히스토그램 노드


히스토그램

히스토그램 노드는 연속형 변수의 분포를 가늠해 볼 수 있는 히스토그램을 그리는 노드입니다. **연속형 변수**일 경우만 가능하며 정규분포와 곡선을 동시에 그릴 수도 있습니다.

사용법

- 히스토그램을 그릴 연속형 변수를 지정합니다.
- 구간화 방법을 지정한 뒤 각 방법에 따른 값(구간 간격 혹은 구간 개수)를 입력합니다.
- 범위설정의 자동설정을 **'예'**로 선택하면 선택된 변수의 최대 / 최소값이 범위가 되며 아니오로 선택하였을 경우 적당한 최소 / 최대값을 입력하여야 합니다.
- 만약, 최대 / 최소값이 잘못 되었을 경우 히스토그램에 데이터가 나타나지 않거나 일부만 나타날 수 있습니다.
- 차트로 그릴 변수를 선택합니다. 변수를 선택하기 위해서는 변수 앞의 체크박스(위 이미지의 2 번 그림)를 선택하면 됩니다. 모든 변수를 선택 또는 해제할 때는 위 이미지 1 번 그림의 버튼을 이용하면 됩니다.
- 조건부 변수(이미지의 3 번)을 이용하여 조건에 맞는 변수를 선택할 수 있습니다.
- 선택사항

  그룹변수 옵션 속성 변경

  Y 축 지정 속성 변경

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
차트 옵션	변수별로 차트 생성	변수 개수만큼 여러 개의 차트를 그린다.	예, 아니오	

	행 내 차트개수	행에 들어갈 차트 개수를 설정합니다.		
그룹변수 옵션	그룹표시	데이터 내에 그룹 관련 변수가 있을 때 그룹 표시 방법을 선택합니다.	그룹 사용 안함, 시리즈로 구분 중 택일	
	그룹변수 #1	첫번째 그룹변수를 선택합니다	선택	
	그룹변수 #2	두번째 그룹변수를 선택합니다	선택	
	그룹변수 #3	세번째 그룹변수를 선택합니다.	선택	
	그룹변수 #4	네번째 그룹변수를 선택합니다.	선택	
	그룹변수 #5	다섯번째 그룹변수를 선택합니다.	선택	
사용자 지정 최대/최소	Y 축 지정	Y 축 최대/최소값을 사용자가 수동으로 지정할 것인지 결정합니다.	예, 아니오	
	Y 축 최소값	Y 축 지정이 **'예'**일 경우 활성화됩니다. Y 축 최소값을 입력합니다.	실수	
	Y 축 최대값	Y 축 지정이 **'예'**일 경우 활성화됩니다. Y 축 최대값을 입력합니다.	실수	
구간화 방법	구간화 방법	지정된 연속형 변수의 데이터 구간을 결정할 방법을 지정합니다.	구간간격, 구간 개수	
	구간 간격	구간화 방법이 구간 간격일 경우 활성화됩니다. 지정된 간격마다 몇 개의 데이터가 있는지 계산 하여 히스토그램을 그립니다.	실수	
	구간 개수	구간화 방법이 구간 개수일	자연수	

		경우 활성화됩니다. 전체 구간을 지정된 개수만큼 나누고 각 구간에 몇 개의 데이터가 포함되는 지 계산하여 히스토그램을 그립니다.		
범위설정	자동설정	히스토그램을 그릴 범위의 최소, 최대값을 계산할 방법을 지정합니다.	예, 아니오	
	최소값	자동설정이 **'아니오'**일 경우 활성화됩니다. 히스토그램을 그릴 범위 중 최소값을 입력합니다.	실수	
	최대값	자동설정이 **'아니오'**일 경우 활성화됩니다. 히스토그램을 그릴 범위 중 최대값을 입력합니다.	실수	
분포선	정규분포 곡선	정규분포 곡선을 같이 그리고자 한다면 **'예'**를 선택합니다.	예, 아니오	
변수 선택	변수 목록	차트를 그릴 변수를 선택합니다. 다중선택도 가능합니다.		

3.3.12 기본확장 차트

기본확장차트 노드는 기본 차트를 수정 보완하여 보다 다각적으로 차트를 볼 수 있는 기능을 제공합니다.

사용법

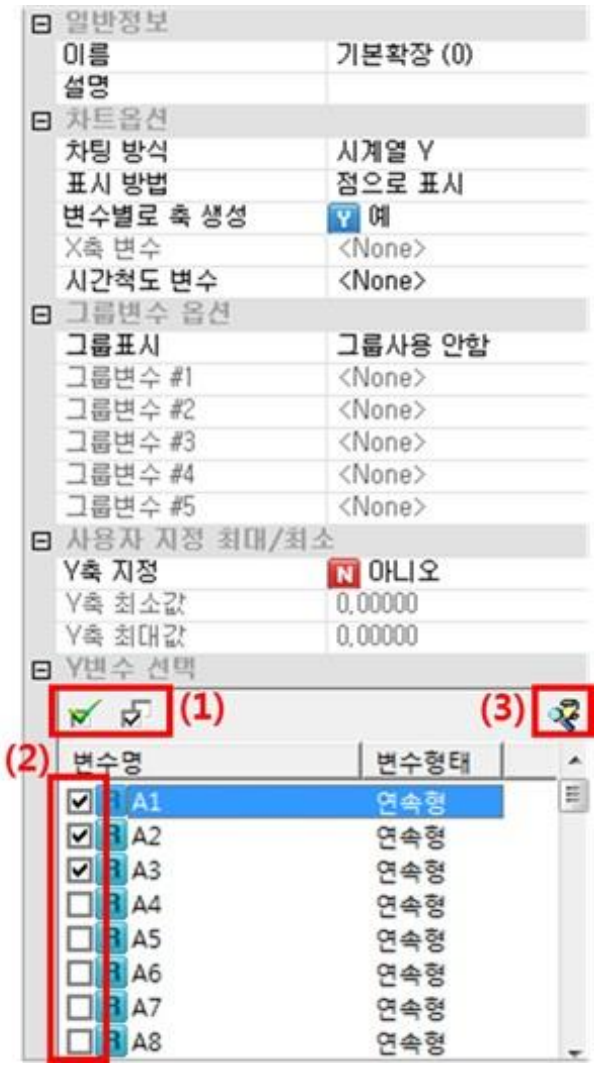


차트 옵션

- 차팅 방식을 선택합니다. 시계열과 X 대 Y 중 하나를 선택할 수 있습니다.
- 표시 방법을 선택합니다. 점으로 표시 혹은 선으로 연결을 선택할 수 있습니다.
- X 축 변수를 선택합니다. 차팅 방식으로 X 대 Y 를 선택하였을 경우 X 축에 어떤 변수를 넣을지를 선택합니다.

시간 척도 변수: 차팅 방식으로 시계열을 선택하였을 때 시간을 나타내는 변수로 어떤 것을 사용할지를 선택합니다.

그룹 변수 옵션

- 그룹 표시 설정을 합니다. 그룹 표시를 하지 않거나 시리즈로 구분하거나 라벨로 표시합니다.
- 그룹 변수를 선택합니다. 어떠한 이산형 변수를 그룹으로 할지를 선택합니다. 최대 5 개까지 선택할 수 있습니다..

Y 변수 선택

- 차트로 그릴 변수를 선택합니다. 변수를 선택하기 위해서는 변수 앞의 체크박스(위 이미지의 2 번 그림)를 선택하면 됩니다. 모든 변수를 선택 또는 해제할 때는 위 이미지 1 번 그림의 버튼을 이용하면 됩니다.
- 조건부 변수(이미지의 3 번)을 이용하여 조건에 맞는 변수를 선택할 수 있습니다.

속성

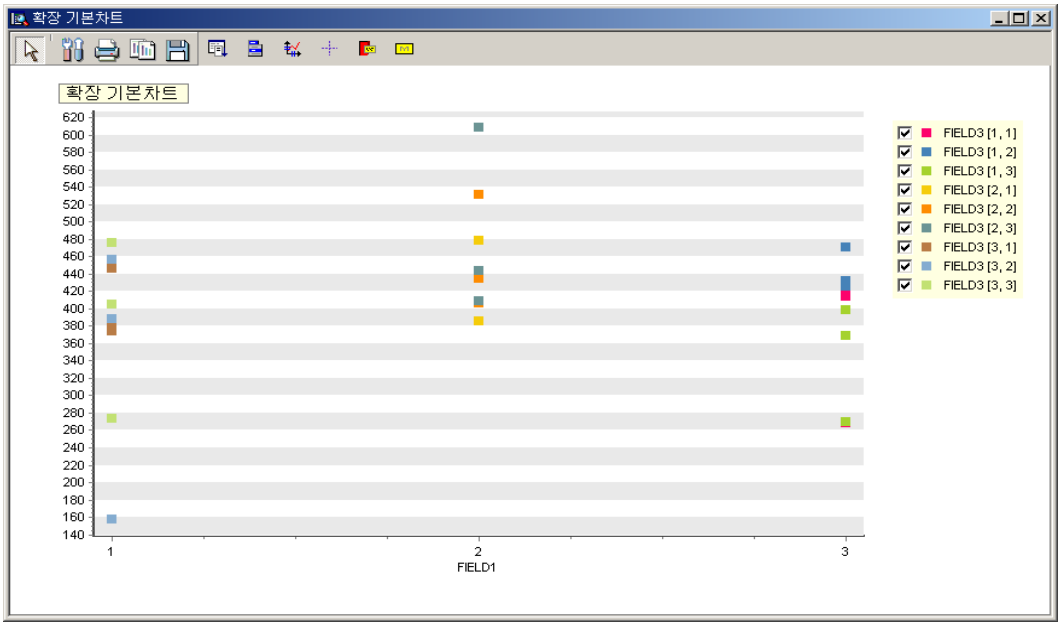
속성 그룹	속성명	설명	기타	비고

일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
차트 옵션	차팅 방식	차팅 방식을 선택합니다.	시계열 Y, X 대 Y 중 택일	
	표시 방법	차트가 표시되는 방식을 선택합니다.	점으로 표시, 선으로 연결 중 택일	
	변수별로 축 생성	다수에 변수를 한 축에 모두 보여줄지 변수별 축으로 생성할지 선택합니다.	예, 아니오	
	X 축 변수	차팅 방식으로 X 대 Y 를 선택했을 시, X 축 변수로 어떤 변수를 사용할지를 선택합니다.	차팅 방식으로 X 대 Y 를 선택했을 시	
	시간척도 변수	차팅 방식으로 시계열 Y 를 선택했을 시, 가로축에 어떠한 변수를 사용할지를 결정합니다.	차팅 방식으로 시계열 Y 를 선택했을 시	
그룹 변수 옵션	그룹표시	데이터 내에 그룹 관련 변수가 있을 때 그룹 표시 방법을 선택합니다.	그룹 사용 안함, 시리즈로 구분, 라벨로 표시 중 택일	
	그룹변수 #1	첫번째 그룹변수를 선택합니다		
	그룹변수 #2	두번째 그룹변수를 선택합니다		
	그룹변수 #3	세번째 그룹변수를 선택합니다.		
	그룹변수 #4	네번째 그룹변수를 선택합니다.		
	그룹변수 #5	다섯번째 그룹변수를 선택합니다.		
사용자 지정 최대/최소	Y 축 지정	Y 축 최대/최소값을 사용자가 수동으로 지정할 것인지 결정합니다.	예, 아니오	
	Y 축 최소값	Y 축 지정이 **'예'**일 경우 활성화됩니다. Y 축 최소값을 입력합니다.	실수	

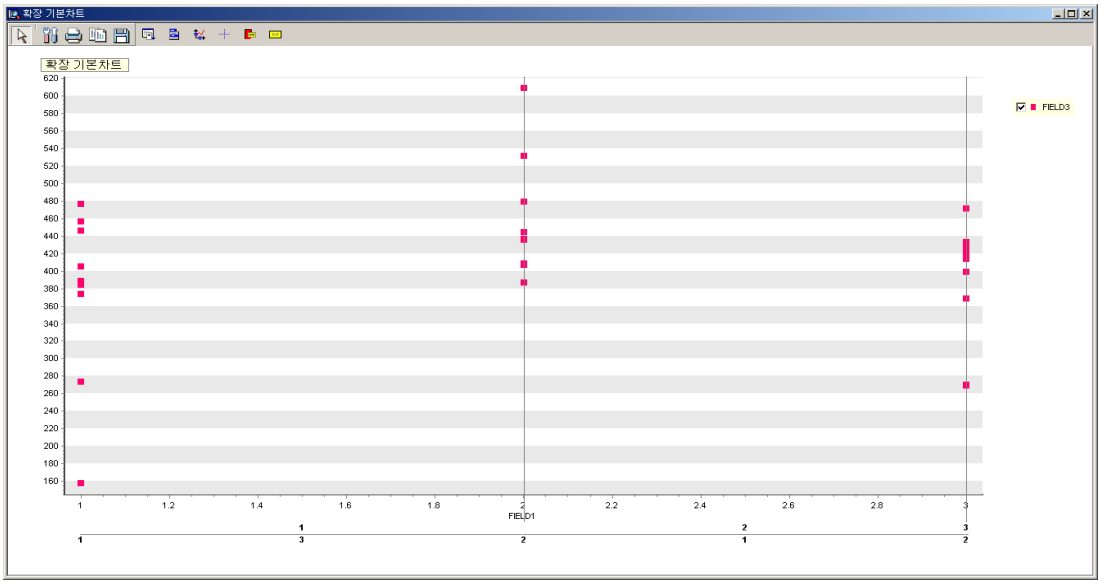
	Y 축 최대값	Y 축 지정이 **'예'**일 경우 활성화됩니다. Y 축 최대값을 입력합니다.	실수	
Y 변수 선택	전체 변수 선택/해제 버튼	전체 변수를 선택하거나, 선택 해제하는 버튼으로 위의 이미지 1 번 그림의 두 버튼 입니다.	버튼	
	변수 목록	차트로 그릴 변수를 지정합니다. 위의 이미지 2 번 그림에 표현된 체크박스를 이용 하여, 변수를 선택할 수 있습니다.		
	조건부 변수 선택	3 번 이미지를 이용하여 변수 목록 중 제시하는 조건에 맞는 변수를 선택할 수 있습니다.		

결과

그룹 표시로 시리즈로 구분을 선택하였을 때(각 그룹별로 서로 다른 색깔의 시리즈로 차팅이 된 것을 알 수 있습니다.)



그룹 표시로 라벨로 표시를 선택하였을 때(하나의 시리즈이지만 라벨로 그룹이 구별되어 있는 것을 확인할 수 있습니다.)



3.3.13 산포 차트


산포차트

산포 차트 노드는 그룹별 산포도를 한눈에 볼 수 있도록 만든 차트입니다. 이를 통해서 최대 3 가지의 이산형 변수에 의해서 나누어지는 그룹들에 대해 산포도를 한눈에 볼 수 있습니다.

사용법

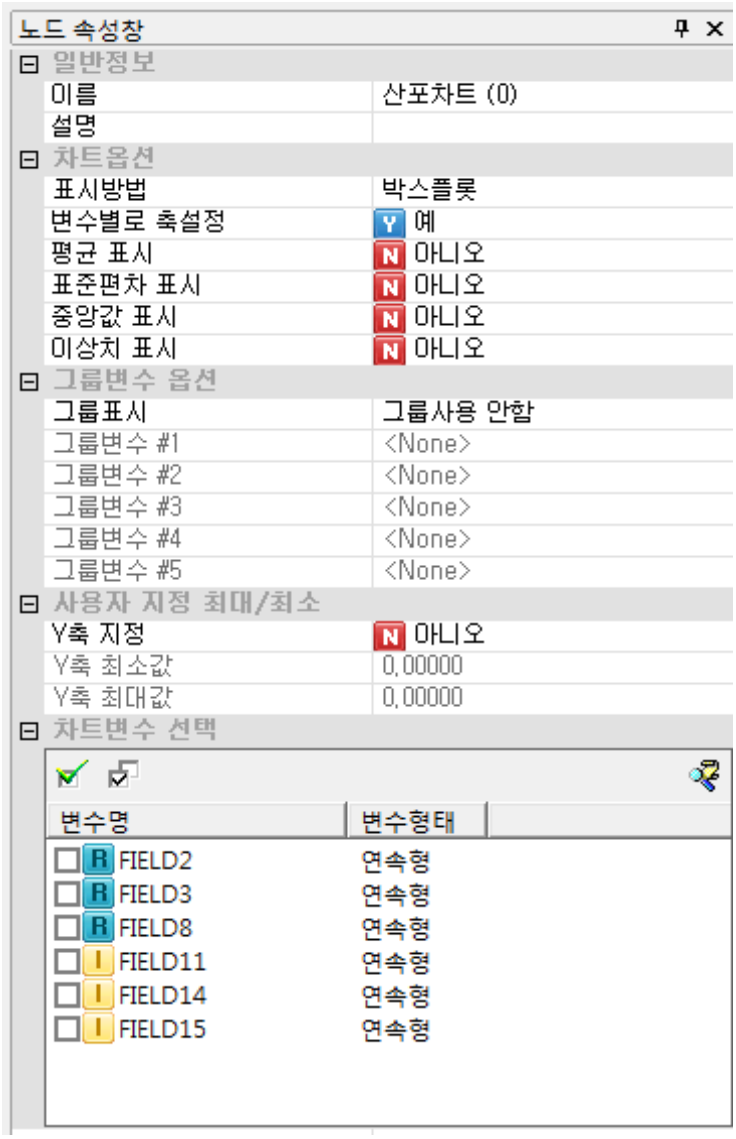


차트 옵션

- 차트 표시 방식을 선택합니다. 산포도와 박스 플롯을 선택할 수 있습니다.

그룹 변수 옵션

- 그룹을 시리즈로 구분할지 구분하지 않을지를 선택합니다.
- 그룹 변수를 선택합니다. 어떠한 이산형 변수를 그룹으로 할지를 선택합니다. 최대 5 개까지 선택할 수 있습니다.
- 선택사항

  통계량 및 이상치 표시 속성 변경

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
차트 옵션	차팅 방식	차팅 방식을 선택합니다.	산포도, 박스 플롯 중 택일	

	변수별로 축설정	변수별로 축을 설정하여 차트를 그릴지 결정합니다.	예, 아니오	
	평균 표시	평균을 차트를 표시할지 결정합니다	예, 아니오	
	표준편차 표시	표준편차을 차트를 표시할지 결정합니다	예, 아니오	
	중앙값 표시	중앙값을 차트를 표시할지 결정합니다	예, 아니오	
	이상치 표시	이상치 차트를 표시할지 결정합니다	예, 아니오	
그룹 변수 옵션	그룹표시	데이터 내에 그룹 관련 변수가 있을 때 그룹 표시 방법을 선택합니다.	그룹 사용 안함, 시리즈로 구분 중 택일	
	그룹변수 #1	첫번째 그룹변수를 선택합니다	선택	
	그룹변수 #2	두번째 그룹변수를 선택합니다	선택	
	그룹변수 #3	세번째 그룹변수를 선택합니다.	선택	
	그룹변수 #4	네번째 그룹변수를 선택합니다.	선택	
	그룹변수 #5	다섯번째 그룹변수를 선택합니다.	선택	
사용자 지정 최대/최소	Y 축 지정	Y 축 최대/최소값을 사용자가 수동으로 지정할 것인지 결정합니다.	예, 아니오	
	Y 축 최소값	Y 축 지정이 **'예'**일 경우 활성화됩니다. Y 축 최소값을 입력합니다.	실수	
	Y 축 최대값	Y 축 지정이 **'예'**일 경우 활성화됩니다. Y 축 최대값을 입력합니다.	실수	
차트 변수 선택	변수명	차트에 사용될 변수를 선택합니다. 복수 선택 가능.	필수	

결과

차트 표시 방식으로 박스 플롯을 선택하고 그룹 표시 방법으로 시리즈로 구분을 선택하였을 경우. 2 개의 그룹변수 FIELD1, FIELD4 에 따라 전체 그룹이 10 개로 나누어지고 이에 따라 10 개의 박스 플롯이 그려진 것을 확인할 수 있습니다.

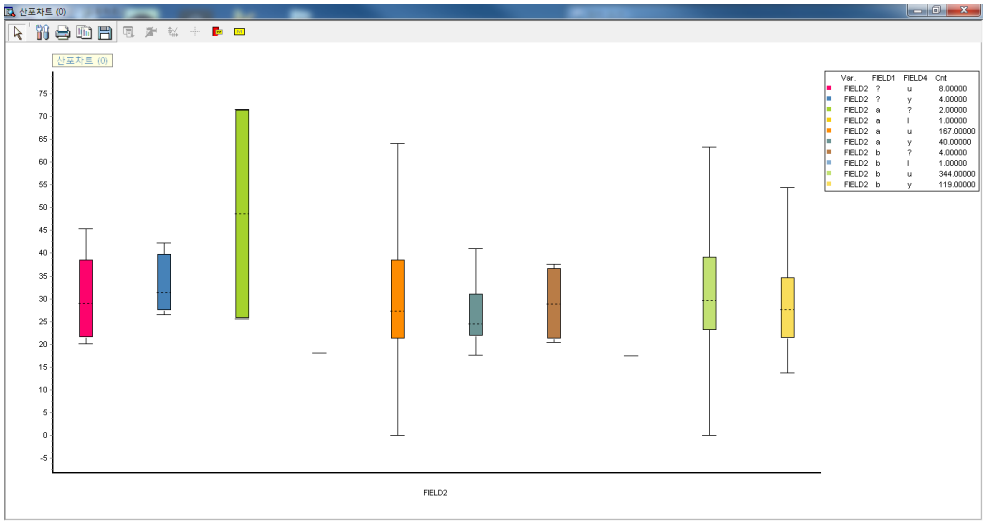
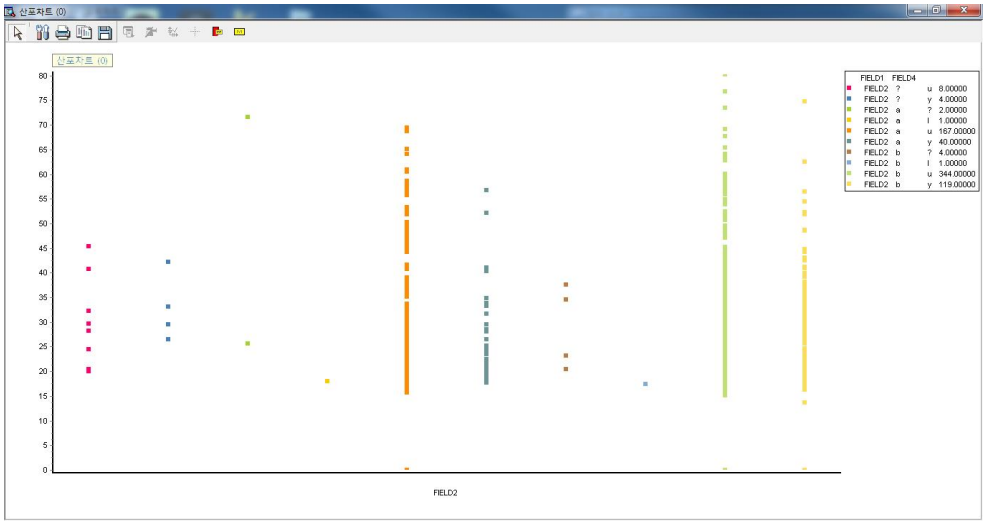


차트 표시 방식으로 산포도를 선택하고 그룹 표시 방법으로 시리즈로 구분을 선택하였을 경우, 2 개의 그룹변수 FIELD1, FIELD4 에 따라 전체 그룹이 10 개로 나누어지고 이에 따라 10 개 시리즈의 산포도가 그려진 것을 확인할 수 있습니다.

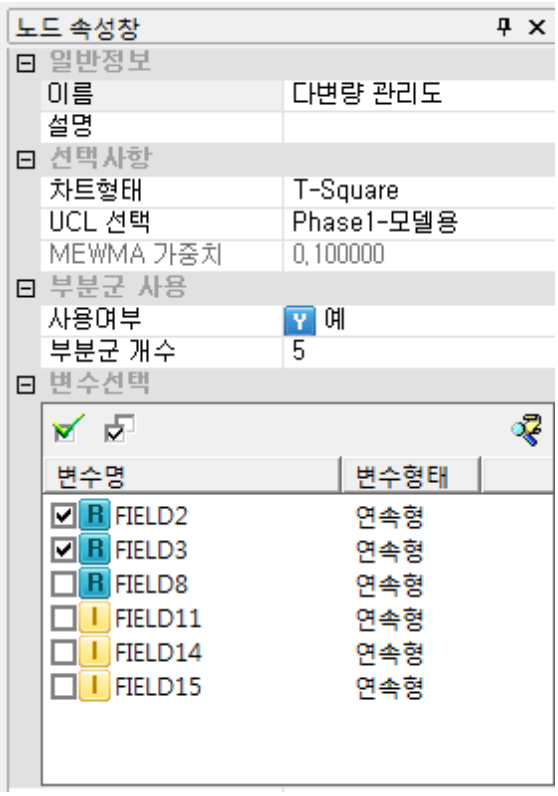


3.3.14 다변량 관리도 차트


다변량 관리도

다변량 관리도 노드는 관리도 차트와 마찬가지로 관리 한계선을 통해 공정이 관리상태에 있는 지 확인할 수 있습니다. 단, 기존의 일반적인 관리도는 하나의 변수가 공정에 영향을 주는 지를 확인하였다면 다변량 관리도는 두 개 이상의 변수가 공정에 어떻게 영향을 주는지 보여줍니다

사용법



차트형태

- 차팅할 다변량 관리도를 선택합니다.
- T^2, MEWMA, 일반화분산, T^2-일반화분산 관리도 중 하나를 선택합니다.

UCL 선택

- T^2 관리도에 사용하는 옵션으로 Phase1 과 Phase2 중 하나를 선택합니다.
- Phase1 은 관리도 모형을 만들 때 사용하여 Phase2 는 새로운 데이터에 대한 모니터링 시 관리한계로 사용합니다.

MEWMA 가중치

- MEWMA 관리도에 사용하는 옵션으로 가중치를 지정합니다. (0~1)

부분군사용

- 부분군 사용여부 및 개수를 지정합니다.

변수선택

- 관리도에 포함할 변수를 선택합니다
- 2 개이상 필수 체크

속성

속성 그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
선택사항	차트형태	차팅 방식을 선택합니다.	T^2 MEWMA 일반화분석 T^2-일반화분석	
	UCL 선택	T^2 관리도 챠팅 시 관리한계선의 기준을 선택합니다.	T^2 선택 시 필수	
	MEWMA 가중치	가중치를 지정합니다.	MEWMA 선택 시 필수	
부분군 사용	사용여부	부분군의 사용여부를 선택합니다.	예, 아니오	
	부분군 개수	사용자가 정한 부분군의 개수를 입력합니다.	조건부 필수	
변수선택	변수명	차트에 사용될 변수를 선택합니다. 복수 선택 가능.	필수	

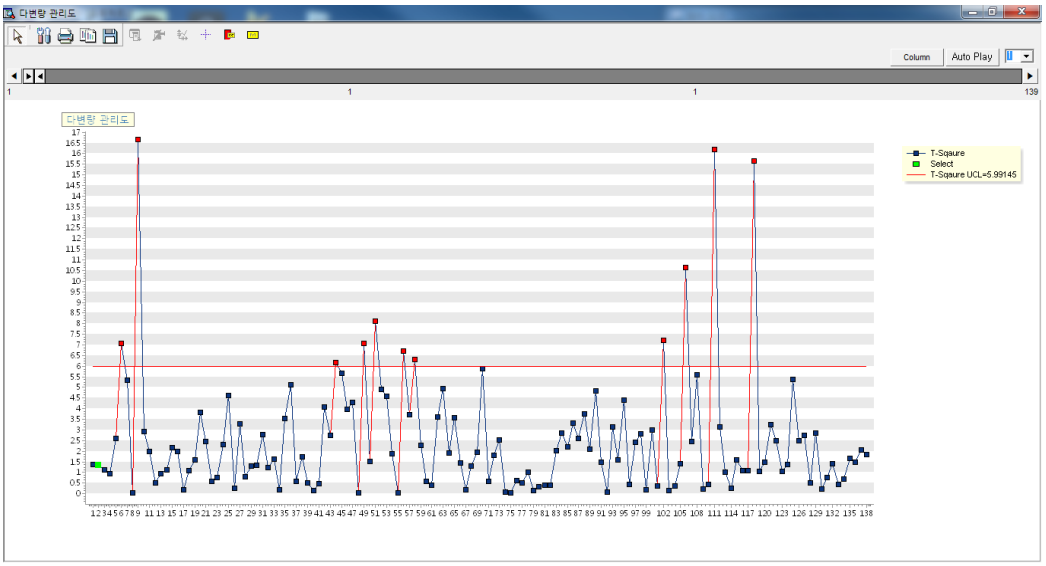
(1) T^2 관리도 (T-square Control Chart)

T^2 관리도는 Hotelling T^2 Control Chart 로서 여러 변수의 평균에 대한 다변량 관리도 입니다. 해석방식은 단변량 관리도의 Shewhart Xbar 관리도와 동일합니다.

사용법

- 차트형태에서 T-square 를 선택합니다.
- 변수선택에서 다변량공정관리를 위한 변수를 체크합니다..
- 부분군의 여부에 따라 부분군의 개수를 입력합니다.
- UCL 선택에서 관리한계선의 종류를 선택합니다.

결과



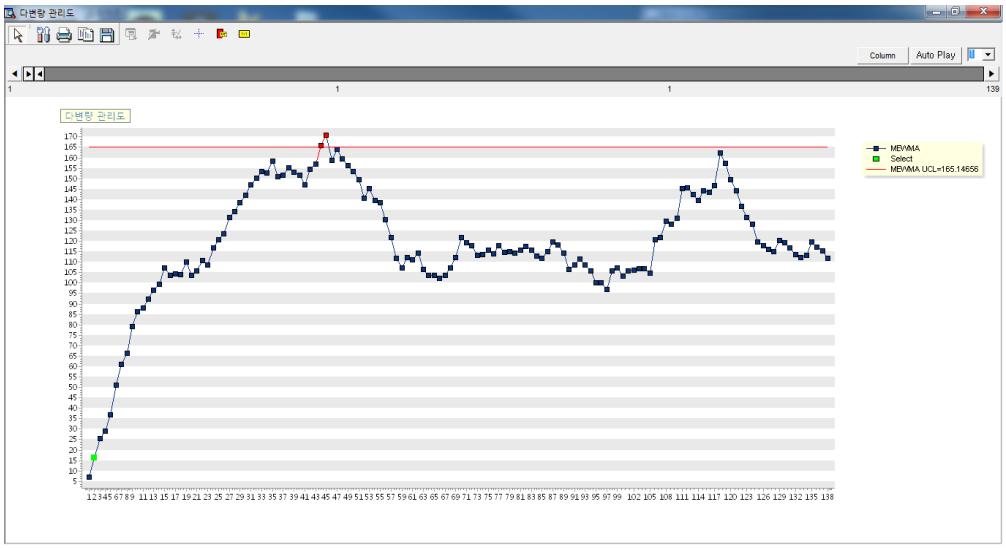
(2) MEWMA 관리도 (Multivariate EWMA Control Chart)

지수가중이동평균에 대한 다변량 관리도입니다. 지수가중이동평균은 공정의 민감한 변동을 탐지할 때 사용됩니다. 최근의 측정치에 더 큰 가중치를 주어 공정변화를 민감하게 만들어 변화를 좀 더 빨리 탐지할 수 있습니다. Weight 가 작을수록 공정평균의 이동을 더 빨리 탐지할 수 있습니다.

사용법

- 차트형태에서 MEWMA 를 선택합니다.
- 가중치를 입력합니다. 가중치는 0~1 까지 입력할 수 있습니다. (default : 0.1)
- 변수를 선택합니다.
- 부분군의 여부에 따라 부분군의 개수를 입력합니다.

결과



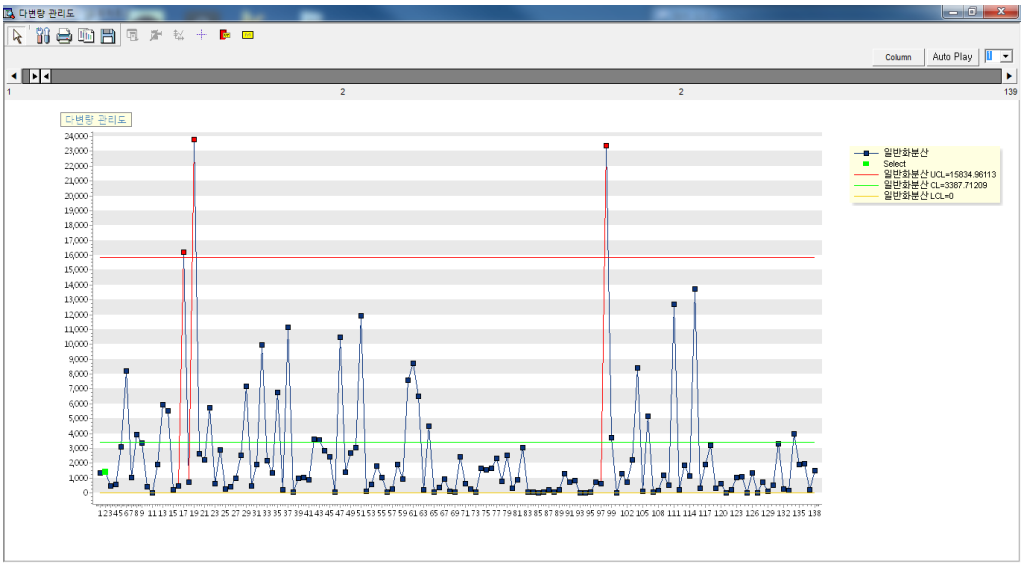
(3) 일반화분산 관리도

일반화분산 관리도는 서로 관련된 두 개 이상의 공정 특성으로 인한 공정 변동을 파악합니다. 일변량 관리도의 S 관리도에 해당되며 해석은 동일합니다.

사용법

- 차트형태에서 일반화분산을 선택합니다.
- 일반화분산은 부분군이 존재하여야만 관리도를 그릴 수 있습니다.
- 변수를 선택합니다.
- 부분군의 개수를 입력합니다.

결과



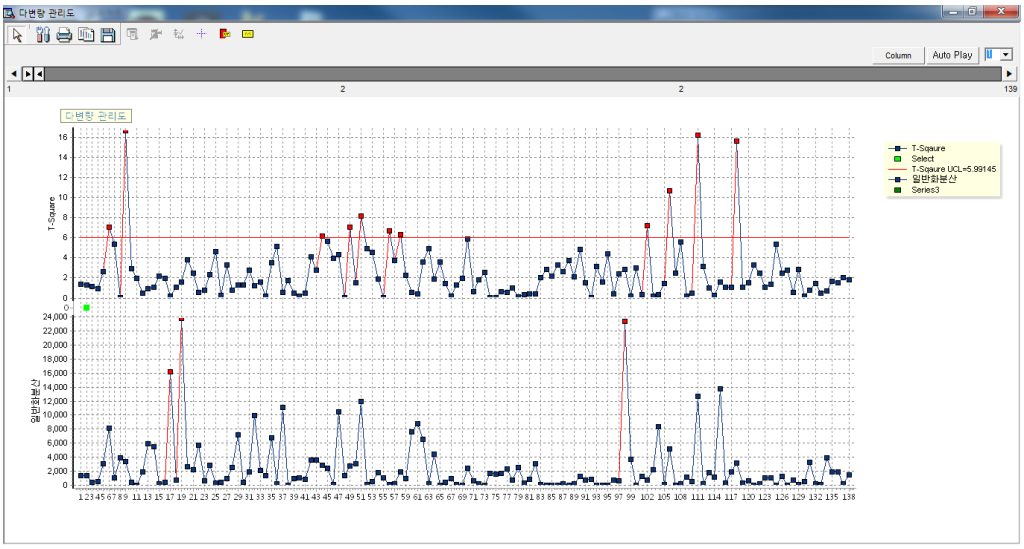
(4) T^2-일반화분산 관리도

T^2-일반화분산 관리도는 공정평균과 변동을 한 화면에서 확인합니다. 일변량 관리도의 Xbar - S, I-MR 에 해당하는 관리도입니다.

사용법

- 차트형태에서 T^2-일반화분산을 선택합니다.
- 일반화분산은 부분군이 존재하여야만 관리도를 그릴 수 있습니다.
- 변수를 선택합니다.
- 부분군의 개수를 입력합니다.

결과



3.4 모델링 노드

실제로 데이터를 분석하는 부분이 모델링 노드입니다. **ECMiner™**에서 지원하는 분석 알고리즘은 다음과 같습니다.

- **연관성 분석**

변수들 간의 연관규칙을 찾아냅니다. (Apriori 알고리즘)

- **CART**

  의사결정나무를 통해 분류 및 회귀분석을 합니다. (Classification and Regression Tree)

- **HIERARCHICAL**

유사성에 따른 군집분석을 합니다.

- **KMEANS**

  입력 받은 군집 수만큼 유사성에 따라 군집을 나눕니다.

- **KNN**

Non-parametric 분류 방법으로써, 이미 분류된 데이터에 새로운 관측치를 유사도에 따라 분류합니다.

- **LDA**

  각 클래스를 구분하는 공간상의 선형평면으로 관측치를 분류합니다.

- **LOGISTIC**

  종속변수가 범주형일 때 이를 효과적으로 분류하는 분석법입니다. (Logistic Regression)

- **MANUAL CART**

  사용자가 원하는 조건에 따라 의사결정나무를 그립니다.

- **MLP**

  다층신경망으로 자가학습을 통해 분류 및 회귀분석을 합니다. (Multi Layer Perception)

- **MLR**

  다변량 선형회귀분석을 합니다.

- **PCA**

독립변수를 분산 설명도에 따라 주성분으로 변환합니다.

- **PCR**

  PCA 를 거친 주성분으로 회귀분석을 합니다.

- **PLS**

  동시에 여러 종속변수들을 회귀분석 합니다.

- **QDA**

  각 클래스를 공간상의 비선형평면으로 구분하여 관측치를 분류합니다.

- **RBF**

  입력데이터를 신경망을 통해 고차원 공간(High-dimensional Space)으로 Feature mapping 을 하여 분류 및 회귀분석을 합니다.

- **순차 연관성**

  시간의 순서나 구매 순서 등 여러 가지 순차적으로 일어나는 사건에서 상관관계를 도출합니다.

- **SOM**

고차원 데이터를 저차원으로 투영, Mapping 시키는 방법으로 클러스터링의 기능 또한 수행합니다.

- **Factor Analysis**

  여러 변수들 상에서 변수들을 설명하는 공통 요인을 추출하는 방법론입니다.

- **RBF DDA**

  기존 RBF 모델링과는 다르게 RBF 중심 수를 자동적으로 정하면서 Classification 을 수행하는 방법입니다.

- **ScoreCard**

Logistic 회귀분석을 방법론을 이용하여 구간화된 데이터의 중요 정도를 Score 로 표시하여 사용자들의 변수의 중요성을 쉽게 알아볼 수 있도록 한 방법론입니다. ScoreCard 는 금융 관련 기관에서 많이 사용되고 있습니다.

- **SVM**

다수의 속성(attribute) 또는 변수를 갖는 객체(object)를 사전에 정해진 그룹 또는 범주(class, category) 중의 하나로 분류하는 분류분석 기법 중 하나로 Support Vector Machine 알고리즘을 사용합니다.

- **SVR**

Support Vector Machine 의 아이디어를 사용하여 회귀분석을 수행합니다.

- **One Class SVM**

One class SVM 은 Support Vector Machine 을 활용하여 이상치 판별을 하는 알고리즘입니다.

- **LOF**

LOF 는 Local Outlier Factor 의 약자로, 데이터가 입력되었을 때 입력된 데이터와 가까운 곳에 위치한 기존 데이터들의 지역적인 밀도를 반영하여 이상치를 판별하는 알고리즘입니다.

3.4.1 연관성 분석 노드

연관성 분석은 상품 혹은 서비스간의 관계를 살펴보고 이로부터 **유용한 규칙**을 찾아내고자 할 때 이용될 수 있는 기법입니다. **ECMiner™**에서는 연관성 분석을 위해 **Apriori 알고리즘**을 제공합니다.

개요

연관성 분석은 하나의 거래나 사건에 포함되어 있는 항목들의 관련성을 파악하는 탐색적 자료분석 방법입니다.

주로 동시에 구매될 가능성이 많은 상품을 분석함으로써 **시장 바구니 분석**에 사용됩니다.

연관성 분석에서의 연관규칙은 "상품 A 를 구매하는 경우는 상품 B 도 구매된다." 라고 해석됩니다.

연관성 분석은 제품이나 서비스의 **교차판매(cross-sell), 매장진열(display), 첨부우편(attached mailings), 사기적발(fraud detection)** 등에 사용됩니다.

연관성 분석은 결과 이해가 쉽고 거래내용에 대한 데이터를 변환 없이 그 자체로 이용할 수 있는 간단한 자료구조를 갖는 분석 방법입니다. 그러나 너무 세분화된 품목을 가지고 **연관성 규칙**을 찾으면 의미 없는 분석이 될 수도 있고 거래량이 적은 품목은 포함된 거래수가 적기 때문에 규칙 발견 시 제외되기가 쉽습니다.

예) " 신발을 구매하는 고객의 10%는 양말을 동시에 구입한다."

고려사항

- **연관성 분석**을 이용할 수 있는 데이터는 판매시점에서 기록된 거래와 품목에 관한 정보를 담고 있어야 합니다.
- 데이터의 형태는 **결과변수(target)**를 갖지 않는 **Unsupervised data** 입니다. 분석 시 고객에 대한 정보 및 기타 정보들을 필요로 하지는 않습니다.
- **DATA TYPE** 은 아래와 같은 형태로 거래는 **1**, 비거래는 **0** 이 되어야 합니다.

	1	2	3	4	5	6
	사과	배	귤	감	바나나	수박
1	1	1	1	0	0	
2	1	1	1	1	1	
3	0	1	1	0	0	
4	0	0	0	1	1	
5	0	0	1	0	0	
6	0	1	1	0	1	
7	1	0	1	0	1	
8	0	1	0	1	0	
9	1	0	0	1	0	
10	0	0	1	0	0	
11	1	1	1	0	1	

사용법

- **입력 노드**를 통해 데이터를 읽어 들입니다.(모든 변수는 정수형태입니다)
- **연관성 분석 노드**를 연결하고 **최소 지지도, 아이템 셋 크기, 최소 신뢰도**를 입력합니다.
- **연관성 분석 diagram** 예시는 아래와 같습니다.(필터 노드는 데이터에 따라 필요하지 않을 수도 있음)

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	

모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	최소지지도(개)	항목들 간에 연관성을 정의하기 위한 최소한의 지지도 값을 설정하는 것으로 Default 는 5 입니다.	필수	자연수입력
	아이템 셋 크기 (개)	분석 시 고려할 최대 연관 아이템 수를 정합니다. Default 는 5 입니다.	필수	자연수입력
	최소신뢰도(%)	연관성 정의 시 최소한의 신뢰도 값을 설정하는 것으로 Default 값은 50(%) 입니다.	필수	0~100
	결과 아이템셋	결과 아이템 셋을 입력합니다. 단일 결과 혹은 다중 결과를 선택할 수 있습니다.	필수	단일 결과 다중 결과
	트랜잭션(TID) 옵션	트랜잭션 아이디로 정렬해야 하는지 혹은 이미 정렬되어 있는지 혹은 하나의 행을 아이디로 보면 되는지를 선택합니다.	필수	TID 로 정렬해야 함, TID 로 정렬되어 있음, 행이 TID 임
	아이템 옵션	현재의 데이터 형태가 아이템이 개수로 적혀 있는지 혹은 하나의 트랜잭션이 아이템명으로 표시되어 있는지를 선택합니다.	필수	아이템명으로 존재, 개수로 존재
	지지도 아이템	지지도 아이템을 입력합니다.	필수	

예시파일

- **연관석 분석.ecm 실행**

3.4.2 CART 노드

의사결정나무(Decision Tree)는 의사결정 진행과정을 나무형태로 표현한 것으로 중요 관심변수와 기타변수들 사이의 연관성 정도에 따라 중요 변수를 선별, 의사결정 조건을 생성하여 효율적인 의사결정을 지원합니다. **ECMiner™**에서는 **CART 알고리즘**을 제공합니다.

개요

의사결정나무는 누구나 이해할 수 있고, 쉽게 설명되는 결과의 간결함으로 인해 많은 분야에서 선호하고 있습니다. 또한 누락된 관측값에 대한 처리가 다른 모델보다 우수하고 변수간의 교호작용의 설명과 처리가 용이하다는 장점이 있습니다.

의사결정나무는 하나의 나무구조를 이루고 있으며, **마디(node)**라고 불리는 구성 요소들로 이루어져 있습니다.

(1) **트리의 형성(Growing the Tree)**
학습표본을 바탕으로 트리를 형성하는 과정으로 **변수의 선정**, **분리기준** (Splitting Criterion), **정지규칙(Stopping Rule)**등의 결정이슈가 있습니다.

(2) **가지치기(Pruning)**
트리가 너무 복잡하면 **분류규칙** 또한 복잡해지기 때문에 일부의 가지를 절단하여 그 가지 이하에서 더 이상 분리하지 않도록 하는 과정입니다.

(3) **분류**
최종적으로 완성된 트리를 바탕으로 **분류규칙**을 도출하고, 기존 또는 새로운 데이터에 대하여 분류를 시행하는 과정입니다.

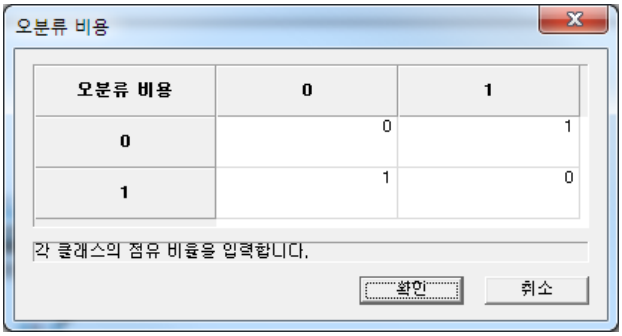
고려사항

- CART 노드 자체가 종속변수의 데이터 타입에 맞는 수행기법을 선택해 종속변수가 연속형일 때는 **예측분석**을 이산형일 때는 **분류분석**을 수행합니다.
- CART 노드에서의 불순도 함수는 종속변수가 연속형일 때(예측분석 시)는 LSD(Least Squared Deviation)을 사용하며, 종속변수가 이산형일 때(분류분석 시)는 Gini, Twoing, Entropy 중 선택할 수 있습니다.

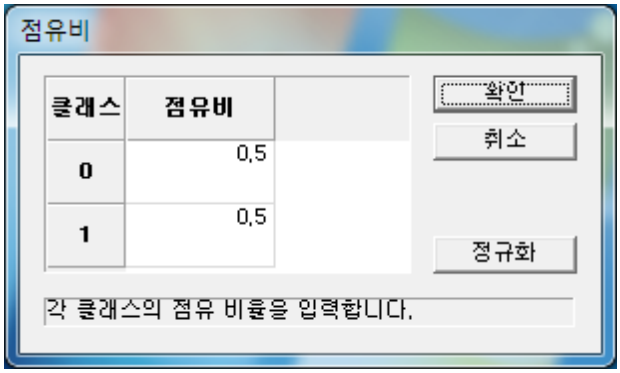
사용법

- **입력노드**를 통해 데이터를 읽어 들입니다.
- **형태 변경 노드**를 통해 읽어 들인 데이터의 타입을 지정합니다. (**독립변수, 종속변수**를 지정)
- **CART 노드**를 **형태 변경 노드**에 연결하고 옵션들을 선택합니다.

▪ **오분류비용 입력**을 **기본값**으로 선택하면 동일한 비율로 분석하고 **직접입력**을 선택하면 아래와 같은 대화창이 떠서 사용자가 각 클래스 별 오분류 비용을 입력합니다.



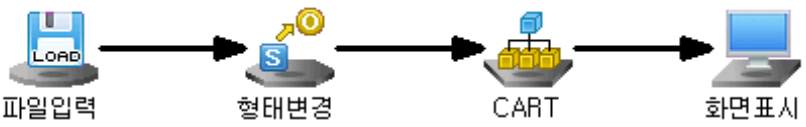
점유도 입력을 **기본값**으로 선택하면 동일한 비율로 분석하고 **직접입력**을 선택하면 아래와 같은 대화창이 떠서 사용자가 각 클래스 별 점유비를 입력합니다.



오분류 비용과 **점유비**는 **분류분석**을 할 때만 사용됩니다.

화면표시 노드를 **CART 노드**에 연결합니다.

CART diagram 예시는 아래와 같습니다.



속성

	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	최소분리 데이터수	트리 형성 시 더 이상 분리를 하지 않을 최소 분리 데이터 수를 말합니다.	필수	자연수.
	불순도함수	트리 형성 시 분리기준으로서의 복잡도의 척도를 나타냅니다.	필수	Gini, Twoing, Entropy, LSD
	가지치기여부	생성된 Tree 의 가지치기 여부를 선택합니다.	필수	예, 아니오
	최대트리깊이	트리가 가질 수 있는 최대 깊이를 입력합니다. 종료 조건 시 사용되며 0 일때 무시됩니다.	필수	
	불순도조건	트리의 각 노드가 불순도 조건 이하의 불순도를 가지면 정지노드가 됩니다.	필수	
	오분류비용	오분류 비용(i,j)는 i-클래스를 j-클래스로 오분류 했을 시의 비용을 나타냅니다. Default 는 diagonal=0, 나머지는 1. 즉, 제대로 분류하면 0, 오분류 하면 1 이 됩니다.	조건부 필수	기본값, 직접입력
	점유도	점유도는 i-class 가 차지하는 비율을 나타냅니다. Default 는 받아들인 데이터 내에서의 각 클래스 비율을 사용합니다.	조건부 필수	기본값, 직접입력

결과

화면표시 노드에서 종속변수의 형태에 따라 분류/예측 결과를 확인할 수 있습니다.

	32	33	34	35
	RESPONSE	CART32_YHAT	CART32_PROB_0	CART32_PROB_1
1	1	1	0,13458	0,86542
2	0	0	0,68342	0,31658
3	1	1	0,13458	0,86542
4	1	0	0,68342	0,31658
5	0	0	0,68342	0,31658
6	1	1	0,13458	0,86542
7	1	1	0,13458	0,86542
8	1	1	0,40625	0,59375
9	1	1	0,13458	0,86542

원래 범주 예측 범주 범주가 0 일 확률 범주가 1 일 확률

예시파일

- **CART.ecm** 실행

3.4.3 HIERARCHICAL 노드

군집분석은 데이터를 비슷한 특성끼리 묶는 것으로 개체들 사이의 상사성 또는 비상사성에 근거하여 군집을 찾고 자료를 요약하는 탐색적 자료분석 방법입니다. 군집분석은 군집의 개수, 내용, 구조 등을 완전히 모르는 상태에서 특성을 파악하기 위한 분석 방법입니다. 군집 분석 중 **계층적 방법**은 사전에 군집 수 k 를 정하지 않고 단계적으로 서로 다른 군집결과를 제공하는 것입니다. **ECMiner™**에서는 계층적 군집분석 중 **가까운 객체끼리 군집화**시키는 **agglomerative** 알고리즘을 제공합니다.

개요

Agglomerative 알고리즘에 의한 계층적 군집방법은 각 객체를 하나의 군집으로 시작하여 군집들을 묶어가는 과정을 반복하여 결국 모든 객체가 하나의 군집이 되도록 하는 것이다.

계층적 군집분석은 개체의 수가 적을 때 유용한 방법입니다.

계층적 군집분석이 수행되는 과정은 다음과 같습니다.

(1) 처음에는 각 개체를 하나의 군집으로 생각합니다. (따라서, n 개의 군집이 있습니다.)

(2) 각 군집간의 거리 또는 유사성을 구합니다.

(3) 가장 가까운 한 쌍의 군집을 선택하여 이 두 군집을 합하여 한 개의 새로운 군집을 형성합니다. (따라서, 군집의 수가 하나 줄어듭니다.)

(4) 새로 형성된 군집의 모임에 대하여 (2)와 (3)을 반복합니다.

계층적 군집분석은 거리 계산 방법이나 연결(Linkage) 방법에 따라 여러 결과가 도출될 수 있습니다.

고려사항

- 데이터들의 단위가 많이 다를 경우에는 전처리 옵션을 이용하여 데이터를 표준화 시킨 후 군집분석을 수행합니다.

사용법

- **입력 노드**를 통해 데이터를 읽어 들입니다.
- **HIERARCHICAL 노드**를 **입력 노드**에 연결하고 옵션들을 선택합니다.
- **화면표시 노드**를 **HIERARCHICAL 노드**에 연결합니다.
- **HIERARCHICAL diagram** 예시는 아래와 같습니다.

속성

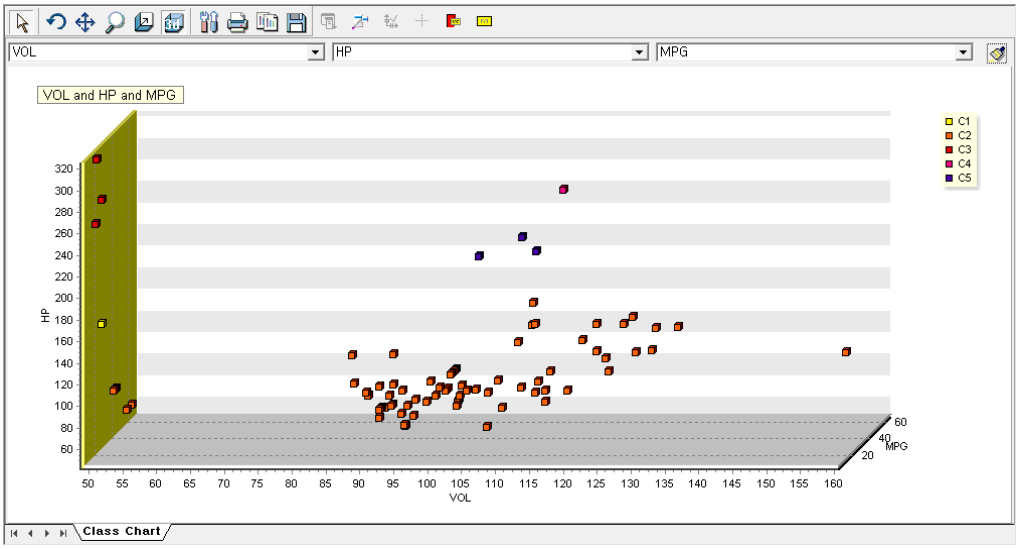
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 입력합니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	결과보기	모델링 후 Class Chart 를 보여줍니다.	클릭	모델링 후 클릭
	절단기준	전체 데이터들을 그룹화하는 기준으로 클러스터 개수는 입력된 클러스터 개수로 전체 데이터를 그룹화합니다. 그리고 절단거리를 선택한 경우, 클러스터 간의	필수	클러스터 개수, 절단거리

		절대적 거리를 바탕으로 전체 데이터를 그룹화합니다.		
	클러스터 개수	전체 데이터를 그룹화하려는 그룹의 수를 지정합니다.	필수	자연수입력
	절단거리	그룹화의 기준이 되는 클러스터 간의 절대적 거리로서 0 이상의 실수값을 가집니다.	필수	>= 0
	거리측정	거리측정에 사용되는 사용되는 방법을 선택합니다.	필수	Euclidean, Squared Euclidean, Manhattan, Minkowski Hamming, Jaccard
	연결방법	연결방법을 선택합니다.	필수	Single, Complete, Average, Centroid, Ward
	Minkowski 계수	Minkowski 거리를 이용할 때 사용되는 계수를 입력합니다.	조건부 필수	
	전처리방법	입력자료 전처리방법을 결정합니다.	필수	전처리 안함, 표준화, 평균보정

결과

- **클래스차트**

옵션창의 결과보기 버튼을 클릭하면 뜨게 됩니다. 상단의 변수 선택창을 통해 최대 3 개 변수에 대하여 차트 결과를 보여줍니다. 차트를 통해 각 관측치의 값 및 군집결과를 볼 수 있습니다.



- **군집할당정보**

화면표시 노드에서 각 관측치의 군집정보 즉, 어떤 Cluster 에 할당되었는지를 알 수 있습니다.

	1	2	3	4	5	6	7
	MAKE	VOL	HP	MPG	sp	wt	HC16_YHAT
1	GM/GeoMetr	89	49	65,40000	96	17,50000	2
2	GM/GeoMetr	92	55	56,00000	97	20,00000	2
3	GM/GeoMetr	92	55	55,90000	97	20,00000	2
4	SuzukiSwift	92	70	49,00000	105	20,00000	2
5	DaihatsuCha	92	53	46,50000	96	20,00000	2
6	GM/GeoSpri	89	70	46,20000	105	20,00000	2
7	GM/GeoSpri	92	55	45,40000	97	20,00000	2
8	HondaCivicC	50	62	59,20000	98	22,50000	2
9	HondaCivicC	50	62	53,30000	98	22,50000	2
10	DaihatsuCha	94	80	43,40000	107	22,50000	2
11	SubaruJustv	89	73	41,10000	103	22,50000	2
12	HondaCivicC	50	92	40,90000	113	22,50000	2
13	HondaCivic	99	92	40,90000	113	22,50000	2
14	SubaruJustv	89	73	40,40000	103	22,50000	2
15	SubaruJustv	89	66	39,60000	100	22,50000	2
16	SubaruJustv	89	73	39,30000	103	22,50000	2
17	ToyotaTerce	91	78	38,90000	106	22,50000	2
18	HondaCivicC	50	92	38,80000	113	22,50000	2
19	ToyotaTerce	91	78	38,20000	106	22,50000	2
20	FordEscort	103	90	42,20000	109	25,00000	2
21	HondaCivic	99	92	40,90000	110	25,00000	2

예측 범주

참조

참조 그룹	참조명	설명
거리측정법	Euclidean	일반적인 거리지표인 유클리드 거리입니다.
	Squared Eculidean	표준화된 유클리드 거리입니다.
	Manhattan	coordinate 의 절대차이의 합을 바탕으로 한 거리입니다.
	Minkowski	Manhattan distance 에서 minkowski 계수가 지수에 들어간 거리입니다.
	Hamming	서로 다른 coordinate 의 수를 바탕으로 한 거리입니다.
	Jaccard	서로 다른 nonzero coordinate 의 수를 바탕으로 한 거리입니다.
연결방법	Single	단일 연결법에서는 군집 i 와 군집 j 의 유사성 척도로 두 군집의 모든 객체 쌍의 거리 중 가장 가까운 거리를 사용합니다.
	Complete	완전 연결법에서는 군집 i 와 군집 j 의 유사성 척도로 두 군집의 모든 객체 쌍의 거리 중 가장 먼 거리를 사용합니다.
	Average	평균 연결법은 Sokal and Michener (1958)에 의하여 비롯된 것으로, 군집 i 와 군집 j 의 유사성 척도로 두 군집의 모든 객체 쌍의 평균거리를 사용합니다.
	Centroid	두 군집의 중심간 거리를 사용합니다.
	Ward	모든 쌍을 묶은 경우 각각에 대하여 전체 군집 내 제곱합을 산출한 후 가장 이 값이 작게 되는 군집 쌍을 묶습니다.

예시파일

- **Hierachial_KMEANS.ecm** 실행

3.4.4 KMEANS 노드

군집분석은 데이터를 비슷한 특성끼리 묶는 것으로 개체들 사이의 상사성 또는 비상사성에 근거하여 군집을 찾고 자료를 요약하는 탐색적 자료분석 방법입니다. 군집분석은 군집의 개수, 내용, 구조 등을 완전히 모르는 상태에서 특성을 파악하기 위한 분석 방법입니다. 군집분석 중 **KMEANS** 는 비계층적 군집분석 방법입니다. 비계층적 군집분석은 사전에 군집 수 k 를 정한 후 각 군집의 대표값 또는 대표객체를 정한 후 각 객체를 k 개 중의

한 군집에 배정하는 것입니다. **ECMiner™**는 비계층적 군집분석을 위해 KMEANS 알고리즘을 제공합니다.

개요

군집의 수를 사전에 지정하고 대상객체들을 적절한 군집에 배정하는 **비계층적 군집방법(Non-HIERARCHICAL Cluster Method)**중 가장 대표적인 방법으로, 이 방법은 n 개의 객체를 사전에 결정된 k 개의 군집에 할당하는 것이라고 할 수 있습니다.

군집분석이 수행되는 과정은 다음과 같습니다.

(1) (초기 객체 선정) 어떤 규칙에 의하여 k 개의 객체의 좌표를 초기군집의 **중심좌표(centroid)**로 선정합니다.

(2) (객체의 군집 배정) 각 객체에 대하여 k 개의 군집 **중심좌표**와의 거리를 산출한 후 가장 가까운 군집에 그 객체를 배정합니다.

(3) (군집 중심좌표의 산출) 새로운 군집에 대한 **중심좌표**를 산출합니다.

(4) (수렴 조건 점검) 새로 산출된 중심 좌표값과 이전 좌표값을 비교하여 변화가 없으면 마치며, 그렇지 않으면 단계 1 을 반복합니다.

K-Means 군집방법은 계산효율이 대체로 양호하나 이상치(outliers)가 존재할 때 결과가 좋지 않을 수 있는 것으로 알려져 있습니다.

고려사항

- **K-Means Clustering** 을 수행하기 위해 사용되는 거리 계산법은 **유클리디안 거리 계산법**입니다.

사용법

- 입력노드를 통해 데이터를 읽어 들입니다.
- KMEANS 노드를 입력 노드에 연결하고 옵션들을 선택합니다.
- 화면표시 노드를 KMEANS 노드에 연결합니다.
- KMEANS diagram 예시는 아래와 같습니다.

속성

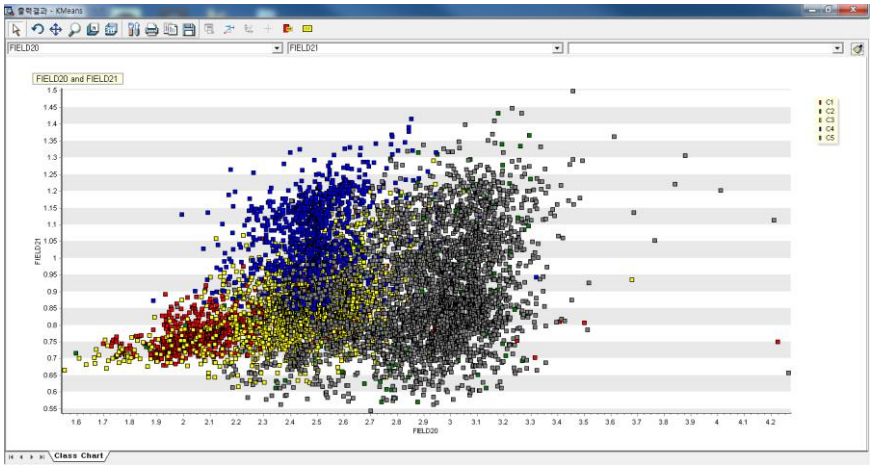
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	결과보기	모델링 후 Class Chart 를 보여줍니다.	클릭	모델링 후에 클릭
	군집수	군집의 수를 결정하는 옵션입니다. K-Means 수행 시 군집의 수 K 를 정해야 하므로 분석 전에 plot 등을 통해 군집의 수를 대략적으로 파악해 놓는 것이 좋습니다.	필수	자연수
	최대 시행수	K-Means 분석 시 시행횟수를 결정합니다.	필수	자연수
	초기화 수	초기 centorid 가 랜덤하게 정해지기 때문에 가장 좋은 결과가 언제 나올지 모릅니다. 따라서 총 루프를 여러 번 돌아야 합니다. 군집중심을 무작위로 구하는 횟수가 초기화 수입니다.	필수	자연수
	거리 측정법	Euclidean: 일반적인 거리지표인 유클리드 거리입니다 Squared Euclidean: 표준화된 유클리드 거리입니다. Manhattan: coordinate 의 절대차이의 합을 바탕으로 한 거리입니다 Minkowski: Manhattan distance 에서 minkowski 계수가 지수에 들어간 거리입니다.	필수	
	Minkowski 계수	Minkowski 를 선택하였을 경우 계수를 입력합니다.	Minkowski 를 입력하였을	

			경우	
	전처리 방법	정규화: 데이터를 정규화 합니다. ➔ X - X(min) / X(max) - X(min) 표준화: 데이터를 표준화 합니다. ➔ (X-Mean) / Std 없음: 전처리 없이 원본데이터를 이용하여 실행합니다.	필수	

결과

- **클래스차트**

옵션창의 결과보기 버튼을 클릭하면 뜨게 됩니다. 상단의 변수 선택창을 통해 최대 3 개 변수에 대하여 차트 결과를 보여줍니다. 차트를 통해 각 관측치의 값 및 군집결과를 볼 수 있습니다.



- **군집할당정보**

화면표시 노드에서 각 관측치의 군집정보 즉, 어떤 Cluster 에 할당되었는지, 해당 그룹의 중심까지의 거리를 알 수 있습니다.

	1	2	3	4	5	6	7
	VOL	HP	MPG	sp	wt	KM1_YHAT	KM1_DISTANCE
1	89	49	65,40000	96	17,50000	1	0,13102
2	92	55	56,00000	97	20,00000	1	0,06463
3	92	55	55,90000	97	20,00000	1	0,06639
4	92	70	49,00000	105	20,00000	3	0,17923
5	92	53	46,50000	96	20,00000	3	0,21045
6	89	70	46,20000	105	20,00000	3	0,13482
7	92	55	45,40000	97	20,00000	3	0,18949
8	50	62	59,20000	98	22,50000	2	0,05651
9	50	62	53,30000	98	22,50000	2	0,05651
10	94	80	43,40000	107	22,50000	3	0,06956
11	89	73	41,10000	103	22,50000	3	0,04633
12	50	92	40,90000	113	22,50000	3	0,37053
13	99	92	40,90000	113	22,50000	3	0,14560
14	89	73	40,40000	103	22,50000	3	0,04885
15	89	66	39,60000	100	22,50000	3	0,09598
16	89	73	39,30000	103	22,50000	3	0,05907
17	91	78	38,90000	106	22,50000	3	0,05389

예측 범주 중심까지의 거리

▪ **Hierachial_KMEANS.ecm** 실행

3.4.5 KNN 노드

다수의 속성(attribute) 또는 변수를 갖는 객체(object)를 사전에 정해진 **그룹 또는 범주(class, category)** 중의 하나로 분류하는 **비모수적 분류분석 기법** 중 하나로 모집단에 대한 가정 없이 간단히 사용할 수 있는 기법입니다.

개요

k-Nearest Neighbor 는 표본의 분포상태에 영향을 받지 않는 **non-parametric 학습방법**의 하나로서, 모든 표본이 n-차원 공간상의 점들로 대응된다고 가정합니다. 즉, n-차원 공간에서 자신과 가장 가깝게 위치하는 k 개의 다른 표본들의 클래스 중에서 가장 많은 것으로 분류되는 것이 **K-Nearest Neighbor 분류방법**입니다.

즉, K-NN 은 학습 데이터를 활용하지만 학습 데이터를 활용하여 규칙을 도출하는 기법이 아니며, 분류하고자 하는 새로운 객체에 대하여 학습 데이터에 있는 가장 가까운 몇 개의 객체들을 찾은 후 이들 인접 객체들이 가장 많이 속하는 범주로 분류하는 기법입니다.

분류분석이 수행되는 과정은 다음과 같습니다.

(1) n 개의 객체를 가진 표본행렬을 입력 받습니다.

(2)'새로운 객체'를 입력 받아 n 개의 객체 사이의 거리를 구하여 가장 거리가 짧은 k 개의 객체를 정합니다.

(3) 고려군에 포함되어 있는 객체가 속한 그룹 중 가장 다수를 차지하고 있는 그룹을 '새로운 객체'에 부여합니다.

(4) 다시 '새로운 객체'를 입력 받아 입력이 끝날 때까지 단계 2 부터 단계 4 를 반복합니다.

고려사항

- 거리가 같은 객체가 k 개 이상 발생하면, 이를 모두 고려군에 포함시킵니다.
- 최다의 그룹이 두 개 이상 존재하면, 구해놓은 거리의 합이 가장 짧은 그룹으로 결정합니다.
- KNN 을 수행하기 위해 **종속변수**는 이산형이어야 하며, **독립변수**는 연속형이어야 합니다.

사용법

- **입력 노드**를 통해 데이터를 읽어 들입니다.
- **형태 변경 노드**를 통해 읽어 들인 데이터의 타입을 지정합니다.(**독립변수, 종속변수**를 지정)
- **KNN 노드**를 **형태 변경 노드**에 연결하고 옵션들을 선택합니다.
- **KNN diagram** 예시는 아래와 같습니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	K 수	몇 번째 가까운 관측치까지의 클래스를	필수	자연수

		고려할 지 정합니다.		
	전처리방법	KNN 수행 전 데이터에 대한 전처리 방법을 설정하는 옵션입니다.	필수	비처리, 표준화, 평균보정
	Leave One Out	Test data 가 없을 시 Leave one out 으로 검정합니다.	예 아니오	

예시파일

- **KNN.ecm** 실행

3.4.6 LDA 노드

다수의 속성(attribute) 또는 변수를 갖는 객체(object)를 사전에 정해진 **그룹 또는 범주(class, category)** 중의 하나로 분류하는 **분류분석 기법** 중 하나로 각 그룹의 **분산-공분산 행렬**이 동일할 때 사용합니다.

개요

각 개체는 p 개의 변수 $x'=(X_1,...,X_p)$ 로 이루어지는데, 범주 1 또는 범주 2 중 하나를 취한다고 가정합니다. 그리고 객체 x 의 **평균벡터**와 **분산-공분산행렬(variance-covariance matrix)**을 아래와 같다고 가정합니다.

$$E[x]=\begin{cases}\mu_1 & x\text{가 범주 1에 속할 때}\\ \mu_2 & x\text{가 범주 2에 속할 때}\end{cases}$$

$$Var[x]=\Sigma \quad \text{(범주에 관계없이 동일)}$$

한편, 학습표본으로 총 n 개의 객체가 다음과 같이 n1 개는 범주 1 에, n2 개는 범주 2 로 구분된다고 가정합니다.

범주 1: $x_i^{(1)}, \quad i=1,...,n_1$

범주 2: $x_i^{(2)}, \quad i=1,...,n_2$

피셔의 방법은 우선 원래 변수의 선형조합으로 새로운 변수를 형성한 후 이를 바탕으로 분류규칙을 만드는 것입니다. 다음과 같은 원래 변수들의 선형조합으로 새로운 변수를 형성하는 함수를 **판별함수(Discriminant function)** 이라 합니다.

$$Z=w_1X_1+w_2X_2+...+w_pX_p=w'x$$

위 **판별함수**의 계수 w 를 결정하여, 아래 분류규칙에 따라 객체 x 를 각 범주로 분류합니다.

$$|\hat{w}'(x-\bar{x}^{(1)})| \leq |\hat{w}'(x-\bar{x}^{(2)})| \text{ 이면, x 를 범주 1 로 분류}$$

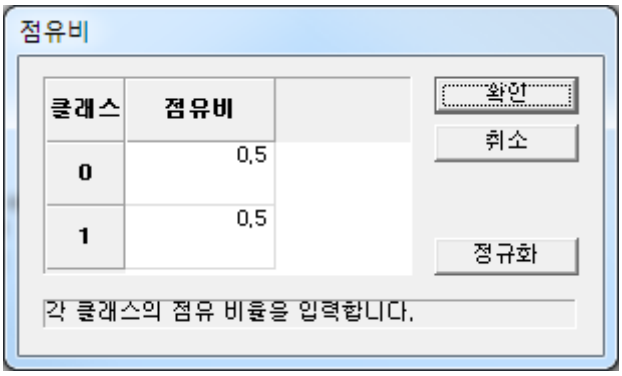
$$|\hat{w}'(x-\bar{x}^{(1)})| > |\hat{w}'(x-\bar{x}^{(2)})| \text{ 이면, x 를 범주 2 로 분류}$$

고려사항

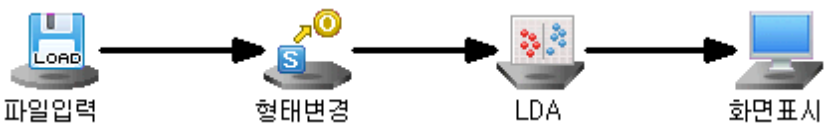
- **종속변수**는 이산형이어야 합니다.
- 벡터 x 가 **다변량 정규분포**를 따르는 것을 가정합니다.
- 각 범주에 관계없이 **분산-공분산 행렬**이 동일함을 가정합니다.k

사용법

- **입력 노드**를 통해 데이터를 읽어 들입니다.
- **형태 변경 노드**를 통해 읽어 들인 데이터의 타입을 지정합니다. (**독립변수, 종속변수**를 지정)
- **LDA 노드**를 **형태 변경 노드**에 연결하고 옵션들을 선택합니다.
- **점유도 입력**을 **기본값**으로 선택하면 동일한 비율로 분석하고 **직접 입력**을 선택하면 아래와 같은 대화창이 떠서 사용자로부터 각 클래스 별 점유비를 입력 받습니다. 대화창에서 취소를 선택하면 기본값으로 분석합니다.



- **화면표시 노드**를 **LDA 노드**에 연결합니다.
- **LDA diagram** 예시는 아래와 같습니다.



속성

속성 그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	점유도	점유비율 입력여부를 결정합니다. 점유도는 데이터에서 각 class 의 비율을 의미합니다.	필수	기본값, 직접입력

결과

▪ **분석결과정보**

화면표시 노드에서 분류분석 결과를 확인할 수 있습니다.

	3	4	5	6	7	8	9	10	11	12	13	14
	A3	A4	A5	A6	A7	A8	A9	Y	LDA8_YHAT	LDA8_POS	LDA8_PROB_0	LDA8_PROB_1
1	M	B	181,39500	42,00000	213	921,00000	A	0	0	0,99476	0,99476	0,00524
2	M	A	218,38000	191,40000	554	1,082,10000	A	1	1	0,91958	0,08042	0,91958
3	M	A	168,69900	37,80000	33	909,00000	A	1	1	0,95474	0,04526	0,95474
4	M	G	166,95200	128,40000	215	990,00000	A	0	0	0,63500	0,63500	0,36500
5	M	H	137,70700	102,00000	421	768,60000	A	1	1	0,98619	0,01381	0,98619
6	M	D	131,15100	36,50000	515	947,00000	A	1	1	0,99198	0,00802	0,99198
7	M	G	210,27900	93,50000	355	1,067,60000	A	1	0	0,60512	0,60512	0,39488
8	M	G	181,52000	70,80000	332	881,40000	A	0	0	0,66648	0,66648	0,33352
9	M	A	177,08500	0,00000	144	750,60000	A	1	1	0,94706	0,05294	0,94706
10	M	B	80,23190	81,60000	161	913,20000	A	1	0	0,98794	0,98794	0,01206
11	M	H	199,45000	78,00000	478	1,069,50000	A	1	1	0,97000	0,03000	0,97000
12	MH	H	205,54300	22,80000	406	1,525,50000	A	1	1	0,98011	0,01989	0,98011
13	M	G	182,79500	24,00000	485	983,10000	A	1	0	0,54521	0,54521	0,45479
14	M	A	156,26800	28,80000	202	913,20000	A	1	1	0,96790	0,03210	0,96790
15	M	K	100,79200	77,50000	527	890,80000	A	1	1	0,95042	0,04958	0,95042
16	M	B	99,35700	1,80000	531	867,00000	A	0	0	0,98277	0,98277	0,01723
17	MH	F	96,22250	0,00000	679	1,455,60000	A	1	1	0,99289	0,00711	0,99289
18	M	B	92,01710	3,50000	361	878,00000	A	0	0	0,98806	0,98806	0,01194
19	M	A	157,76400	49,80000	370	843,30000	A	1	1	0,95667	0,04333	0,95667
20	M	D	127,64100	94,00000	544	976,30000	A	1	1	0,99559	0,00441	0,99559

원래 범주 예측 범주

예시파일

▪ **LDA_QDA.ecm** 실행

3.4.7 LOGISTIC 노드

하나 또는 둘 이상의 변수들이 다른 하나의 변수에 미치는 영향의 정도와 방향을 파악하기 위해 회귀분석을 사용합니다. 로지스틱 회귀분석은 회귀분석과 같이 **하나의 종속변수**와 **한 개 이상의 독립변수 사이의 관계**를 표현하기 위해, 가장 잘 적합되고, 모수의 수를 절약한 모형을 찾는 것입니다. 로지스틱 회귀분석은 기본적으로 여러 설명 변수들로부터 **두 범주**만을 가지는 반응변수를 예측하는데 사용하며, 3 개 이상의 범주의 경우도 사용 가능합니다. 또한 회귀분석과는 달리 로지스틱 회귀분석은 종속변수와 독립변수들 사이의 함수관계를 선형이 아닌 비선형 관계로 가정하며, 회귀분석의 기본가정(오차의 정규성, 오차의 독립성, 오차의 등분산성)이 가정되지 않습니다.

개요

Binary Logistic Regression 은 종속변수가 2 가지 값의 범주를 취하는 경우에 사용되는 분류 방법입니다. **베르누이(Bernoulli) 분포**를 따르는 경우 **Log-likelihood Function** 은 아래와 같습니다.

$$\log L = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \log(1 + e^{\beta_0 + \beta_1 X_i})$$

한편, 이항분포(binomial distribution)를 따르는 경우의 Likelihood Function 은 아래와 같습니다.

$$L = \prod_{i=1}^{n} \binom{r_i}{y_i} P_i^{y_i} (1 - P_i)^{r_i - y_i}$$

위와 같은 likelihood function 들로부터 beta 들을 추정하여 아래 model 을 fitting 하도록 합니다.

$$Logit(p) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$$

이 로지스틱 회귀모형에서 기본적인 분류규칙은 Y 가 1 일 확률이 0 일 확률보다 크면 1 로 분류하고 그렇지 않으면 0 으로 분류하는 것입니다. 즉,

$P_i > 0.5$이면 객체 i를 '1'로 분류

$P_i \le 0.5$이면 객체 i를 '0'으로 분류

이 규칙은 로짓에 대하여서는 다음과 같습니다.

$Logit(P_i) > 0$ 이면 객체 i를 '1'로 분류

$Logit(P_i) \le 0$ 이면 객체 i를 '0'으로 분류

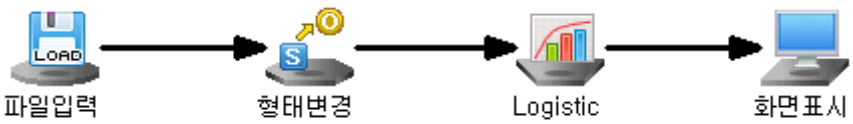
종속변수의 값이 3 개 이상일 경우는 Nominal Logistic Regression 을 사용합니다.

고려사항

- **종속변수**는 이산형이어야 합니다.

사용법

- **입력 노드**를 통해 데이터를 읽어 들입니다.
- **형태 변경 노드**를 통해 읽어 들인 데이터의 타입을 지정합니다.(**독립변수, 종속변수**를 지정)
- **Logistic** 를 **형태 변경 노드**에 연결하고 옵션들을 선택합니다.
- **화면표시 노드**를 **Logistic 노드**에 연결합니다.
- **Logistic diagram** 예시는 아래와 같습니다.



속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 입력합니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	최대시행수	모델 생성시 최대 시행수를 입력합니다.	필수	자연수
	변수 선택기법	None 과 Stepwise 중 하나를 선택할 수 있습니다. None 은 현재 선택된 변수에서 더하거나 빼지 않고 모두 사용하겠다는 것이고 Stepwise 는 Stepwise 기법을 이용하여 변수를 선별하겠다는 뜻입니다.	필수	None 과 Stepwise 중 택일

	입력 유의수준	Stepwise 선택 시 입력 유의수준을 입력합니다.	Stepwise 선택 시 필수	
	제거 유의수준	Stepwise 선택 시 제거 유의수준을 입력합니다.	Stepwise 선택 시 필수	
	(이항분리) 분리 기준점	이항 분리 시에 사용할 분리 기준점(cutoff)을 입력합니다.		

결과

▪ **분석결과정보**

화면표시 노드에서 각 관측치의 예측 클래스를 알 수 있습니다.

	1	2	3	4	5
	FIELD1	FIELD2	FIELD3	LRN_RES	LRN_POS
1	2,400	1,700	1	1	0,275
2	1,800	0,900	1	2	0,723
3	2,400	1,600	1	1	0,341
4	3,000	1,900	1	1	0,262
5	2,000	0,500	2	2	0,916
6	1,200	0,600	2	2	0,792
7	2,000	1,100	2	2	0,628
8	2,700	2,000	2	1	0,164
9	2,000	2,000	1	1	0,093
10	4,000	0,700	2	2	0,974
11	2,200	1,600	1	1	0,301
12	1,900	1,200	2	2	0,530

원래 범주 예측 범주 예측 확률

예시파일

▪ **logistic.ecm** 실행

3.4.8 MANUAL CART 노드

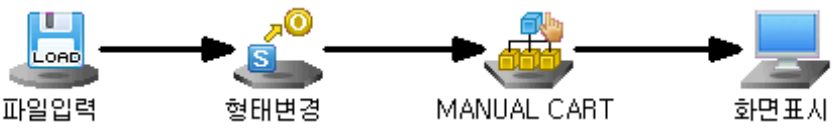
MANUAL CART 노드는 사용자 CART(Decision Tree)로써 본인이 원하는 조건에 따라 의사결정 나무를 그려 가는 노드입니다.

개요

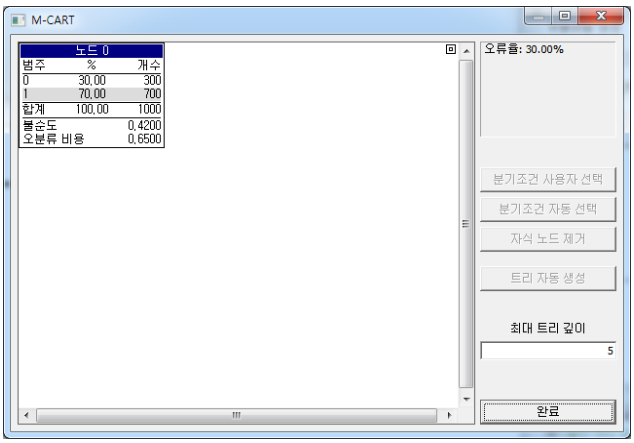
사용자 CART 로 사용자가 임의로 분기조건을 주어 분석할 수 있는 노드입니다. 일반적인 의사결정나무 분석과 같지만 사용자가 임의로 분기조건을 주기도 하고 자동으로 줄 수도 있습니다. 그리고 트리완성 옵션이 있어서 일반 CART 노드에서 사용자 정의 형태 노드라고 볼 수 있습니다.

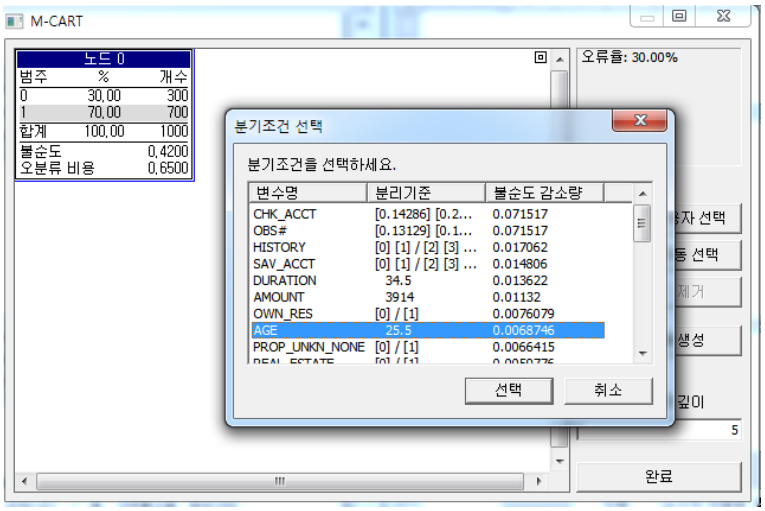
고려사항

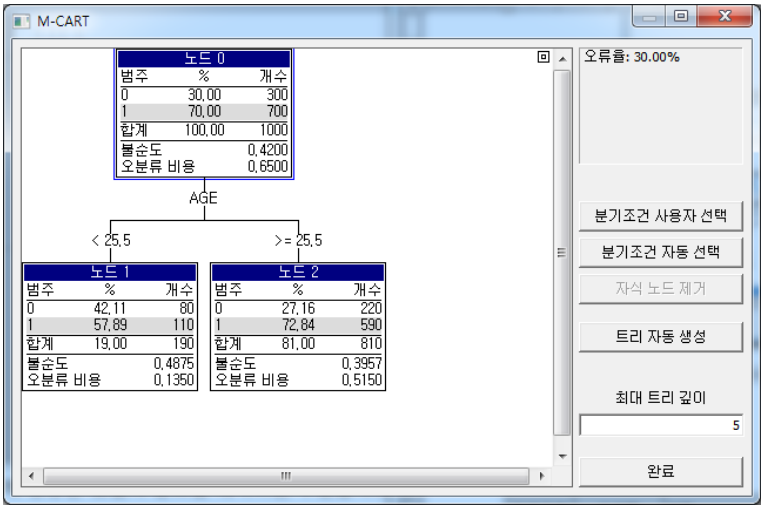
- CART 노드 자체가 종속변수의 데이터 타입에 맞는 수행기법을 선택해 종속변수가 연속형일 때는 **예측분석**을 이산형일 때는 **분류분석**을 수행합니다.
- CART 노드에서의 불순도 함수는 종속변수가 연속형일 때(예측분석 시)는 LSD(Least Squared Deviation)을 사용하며, 종속변수가 이산형일 때(분류분석 시)는 Gini, Twoing, Entropy 중 선택할 수 있습니다.
- 사용법
  - 일반적인 CART 알고리즘에 **MANUAL CART 노드**를 넣고 분석을 수행합니다.



  - 실행 후 다음과 같은 화면이 뜹니다. 이때 개인이 원하는 분기조건을 선택합니다.







속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 입력합니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	

선택사항	불순도 함수	분류 분석의 경우 Gini, Twoing, Entropy 중 하나의 불순도 함수를 선택합니다. 예측 분석의 경우 LSD 만 선택가능 합니다.	필수	Gini, Twoing, Entropy, LSD
	오분류비용	오분류 비용(i,j)는 i-클래스를 j-클래스로 오분류 했을 시의 비용을 나타냅니다. Default 는 diagonal=0, 나머지는 1. 즉, 제대로 분류하면 0, 오분류하면 1 이 됩니다.	조건부 필수	기본값, 직접입력
	점유도	점유도는 i-class 가 차지하는 비율을 나타냅니다. Default 는 받아들인 데이터 내에서의 각 클래스 비율을 사용합니다.	조건부 필수	기본값, 직접입력

결과

▪ **분석결과정보**

화면표시 노드에서 각 관측치의 예측 클래스를 알 수 있습니다.

	32	33	34	35
	RESPONSE	CART32_YHAT	CART32_PROB_0	CART32_PROB_1
1	1	1	0,13458	0,86542
2	0	0	0,68342	0,31658
3	1	1	0,13458	0,86542
4	1	0	0,68342	0,31658
5	0	0	0,68342	0,31658
6	1	1	0,13458	0,86542
7	1	1	0,13458	0,86542
8	1	1	0,40625	0,59375
9	1	1	0,13458	0,86542

원래 범주 예측 범주 범주가 0 일 확률 범주가 1 일 확률

예시파일

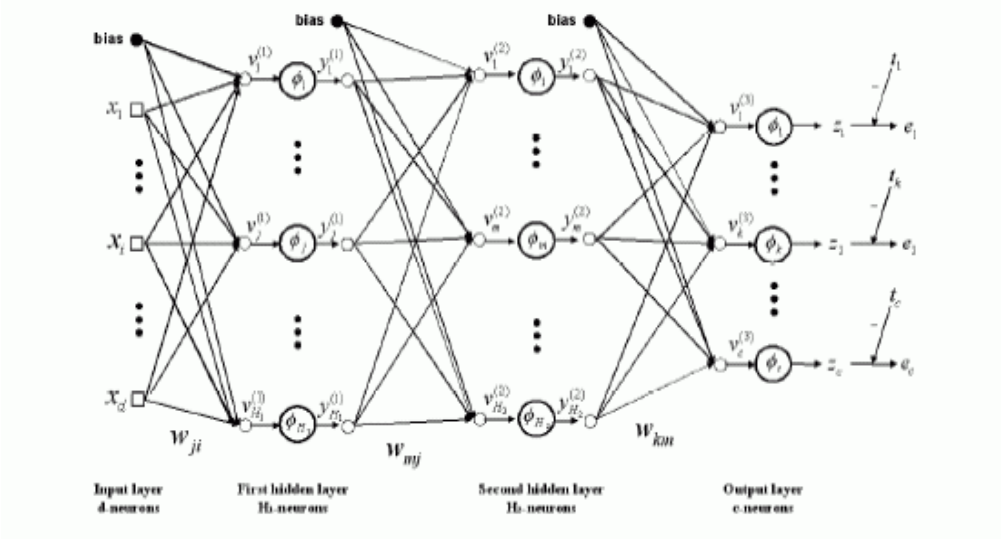
▪ CART.ecm 실행

3.4.9 MLP 노드

인간 두뇌의 신경망을 모방한 모델링 기법으로 실제 자신이 가진 데이터를 이용한 반복적인 학습과정을 거쳐 데이터에 숨어있는 패턴을 찾아냅니다.

개요

Multilayer feedforward network 의 구조는 크게 input layer, hidden layer, output layer 로 이뤄져 있습니다. 이때, input signal 은 layer-by-layer 로 전진해 나갑니다. 이러한 신경망을 **MLP(Multi-Layer Perceptron)**이라고 합니다. **Function approximation, Pattern classification** 에 널리 이용되는 MLP 는 RBF, SVM 등과 함께 대표적인 **Supervised neural network** 입니다. 각 layer 의 neuron(or node)들은 bias 와 다른 layer 의 neuron 들과 adjustable weight 에 의해 연결되어 있습니다. 아래 그림은 hidden layer 가 2 개인 MLP 의 구조를 나타낸 것입니다.



이러한 MLP 의 weight 를 learning 하는 대표적인 algorithm 이 Error back-propagationAlgorithm (EBP)입니다. EBP 는 다음과 같은 2 단계로 구성되어 있습니다.

(1) Forward pass

MLP 의 input layer 에 input pattern 이 입력되면 forward 방향으로 input, hidden layer 를 거치면서 최종적으로 output layer 를 통해 output 의 추정치인 network 의 response 가 구해집니다. Forward pass 동안은 network 의 output 만 구해지고 weight 는 고정되어 있습니다.

(2) Backward pass

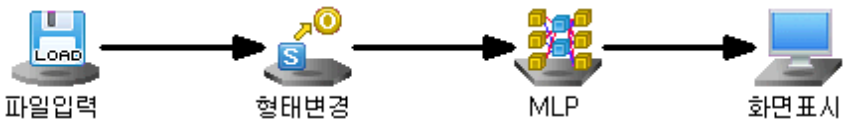
Error correction rule(back propagation algorithm)에 의해 backward 방향으로 weight 가 update 됩니다. weight 는 network 의 actual response 가 좀더 desired response 에 가까워 지도록 update 됩니다.

고려사항

- MLP 노드 자체가 종속변수의 데이터 타입에 맞는 수행기법을 선택해 종속변수가 연속형일 때는 예측분석을 이산형일 때는 **분류분석**을 수행합니다.
- 구조 최적화 옵션 사용을 위해서는 전처리 노드 중 '분할' 노드가 필수적으로 필요합니다.

사용법

- **입력 노드**를 통해 데이터를 읽어 들입니다.
- **형태 변경 노드**를 통해 읽어 들인 데이터의 타입을 지정합니다.(**독립변수, 종속변수**를 지정)
- **MLP 노드**를 **형태 변경 노드**에 연결하고 옵션들을 선택합니다.
- **화면표시 노드**를 **MLP 노드**에 연결합니다.
- **MLP diagram** 예시는 아래와 같습니다.



속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	Optimizing	목적함수를 최적화할 Optimizing	필수	Steepest

	Method	Method 를 선택합니다. Steepest Descent 와 Conjugate Gradient, Levenberg Marquardt 를 선택할 수 있습니다.		Descent. Conjugate Gradient. Levenberg Marquardt.
	은닉층수	은닉층수를 선택합니다.	필수	1 or 2
	은닉 1 층노드수	은닉층의 노드수를 입력합니다.	필수	자연수
	은닉 2 층노드수	은닉층의 노드수를 입력합니다.	필수	자연수
	은닉 1 층 활성함수	첫번째 은닉층의 활성함수를 지정합니다.	필수	Linear, TanSig, LogSig
	출력 1 층 활성함수	출력층의 활성함수를 지정합니다.	필수	Linear, TanSig, LogSig
	은닉 2 층 활성함수	두번째 은닉층의 활성함수를 지정합니다.	필수	Linear, TanSig, LogSig
	모멘텀	최적화 문제에서 어떤 iteration 에서의 gradient 방향에 이전 단계의 방향을 반영하는 정도를 나타내는 것입니다.	필수	Steepest Descent 일 경우 활성화
	학습률	Weight 를 구할 때 step length 를 말합니다.	필수	Steepest Descent 일 경우 활성화
	시행수	Algorithm 이 수렴하지 않을 경우의 최대 반복수를 입력합니다.	필수	자연수,
	최종목표에러	Algorithm 이 수렴하기 위한 actual response 와 desired response 사이의 에러로 이 목표에러에 도달하면 iteration 을 중지합니다.	필수	
	Epsilon 1	Epsilon 1 을 입력합니다.	필수	Levenberg Marquardt 일 경우 활성화
	Epsilon 2	Epsilon 2 를 입력합니다.	필수	Levenberg

				Marquardt 일 경우 활성화
	구조 최적화 여부	분할 노드를 앞에 연결하였을 때 구조최적화를 할 수 있는데 이를 시행할 지 여부를 선택합니다.	필수	예, 아니오
	첫번째 은닉층에서의 시작노드 수	구조 최적화 시 MLP Network 의 첫번째 은닉층에서 시작 노드 수를 설정합니다.	구조 최적화 시 필수	자연수
	첫번째 은닉층에서의 증가노드 수	구조 최적화 시 MLP Network 의 첫번째 은닉층에서 증가 노드 수를 설정합니다.	구조 최적화 시 필수	자연수
	첫번째 은닉층에서의 Test 경우 수	구조 최적화 시 MLP Network 의 첫번째 은닉층에서 Test 경우 수를 설정합니다. 예를 들어 시작 노드 수 1, 증가 노드 수 2, Test 경우의 수 4 이면 1, 3, 5, 7 인 케이스에 대해서 Test 합니다.	구조 최적화 시 필수	자연수
	두번째 은닉층에서의 시작노드 수	구조 최적화 시 MLP Network 의 두번째 은닉층에서 시작 노드 수를 설정합니다.	구조 최적화 시 필수	자연수
	두번째 은닉층에서의 증가노드 수	구조 최적화 시 MLP Network 의 두번째 은닉층에서 증가 노드 수를 설정합니다.	구조 최적화 시 필수	자연수
	두번째 은닉층에서의 Test 경우 수	구조 최적화 시 MLP Network 의 두번째 은닉층에서 증가 노드 수를 설정합니다. 예를 들어 시작 노드 수 1, 증가 노드 수 2, Test 경우의 수 4 이면 1, 3, 5, 7 인 케이스에 대해서 Test 합니다. 만약 첫번째 은닉층에서의 Test 수가 4 이고 두번째 은닉층에서의 Test 수가 5 이면 총 20 가지 경우에 대해서 Test 를 하게 되는 것입니다.	구조 최적화 시 필수	자연수

결과

▪ **분석결과정보**

화면표시 노드에서 분석 결과를 확인할 수 있습니다. 아래는 분류분석의 예시입니다.

	1	2	3	4
	FIELD1	FIELD2	FIELD3	MLPC_YHAT
1	2,400	1,700	1	1
2	1,800	0,900	1	2
3	2,400	1,600	1	1
4	3,000	1,900	1	1
5	2,000	0,500	2	2
6	1,200	0,600	2	2
7	2,000	1,100	2	2
8	2,700	2,000	2	1
9	2,000	2,000	1	1
10	4,000	0,700	2	2
11	2,200	1,600	1	1

원래 범주 예측 범주

예시파일

▪ **MLP.ecm** 실행

3.4.10 MLR 노드

다변량회귀분석(Multi Linear Regression)이란 관찰된 연속형 변수들에 대해 독립변수와 종속변수 사이의 상관관계를 나타내는 **선형 관계식**을 구하는 기법 및 이렇게 얻은 모형의 적합도를 측정하는 분석 방법을 말합니다.

상관계수를 이용하여 두 변수들 간의 관계의 정도는 파악할 수 있지만 변수들 간의 정확한 관계를 알기는 어렵습니다. 보통 하나 또는 둘 이상의 변수들이 다른 하나의 변수에 미치는 **영향의 정도와 방향**을 파악하기 위해 **회귀분석**을 사용합니다.

개요

어떤 변수가 다른 변수에 영향을 주는 경우에, 영향을 주는 변수를 독립변수 (independent variable, 설명변수)라 하고, 영향을 받는 변수를 종속변수(dependent variable, 반응변수)라 합니다.

회귀분석의 목적은 독립변수와 종속변수의 관계를 파악하고, 종속변수를 예측하는 것입니다.

일반적으로 독립변수에 각기 다른 가중치를 곱한 선형함수로 종속변수를 추정하는 모형을 회귀모형이라고 합니다.

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

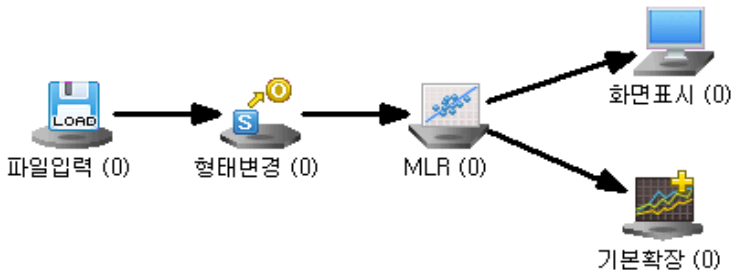
위의 식에서 X_i 는 독립변수, Y_i 는 종속변수로써 자료에서 주어집니다. b_i 는 회귀계수로 모형에서 추정되고 ε_i 는 모형 예측 시 나오는 잔차입니다. 여기서 각 변수의 회귀계수는 다른 변수들이 고정되어 있다고 가정할 때, 그 변수가 1 단위 증가 시 예측된 값의 증가량을 나타냅니다.

고려사항

- **종속변수**와 **독립변수**는 연속형이어야 합니다.
- **독립변수** X 와 **종속변수** Y 사이의 관계는 선형입니다.
- **오차(error)**는 기대값은 0, **분산**은 등분산인 **정규분포**를 따릅니다.(Normality: 정규성)
- 예측 오차 값들은 서로 독립적이어야 합니다. (Independece)
- 각 **독립변수** X 에 대한 **종속변수** Y 값의 변동성은 독립변수 X 의 값에 상관없이 **동일**해야 합니다.(Homoskedasticity)
- **다중 공선성** 문제가 있을 경우 회귀계수 추정이 어렵게 되며, 부정확한 결과를 도출해 냅니다.

사용법

- **입력 노드**를 통해 데이터를 읽어 들입니다.
- **형태 변경 노드**를 통해 읽어 들인 데이터의 타입을 지정합니다. (**독립변수, 종속변수**를 지정)
- **MLR** 를 **형태 변경 노드**에 연결하고 옵션들을 선택합니다.
- **화면표시 노드**를 **MLR 노드**에 연결합니다.
- **MLR diagram** 예시는 아래와 같습니다.



속성

속성그룹	속성명	설명	기타	비고

일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	기법	분석기법을 선택합니다. General 기법은 입력된 모든 독립변수를 모델에 넣고 모델링하는 기법입니다. Stepwise 기법은 입력된 독립변수 모두를 사용하지 않고 stepwise 변수선택 방법에 의해 유의한 변수만을 넣어 모델링하는 기법입니다.	필수	General, StepWise
	입력 유의수준	변수추가를 위한 유의수준입니다. 유의수준이 작을수록 모형에 추가되는 독립변수 수는 적어집니다.	필수	Stepwise 기법, 제거유의수준 이하
	제거 유의수준	변수제거를 위한 유의수준입니다. 유의수준이 클수록 모형에 추가되는 독립변수 수는 적어집니다.	필수	Stepwise 기법, 입력유의수준 이상
	VIFs 구하기	VIFs 값을 구하려면 True, 구하지 않으려면 False 값을 선택합니다. 수행시간이 길어지기 때문에 필요가 없다면 False 로 합니다. VIF 값이 10 을 넘으면 변수간 다중 공선성이 있다고 판단합니다.	필수	True, False
	신뢰구간 저장	지정한 종속변수에 대한 95% 신뢰구간 저장여부를 결정합니다.		True, False
	예측구간 저장	예측값에 대한 95% 예측구간 저장여부를 결정합니다.		True, False

결과

▪ **분석결과정보**

화면표시 노드에서 Y 추정치와 잔차, Leverage 값을 볼 수 있습니다.

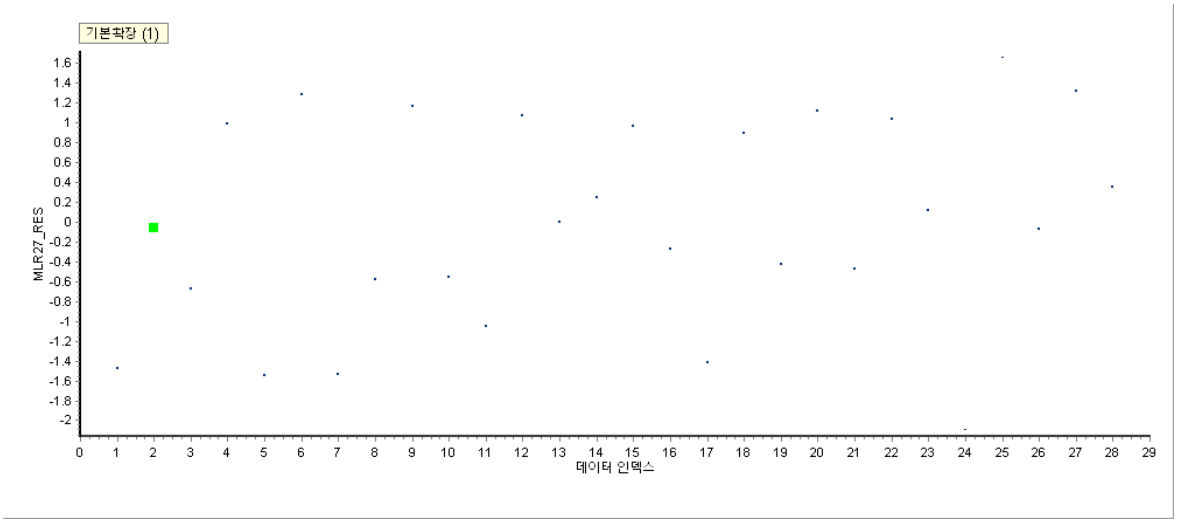
	1	2	3	4	5	6
	x1	x2	y	MLR27_YHAT	MLR27_RES	MLR27_LEV
4	11,30000	30	46,80000	45,80652	0,99348	0,03761
5	12,90000	28	45,90000	47,44290	-1,54290	0,11280
6	11,60000	35	49,20000	47,91737	1,28263	0,04251
7	11,50000	45	49,60000	51,12947	-1,52947	0,10133
8	8,90000	3	32,70000	33,28295	-0,58295	0,26205
9	10,80000	20	42,90000	41,72905	1,17095	0,05826
10	10,90000	28	44,00000	44,55834	-0,55834	0,04684
11	13,00000	27	46,20000	47,25150	-1,05150	0,12940
12	13,10000	28	48,80000	47,73136	1,06864	0,13525
13	12,10000	34	48,30000	48,30287	-0,00287	0,04556
14	11,60000	50	53,20000	52,95186	0,24814	0,14466
15	10,60000	3	36,70000	35,73483	0,96517	0,18557
16	11,60000	31	46,30000	46,57484	-0,27484	0,03615
17	12,20000	36	47,70000	49,11837	-1,41837	0,05050
18	12,90000	24	47,00000	46,10037	0,89963	0,13782
19	11,40000	32	46,20000	46,62201	-0,42201	0,03854
20	11,30000	35	48,60000	47,48468	1,11532	0,04880
21	11,70000	55	54,30000	54,77426	-0,47426	0,19938
22	10,20000	20	41,90000	40,86368	1,03632	0,08422
23	12,00000	32	47,60000	47,48738	0,11262	0,04231
24	11,90000	28	43,90000	46,00062	-2,10062	0,04381

데이터

원래 값 예측 값

잔차 PLOT

기본확장차트 노드를 연결해서 볼 수 있습니다.



회귀모형이 잘 적합 되었다면 오차는 평균이 0 이고 분산이 등분산인 가정을 만족해야 합니다. 기본차트 노드에서 X 축은 Data index, Y 축은 Residual(MLR_RES)로 하여 수행하면 위와 같은 plot 이 생성됩니다.

예시파일

- **MLP.ecm** 실행

3.4.11 PCA 노드

주성분 분석(principal component analysis)은 차원축소를 통하여 저차원상에서 변수의 관계를 규명하는 다변량 자료 분석기법입니다. 주성분 분석의 구체적 목적은 변수들의 선형결합을 결정하는 것, 변수의 수를 줄이는 것, 의미 있는 새로운 잠재적 변수를 발견하는 것, 자료세트 행렬의 차원을 축소하는 것 등이 포함될 수 있습니다.

개요

PCA 를 통해 자료가 펼쳐진 방향 또는 경향을 발견함으로써 자료가 내포하고 있는 정보를 발견할 수 있습니다. 발견된 정보에 근거해서 압축하여 저장할 수 있고 또 차후에 자료를 재구성하더라도 손실을 최소화할 수 있습니다. 다변량 자료에서 자료간의 변동을 변수들의 선형결합들로 이루어진 새로운 변수(주성분)로 나타내려는 생각이 이 기법의 주된 내용입니다.

공정 데이터에서 PCA 를 사용할 때는 **이상치 발견, 조업편차 분석, 공정 모니터링** 등에 사용됩니다.

관측치 n 개, 독립변수 k 개인 input matrix X 를 A(≤k)개의 **PCs(Principal Components, 주요성분)**으로 분해합니다.

$$X = t_1 p_1^T + t_2 p_2^T + \mathrm{K} + t_T p_T^T$$

: t_h 는 score, p_h^T 는 weight (h 는 임의의 수)

고려사항

- **독립변수**만 입력되어야 하며 연속형이어야 합니다.
- **PCA 노드** 자체에 표준화 전처리 과정을 포함하고 있습니다.

사용법

- **입력 노드**를 통해 데이터를 읽어 들입니다.
- **PCA 노드**를 **입력 노드**에 연결하고 옵션들을 선택합니다.
- **PCA diagram** 예시는 아래와 같습니다.

속성

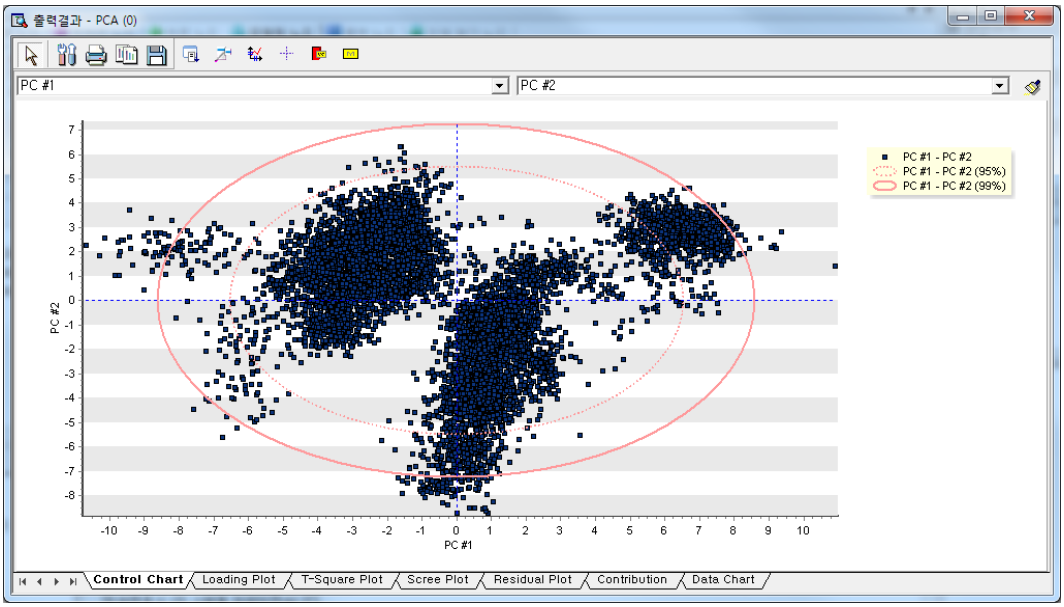
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 입력합니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	결과보기	Control Chart, Loading Plot, T-Square Plot, Scree Plot, Residual Plot, Contribution Plot 출력됩니다.	클릭	모델링 후 클릭
	주성분수	주성분수를 입력합니다.	필수	독립변수 수보다 작은 자연수
	전처리방법	입력자료 전처리 방법을 결정합니다.	필수	표준화, 평균보정

결과

(1) 분석을 실행한 Diagram 상에서 PCA 노드를 클릭해서 노드 속성창의 결과보기 버튼을 누르면 다음 결과들이 출력됩니다.

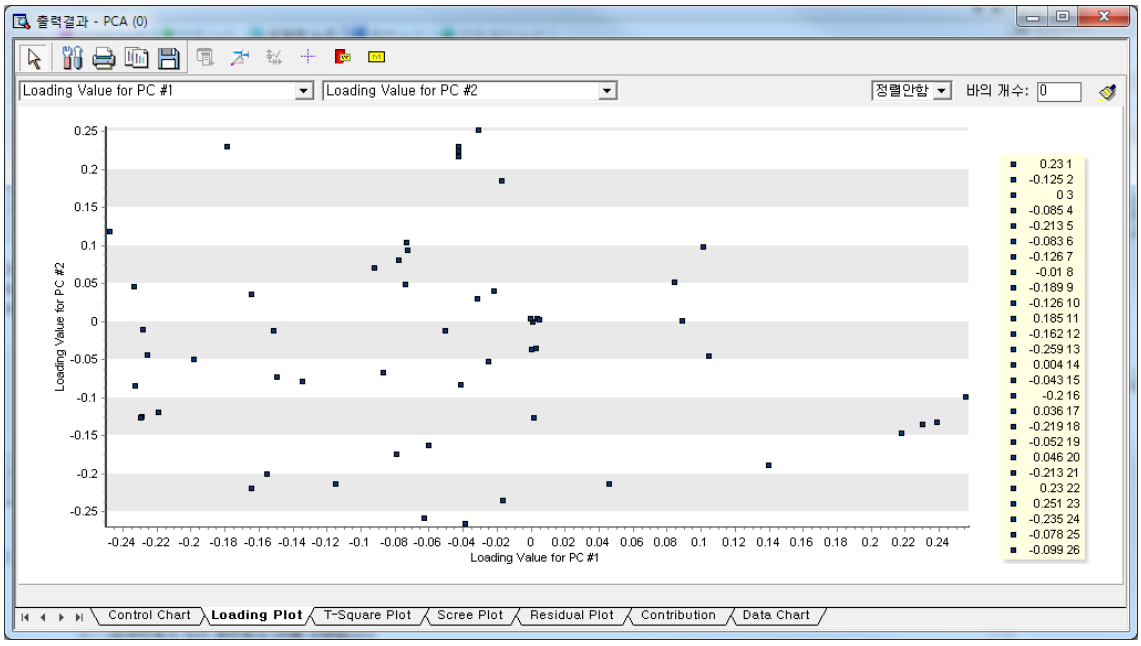
▪ **Control Chart**

Control Chart 를 통해 이상치 및 조업편차를 확인할 수 있습니다.



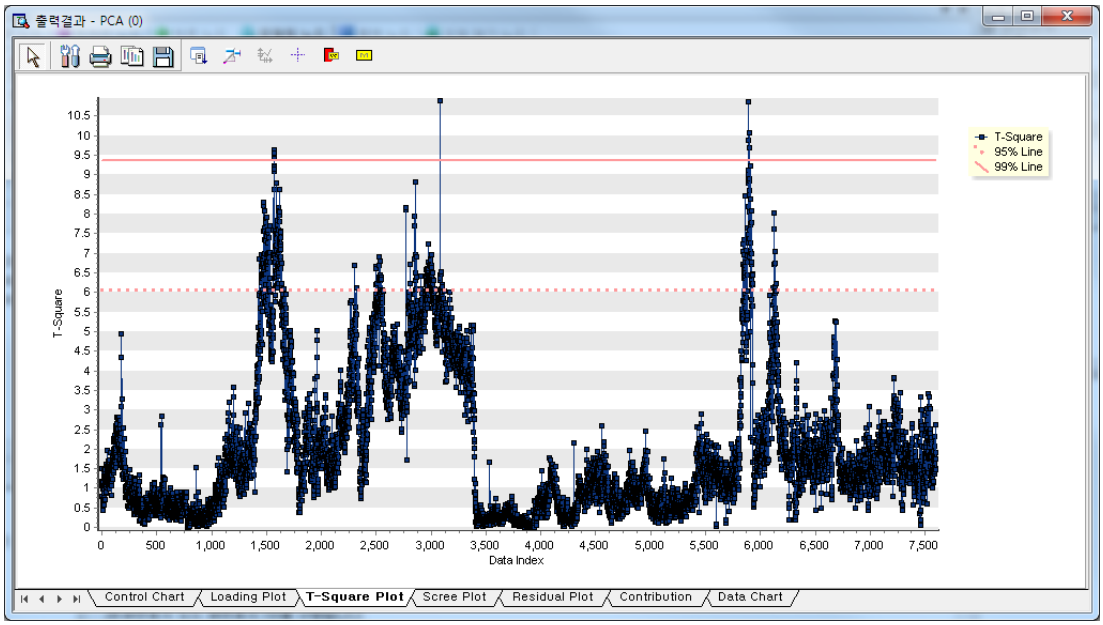
▪ **Loading Plot**

Loading plot 은 선택된 두 개의 원 데이터 혹은 주성분의 loading 값에 대한 정보를 제공합니다.



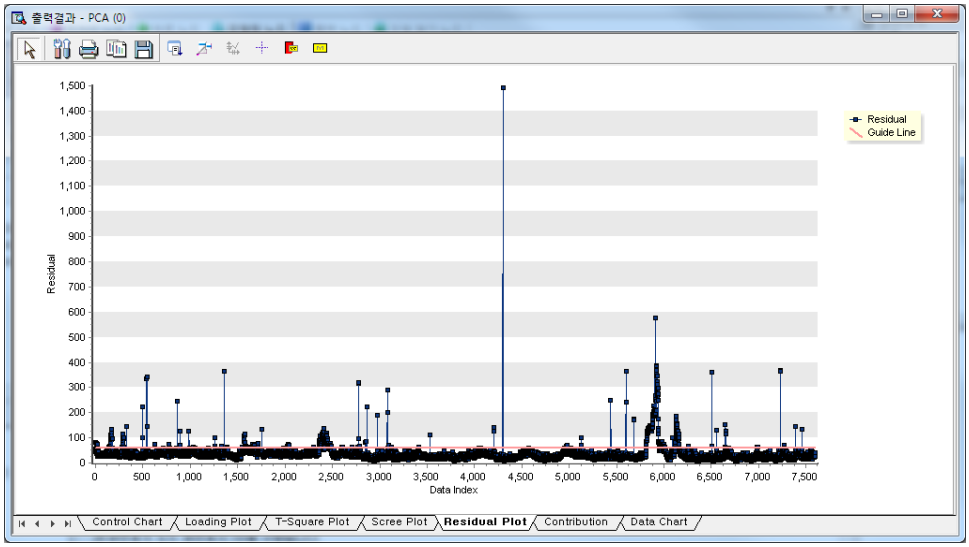
▪ **T-Square plot**

T-Square 값은 현재의 상태가 얼마나 극단적인 값을 갖는지를 나타내는 척도입니다. 즉, 변수 중에서 특이하게 큰 값이 있을 경우 T-Square 값이 커지게 되며, 표시된 Guide Line 을 넘는 데이터는 이상치로 판단할 수 있습니다.



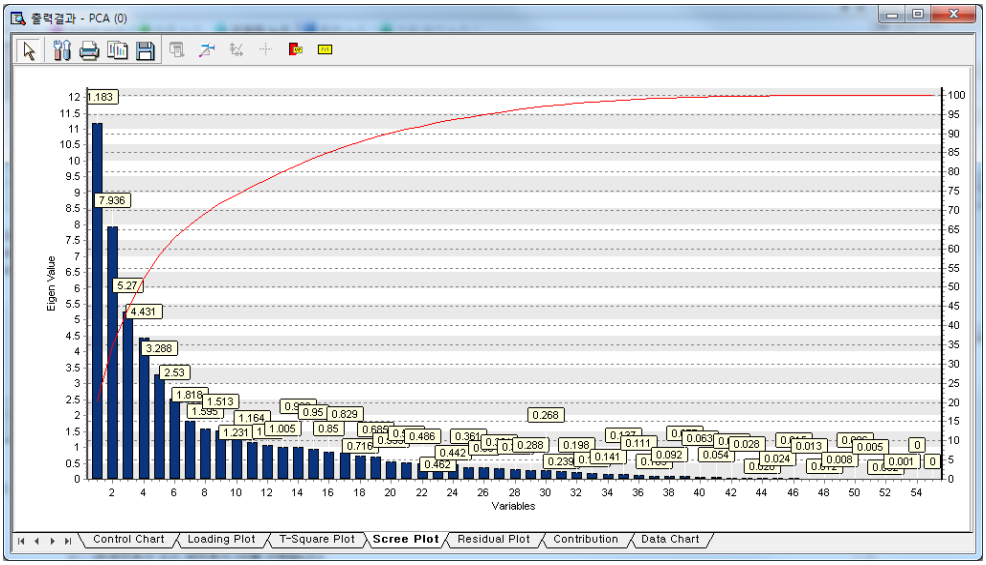
▪ **Residual plot**

Residual 값은 데이터가 축소 공간과 얼마나 떨어져 있는 가를 나타내는 척도입니다. 이 값이 커질수록 현재의 데이터와 축소 공간 사이의 거리가 멀다는 것을 나타내며, 표시된 Guide Line 을 넘는 데이터는 이상치로 판단할 수 있습니다.



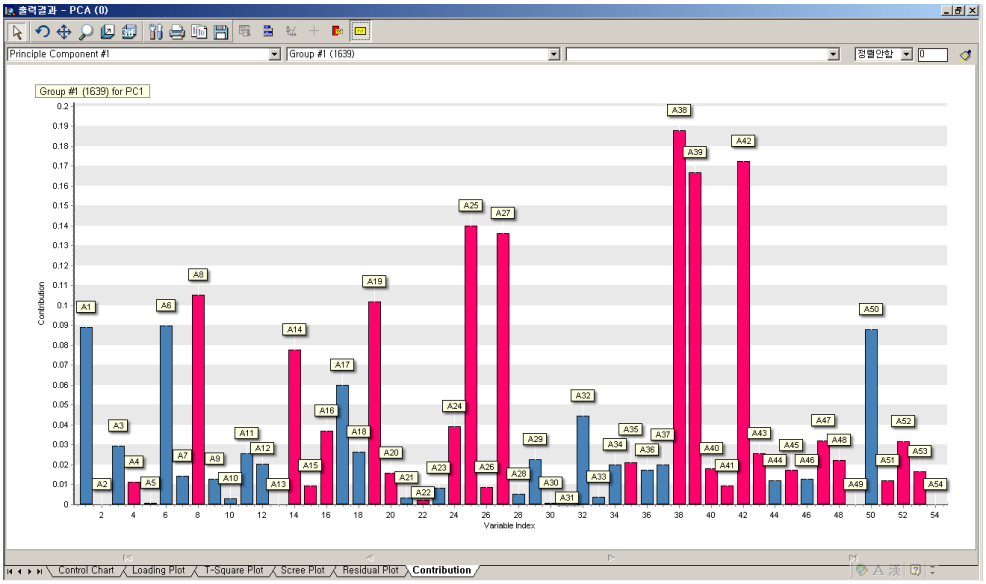
▪ **Scree Plot**

Scree plot 은 각 주성분에 대한 Eigenvalue 값을 표시한 것입니다. Scree plot 은 주성분의 수를 결정하는 하나의 방법으로 Eigenvalue 값의 변화가 급격히 이루어지는 에서 주성분의 수를 결정할 수 있습니다.

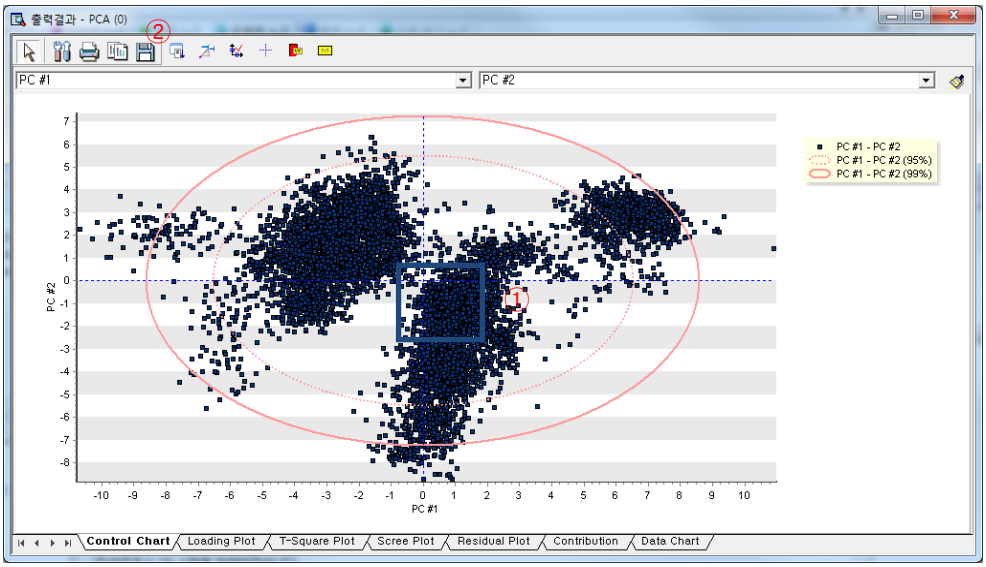


▪ **Contribution plot**

사용자가 추출한 데이터의 편차에 영향을 주는 변수의 공헌도를 보여줍니다.



▪ **Contribution plot 도출 방법 및 해석**



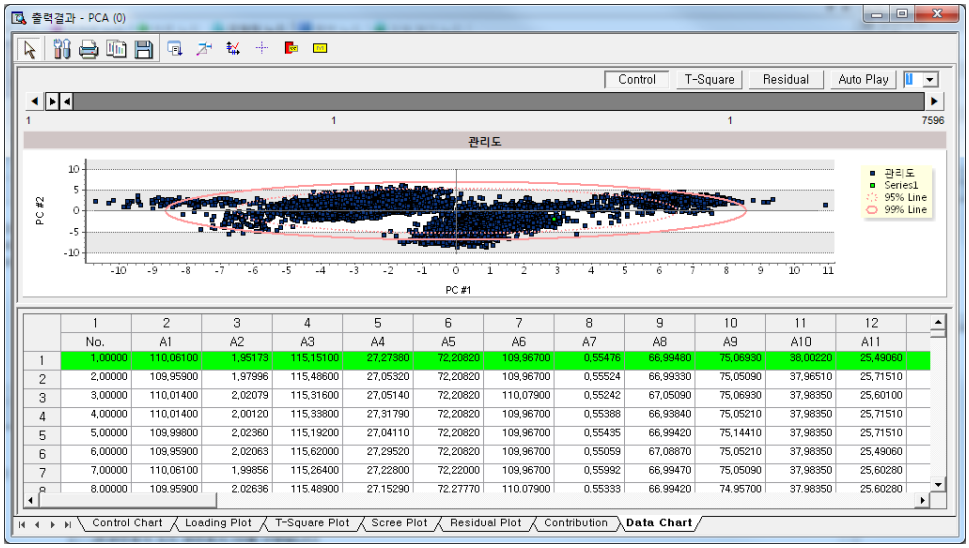
1. control chart 에서 편차가 커 보이는 두 그룹을 분석 목표로 정한 후, 우선 첫번째 그룹을 위 그림에서 보이는 것과 같이 ①처럼 드래그 한 뒤, ②버튼을 클릭해서 앞서 드래그한 데이터를 추출합니다. 두 번째 그룹에도 같은 과정을 반복합니다.

2. 데이터 선택이 끝났으면 Contribution Plot 부분으로 가서 공헌도를 파악하고 싶은 주성분과, 그룹 데이터를 선택합니다. 선택 후 그리기 버튼을 누르면 아래와 같은 Contribution Plot 이 출력됩니다.

위의 Contribution Plot 을 보면 주성분 1 에서 group 1 편차에 많은 영향을 주는 변수는 25, 27, 38, 39, 42 번 변수임을 알 수 있습니다. 이 plot 에서 빨간색으로 표시되는 변수는 양의 관계를 가지는 변수이며, 파란색으로 표시되는 변수는 음의 관계를 가지는 변수입니다. 즉 위에서 추출된 25, 27, 38, 39, 42 번 변수는 전부 빨간색으로 표시되어 있으며, 이는 양의 공헌도를 가짐을 의미합니다.

▪ Data chart

Data chart 는 chart 와 분석에 이용된 data 를 동시에 보여줌으로써, data 와 chart 의 매칭 기능을 제공합니다. 또한 여러 차트를 한번에 볼 수 있어 비교도 가능합니다.



	1	2	3	4	5	6	7	8	9	10	11	12
	No.	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
1	1,00000	110,06100	1,95173	115,15100	27,27380	72,20820	109,96700	0,55476	66,99480	75,06930	38,00220	25,49060
2	2,00000	109,95900	1,97996	115,48600	27,05320	72,20820	109,96700	0,55524	66,99330	75,05090	37,96510	25,71510
3	3,00000	110,01400	2,02079	115,31600	27,05140	72,20820	110,07900	0,55242	67,05090	75,06930	37,98350	25,60100
4	4,00000	110,01400	2,00120	115,33800	27,31790	72,20820	109,96700	0,55388	66,93840	75,05210	37,98350	25,71510
5	5,00000	109,99800	2,02360	115,19200	27,04110	72,20820	109,96700	0,55435	66,99420	75,14410	37,98350	25,71510
6	6,00000	109,95900	2,02063	115,62000	27,29520	72,20820	109,96700	0,55059	67,08870	75,05210	37,98350	25,49060
7	7,00000	110,06100	1,99856	115,26400	27,22800	72,22000	109,96700	0,55992	66,99470	75,05090	37,98350	25,60280
8	8,00000	109,95900	2,02636	115,48900	27,15290	72,27770	110,07900	0,55333	66,99420	74,95700	37,98350	25,60280

상단의 데이터 슬라이드바를 이용하여 데이터와 차트를 매칭시킬 수 있습니다. 이를 이용하여 데이터의 진행 방향을 파악할 수 있습니다. 또한 우측 상단에 차트별로 버튼이 구성되어 있어서 버튼을 누르면 여러 개의 차트를 한눈에 파악할 수 있습니다.

(2) 분석결과정보

화면표시 노드에서 각 주성분의 Score 값과 T-square 값을 파악할 수 있습니다.

	51	52	53	54	55	56	57	58	59
	A50	A51	A52	A53	A54	PCA15_T1	PCA15_T2	PCA15_Tsqs	PCA15_Residual
1	2801,03	8,84598	45,3797	114,402	1,61773	2,88984	-1,82577	1,16683	45,56767
2	2840,81	8,84598	47,3984	124,647	3,46919	2,05845	-1,79775	0,78616	57,29965
3	2839,77	8,95819	44,3114	112,848	0,31438	2,16279	-2,28003	1,07337	48,91753
4	2839,42	8,95819	50,4884	107,954	0,64159	2,28866	-2,87604	1,51072	81,18412
5	2830,4	8,95819	46,7822	148,394	0,05894	1,16255	-1,60215	0,44432	78,75799
6	2840,47	8,95819	52,3942	152,027	0	0,83303	-1,82455	0,48155	48,5948
7	2831,08	9,06853	48,1532	141,114	0	1,53762	-1,48815	0,49049	38,88499
8	2825,47	9,05201	51,202	132,141	0	1,29031	-1,52443	0,44172	46,85754
9	2827,72	9,0704	51,3208	120,377	0	1,9559	-2,26566	0,98894	45,60716
10	2850,33	9,0517	49,3681	125,514	0	1,28379	-2,12453	0,71616	61,67017
11	2816,16	9,12681	51,9721	128,326	0,00092	2,06882	-1,41626	0,63548	46,09202
12	2822,65	9,14521	49,8839	111,144	0	1,49006	-1,96058	0,68292	44,31256
13	2841,68	9,14489	47,156	104,482	0	2,68496	-1,71382	1,01477	46,82916
14	2847,34	9,0704	52,3523	106,721	0	1,81406	-2,58779	1,13814	37,13715
15	2831,26	8,95819	48,0716	100,183	3,46828	2,79587	-2,1763	1,29583	56,17555
16	2898,41	8,84598	50,3756	145,219	0,71128	1,87672	-1,43976	0,57616	73,77315
17	2848,71	8,73377	45,6932	121,033	0,00008	2,73145	-1,80885	1,07946	58,01534
18	2916,91	8,65896	51,0543	143,556	0,03767	1,82967	-1,78733	0,70191	54,15146
19	2855,83	8,54675	50,7512	112,135	0,02723	2,3467	-2,11831	1,0579	40,92285
20	2844,11	8,5455	46,365	141,96	0,0565	2,03773	-1,91597	0,8339	39,3847
21	2863,72	8,43453	51,5514	135,338	0,00163	2,09565	-2,1938	0,99919	35,61428

예시파일

- **조업편차분석.ecm** 실행

3.4.12 PCR 노드

보통 종속변수 예측을 위해 많이 사용되는 방법이 MLR 입니다. 그러나 독립변수 x 간의 **다중공선성(multicollinearity)**이 크거나 변수의 수가 관측치의 수보다 많을 때 MLR 을 사용하는 것은 효율적이지 않습니다. 이럴 경우 종속변수 예측을 위해 **PCR(Principal Component Regression)**을 사용합니다.

개요

종속변수 y 와 독립변수 x 가 있을 때 y 에 대한 회귀모형은 다음과 같습니다.

$$Y = X\beta + \varepsilon$$

여기서 **β** 는 **회귀계수벡터**이며, **ε** 은 **오차벡터**입니다. x 의 벡터들간 **다중공선성(multicollinearity)**이 높으면 최소자승법에 의한 β 추정치의 분산이 커지는 문제가 있으며, 관측치 수보다 변수 수가 많을 때는 위의 모형으로는 올바른 예측을 할 수 없습니다. 이러한 문제가 있을 때 MLR 대신 PCR 을 사용합니다.

$$X = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \ldots + \mathbf{t}_A\mathbf{p}_A^T + E$$
$$\mathbf{y} = T\mathbf{q}^T + \mathbf{f} = q_1\mathbf{t}_1 + q_2\mathbf{t}_2 + \ldots + q_A\mathbf{t}_A + \mathbf{f}$$

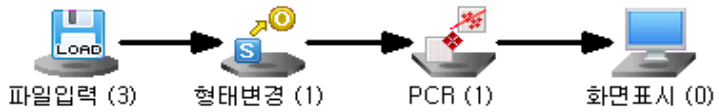
여기서는 PCA 를 적용해 PC score 행렬인 T 를 먼저 구한 후, 이 값을 토대로 회귀분석을 실시합니다.

고려사항

- **PCR** 을 수행하기 위해 **종속변수**는 연속형이어야 하며, **독립변수**도 연속형이어야 합니다.
- **PCR 노드** 자체에 표준화 전처리 과정을 포함하고 있습니다.

사용법

- **입력 노드**를 통해 데이터를 읽어 들입니다.
- **형태 변경 노드**를 통해 읽어 들인 데이터의 타입을 지정합니다.(**독립변수, 종속변수**를 지정)
- **PCR 노드**를 **형태 변경 노드**에 연결하고 옵션들을 선택합니다.
- **PCR diagram** 예시는 아래와 같습니다.



속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 입력합니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	주성분수	전체 변수를 축약할 주성분의 수를 입력합니다. 주성분 수는 독립 변수의 수보다는 작아야 합니다.	필수	독립변수 수보다 작은 자연수
	신뢰구간 저장	지정한 종속변수에 대한 95% 신뢰구간을 저장합니다.	필수	True, False
	예측구간 저장	예측값에 대한 95% 예측구간을 저장합니다.	필수	True, False
변수정보	변수정보	읽을 데이터의 변수명 및 형태를 나타냅니다.		

결과

▪ **분석결과정보**

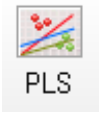
PCR 노드에 화면표시 노드를 연결하면 추출된 주성분값과 Y 추정치와 잔차값을 볼 수 있습니다. 선택에 따라서 95%신뢰구간과 95%예측구간도 확인 할 수 있습니다.

	14	15	16	17	18	19
	PCR_T4	PCR_T5	PCR_T6	PCR_T7	PCR_YHAT	PCR_Res
1	0,060	-0,251	-0,328	-0,627	1,210	-0,060
2	0,198	-0,361	-0,521	-0,563	1,199	-0,129
3	0,378	-0,868	0,195	-1,013	1,214	-0,034
4	0,164	0,536	-1,333	-0,181	1,198	0,042
5	0,651	-0,146	-1,249	-0,125	1,175	-0,095
6	-0,811	0,057	0,015	-0,856	1,176	-0,056
7	0,694	-0,612	-0,568	1,209	1,167	-0,017
8	0,646	-0,704	-0,461	1,115	1,182	-0,072
9	0,400	-0,775	-0,187	0,259	1,181	-0,091
10	-1,151	0,845	-0,272	-0,586	1,196	0,064
11	-0,348	-0,022	-0,502	-0,531	1,202	-0,072

예시파일

- **PCR.ecm** 실행

3.4.13 PLS 노드



PLS(Partial least squares)는 여러 개의 반응변수를 예측하는데 사용되는 분석방법입니다. PLS 는 독립변수간 상관성이 크거나 변수의 수가 관측치의 수보다 큰 데이터를 분석하는데 사용됩니다. PLS 는 x 와 y 사이의 covariance 에 기초하여 상관성이 없는 성분을 찾아내고 그 성분들을 이용하여 **최소 제곱 회귀분석**을 수행하는 방법입니다. 즉, 어떤 종속변수를 예측하고자 하는데 독립변수들의 개수가 너무 많고 다중공선성 문제까지 있을 경우 차원을 축소하여 독립변수의 수를 줄이고, 다중 공선성 문제 또한 해결하여 종속변수를 예측하는 회귀모델을 만드는 분석 방법입니다.

개요

PLS 는 특히 공정 변수수(k)와 품질 특성치수(m)가 매우 클 때 [X(n by k), Y(n by m)]의 정보를 작은 수의 latent variable 들을 갖는 저차원의 공간에 투영시키는 방법입니다.

A 개의 latent variable 을 $\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_A$ 이라 하고 이로 이루어지는 (n x A) 행렬을 $T=[\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_A]$ 라 가정하면 T 는 X 가 축소된 형태로서 다음과 같이 쓸 수 있습니다.

$$T = XP$$

마찬가지로 행렬 Y 에 대하여서도 dimension 을 축소시킨 score matrix U 를 산출할 수 있습니다.

$$U = YQ$$

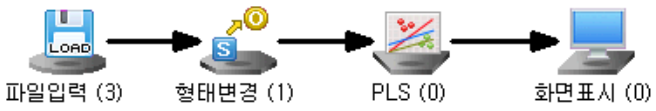
여기서 (m x A) 행렬 Q 는 U_a 에 대응하는 loading vector q_a (a=1,2,…,A)로 이루어진 것입니다. PLS 의 주 아이디어는 T 와 U 의 관계를 설정함으로써 X 로 Y 를 설명하고자 하는 것입니다.

고려사항

- **PLS** 을 수행하기 위해 **종속변수**는 연속형이어야 하며, **독립변수**도 연속형이어야 합니다.
- **PLS 노드**는 여러 개의 **종속변수**를 설정할 수도 있습니다.
- **PLS 노드** 자체에 표준화 전처리 과정을 포함하고 있습니다.

사용법

- **입력 노드**를 통해 데이터를 읽어 들입니다.
- **형태변경 노드**를 통해 읽어 들인 데이터의 타입을 지정합니다.(**독립변수, 종속변수**를 지정)
- **PLS 노드**를 **형태변경 노드**에 연결하고 옵션들을 선택합니다.
- **PLS diagram** 예시는 아래와 같습니다.



속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 입력합니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	결과보기	Scree Plot of X, Scree Plot of Y, Coefficient Plot 을 보여줍니다.	클릭	모델링 후 클릭.
	잠재변수 수	잠재변수 수를 입력합니다.	필수	독립변수 수보다 작은 자연수.
	신뢰구간	지정한 종속변수에 대한 95% 신뢰구간을	필수	True, False

	저장	저장합니다.		
	예측구간 저장	예측값에 대한 95% 예측구간을 저장합니다.	필수	True, False

결과

(1) 분석결과정보

PLS 노드에 **화면표시 노드**를 연결하면 추출된 각 종속변수에 대한 추정치와 각각의 잔차값을 볼 수 있습니다.

	53	54	55	56	57	58	59	60	61	62	63	64
	A53	A54	PLS26_T1	PLS26_T2	PLS26_T3	PLS26_T4	PLS26_U1	PLS26_U2	PLS26_U3	PLS26_U4	PLS26_YHAT1	PLS26_Res1
1	114,40200	1,61773	-0,72272	4,31122	-0,41162	-0,15921	0,40561	0,61886	0,03962	0,06753	109,92790	0,13310
2	124,64700	3,46919	0,01869	3,86745	-1,00675	-1,09374	0,34709	0,34157	-0,17805	-0,10976	110,04463	-0,08563
3	112,84800	0,31438	-0,06024	3,54473	-2,31820	-1,76636	0,37864	0,39642	-0,07984	0,07739	109,70841	0,30559
4	107,95400	0,64159	0,09994	3,87348	-3,17209	-2,40308	0,37864	0,34915	-0,17128	0,04387	109,70530	0,30870
5	148,39400	0,05894	0,38879	2,38749	-2,04028	-2,15041	0,36946	0,25474	-0,06603	0,07235	109,66408	0,33392
6	152,02700	0,00000	0,81048	2,61071	-1,36915	-1,40623	0,34709	0,10794	-0,24283	-0,14996	110,08450	-0,12550
7	141,11400	0,00000	0,22617	3,27354	-0,59209	-0,86161	0,40561	0,33887	-0,10095	-0,06079	110,08370	-0,02270
8	132,14100	0,00000	0,60226	3,57687	-0,29081	-0,72458	0,34709	0,16938	-0,31120	-0,29147	110,39703	-0,43803
9	120,37700	0,00000	0,28944	3,96815	-1,88871	-1,96200	0,37004	0,28463	-0,24852	-0,12041	110,01929	-0,02029
10	125,51400	0,00000	0,63511	3,01983	-2,51012	-2,35132	0,36545	0,17804	-0,22769	-0,05744	109,86389	0,12711
11	128,32600	0,00092	-0,02907	4,05802	-0,54465	-1,39705	0,39700	0,40558	-0,13965	-0,10270	110,09001	-0,04401
12	111,14400	0,00000	0,43788	3,26067	-1,67934	-1,27053	0,36487	0,23567	-0,20243	-0,08852	110,02152	-0,03152
13	104,48200	0,00000	-0,30293	4,93039	0,05918	-0,24590	0,37864	0,46803	-0,19441	-0,19842	110,33609	-0,32209
14	106,72100	0,00000	0,43469	3,79604	-1,74108	-0,72391	0,33389	0,20563	-0,30440	-0,18631	110,19078	-0,25478
15	100,18300	3,46828	-0,41979	4,60206	-0,67033	-0,09446	0,38782	0,51169	-0,10663	-0,06117	110,12749	-0,09749
16	145,21900	0,71128	-0,01203	3,73508	0,42834	0,39713	0,39585	0,39940	-0,10243	-0,13149	110,31156	-0,26756
17	121,03300	0,00008	-0,57945	4,22863	-0,34317	-0,19920	0,32988	0,50085	-0,06729	-0,04402	109,98648	-0,05748
18	143,55600	0,03767	0,04460	3,67808	0,71663	1,35241	0,42741	0,41425	-0,07993	-0,12854	110,45374	-0,35474
19	112,13500	0,02723	-0,07899	4,30312	-0,51062	-0,11296	0,31668	0,33999	-0,23817	-0,20353	110,24985	-0,34385
20	141,96000	0,05650	-0,21464	3,35688	-0,15249	0,67295	0,38725	0,45058	-0,00044	0,00990	110,07677	-0,04777

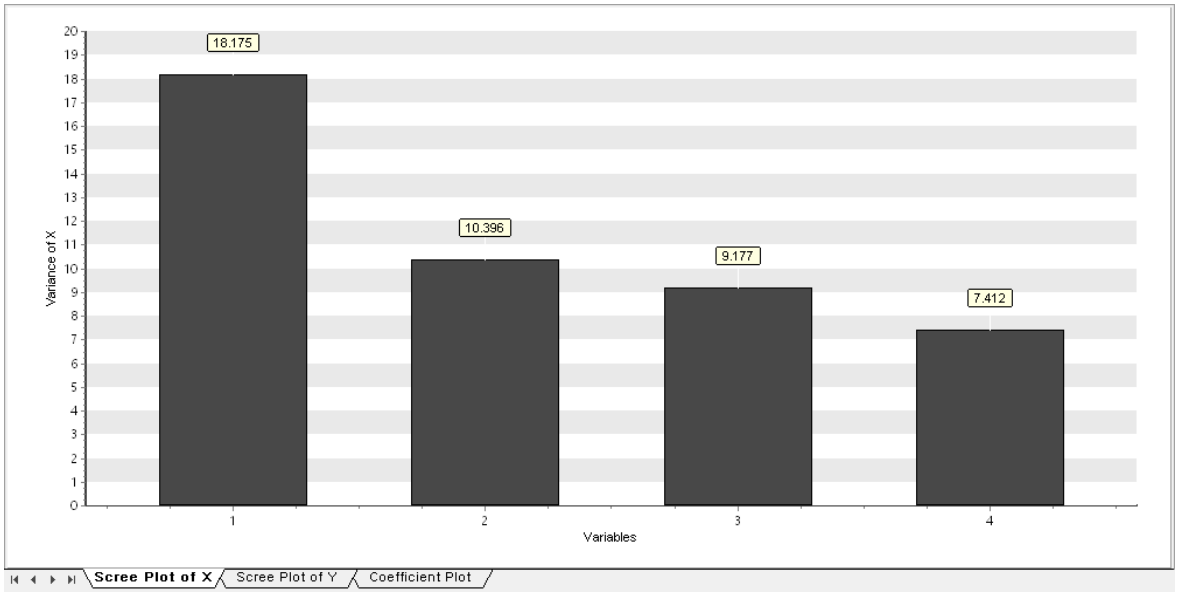
데이터

추정치 잔차

(2) 분석을 실행한 Diagram 상에서 PLS 노드를 클릭해서 노드 속성창의 결과보기 버튼을 누르면 다음 결과들이 출력됩니다.

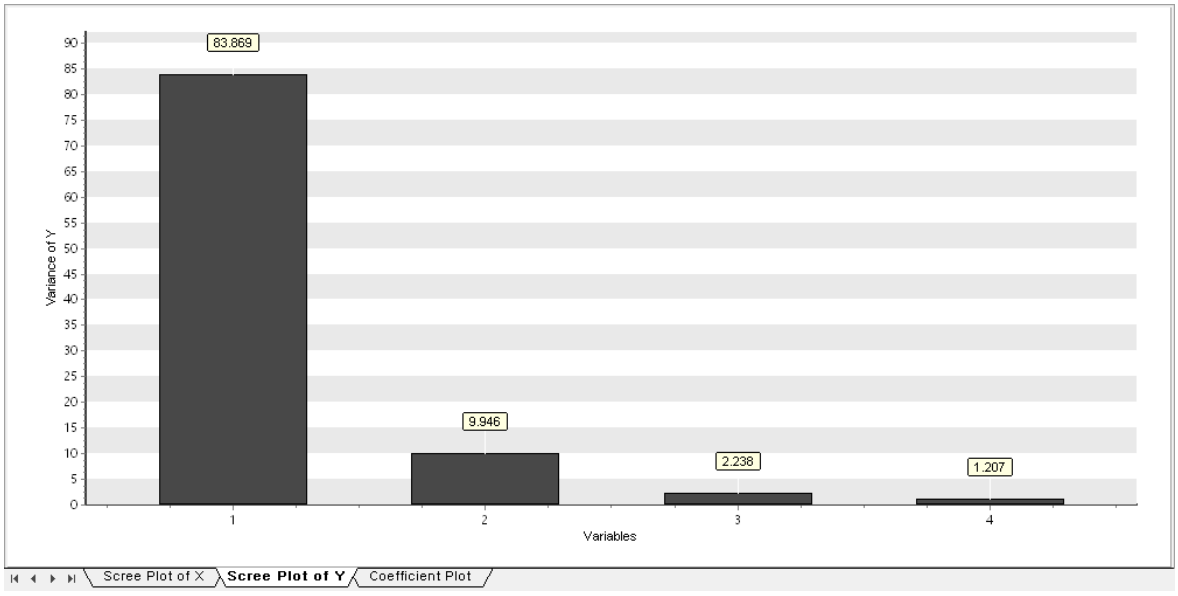
▪ **Scree Plot of X**

각 잠재변수가 X 에 대한 설명력을 얼마나 가지고 있는지를 plot 으로 나타내고 있습니다.



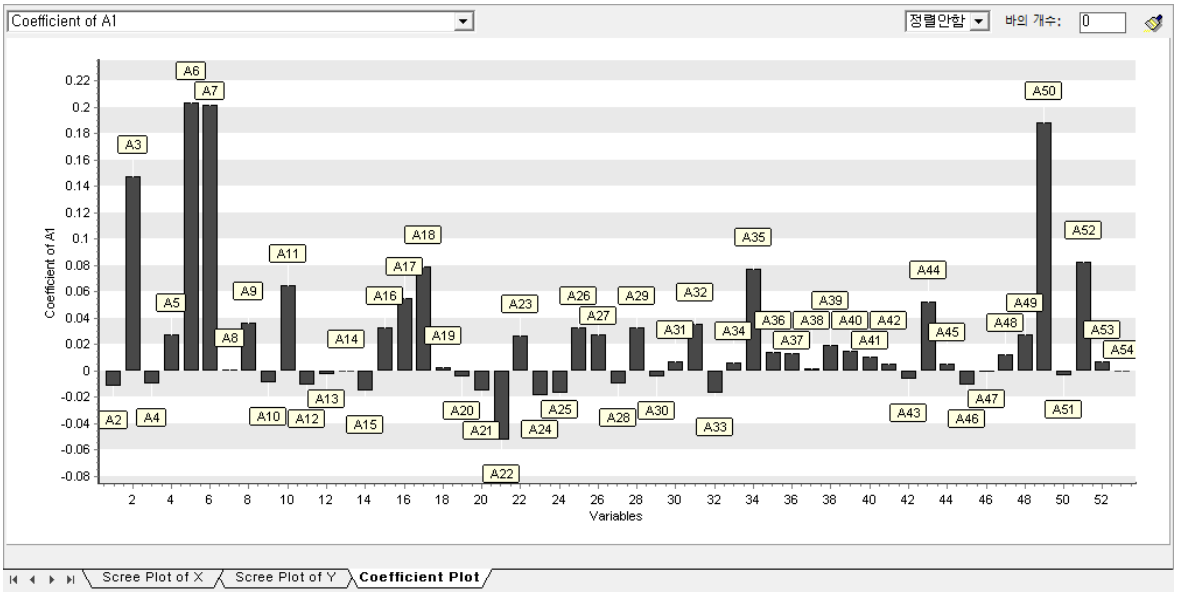
▪ **Scree Plot of Y**

각 잠재변수가 Y 에 대한 설명력을 얼마나 가지고 있는지를 plot 으로 나타내고 있습니다.



▪ **Coefficient Plot**

각 종속변수에 대해 어떤 변수가 가장 영향을 많이 미치는지를 파악할 수 있습니다. 또한, Coefficient 크기에 따라 정렬이 가능합니다.



예시파일

- **PLS.ecm** 실행

3.4.14 QDA 노드

다수의 속성(attribute) 또는 변수를 갖는 객체(object)를 사전에 정해진 **그룹 또는 범주(class, category)** 중의 하나로 분류하는 **분류분석기법** 중 하나로 각 그룹의 **분산-공분산 행렬**이 다를 때 이차식의 판별함수를 사용하는 **QDA** 를 사용합니다.

개요

분산-공분산 행렬이 범주에 관계없이 동일하다고 가정했을 때 판별함수는 선형으로 도출되지만(LDA), 분산-공분산 행렬이 범주별로 다르다고 가정하면 이차식의 판별 함수가 유도되는데 이 경우를 **이차판별분석(Quadratic Discriminant Analysis: QDA)** 이라고 합니다.

(1) 각 범주별 평균벡터를 계산합니다.

$$\hat{\mu}_j \leftarrow \bar{x}^{(j)} \leftarrow \frac{1}{n_j}\sum_{i=1}^{n_j} x_i^{(j)} \quad , j = 1,2$$

(2) 범주별 표본분산-공분산행렬을 계산합니다.

$$\hat{\Sigma}_j \leftarrow S_j \leftarrow \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(x_i - \bar{x}^{(j)})(x_i - \bar{x}^{(j)})' \quad j = 1,2$$

(3) 범주 k 에 대한 판별함수가 아래와 같습니다.

$$U_k = \ln \pi_k - \frac{1}{2}\ln |\Sigma_k| - \frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k)$$

(4) 분류규칙은 아래와 같습니다.

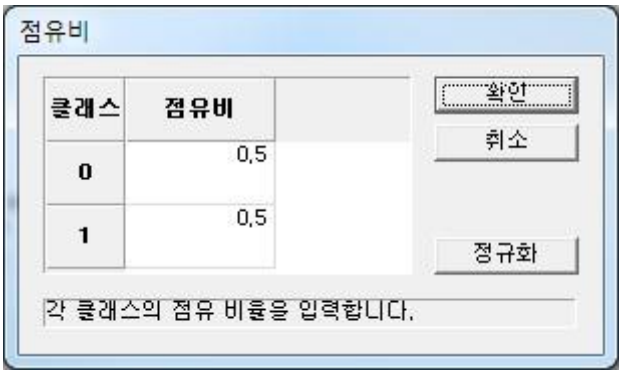
$U_k = \max_j U_j$ 일 때, x 를 범주 k 에 분류

고려사항

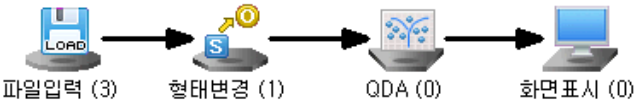
- **종속변수**는 이산형이어야 합니다.
- 벡터 x 가 **다변량 정규분포**를 따르는 것을 가정합니다.
- 각 범주에 따라 **분산-공분산 행렬**이 다름을 가정합니다.

사용법

- **입력 노드**를 통해 데이터를 읽어 들입니다.
- **형태변경 노드**를 통해 읽어 들인 데이터의 타입을 지정합니다.(**독립변수, 종속변수**를 지정)
- **QDA 노드**를 **형태변경 노드**에 연결하고 옵션들을 선택합니다.
- **점유도 입력**을 **기본값**으로 선택하면 동일한 비율로 분석하고 **직접 입력**을 선택하면 아래와 같은 대화창이 떠서 사용자로부터 각 클래스 별 점유비를 입력 받습니다.



- **화면출력 노드**를 **QDA 노드**에 연결합니다.
- **QDA diagram** 예시는 아래와 같습니다.



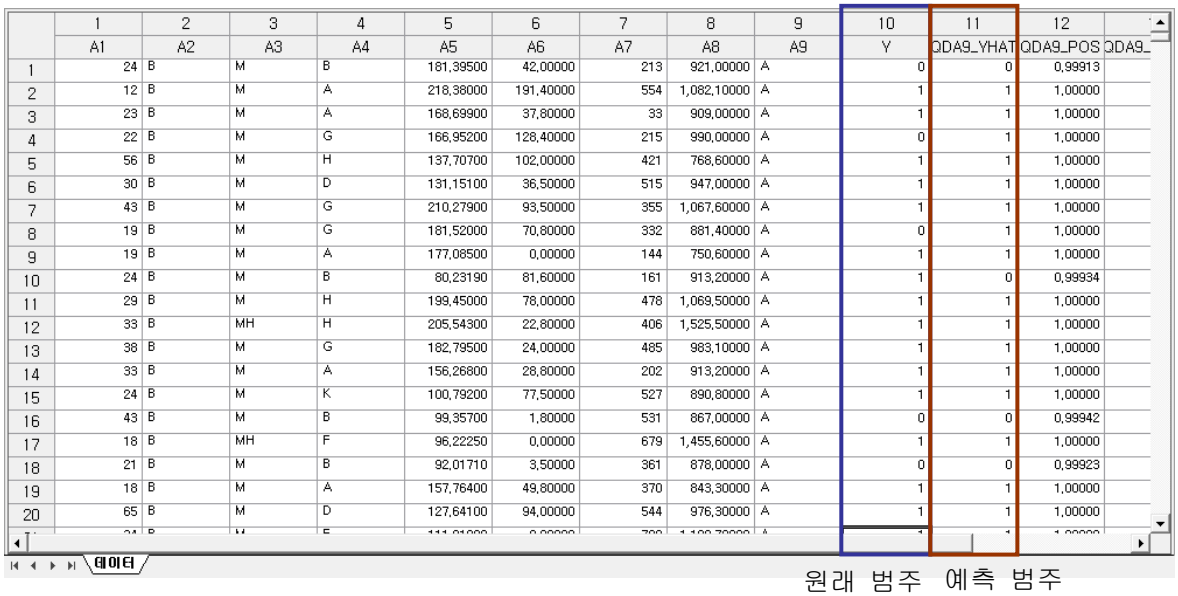
속성

속성 그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	점유도	점유비율 입력여부를 결정합니다. 점유도는 데이터에서 각 class 의 비율을 의미합니다.	필수	기본값, 직접입력

결과

▪ **분석결과정보**

화면표시 노드에서 분류예측 결과 및 예측값에 대한 확률을 확인할 수 있습니다.



	1	2	3	4	5	6	7	8	9	10	11	12	
	A1	A2	A3	A4	A5	A6	A7	A8	A9	Y	QDA9_YHAT	QDA9_POS	QDA9_
1	24	B	M	B	181,39500	42,00000	213	921,00000	A	0	0	0,99913	
2	12	B	M	A	218,38000	191,40000	554	1,082,10000	A	1	1	1,00000	
3	23	B	M	A	168,69900	37,80000	33	909,00000	A	1	1	1,00000	
4	22	B	M	G	166,95200	128,40000	215	990,00000	A	0	1	1,00000	
5	56	B	M	H	137,70700	102,00000	421	768,60000	A	1	1	1,00000	
6	30	B	M	D	131,15100	36,50000	515	947,00000	A	1	1	1,00000	
7	43	B	M	G	210,27900	93,50000	355	1,067,60000	A	1	1	1,00000	
8	19	B	M	G	181,52000	70,80000	332	881,40000	A	0	1	1,00000	
9	19	B	M	A	177,08500	0,00000	144	750,60000	A	1	1	1,00000	
10	24	B	M	B	80,23190	81,60000	161	913,20000	A	1	0	0,99934	
11	29	B	M	H	199,45000	78,00000	478	1,069,50000	A	1	1	1,00000	
12	33	B	MH	H	205,54300	22,80000	406	1,525,50000	A	1	1	1,00000	
13	38	B	M	G	182,79500	24,00000	485	983,10000	A	1	1	1,00000	
14	33	B	M	A	156,26800	28,80000	202	913,20000	A	1	1	1,00000	
15	24	B	M	K	100,79200	77,50000	527	890,80000	A	1	1	1,00000	
16	43	B	M	B	99,35700	1,80000	531	867,00000	A	0	0	0,99942	
17	18	B	MH	F	96,22250	0,00000	679	1,455,60000	A	1	1	1,00000	
18	21	B	M	B	92,01710	3,50000	361	878,00000	A	0	0	0,99923	
19	18	B	M	A	157,76400	49,80000	370	843,30000	A	1	1	1,00000	
20	65	B	M	D	127,64100	94,00000	544	976,30000	A	1	1	1,00000	

데이터

원래 범주 예측 범주

예시파일

▪ **LDA_QDA.ecm** 실행

3.4.15 RBF 노드

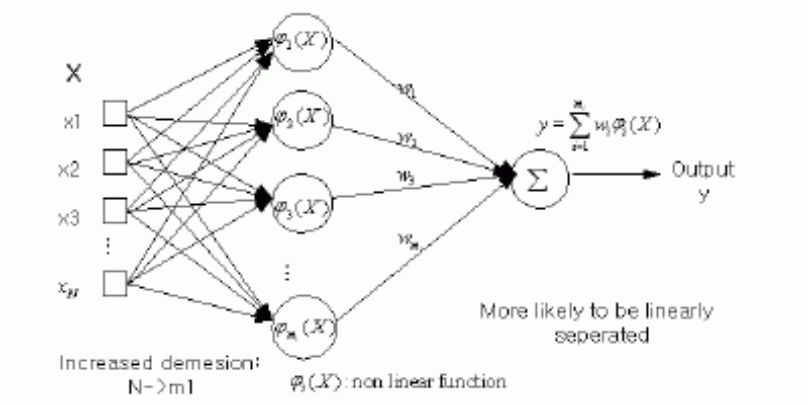
RBF 신경망은 은닉층이 한 개만 있으며 수학적 형태는 **MLP 신경망**과 유사하나, 은닉층의 결합함수로 원형기준함수를 사용합니다.

개요

RBF(Radial-Basis Function)네트웍은 **교사신경망(Supervised Neural Network)**으로서 고차원공간(High-dimensional Space)으로의 Feature mapping 을 통한 Curve fitting 에 주로 쓰입니다. 따라서 RBF 네트웍의 관점에서 학습(Learning)이란 고차원공간에서 Training example 을 가장 잘 fitting 하는 surface 를 찾는 것을 말합니다. RBF 네트웍의 hidden layer 는 input pattern 이 hidden space 로 확장될 때, basis 를 구성하는 function set 을 제공하며 이러한 function 들을 **radial-basis function** 이라 합니다. 따라서 RBF 네트웍은 이러한 radial-basis function 들의 weighted combination 의 형태를 띕니다. RBF 네트웍은 서로 다른 역할을 수행하는 세 개의 Layer 로 구성되어 있습니다.

Input layer 는 외부환경과 네트웍을 연결하는 source(sensory) units 로 구성됩니다. Hidden layer 에서는 input space 에서 hidden space 로 nonlinear transformation 을 수행합니다. 대개의 경우, Hidden space 는 high dimension 입니다. Output layer 는 선형이며, Input layer 에 적용된 pattern 에 대한 결과를 제공합니다.

RBF 네트웍의 구조는 다음 그림과 같습니다.



RBF 네트웍은 주로 복잡한 **Pattern classification** 문제들에 적용되며 이러한 문제들은 비선형형태로 high dimensional space 으로의 transforming 을 통해 풀려집니다. high

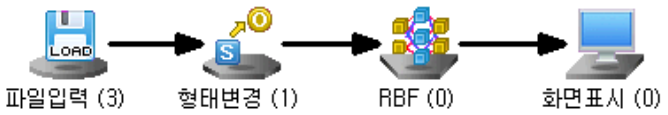
dimensional space 로 transform 된 pattern 이 low-dimensional space 의 그것보다 좀 더 linearly separable 하기 때문에 RBF 네트웍은 대개 hidden space 의 dimension 을 크게 하고 있습니다. 또 다른 중요한 관점은 hidden space 의 dimension 은 input-output mapping function 의 추정의 정확도를 판가름한다는 것입니다. 즉, hidden space 의 dimension 을 더 높일수록 training data 에 대해서는 좀 더 정확한 추정을 할 수 있게 됩니다.

고려사항

- RBF 노드 자체가 종속변수의 데이터 타입에 맞는 수행기법을 선택해 종속변수가 연속형일 때는 **예측분석**을 이산형일 때는 **분류분석**을 수행합니다.

사용법

- **입력 노드**를 통해 데이터를 읽어 들입니다.
- **형태 변경 노드**를 통해 읽어 들인 데이터의 타입을 지정합니다.(**독립변수, 종속변수**를 지정)
- **RBF 노드**를 **형태 변경 노드**에 연결하고 옵션들을 선택합니다.
- **화면표시 노드**를 **RBF 노드**에 연결합니다.
- **RBF diagram** 예시는 아래와 같습니다.



속성

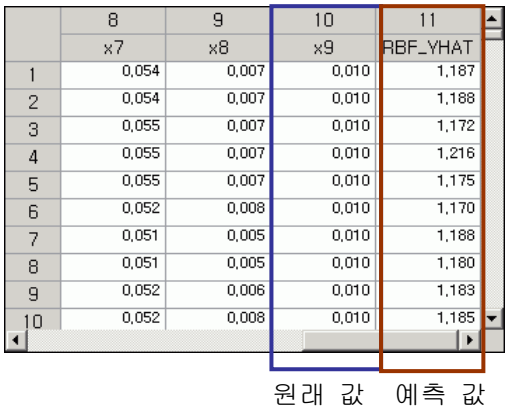
속성 그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	중심수	Radial Basis Function 의 center 의 수입니다.	필수	데이터수 보다

				작은 자연수
	람다	Regularization parameter	필수	

결과

▪ **분석결과정보**

화면표시 노드에서 분석 결과를 확인할 수 있습니다. 예측분석일 경우, 예측값과 잔차를 볼 수 있으며, 분류분석일 경우 분류된 결과를 볼 수 있습니다.



	8	9	10	11
	x7	x8	x9	RBF_YHAT
1	0,054	0,007	0,010	1,187
2	0,054	0,007	0,010	1,188
3	0,055	0,007	0,010	1,172
4	0,055	0,007	0,010	1,216
5	0,055	0,007	0,010	1,175
6	0,052	0,008	0,010	1,170
7	0,051	0,005	0,010	1,188
8	0,051	0,005	0,010	1,180
9	0,052	0,006	0,010	1,183
10	0,052	0,008	0,010	1,185

원래 값 예측 값

예시파일

▪ **RBF.ecm** 실행

3.4.16 순차 연관성 노드

순차 연관성 분석은 시간적인 흐름에 따라 상품 혹은 서비스간의 관계를 살펴보고 이로부터 **유용한 규칙**을 찾아내고자 할 때 이용될 수 있는 기법입니다. ECMiner™에서는 순차 연관성 분석을 위해 순차 연관성 분석 알고리즘을 제공합니다.

개요

순차 즉, 시간의 순서 또는 구매 순서 등 여러 가지 순차적으로 일어나는 사건에서 상관관계를 도출해내는 노드입니다.

고려사항

- **변수**는 고객 ID(CID), 시간(TID), 상품(ITEMID) 순서로 **3 개의 변수만**으로 구성되어야 합니다.

  예시)

	1	2	3
	고객No.	시간	과일
1	1	1	귤
2	1	2	귤
3	2	3	감
4	2	4	바나나
5	2	5	감
6	3	6	귤
7	3	7	귤
8	3	8	대추
9	4	9	감
10	4	10	감
11	4	11	바나나
12	4	12	사과
13	5	13	사과
14	5	14	감
15	5	15	감
16	6	16	대추
17	6	17	귤
18	6	18	귤

- 시간 변수의 경우 날짜형, 연속형 모두 가능합니다.

사용법

- **입력 노드**를 통해 데이터를 읽어 들입니다.
- **입력 노드**에 **순차연관성** 노드를 연결하고 옵션들을 선택합니다.
- **화면표시 노드**를 **순차연관성 노드**에 연결합니다.
- **순차연관성 diagram** 예시는 아래와 같습니다.

속성

속성 그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
모델파일	모델파일	모델링 후에 모델 파일을 생성할 지 여부를	필수	예,

	생성	선택합니다.		아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	최소 지지도(개)	항목들 간에 연관성을 정의하기 위한 최소한의 지지도 값을 설정하는 것으로 Default 는 5 입니다.	필수	자연수
	최대 패턴길이	패턴의 최대 길이를 입력 합니다.	필수	자연수
	최소 신뢰도(%)	연관성 정의 시 최소한의 신뢰도 값을 설정하는 것으로 Default 값은 50(%) 입니다.	필수	

예시파일

- **순차연관성 분석.ecm** 실행

3.4.17 ScoreCard 노드

Scorecard 노드는 Logistic 회귀분석의 알고리즘을 사용하여 종속변수(이산형)에 미치는 독립변수들의 영향도를 Score 형태로 계산해 줌으로써 사용자가 독립변수의 영향도를 더욱 쉽게 파악할 수 있도록 만든 모델링 방법입니다. 이는 금융 분야에서 고객의 신용 등급을 평가할 때 유용하게 사용되고 있습니다.

개요

Scorecard 는 기본적으로 Logistic 회귀분석의 방법론을 그대로 이용합니다. 쉽게 말해서 Scorecard 는 Logistic 회귀분석을 더욱 쉽게 이해하기 위해서 만든 방법이라고 할 수 있습니다. Scorecard 의 기능을 수행하기 위해서는 일단 구간화 노드를 이용하여 변수를 구간화 하는 과정이 필요합니다. 이렇게 구간화된 변수를 가지고 Scorecard 노드를 이용해 Scorecard 를 만들 수 있습니다.

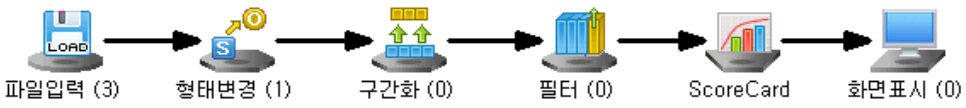
고려사항

- **구간화 노드**를 이용하기 위해서는 종속변수가 설정되어 있는 ecl 파일이 필요합니다. 이때 종속변수는 이산형이면서 클래스가 2 개이어야 합니다.

- **ScoreCard 알고리즘**은 이산형 변수만 입력될 수 있습니다. 즉, 구간화를 통해 기존의 연속형 독립변수를 모두 이산형으로 바꿔주어야 합니다. 종속변수는 이산형이면서 클래스가 2 개이어야 합니다.

사용법

- **입력노드**를 통해 데이터를 읽어 들입니다.
- **구간화 노드를** 이용해 독립변수를 가지고 구간 변수를 만듭니다.
- **필터 노드**를 이용해 구간 변수와 종속변수를 제외한 다른 변수를 제외시킵니다.
- **형태변경 노드**를 이용해 독립변수, 종속 변수를 지정합니다.
- **ScoreCard 노드**를 이용해 ScoreCard 를 만듭니다.
- **ScoreCard diagram** 예시는 아래와 같습니다.



속성

속성 그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	최대시행수	Logistic 회귀분석을 위해 Maximum Likelihood Estimator 를 추정할 때 최대 몇 번의 반복을 통해서 Estimator 를 추정할지를 결정합니다.	필수	
	최대값	Score 의 최대값을 입력합니다.	필수	
	최소값	Score 의 최소값을 입력합니다.	필수	

결과

▪ **분석결과정보**

화면표시 노드에서 분석 결과를 확인할 수 있습니다. 모델이 예측한 클래스와 예측값에 대한 확률, Score 의 값을 확인할 수 있습니다.

	1	2	3	4	5	6	7	8	9	10
	A2	A3	A4	A9	이탈여부	SC41_YHAT	SC41_SCORE	SC41_POS	SC41_PROB_0	SC41_PROB_1
1	D	ML	C	A	1	0	82,35677	0,50032	0,50032	0,49968
2	D	ML	C	A	1	0	82,35677	0,50032	0,50032	0,49968
3	D	ML	C	A	0	0	82,35677	0,50032	0,50032	0,49968
4	D	ML	C	A	1	0	82,35677	0,50032	0,50032	0,49968
5	D	ML	C	A	1	0	82,35677	0,50032	0,50032	0,49968
6	D	ML	C	A	0	0	82,35677	0,50032	0,50032	0,49968
7	D	ML	C	A	1	0	82,35677	0,50032	0,50032	0,49968
8	D	ML	C	A	1	0	82,35677	0,50032	0,50032	0,49968
9	B	L	G	A	1	1	82,49014	0,51244	0,48756	0,51244
10	D	ML	G	A	1	0	82,02380	0,53212	0,53212	0,46788
11	D	ML	G	A	1	0	82,02380	0,53212	0,53212	0,46788
12	D	ML	G	A	1	0	82,02380	0,53212	0,53212	0,46788
13	D	ML	G	A	1	0	82,02380	0,53212	0,53212	0,46788
14	D	ML	G	A	1	0	82,02380	0,53212	0,53212	0,46788
15	B	MH	C	A	0	0	82,00926	0,53351	0,53351	0,46649
16	B	MH	C	A	0	0	82,00926	0,53351	0,53351	0,46649
17	B	MH	C	A	1	0	82,00926	0,53351	0,53351	0,46649

원래 범주 스코어 범주가 1 일 확률

예측 범주 범주가 0 일 확률

예시파일

▪ **ScoreCard.ecm** 실행

3.4.18 SOM 노드

SOM 신경망은 인간의 대뇌 피질이 각 구역마다 맡은 기능이 존재하는 특징에 기반한 신경망 알고리즘입니다. **SOM 신경망**을 이용하여 고차원 데이터를 저차원으로 시각화하여 볼 수 있습니다.

개요

Self Organizing Map(혹은 Self Organizing Feature Map)은 대뇌 피질의 시각피질을 모델화한 인공신경망의 일종입니다. 이는 **비교사 학습(unsupervised learning)**의 일종으로 클러스터링 방법의 하나임과 동시에 차원을 줄여서 데이터를 가시화하는 방법 중의 하나입니다.

인공신경망의 일종인 자기 조직화 지도는 자율학습의 방법으로 훈련되며 저차원(보통 2 차원)의 지도를 생성합니다. 이 지도는 입력 공간에서 주어진 훈련 샘플에 대한 이산적인 표현을 나타내며 입력 공간에 대한 위상 속성을 보존하려고 합니다.

SOM 은 고차원으로 표현된 데이터를 저차원으로 변환해서 보는데 유용합니다. 이 모델은 이 모델을 처음 고안한 핀란드 과학자인 Teuvo Kohonen 의 이름을 본따 Kohonen map 이라고 불리기도 합니다. 다른 인공신경망과 같이, SOM 은 훈련과 매핑의 두 가지 모드로 동작합니다. 훈련은 입력 샘플을 이용해서 지도를 만드는 과정으로, 경쟁적이며, vector quantization 이라고 불립니다. 매핑 과정에서는 새로운 입력을 훈련 결과에 따라 자동적으로 분류합니다.

학습과정은 다음과 같습니다.

(1) Initialization : 가중치(Weight Vector)를 0 과 1 사이의 임의의 값으로 초기화합니다.

(2) Sampling : 입력데이터 중 하나를 선택합니다.

(3) Matching : 선택된 입력 데이터와 가장 가까운 노드를 찾습니다.(경쟁에 의한 승자전취)

(4) Updating : 승자노드와 승자노드에 이웃하는 노드들을 입력데이터 쪽으로 이동시킵니다.(협동학습)

(5) Continuation : Updating 이 미미해질 때까지 Sampling 부터 Updating 을 반복합니다.

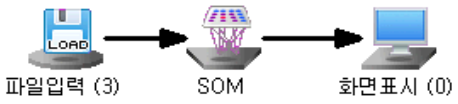
(적응과정)

고려사항

- **독립변수**만 입력되어야 하며 **입력 변수는** 연속형이어야 합니다.
- **비교사 학습(unsupervised learning)** 이므로 형태변경이 필요 없습니다.
- **SOM 노드** 자체에 표준화 전처리 과정을 포함하고 있습니다.

사용법

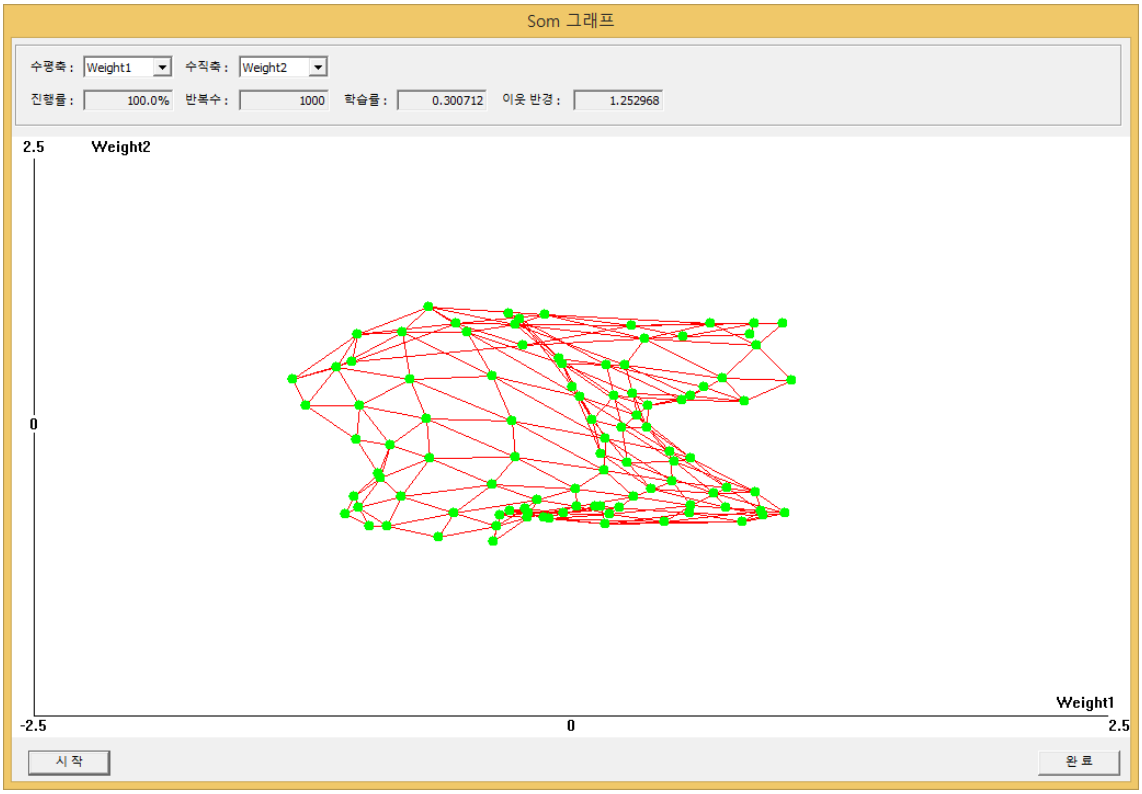
- **입력노드**를 통해 데이터를 읽어 들입니다.
- **SOM 노드**를 **입력노드**에 연결하고 옵션들을 선택합니다.
- **SOM diagram** 예시는 아래와 같습니다.



속성

속성 그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	가로방향 격자점 수	Topology 의 가로 방향 격자수를 입력합니다.	필수	자연수
	세로방향 격자점 수	Topology 의 세로 방향 격자수를 입력합니다.	필수	자연수
	최대 반복수	알고리즘의 최대 반복수를 입력합니다.	필수	
	Topology 선택	사용할 수 있는 Topology 는 두 가지가 있습니다. Hexagonal 과 Grid 중 하나의 Topology 를 선택하도록 합니다.	필수	Grid, Hexagonal
	학습률 초기값	알고리즘이 반복되면서 학습률은 지속적으로 작아집니다. Training 이 시작될 때 사용하는 학습률의 값을 입력하도록 합니다.	필수	
	이웃 반경 초기값	알고리즘이 반복되면서 이웃 반경의 값은 지속적으로 작아집니다. Training 이 시작될 때 사용하는 이웃 반경의 값을 입력하도록 합니다.	필수	

SOM 을 실행할 때에는 Training 과정을 시각적, 동적으로 보여줍니다. 이를 통해서 사용자는 Training 의 과정을 통해 Topology 가 어떻게 변하는지를 생생하게 볼 수 있습니다.



결과

▪ **분석결과정보**

화면표시 노드에서 분석 결과를 확인할 수 있습니다. 결과로 각 데이터가 어떤 group 에 해당하는지를 볼 수 있습니다.

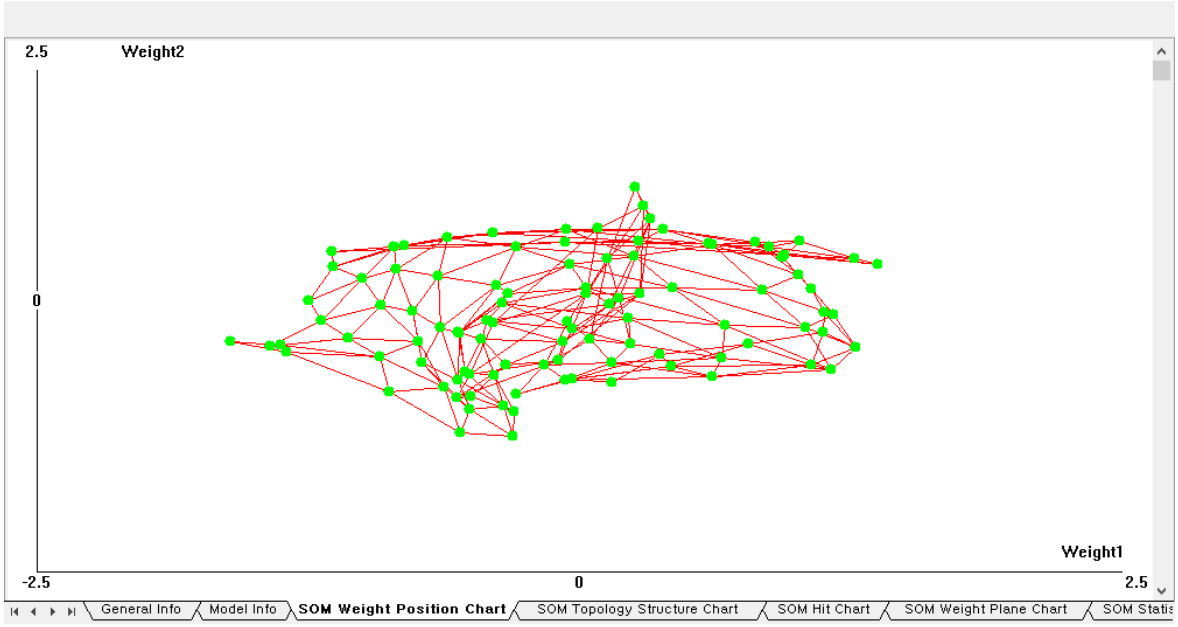
29	30	31	32	33	34
hnique	Aggression	Mentality	GK Skill	Team work	SOM28YHAT
97	85	72	50	82	1
87	83	89	50	90	7
87	94	78	50	76	8
91	80	83	50	85	7
85	70	98	50	95	7
94	82	76	50	82	1
77	62	72	50	84	28
85	73	82	50	88	7
85	95	85	50	82	8
95	85	85	50	87	6
43	57	90	91	80	80
83	95	67	50	72	9
97	85	73	50	83	1
92	86	73	50	78	1
88	92	68	50	75	8
83	80	83	50	80	1

예측 범주

분석을 실행한 Diagram 상에서 SOM 노드를 클릭해서 노드 속성창의 결과보기 버튼을 누르면 다음 결과들이 출력됩니다.

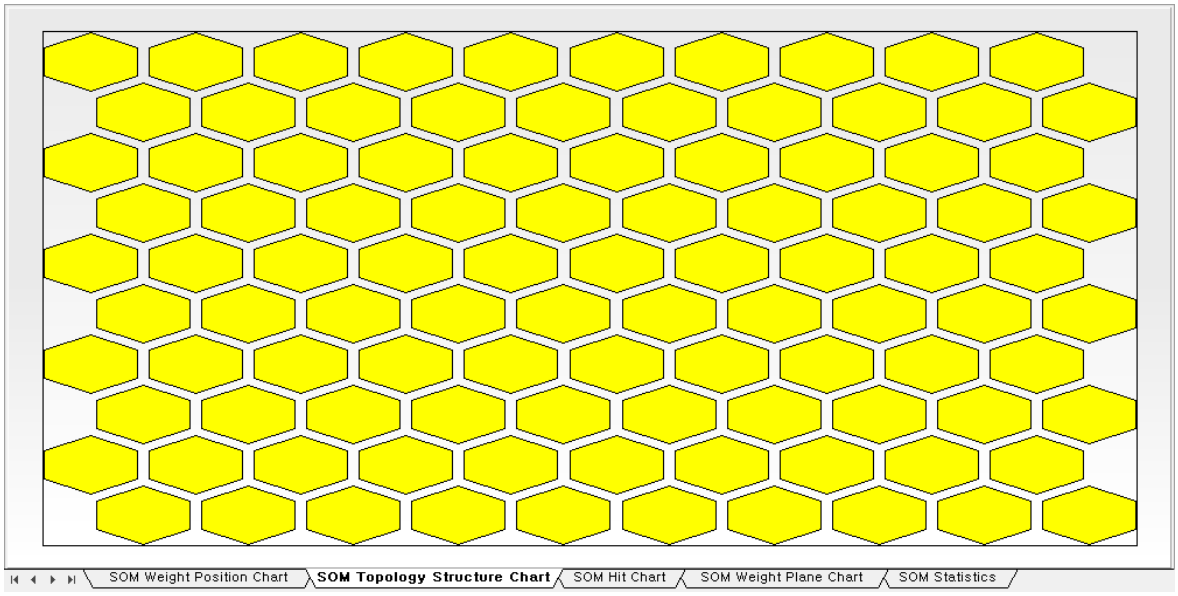
▪ **SOM Weight Position Chart**

본 Chart 를 통해서 Weight 가 공간상에 어떻게 배치되어 있는 지와 각 Weight 간 어떠한 연결관계가 있는지를 알 수 있습니다.



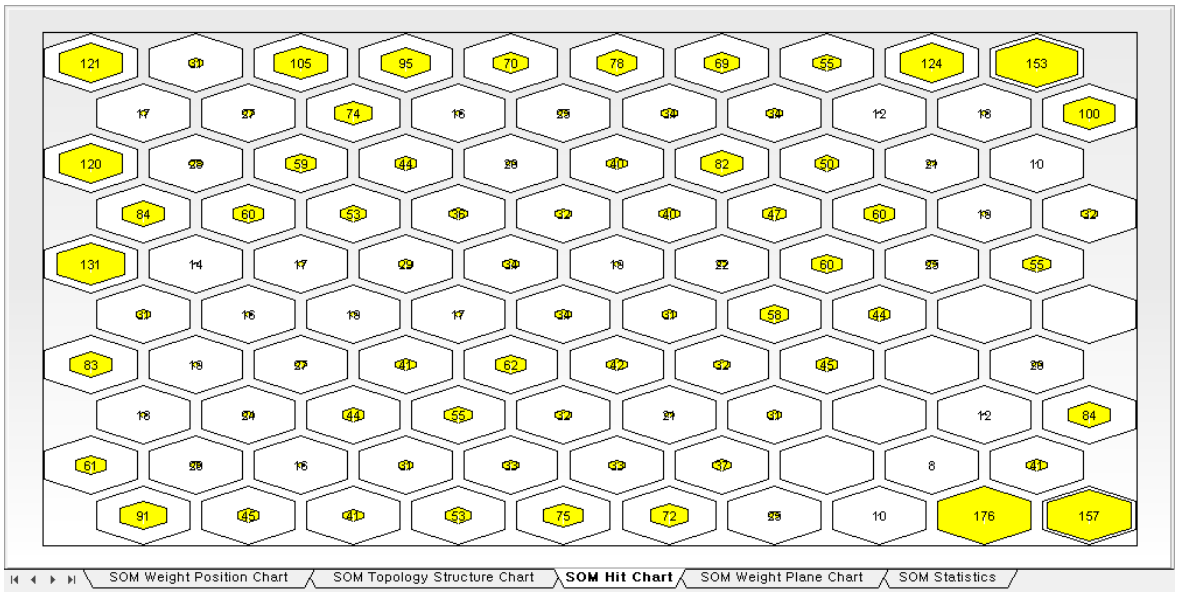
▪ **SOM Topology Structure Chart**

본 Chart 를 통해서 Topology 의 Structure 를 시각적으로 볼 수 있습니다. 아래의 예시는 Hexagonal Structure 를 시각화한 것입니다.



▪ **SOM Hit Chart**

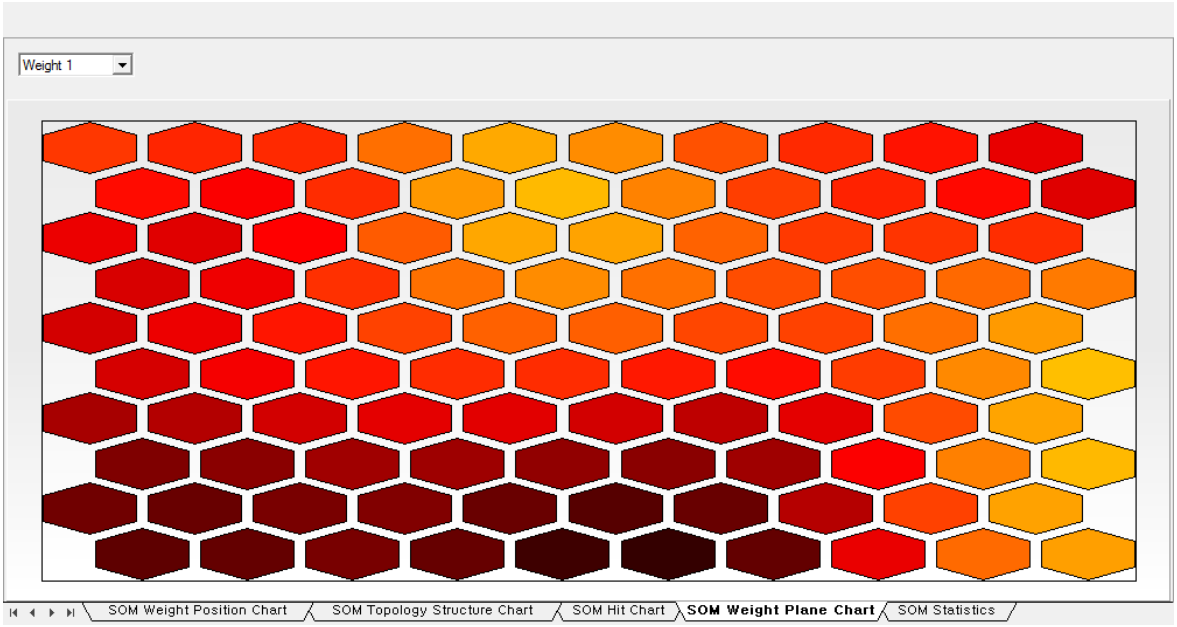
본 Chart 를 통해서 각 Group 당 가장 가까운 데이터가 몇 개씩 있는지를 시각적으로 확인할 수 있습니다.



▪ **SOM Weight Plane Chart**

본 Chart 를 통해서 각 Weight 의 성분의 값을 시각적으로 알 수 있습니다. Weight 1 이라는 것은 Weight Vector 의 첫 번째 성분을 나타냅니다. Topology 상에서 각 Weight 의

성분(여기서는 첫 번째 성분)의 값이 클수록 밝은 색이 되고 작을수록 짙은 색이 됩니다. ECMiner™ SOM 에서는 빨강색을 기준으로 Weight 의 성분 값이 0 보다 크면 노란색에 가까워지고, 0 보다 작으면 검은색에 가까워지도록 하였습니다.



▪ **SOM Statistics**

본 Statistics 를 통하여 Topology 상에 위치한 Weight Vector 의 성분 값을 구체적으로 알 수 있고 화면표시에서 나타나는 SOM_YHAT 이 나타내는 숫자에 해당하는 가로, 세로 격자 순서를 알 수 있습니다.

	1	2	3	4	5	6	7	8	9	10	11	
	세로 격자 순서	가로 격자 순서	Group Numbe	Weight1	Weight2	Weight3	Weight4	Weight5	Weight6	Weight7	Weight8	W
1	1	1	1	0,327905	-0,634704	0,91997	-0,642454	0,079498	0,747604	0,88671	1,365831	
2	1	2	2	0,217939	-0,622986	0,863317	-0,727052	-0,026231	0,560796	0,658245	1,179924	
3	1	3	3	0,244718	-0,626727	0,749674	-0,614102	-0,257809	0,322861	0,365286	0,763721	
4	1	4	4	0,657363	-0,687467	0,562843	-0,105627	-0,428812	0,253078	0,027108	0,155779	
5	1	5	5	0,999175	-0,699603	0,42922	0,401842	-0,356041	0,446084	-0,156141	-0,208089	
6	1	6	6	0,8242	-0,638509	0,476523	0,478725	-0,072834	0,715425	-0,035742	-0,094914	
7	1	7	7	0,478259	-0,466853	0,645471	0,217163	0,3734	0,837038	0,276125	0,329515	
8	1	8	8	0,241775	-0,149932	0,795897	-0,18843	0,700178	0,730627	0,549686	0,642798	
9	1	9	9	0,111396	0,445023	0,814863	-0,576172	0,770161	0,440623	0,640227	0,603598	
10	1	10	10	-0,132796	1,062177	0,771284	-0,930434	0,677138	0,170086	0,495218	0,300125	
11	2	1	11	0,069854	-0,683937	0,819564	-0,817509	-0,095698	0,615802	0,683023	1,23724	
12	2	2	12	-0,021141	-0,720363	0,722217	-0,757593	-0,309132	0,432505	0,441571	0,954965	
13	2	3	13	0,263244	-0,784021	0,57338	-0,350063	-0,49491	0,339862	0,195022	0,484964	
14	2	4	14	0,899578	-0,766051	0,373317	0,280549	-0,492194	0,403032	-0,076526	-0,036001	
15	2	5	15	1,097042	-0,687129	0,316159	0,587476	-0,267559	0,571207	-0,123923	-0,207018	
16	2	6	16	0,766867	-0,50324	0,395001	0,503337	0,136142	0,707191	0,074768	0,012945	
17	2	7	17	0,364988	-0,231756	0,521502	0,222341	0,58031	0,706224	0,336792	0,362268	
18	2	8	18	0,201268	0,155549	0,540954	-0,05337	0,744875	0,498605	0,426953	0,422433	
19	2	9	19	0,048289	0,70684	0,503387	-0,375861	0,687471	0,22506	0,316614	0,170127	
20	2	10	20	-0,186141	1,106997	0,529201	-0,737061	0,546156	0,066263	0,123955	-0,128902	
21	3	1	21	-0,110838	-0,640637	0,790196	-0,98708	-0,311946	0,5698	0,780287	1,256771	

SOM Weight Position Chart | SOM Topology Structure Chart | SOM Hit Chart | SOM Weight Plane Chart | **SOM Statistics**

예시파일

- **SOM.ecm** 실행

3.4.19 RBF DDA 노드

RBF DDA

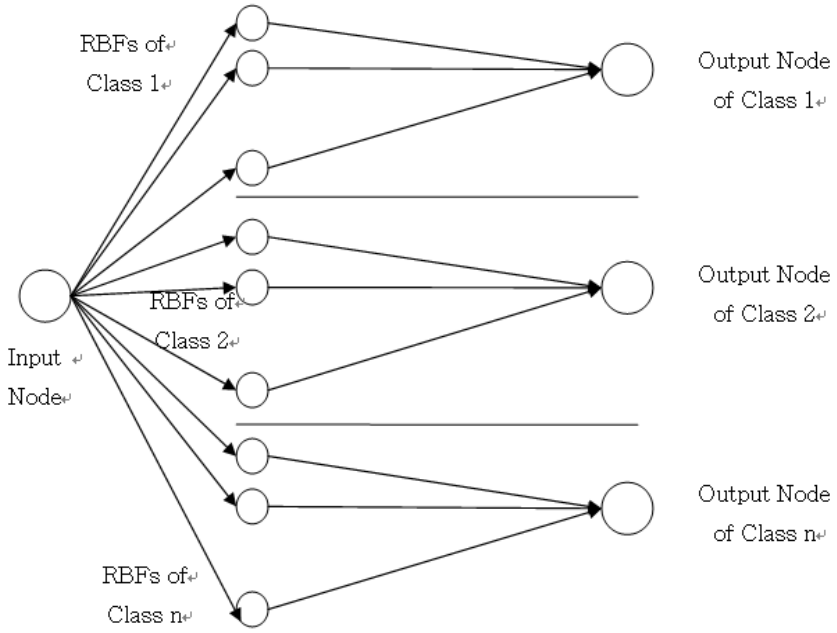
RBF DDA 는 공간상에 배치되는 RBF 의 개수를 자동적으로 정해 줌으로써 사용자가 알고리즘을 실행하기 위해 입력해야 하는 모수를 결정하는데 생기는 어려움을 해결한 알고리즘입니다.

개요

RBF DDA 는 Radial Basis Function with Dynamic Decay Adjustment 의 약자로 기존의 RBF 의 단점을 보완하기 위해서 만들어진 알고리즘입니다. 기존의 RBF Network 알고리즘은 사용자가 RBF 의 개수와 각 RBF 의 σ 값 등을 지정해야 했습니다. 하지만 이는 사용자의 입장에서는 매우 부담스러운 것으로 몇 개의 RBF 를 설정해야 하는지에 대한 문제는 데이터에 대한 기본적인 이해만으로는 해결하기 힘듭니다.

RBF DDA 는 RBF 의 개수를 자동적으로 정해 줍니다. 그리고 σ 와 같이 사용자가 정해주기 어려운 값들을 자동적으로 정해주게 됩니다. 물론 RBF DDA 의 경우 Regression 을 수행할 수는 없고 Classification 만을 수행해 주는 알고리즘이라는 것이 단점일 수 있겠지만 사용자의 부담을 덜어 주고, 또한 기존의 RBF 보다 더 좋은 성능을 나타낸다는 점에서 우위가 있다고 할 수 있습니다.

RBF - DDA 의 알고리즘을 가장 쉽고 간결하게 설명한다면 '매우 새로운 데이터가 들어오면 RBF 를 추가한다' 입니다. 먼저 그림을 통해서 RBF-DDA 를 표현하자면 다음과 같습니다.



i 번째 클래스의 j 번째 RBF 는 세 가지의 특성치를 갖습니다.

$\overrightarrow{r_j^i}$: i 번째 Class 의 j 번째, RBF 의 중심 좌표

σ_j^i : i 번째 Class 의 j 번째, sigma

A_j^i : i 번째 Class 의 j 번째, weight

이 때 i 번째 클래스의 j 번째 RBF 에 x 데이터가 들어가면 다음과 같은 반응 값이 구해집니다.

$$R_j^i(\vec{x}) = \exp\left(-\frac{\| \vec{x} - \vec{r}_j^i \|^2}{(\sigma_j^i)^2}\right)$$

이와 같은 반응 값을 이용하여 각 클래스마다 다음과 같은 output 을 출력하게 됩니다.

$$f^i(\vec{x}) = \sum_{j=1}^{m_i} A_j^i * R_j^i(\vec{x})$$

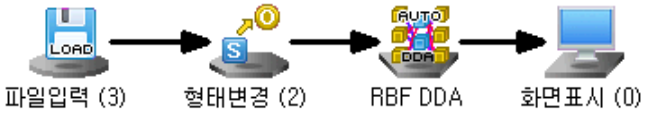
이 때 위의 값이 가장 큰 클래스로 x 데이터가 classification 되는 것입니다.

고려사항

- **독립변수, 종속변수가** 입력되어야 하며 독립변수는 연속형, 종속변수는 이산형이어야 합니다.
- **RBF DDA 노드** 자체에 표준화 전처리 과정을 포함하고 있습니다.

사용법

- **입력노드**를 통해 데이터를 읽어 들입니다.
- **형태 변경 노드**를 이용해 독립변수와 종속변수를 지정합니다.
- **RBF DDA 노드**를 **형태변경노드**에 연결하고 옵션들을 선택합니다.
- **RBF DDA diagram** 예시는 아래와 같습니다.



속성

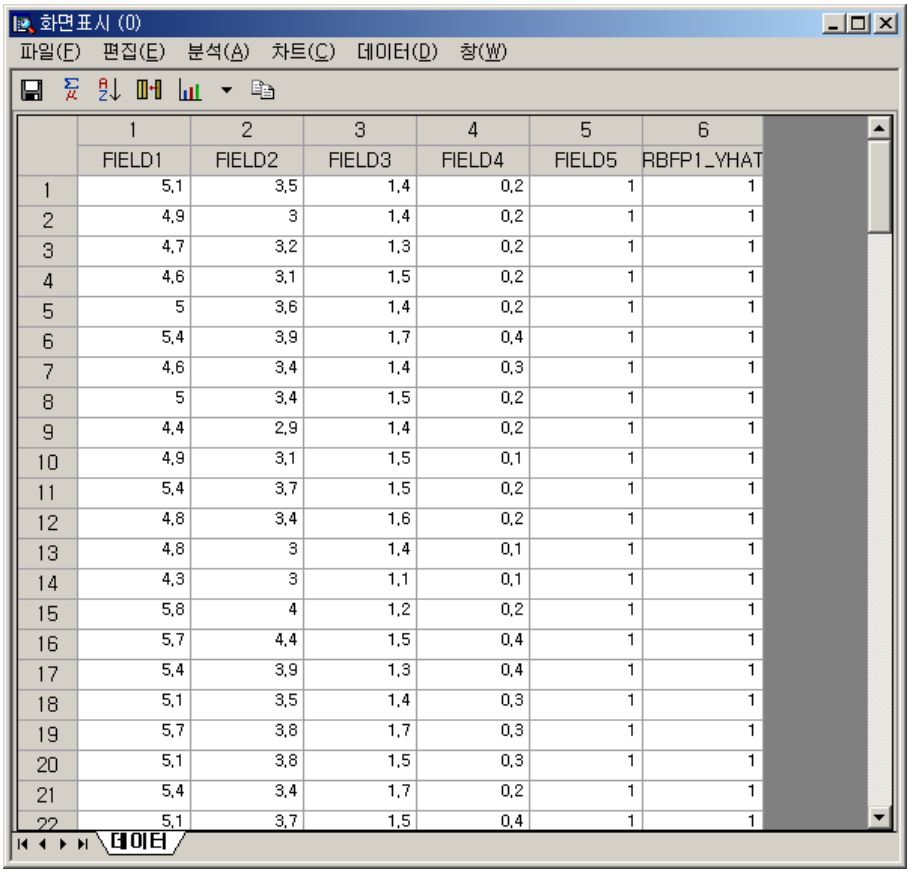
속성 그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	θ^+	RBF DDA Training 과정 중 새로운 데이터에 대해 데이터가 속해 있는 클래스의 RBF 의 Response 가 모두 이 값보다 작으면 새로운 RBF 를 추가합니다.	필수	
	θ^-	RBF DDA Training 과정에서 RBF 의 σ 값을 조정하기 위해서 사용되는 기준 값입니다.	필수	
	Min 추출 데이터	이는 알고리즘의 속도를 높이기 위해 도입한 parameter 입니다. RBF DDA 는 알고리즘의	필수	

	백분율	특성상 한번 새로운 데이터가 들어올 때마다 그리고 RBF 가 새로 생성될 때마다 기존의 RBF 를 모두 탐색하면서 σ 의 값을 변경 혹은 추가합니다. 이 과정에서 일부의 RBF 만을 선택함으로써 Training Time 을 줄일 수 있습니다. 따라서 본 parameter 를 통해서 몇 Percent 의 RBF 만을 탐색할 것인지를 설정하여 Training Time 을 줄일 수 있습니다.		

결과

▪ **분석결과정보**

화면표시 노드에서 분석 결과를 확인할 수 있습니다. 모델이 예측한 클래스를 볼 수 있습니다.



	1	2	3	4	5	6
	FIELD1	FIELD2	FIELD3	FIELD4	FIELD5	RBFP1_YHAT
1	5,1	3,5	1,4	0,2	1	1
2	4,9	3	1,4	0,2	1	1
3	4,7	3,2	1,3	0,2	1	1
4	4,6	3,1	1,5	0,2	1	1
5	5	3,6	1,4	0,2	1	1
6	5,4	3,9	1,7	0,4	1	1
7	4,6	3,4	1,4	0,3	1	1
8	5	3,4	1,5	0,2	1	1
9	4,4	2,9	1,4	0,2	1	1
10	4,9	3,1	1,5	0,1	1	1
11	5,4	3,7	1,5	0,2	1	1
12	4,8	3,4	1,6	0,2	1	1
13	4,8	3	1,4	0,1	1	1
14	4,3	3	1,1	0,1	1	1
15	5,8	4	1,2	0,2	1	1
16	5,7	4,4	1,5	0,4	1	1
17	5,4	3,9	1,3	0,4	1	1
18	5,1	3,5	1,4	0,3	1	1
19	5,7	3,8	1,7	0,3	1	1
20	5,1	3,8	1,5	0,3	1	1
21	5,4	3,4	1,7	0,2	1	1
22	5,1	3,7	1,5	0,4	1	1

예시파일

▪ **RBF.ecm** 실행

3.4.20 Factor Analysis 노드

Factor Analysis 는 소수의 의미 있는 변수로 다수의 관측된 변수를 설명하는 방법론입니다. 이를 통해 얻어진 요인들을 회귀분석이나 판별분석에 이용 할 수도 있습니다.

개요

Factor Analysis 는 적은 수의 variable 로 관측된 variable 을 설명하는 방법론입니다. 이러한 적은 수의 variable 을 common factor 라고 하는데 이것이 관측되지 않기 때문에 어려움이 있습니다. Factor Analysis 는 관측되지 않는 이 common factor 들을 찾고 이 common factor 들이 각 관측된 variable 에 어떠한 영향을 미치는지를 알고자 하는 것이 목적입니다. 예를 들어 여러 학생의 수학, 영어, 국어, 사회, 과학, 기술 성적이 있다고 합니다. 경험적으로 수학을 잘하는 학생들은 과학, 기술을 잘하는 경우가 많고 국어를 잘하는 학생들은 영어, 사회를 잘하는 경우가 많습니다. 이는 아마도 수학, 과학, 기술 성적은 인간의 계량적인 능력에 의해 영향을 많이 받고, 국어, 영어, 사회는 인간의 언어적인 능력에 영향을 많이 받기 때문이라 추측해 볼 수 있습니다.

$$x_1 = \mu_1 + \lambda_{11}F_1 + \lambda_{12}F_2 + \epsilon_1$$

$$x_2 = \mu_2 + \lambda_{21}F_1 + \lambda_{22}F_2 + \epsilon_2$$

$$x_3 = \mu_3 + \lambda_{31}F_1 + \lambda_{32}F_2 + \epsilon_3$$

$$x_4 = \mu_4 + \lambda_{41}F_1 + \lambda_{42}F_2 + \epsilon_4$$

$$x_5 = \mu_5 + \lambda_{51}F_1 + \lambda_{52}F_2 + \epsilon_5$$

$$x_6 = \mu_6 + \lambda_{61}F_1 + \lambda_{62}F_2 + \epsilon_6$$

기술된 내용을 수식으로 표현하면 위와 같습니다. 이 때 X_i 는 각 과목의 성적을 나타내고, μ_i 는 각 과목 성적의 모평균을 말합니다. 그리고 F_1 을 common factor1, F_2 을 common factor2 라고 하고 ε_i 를 common factor 들로 설명되지 않는 부분이라고 합니다. 국어, 영어 사회 성적에 대해서 F_1 의 값이 크고, F_2 은 값은 작고, 또한 수학, 과학, 기술 성적에 대해 F_2 의 값이 크고, F_1 의 값이 작다면 F_2 을 계량적 능력을 나타내는 하나의 지표가 되는 요인으로, F_1 를 언어적 능력을 나타내는 하나의 지표가 되는 요인으로 **해석**할 수 있습니다. Factor analysis 를 통해서 위와 같은 분석을 하게 되는 것입니다.

고려사항

- **독립변수**만 알고리즘 실행에 적용되며, 독립변수는 연속형이어야 합니다.
- **Factor Analysis 노드** 자체에 표준화 전처리 과정을 포함하고 있습니다.

사용법

- **입력노드**를 통해 데이터를 읽어 들입니다.
- **Factor Analysis 노드**를 **입력노드**에 연결하고 옵션들을 선택합니다.
- **Factor Analysis diagram** 예시는 아래와 같습니다.

속성

속성 그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	추출할 요인	데이터에서 추출할 요인의 수를 입력합니다.	필수	자연수

	수	추출방법이 Maximum Likelihood 일 경우 입력변수의 수보다 요인수가 적어야 하며, 추출방법이 주성분일 경우 추출 요인수가 입력변수 수보다 작거나 같아야 합니다.		
	추출 방법	요인을 추출할 방법을 선택합니다. 주성분과 Maximum Likelihood 에서 하나를 선택할 수 있습니다.	필수	주성분, Maximum Likelihood
	회전 유형	추출한 요인의 해석을 용이하게 하기 위해 어떠한 회전을 사용할지 선택합니다. 없음, Equimax, Varimax, Quartimax 중 하나를 선택할 수 있습니다.	필수	없음, Equimax, Varimax, Quartimax
	알고리즘 최대 반복 수	최적화 알고리즘의 최대 반복 수를 지정합니다.	필수	

▪ 요인추출방법

- 공통분산을 통한 잠재변수의 추정오차를 고려하여 추출합니다.
- Principal Component 방법으로 측정 변수의 분산을 측정합니다. 변수들의 선형결합으로 주성분을 도출하는 모형입니다.
- Maximum Likelihood (ML)는 우도함수를 최대화(목적함수의 최소화)하는 추정량(Λ, Ψ2)을 구하는 방법입니다. CV/TV 비율이 크거나 자료의 특성에 대하여 아는 바가 없을 때 사용합니다.
- ECMiner™에서는 주성분과 ML 방법 제공하며 그 외 추정방법으로는 Common Factor Model, ULS, GLS 가 있습니다.

▪ 요인회전방법

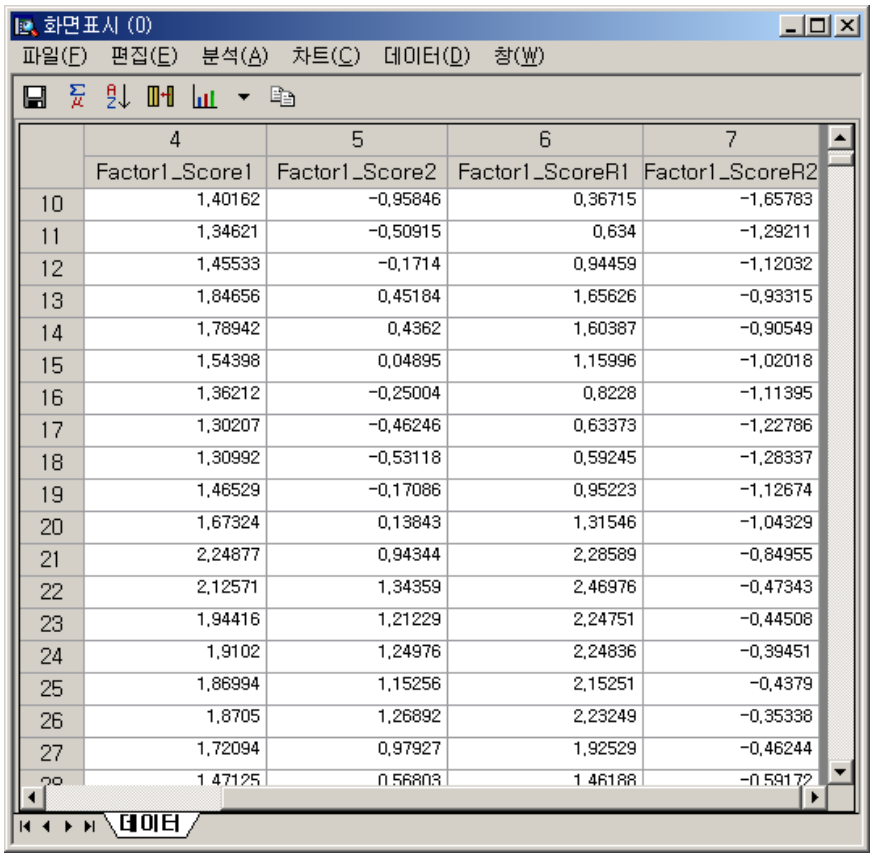
- 요인 적재치가 큰 값은 더욱 크게, 작은 값은 더욱 줄여 요인행렬의 해석을 용이하게 합니다.
- 직각회전(orthogonal)은 회전축을 직각으로 유지한 방법이며, 요인이 독립적인 관계를 가집니다.
- Varimax(직각회전)
  각 요인의 적재값이 높은 변수의 수를 최소화하는 직교 회전 방법입니다. 이 방법을 사용하면 요인 해석을 단순화할 수 있습니다.
- Quartimax(직각회전)
  각 변수를 설명하는 데 필요한 요인 수를 최소화하는 회전 방법입니다. 이 방법을 사용하면 관측된 변수의 해석을 단순화할 수 있습니다.

- Equimax(직각회전)
  요인을 단순화하는 베리멕스 방법과 변수를 단순화하는 쿼티멕스 방법을 조합한 회전 방법입니다. 요인에 읽어 들인 변수의 수와 변수 설명에 사용할 요인 수는 최소화됩니다.
- 비직각회전(oblique)은 요인 축 간의 관계(상관)를 어느 정도 허용한 방법입니다.

결과

▪ **분석결과정보**

화면표시 노드에서 분석 결과를 확인할 수 있습니다. Factor Score 와 Rotated Factor Score 를 볼 수 있습니다.



	4	5	6	7
	Factor1_Score1	Factor1_Score2	Factor1_ScoreR1	Factor1_ScoreR2
10	1,40162	-0,95846	0,36715	-1,65783
11	1,34621	-0,50915	0,634	-1,29211
12	1,45533	-0,1714	0,94459	-1,12032
13	1,84656	0,45184	1,65626	-0,93315
14	1,78942	0,4362	1,60387	-0,90549
15	1,54398	0,04895	1,15996	-1,02018
16	1,36212	-0,25004	0,8228	-1,11395
17	1,30207	-0,46246	0,63373	-1,22786
18	1,30992	-0,53118	0,59245	-1,28337
19	1,46529	-0,17086	0,95223	-1,12674
20	1,67324	0,13843	1,31546	-1,04329
21	2,24877	0,94344	2,28589	-0,84955
22	2,12571	1,34359	2,46976	-0,47343
23	1,94416	1,21229	2,24751	-0,44508
24	1,9102	1,24976	2,24836	-0,39451
25	1,86994	1,15256	2,15251	-0,4379
26	1,8705	1,26892	2,23249	-0,35338
27	1,72094	0,97927	1,92529	-0,46244
28	1,47125	0,56803	1,46188	-0,59172

예시파일

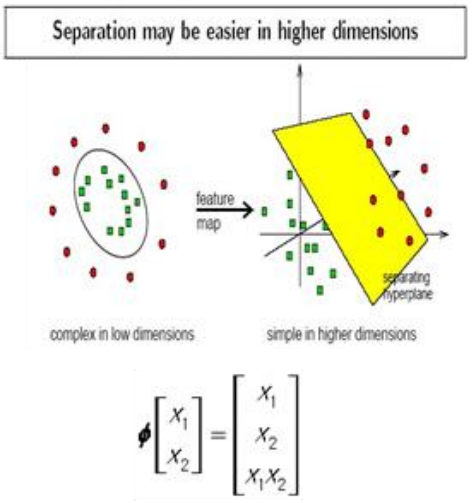
▪ **Factor Analysis.ecm** 실행

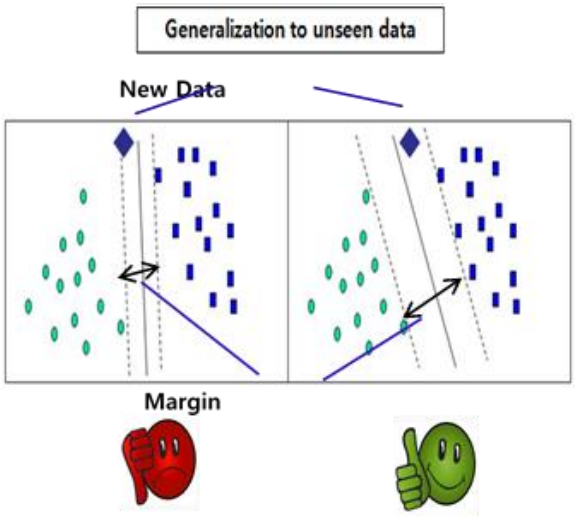
3.4.21 SVM 노드

SVM 노드는 다수의 속성(attribute) 또는 변수를 갖는 객체(object)를 사전에 정해진 **그룹 또는 범주(class, category)** 중의 하나로 분류하는 **분류분석 기법** 중 하나로 Support Vector Machine 알고리즘을 사용합니다.

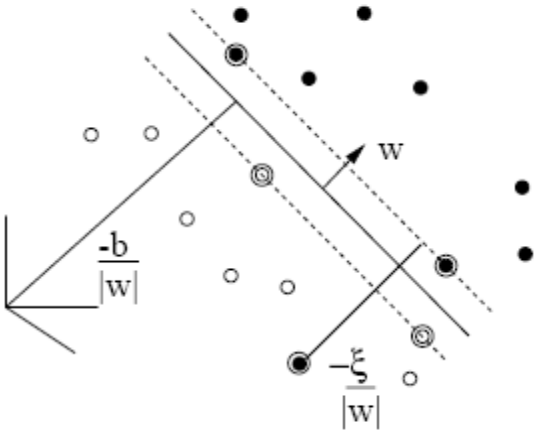
개요

SVM 는 비모수적인 분류 방법의 하나로, 고차원으로 매핑된 Feature 공간에서, Margin 이 최대가 되는 선형 분류기를 학습하는 알고리즘을 말합니다.왼쪽 아래 그림에서 보시는 것처럼, 입력 공간에서 두 집단을 선형적으로 나눌 수 없을 경우가 존재하는데, SVM 에서는 저차원 데이터를 선형 분류가 가능한 고차원으로 매핑한 후, 고차원에서 선형 분류를 진행합니다.오른쪽 아래의 그림은 Margin 최대화라는 SVM 의 특징을 보여줍니다.SVM 는 데이터와 분류 평면이 되도록 멀리 떨어지도록 하여, 새로운 데이터에 대해서도 판별이 잘 되도록 합니다.





SVM 은 Support Vector Machine(SVM)을 사용한 분류 분석 기법입니다. C-SVM 과 nu-SVM 두 가지 방식이 존재합니다. C-SVM 의 아이디어는 다음의 그림이 가장 잘 표현한다고 할 수 있습니다.



위와 같이 두 개의 클래스를 갖는 데이터가 존재할 때, 두 클래스를 가를 수 있는 평면은 매우 많습니다. 하지만 그 중에서 가장 좋은 평면은 바로 위의 두 개의 점선으로 표현된 평면 사이의 실선으로 표현된 평면이라고 할 수 있습니다. 이 두 평면 사이의 거리가 멀면 멀수록 더욱 안전하게 분류를 할 수 있을 것이기 때문입니다. 두 평면 위의 Vector 를 Support Vector 라 하고, 두 평면 사이의 거리를 Margin 이라고 하는데 최대 마진을 갖는 평면을 찾는 문제가 SVM 의 문제입니다. 이 때 두 평면 사이의 거리는 $d = \frac{2}{\|w\|}$ 이고, 이는 $\frac{1}{2}\|\mathbf{w}\|^2$를 최소화하는 문제를 풂으로써 같은 결과를 얻을 수 있습니다. 단, 하나의 평면으로 완벽하게 분리가 불가능한 데이터(Non-separable case)의 경우에는 변형된 식을 통해 약간의 오차를 허용하면서 최적의 분류를 하기 위해 다음과 같이 Formulate 합니다.

$$\min_{w,b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \quad s.t.\ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \quad \xi_i \geq 0\ i = 1,2,\dots,l$$

위 문제를 Dual problem 을 통하여 풀면,

$$L = -\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j + \sum_{i=1}^{l}\alpha_i \quad s.t.\ 0 \leq \alpha_i \leq C$$

와 같이 나타낼 수 있습니다. 위의 식에 마이너스를 붙여서 최소화 문제로 바꾸고, $\mathbf{x}_i^T\mathbf{x}_j$를 내적한 부분을 Kernel Function 으로 바꾸면 분류 곡면을 구할 수 있습니다. 최종적인 Decision function 은 다음과 같습니다.

$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b\right)$$

Nu-SVM 은 C-SVM 를 고도화하기 위한 방법론으로, ν라는 새로운 parameter 를 support vector 의 수와 training error 를 control 하기 위해서 도입합니다. 데이터의 수가 충분하다고 할 때, 전체 vector 중 support vector 의 개수의 비율과, training error 의 비율이 같아지기 때문에 이와 같은 방법을 사용할 수 있습니다. Nu-SVM 의 primal problem 은

$$\min_{\mathbf{w},b,\xi,\rho} \frac{1}{2}\mathbf{w}^T\mathbf{w} - \nu\rho + \frac{1}{l}\sum_{i=1}^{l}\xi_i \quad \text{s.t. } y_i(w^T\phi(x_i) + b) \geq \rho - \xi_i \quad \xi_i \geq 0, i = 1,2,\ldots,l \ \rho > 0.$$

입니다. C-SVM 와 마찬가지로 Dual problem 을 통해 구한 Decision function 은 다음과 같습니다.

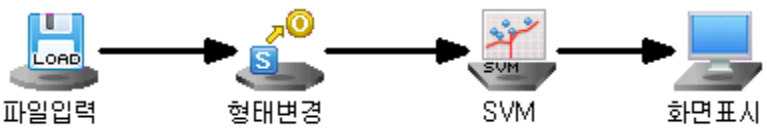
$$\mathrm{sgn}\left(\sum_{i=1}^{l} y_i \left(\frac{\alpha_i}{\rho}\right)\left(K(x_i, x) + b\right)\right)$$

고려사항

- **독립변수, 종속변수가** 입력되어야 하며 독립변수는 연속형, 종속변수는 이산형이어야 합니다.

사용법

- **입력 노드**를 통해 데이터를 읽어 들입니다.
- **형태 변경 노드**를 통해 읽어 들인 데이터의 타입을 지정합니다. (**독립변수, 종속변수**를 지정)
- **SVM 노드**를 **형태 변경 노드**에 연결하고 옵션들을 선택합니다.
- **화면표시 노드**를 SVM **노드**에 연결합니다.
- **SVM diagram** 예시는 아래와 같습니다.



속성

속성 그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	유형	사용할 SVM 알고리즘의 종류를 선택합니다.	필수	C-SVM, NU-SVM
	커널 유형	판별식에 사용할 커널 함수를 선택합니다.	필수	LINEAR, POLY, RBF, SIGMOID
	degree	Polynimial 커널 함수의 차수입니다. 커널 유형 옵션에서 'Polynomial' 선택 시 활성화 됩니다.	선택	
	gamma	각 커널 함수의 계수입니다. 커널 유형 옵션에서 'Polynomial','RBF','Sigmoid' 선택 시 활성화 됩니다.	선택	
	Coef0	각 커널 함수의 상수항입니다. 커널 유형 옵션에서 'Polynomial','Sigmoid' 선택 시 활성화 됩니다.	선택	
	nu	Nu-SVM 판별식의 nu 값입니다. 유형 옵션에서 'Nu-SVM' 선택 시 활성화 됩니다.	선택	
	C	C-SVM 판별식의 C 값입니다. 유형 옵션에서 'C-SVM' 선택 시 활성화 됩니다.	선택	
	최적 모수 추정	최적의 Nu 값 또는 C 값을 추정합니다.		예, 아니오

결과

▪ **분석결과정보**

화면표시 노드에서 분류분석 결과를 확인할 수 있습니다.

27	28	29	30	31	32
JOB	NUM_DEPENDENTS	TELEPHONE	FOREIGN	RESPONSE	SVM29_YHAT
2	1	1	0	1	1
2	1	0	0	0	0
1	2	0	0	1	1
2	2	0	0	1	1
2	2	0	0	0	0
1	2	1	0	1	1
2	1	0	0	1	1
3	1	1	0	1	1
1	1	0	0	1	1
3	1	0	0	0	0
2	1	0	0	0	0
2	1	0	0	0	0
2	1	1	0	1	1
1	1	0	0	0	1
2	1	0	0	1	0
1	1	0	0	0	0
2	1	0	0	1	1
2	1	0	0	1	0
3	1	1	0	0	0

원래 범주 예측 범주

예시파일

- **SVM.ecm** 실행

3.4.22 SVR 노드

하나 또는 둘 이상의 변수들이 다른 하나의 변수에 미치는 **영향의 정도와 방향**을 파악하기 위해 **회귀분석**을 사용합니다. **Support Vector Regression 노드**는 Support Vector Machine 의 아이디어를 사용하여 회귀분석을 수행합니다.

개요

SVR 은 Support Vector Regression 의 약자로, SVM 의 아이디어를 발전시켜 회귀분석에 응용한 방법론입니다. ECMiner™ 은 epsilon-SVR 과 nu-SVR 을 제공합니다. $\epsilon - SVR$은 특히 모든 데이터가 다음과 같은 관계를 만족시키도록 함수를 찾는 것을 목표로 합니다.

$$|z_i - f(\mathbf{x}_i)| \leq \epsilon$$

함수가 x 에 대해 선형이라고 할 때, 찾는 평면이 납작할수록 좋기 때문에 풀어야 할 문제는 다음과 같습니다.

$$\min_{w} \frac{1}{2} \| w \|^2$$

,

그러나 모든 데이터가 epsilon tube 안에 존재할 가능성은 희박하기 때문에, epsilon tube 를 벗어나는 error ξ 를 고려한 완화된 제약식을 통해 문제를 변형합니다. 즉, 다음과 같은 Training data 가 주어졌을 때, $\{(x_1, z_1), \ldots, (x_l, z_l)\}$ $x_i \in R^n$, $z_i \in R$ Support Vector Regression 의 Standard Form 은 다음과 같습니다.

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i + C \sum_{i=1}^{l} \xi_i^*$$

, ,

그리고 위의 문제의 쌍대 문제 (Dual Problem)은 다음과 같습니다.

$$\min_{\alpha,\alpha^*} \frac{1}{2} (\alpha - \alpha^*)^T Q (\alpha - \alpha^*) + \epsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} z_i (\alpha_i - \alpha_i^*)$$

$$\text{s.t.} \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0, \quad 0 \le \alpha_i, \alpha_i^* \le C, \quad i = 1, 2, \ldots, l,$$

여기서 Q 는 $l \times l$ positive semidefinite matrix 로 $Q_{ij} = K(x_i, x_j)$, $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ 입니다.

위의 결과로 얻어지는 approximate function 은 다음과 같습니다.

$$f(x_i) = \sum_{i=1}^{l} (-\alpha_i + \alpha_i^*) K(x_i, x) + b.$$

nu-SVR 은 epsilon-SVR 을 발전시키기 위한 아이디어로, ϵ 의 값을 자동적으로 정해주는 알고리즘을 제시하고자 하는 것이 바로 ν SVR 입니다. ν SVR 를 Formulate 하면 다음과 같습니다.

$$\min_{w,b,\xi,\xi^*,\epsilon} \frac{1}{2} w^T w + C \left(\nu\epsilon + \frac{1}{l} \sum_{i=1}^{l} (\xi_i + \xi_i^*) \right)$$

, ,

이 문제의 dual problem 에서, inequality $e^T(\alpha + \alpha^*) \le Cv$ 를 equality 로 변경하고 scale 을 조정하면 다음과 같은 식을 얻을 수 있습니다.

$$\min_{\alpha,\alpha^*} \frac{1}{2}(\alpha - \alpha^*)^T Q(\alpha - \alpha^*) + z^T(\alpha - \alpha^*)$$

$$s.t.\ e^T,$$

그리고 이를 통해 다음과 같은 Decision function 을 얻을 수 있습니다.

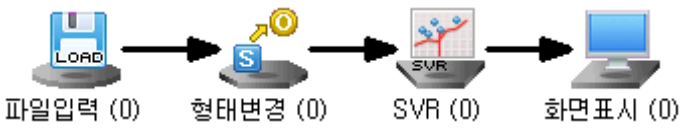
$$f(x) = \sum_{i=1}^{l}(-\alpha_i + \alpha_i)K(x_i, x) + b$$

고려사항

- **독립변수, 종속변수가** 입력되어야 하며 독립변수는 연속형, 종속변수는 이산형이어야 합니다.

사용법

- **입력 노드**를 통해 데이터를 읽어 들입니다.
- **형태 변경 노드**를 통해 읽어 들인 데이터의 타입을 지정합니다. (**독립변수, 종속변수**를 지정)
- **SVR 노드**를 **형태 변경 노드**에 연결하고 옵션들을 선택합니다.
- **화면표시 노드**를 **SVR 노드**에 연결합니다.
- **SVR diagram** 예시는 아래와 같습니다.



속성

속성 그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	

모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	유형	사용할 SVR 알고리즘의 종류를 선택합니다.	필수	EPSILON-SVR NU-SVR
	커널 유형	회귀식에 사용할 커널 함수를 선택합니다.	필수	LINEAR, POLY, RBF, SIGMOID
	degree	Polynimial 커널 함수의 차수입니다. 커널 유형 옵션에서 'Polynomial' 선택 시 활성화 됩니다.	선택	
	gamma	각 커널 함수의 계수입니다. 커널 유형 옵션에서 'Polynomial','RBF','Sigmoid' 선택 시 활성화 됩니다.	선택	
	Coef0	각 커널 함수의 상수항입니다. 커널 유형 옵션에서 'Polynomial','Sigmoid' 선택 시 활성화 됩니다.	선택	
	nu	Nu-SVR 회귀식의 nu 값입니다. 옵션에서 'Nu-SVR' 선택 시 활성화 됩니다.	선택	
	C	SVR 회귀식의 C 값입니다.	선택	
	epsilon	Epsilon-SVR 회귀식의 epsilon 값입니다. 옵션에서 'Epsilon-SVR' 선택 시 활성화 됩니다.	선택	

결과

▪ **분석결과정보**

화면표시 노드에서 회귀분석 결과를 확인할 수 있습니다.

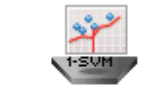
6	7	8	9	10	11	12	13	14	15
RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	SVR33_YHA
6,57500	65,20000	4,09000	1	296	15,30000	396,90000	4,98000	24,00000	29,32735
6,42100	78,90000	4,96710	2	242	17,80000	396,90000	9,14000	21,60000	23,07020
7,18500	61,10000	4,96710	2	242	17,80000	392,83000	4,03000	34,70000	31,22404
6,99800	45,80000	6,06220	3	222	18,70000	394,63000	2,94000	33,40000	28,98575
7,14700	54,20000	6,06220	3	222	18,70000	396,90000	5,33000	36,20000	29,33387
6,43000	58,70000	6,06220	3	222	18,70000	394,12000	5,21000	28,70000	23,91565
6,01200	66,60000	5,56050	5	311	15,20000	395,60000	12,43000	22,90000	21,58031
6,17200	96,10000	5,95050	5	311	15,20000	396,90000	19,15000	27,10000	19,40061
5,63100	100,00000	6,08210	5	311	15,20000	386,63000	29,93000	16,50000	16,60005
6,00400	85,90000	6,59210	5	311	15,20000	386,71000	17,10000	18,90000	19,00012
6,37700	94,30000	6,34670	5	311	15,20000	392,52000	20,45000	15,00000	19,94972
6,00900	82,90000	6,22670	5	311	15,20000	396,90000	13,27000	18,90000	20,28667
5,88900	39,00000	5,45090	5	311	15,20000	390,50000	15,71000	21,70000	21,12690
5,94900	61,80000	4,70750	4	307	21,00000	396,90000	8,26000	20,40000	19,52510
6,09600	84,50000	4,46190	4	307	21,00000	380,02000	10,26000	18,20000	18,55036
5,83400	56,50000	4,49860	4	307	21,00000	395,62000	8,47000	19,90000	19,39884

원래 범주 예측 범주

예시파일

- **SVR.ecm** 실행

3.4.23 One class SVM 노드



One Class SVM

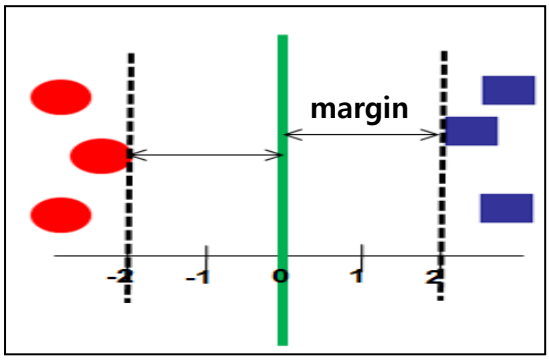
One class SVM 은 Support Vector Machine 을 활용하여 이상치 판별을 하는 알고리즘입니다. 기존의 SVM 과 같이 데이터를 고차원으로 mapping 하여 하나의 군집(class)로 군집화를 진행합니다. LOF 와 더불어 이상치 탐지 알고리즘으로 활용할 수 있습니다.

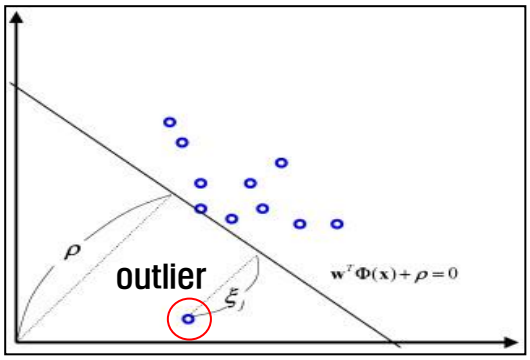
개요

One class SVM 은 기존의 Support Vector Machine 알고리즘을 활용한 방법입니다. 기존의 SVC(Support Vector Machine Classification) 알고리즘의 경우 두 개 이상의 class 를 갖는 데이터를 이용하여 support vector machine 알고리즘을 이용하여 class 를 분류하는 방법이었다면, One class SVM 은 데이터가 하나의 class 를 갖고 있을 때 training 을 통하여 데이터를 군집화 한 후 새로운 데이터에 대해 해당 군집에 속하는지를 확인하는 방법입니다. 만약 해당 군집에 속하지 않을 경우, outlier 라고 판단 할 기준이 될 수 있습니다.

$$\min_{\mathbf{w},\xi,\rho} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu m}\sum_{i=1}^{m}\xi_i - \rho,$$
$$\text{subject to } \langle \mathbf{w}, \Phi(x_i)\rangle \geq \rho - \xi_i,\ \xi_i \geq 0,$$

위의 식은 기존의 SVM 의 알고리즘에서 설명되어 있는 식과 같습니다. 위의 식은 다음과 같은 의미를 가집니다. 첫째로, 기존의 SVM 같이 margin 을 최대화 하기 위한 목적이 있습니다. 두 번째는, 군집을 나누는 선형 분류 평면을 좀 더 tight 하게 조정하는 의미를 갖고 있습니다. 이는 원점으로부터 최대한 멀리 떨어지도록 ρ 인자를 조정합니다. 마지막으로 slack variable 을 통해 나눠진 군집에 속하지 않는 데이터에 대해 penalty 를 주게 됩니다. 세 가지의 식을 통합적으로 고려한 평면이 만들어 지며, 이를 통해 데이터의 군집이 나눠지게 됩니다. 이 과정은 아래의 그림을 참고하시면 됩니다.





추가적으로 kernel 함수의 식에 대해 살펴보겠습니다.

i) Linear : $K(\mathbf{x},\mathbf{x}') = (\mathbf{x}^T\mathbf{x}')$

ii) D-degree Polynomial : $K(\mathbf{x},\mathbf{x}') = (\gamma\mathbf{x}^T\mathbf{x}' + coef.)^d$

iii) Radial Basis Function : $K(\mathbf{x},\mathbf{x}') = \exp\left(-\gamma\|\mathbf{x}-\mathbf{x}'\|^2\right)$

iv) Sigmoid : $K(\mathbf{x},\mathbf{x}') = \tanh(\gamma\mathbf{x}^T\mathbf{x}' + coef.)$

현재 ECMiner™에서의 Kernel function 은 다음과 같이 4 가지의 function 으로 되어있으며, 각각의 parameter 가 존재합니다. 각 function 별 parameter 는 옵션에서 설정 할 수 있습니다.

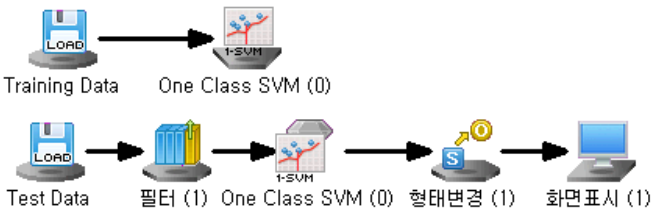
고려사항

- 종속변수를 따로 설정하지 않습니다..
- Traning 데이터의 경우 정상 데이터를 이용해야만 test 결과 시 이상치 판단에 용이합니다.

사용법

- **입력 노드**를 통해 훈련용 데이터를 읽어 들입니다.
- **One Class SVM 노드를 입력 노드에** 연결하고 옵션들을 선택 한 후 분석을 실행합니다.
- 실행 후 생성된 **One Class SVM 모델 노드**를 이용하여 테스트 데이터에 대한 분석을 수행합니다.

- **입력 노드**를 통해 테스트 데이터를 읽어 들입니다.
- **필터 노드**를 통해 훈련용 데이터에 없는 변수를 삭제합니다.
- **One Class SVM 모델 노드**를 연결합니다.
- **형태변경 노드**를 이용하여 **예측 범주 변수(OneClassSVM10_YHAT)**를 이산형으로 변경합니다.(옵션)
- **One Class SVM diagram** 예시는 아래와 같습니다.



속성

속성 그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	필수	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	커널 유형	판별식에 사용할 커널 함수를 선택합니다.	필수	LINEAR, POLY, RBF, SIGMOID
	degree	Polynimial 커널 함수의 차수입니다. 커널 유형 옵션에서 'Polynomial' 선택 시 활성화 됩니다.	조건부 필수	
	gamma	각 커널 함수의 계수입니다. 커널 유형 옵션에서 'Polynomial','RBF','Sigmoid' 선택 시 활성화 됩니다.	조건부 필수	
	Coef0	각 커널 함수의 상수항입니다. 커널 유형 옵션에서 'Polynomial','Sigmoid' 선택 시 활성화 됩니다.	조건부 필수	

	nu	v 값 지정, 0 에서 1 사이의 값을 가지며, boundary 의 tightness 를 정해주는 척도를 나타냄. Default 값으로 0.5 를 가집니다.	필수	

결과

▪ **분석결과정보**

화면표시 노드에서 One class SVM 결과인 decision value 와 그에 따른 이상치 판별 결과를 확인할 수 있습니다.

	1	2	3	4	5	6	7
	FIELD1	FIELD2	FIELD3	FIELD4	FIELD5	OneClassSVM10_DecisionValue	OneClassSVM10_YHA
1	0,07000	0,40000	0,54000	0,35000	0,44000	-1,1794	1
2	0,59000	0,49000	0,52000	0,45000	0,36000	-1,1949	1
3	0,67000	0,39000	0,36000	0,38000	0,46000	-1,4422	1
4	0,21000	0,34000	0,51000	0,28000	0,39000	0,2132	0
5	0,42000	0,40000	0,56000	0,18000	0,30000	-0,4518	1
6	0,25000	0,48000	0,44000	0,17000	0,29000	-0,5129	1
7	0,51000	0,50000	0,46000	0,32000	0,35000	0,1624	0
8	0,25000	0,40000	0,46000	0,44000	0,52000	-0,1424	1
9	0,44000	0,27000	0,55000	0,52000	0,58000	-2,7456	1
10	0,41000	0,57000	0,39000	0,21000	0,32000	-0,9557	1
11	0,31000	0,23000	0,73000	0,05000	0,14000	-7,4902	1
12	0,30000	0,16000	0,56000	0,11000	0,23000	-4,3667	1
13	0,29000	0,37000	0,48000	0,44000	0,52000	-0,0879	1
14	0,21000	0,51000	0,50000	0,32000	0,41000	-0,0594	1
15	0,43000	0,39000	0,47000	0,31000	0,41000	1,0207	0

Decision value 이상치 판별 결과

예시파일

▪ **One Class SVM.ecm** 실행

3.4.24 LOF 노드


LOF

LOF 는 Local Outlier Factor 의 약자로, 데이터가 입력되었을 때 입력된 데이터와 가까운 곳에 위치한 기존 데이터들의 지역적인 밀도를 반영하여 이상치를 판별하는 알고리즘입니다. One Class SVM 과 더불어 이상치 탐지 알고리즘으로 활용할 수 있습니다.

개요

LOF 는 Local Outlier Factor 의 약자로, Test 데이터가 입력되면 Test 데이터와 가까운 곳에 위치한 Training 데이터의 지역적인 밀도를 반영하여 이상치를 판별하는 알고리즘입니다. LOF 의 경우, 학습 단계가 존재하지 않는 lazy learning 의 일종입니다. Lazy learning 이란 training 데이터에 대하여 별도의 학습 과정 없이 새로운 test 데이터가 올 때까지 저장만 해두고 있는 것을 말합니다. 다시 말해, 우선적으로 training 데이터로 LOF 모델링을 진행하고, 그 모델을 test 데이터에 적용시켜서 training 데이터와 떨어져 있는 데이터를 이상치로 판단하는 것입니다. 이러한 lazy learning 알고리즘으로는 K-Nearest Neighbor 알고리즘이 있습니다.

우선, LOF 에서는 다음과 같이 값을 정의합니다.

$\text{k-distance}(A)$= Object A 로부터 k 번째로 가까이 있는 데이터와의 거리

$N_k(A)$= A 에서부터 k-distance(A) 거리 내에 있는 모든 데이터의 집합

$$\text{reachability-distance}_k(A,B) = \max\{\text{k-distance}(B), d(A,B)\}$$

Local reachability density 는 중심점 A 에서 이웃점들까지의 reachability distance 평균의 역수로 정의합니다.

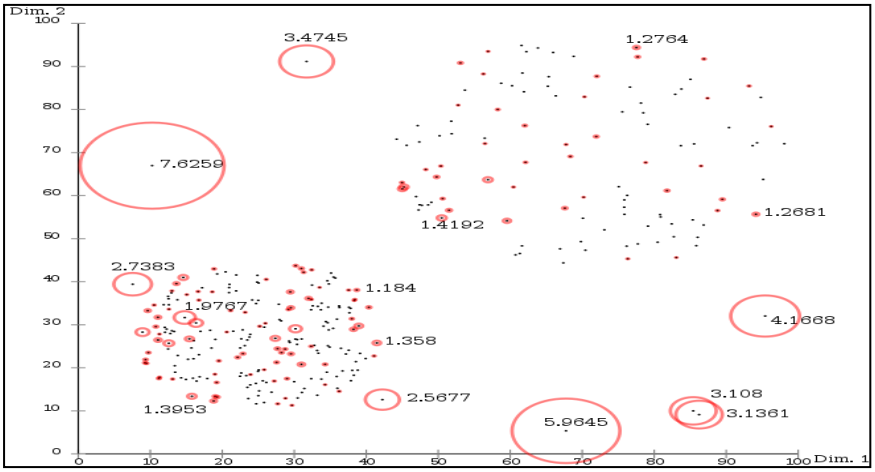
$$\text{lrd}(A) := 1/\left(\frac{\sum_{B\in N_k(A)} \text{reachability-distance}_k(A,B)}{|N_k(A)|}\right)$$

LOF score 는 중심점(A)의 주변 밀도와 이웃 점들의 주변 평균 밀도의 비율을 계산하여 구합니다.

$$\text{LOF}_k(A) := \frac{\sum_{B\in N_k(A)} \frac{\text{lrd}(B)}{\text{lrd}(A)}}{|N_k(A)|} = \frac{\sum_{B\in N_k(A)} \text{lrd}(B)}{|N_k(A)|}/\text{lrd}(A)$$

만약 중심점(A)의 주변 밀도가 주변 점들의 평균 밀도보다 크게 되면 LOF score 값이 1 보다 작은 값을 가지며, 반대의 경우는 1 보다 큰값을 갖습니다. 1 보다 큰 score 를 가질 경우, 해당 중심점(A)가 멀리 떨어져 있다는 것을 의미하므로, 큰 값을 가질수록 이상치에 가깝다고 판단할 수 있습니다.

아래 그림에서처럼 데이터가 모여 있는 집단 내의 데이터는 LOF score 가 1 에 근접한 값을 갖지만, 떨어져 있는 데이터의 경우 LOF score 가 3 이상의 값을 갖는 것을 볼 수 있습니다.

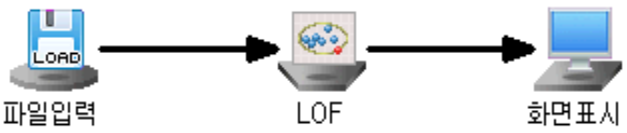


고려사항.

- 입력 데이터는 **연속형**이어야 합니다.
- 종속변수를 따로 설정하지 않으셔도 됩니다.
- Traning 데이터의 경우 정상 데이터를 이용해야만 test 결과 시 이상치 판단에 용이합니다.
- K 값은 최소 1 이상이어야 합니다.

사용법

- **입력 노드**를 통해 데이터를 읽어 들입니다.
- **LOF 노드를 입력 노드에** 연결하고 옵션들을 선택합니다.
- **화면표시 노드**를 **LOF 노드**에 연결합니다.
- **LOF diagram** 예시는 아래와 같습니다.



속성

속성 그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	

	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
모델파일	모델파일 생성	모델링 후에 모델 파일을 생성할 지 여부를 선택합니다.	필수	예, 아니오
	모델파일 경로	모델 파일을 생성할 시 저장할 모델 파일의 경로를 선택합니다.	조건부 필수	
선택사항	이웃의 수	이웃데이터로 할 k 값을 뜻합니다.k 는 최소 1 이상이어야 합니다. 일반적으로 3 으로 설정합니다.	필수	
	이상치 판별 여부	Test 데이터를 모델 통과할 시 '이상치 기준 점수' 와 LOF score 값을 비교하여 이상치 여부를 파생변수로 출력합니다.	예/아니오	
	이상치 기준 점수	이상치로 판단할 LOF score 점수입니다.	조건부 필수	
	거리 측정법	거리측정에 쓰일 방법입니다. Euclidean, Manhattan, Minkowski 방법이 있습니다.	필수	
	Minkowski 계수	거리측정법을 Minkowski 로 설정했을 때 활성화 됩니다.	조건부 필수	

결과

▪ **분석결과정보**

화면표시 노드에서 LOF 결과인 LOF score 값이 나옵니다.

화면표시 노드에서 LOF 결과인 LOF score 값이 나옵니다.

	1	2	3	4	5	6
	FIELD1	FIELD2	FIELD3	FIELD4	FIELD5	LOF8_Score
1	0.49000	0.29000	0.56000	0.24000	0.35000	1.05488
2	0.56000	0.40000	0.49000	0.37000	0.46000	1.03431
3	0.23000	0.32000	0.55000	0.25000	0.35000	1.23066
4	0.29000	0.28000	0.44000	0.23000	0.34000	1.06630
5	0.20000	0.44000	0.46000	0.51000	0.57000	1.20084
6	0.42000	0.24000	0.57000	0.27000	0.37000	1.04733
7	0.39000	0.32000	0.46000	0.24000	0.35000	1.02009
8	0.22000	0.43000	0.48000	0.16000	0.28000	1.25155
9	0.34000	0.45000	0.38000	0.24000	0.35000	1.03326
10	0.23000	0.40000	0.39000	0.28000	0.38000	0.95771
11	0.40000	0.45000	0.38000	0.22000	0.00000	2.18250
12	0.51000	0.54000	0.41000	0.34000	0.43000	1.05800
13	0.36000	0.39000	0.48000	0.22000	0.23000	1.03480
14	0.25000	0.40000	0.47000	0.33000	0.42000	1.11859
15	0.43000	0.37000	0.53000	0.35000	0.44000	1.03240

예시파일

- **LOF.ecm** 실행

3.5 모델 노드

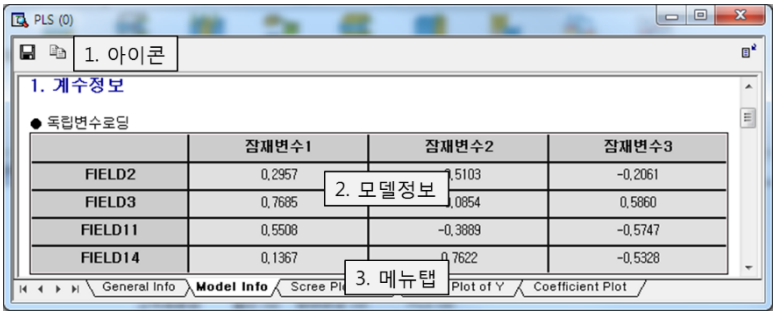
모델 노드는 모델링을 통하여 생성된 노드입니다. 사용자가 어떻게 스트림을 구성하여 모델링을 하였느냐에 따라 여러 가지 형태의 **모델 노드**가 산출될 수 있습니다. **모델 노드**는 크게 두 가지로 나눌 수 있는데 임시로 생성된 **임시모델**과 전역화 시킨 **전역 모델**로 나눌 수 있습니다. 임시로 생성된 모델은 **생성된 모델창**에 나타나며 프로그램 종료 시 이는 사라집니다. 이들 중 잘 만들어진 모델이라고 판단되면 **전역 모델화**할 수 있으며 프로그램이 종료되어도 이는 사라지지 않고 다음에 재사용 할 수 있습니다.
ECMiner™은 22 가지 알고리즘을 지원하며 이에 따라 다른 종류의 모델이 생성됩니다. 그리고 알고리즘에 따라 모델의 결과 (통계량, 출력값 등)가 달라지며 이하 섹션을 참조하시기 바랍니다.

화면 구성

화면 구성은 **모델 노드** 공통 사용자 인터페이스를 설명합니다.

(1) 모델 화면(Model Window)

모델 노드는 다음과 같은 모델 화면을 제공합니다.



1. 아이콘: 모델 주 화면에서 지원하는 기능을 표시합니다.

2. **모델정보**: 해당되는 모델 알고리즘의 정보를 보여줍니다. 알고리즘마다 표시되는 정보는 달라집니다.

3. 메뉴탭: 모델 보고서의 각종 결과 화면으로의 이동을 가능하게 합니다.

(2) 아이콘

모델 주 화면 상단부에는 유용한 기능을 제공하는 아이콘들이 모여있습니다. 각 아이콘의 기능은 다음과 같습니다.

 저장하기

 클립보드로 복사하기

일반 정보

일반 정보 화면에선 독립 변수, 종속 변수의 통계 정보와 모델 생성 정보를 볼 수 있습니다.

독립 변수 정보

Observation 개수: 177

총 13개의 독립변수가 사용됨.

순번	변수명	통계량 정보				
1	alcohol	총합	최소값	최대값	범위	MidRange
		2300	11.03	14.83	3.8	12.93
		평균	분산	표준편차	첨도	변동계수
		12.99	0.6555	0.8096	-0.8743	6.231
2	malic-acid	총합	최소값	최대값	범위	MidRange
		411.8	0.74	5.8	5.06	3.27

종속 변수 정보

총 1개의 종속변수가 사용됨.

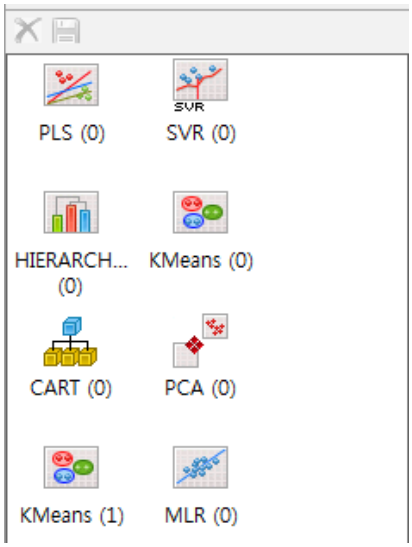
순번	변수명	통계량 정보			
1	type	Value	빈도수	Graph	백분율
		1	59		33.33%
		2	71		40.11%
		3	47		26.55%
		Total Count	177		

모델 생성 정보

Model Type	CART (Classification And Regression Tree)
생성자	Administrator
생성 일자	2004년 04월 09일 16:59
소요 시간	343 ms.

관련항목

모델 데이터 그룹은 모델링 결과인 데이터 마이닝 모델을 데이터 형태로 관리합니다.



프로젝트창 내 스트림 중에 모델링 노드가 포함되어 있다면 모델이 산출되며 새로 생성된 모델은 모델 데이터 그룹에서 관리됩니다. 노드와 같이 더블 클릭 혹은 드래그로 프로젝트창에 새로 생성 후 다른 노드와 연결하여 사용할 수 있습니다. 모델링은 모델을 생성하는 과정 입니다. 목적에 따라 적절한 **데이터 마이닝 알고리즘 (모델링 노드)**과 옵션을 선택하고 실행시키면 해당되는 모델이 생성됩니다. 모델의 정보를 보려면 작업 창에 노드를 생성한 뒤 더블 클릭합니다. 생성정보를 통해 모델의 적합성을 판단할 수 있습니다.
모델그룹 영역에 있는 모델파일을 Local 이나 Server 의 폴더로 드래그하면 파일이 이동하며 **광역모델화** 됩니다. 또한 저장 버튼을 눌려 **광역모델화**가 가능합니다.

3.5.1 연관성 분석 모델 노드

리소스창의 Model 윈도우에서 **연관성 분석 모델 노드**를 더블 클릭하면 다음과 같은 정보를 볼 수 있습니다.

예제데이터 (연관성분석.csv)

어느 과일 가게에 고객이 구입한 과일의 정보가 아래와 같을 때 과일 선호에 따른 규칙 탐사를 연관성 분석 기법을 활용하여 수행하였습니다.

	1	2	3	4	5	6	7	8
	고객	사과	배	감	귤	대추	바나나	파인애플
1	1	1	1	1	0	0	0	0
2	2	1	1	1	1	1	1	1
3	3	0	1	1	0	0	0	0
4	4	0	0	0	1	1	1	0
5	5	0	0	1	0	0	1	1
6	6	0	1	1	0	1	0	1
7	7	1	0	1	0	1	0	1
8	8	0	1	0	1	0	1	0
9	9	1	0	0	1	0	0	0
10	10	0	0	1	0	0	0	0
11	11	1	1	1	0	1	0	1
12	12	1	0	1	0	1	0	0
13	13	1	1	1	1	1	0	0
14	14	1	1	1	0	0	1	0
15	15	0	0	1	1	1	1	1

연관성 분석 수행 기본정보

수행 시 선택된 옵션(최소 지지도, 아이템 셋 크기, 최소 신뢰도)에 대한 정보와 생성 셋의 수를 나타냅니다.

생성된 Sets

연관성분석으로 생성된 Frequent Item Set 으로 전체 거래내역에서 아이템들의 조합을 나타내고 있습니다.

순번	Set	아이템수	지지도
1	귤, 감, 바나나, 수박, 참외	5	3
2	사과, 배, 귤, 감, 바나나	5	2
3	사과, 배, 귤, 바나나, 참외	5	2
4	배, 귤, 감, 바나나	4	3
5	배, 귤, 감, 수박	4	2
6	귤, 감, 바나나, 수박	4	3
7	배, 감, 바나나, 수박	4	2
8	배, 귤, 감, 참외	4	2
9	귤, 감, 바나나, 참외	4	4

생성된 규칙

순번	연관규칙	아이템 수	신뢰도(%)	향상도	지지율(%)
1	[사과] [배] --->[감]	3	100,00	1,43	20,00
2	[사과] [배] --->[대추]	3	50,00	1,07	10,00
3	[사과] [감] --->[배]	3	75,00	1,32	20,00
4	[사과] [감] --->[대추]	3	62,50	1,34	16,67
5	[사과] [귤] --->[배]	3	66,67	1,18	6,67
6	[사과] [귤] --->[감]	3	66,67	0,95	6,67
7	[사과] [귤] --->[대추]	3	66,67	1,43	6,67
8	[사과] [대추] --->[배]	3	60,00	1,06	10,00
9	[사과] [대추] --->[감]	3	100,00	1,43	16,67
10	[사과] [대추] --->[파인애플]	3	60,00	1,64	10,00
11	[사과] [바나나] --->[배]	3	100,00	1,76	6,67
12	[사과] [바나나] --->[감]	3	100,00	1,43	6,67
13	[사과] [바나나] --->[귤]	3	50,00	0,79	3,33
14	[사과] [바나나] --->[대추]	3	50,00	1,07	3,33
15	[사과] [바나나] --->[파인애플]	3	50,00	1,36	3,33
16	[사과] [파인애플] --->[배]	3	66,67	1,18	6,67
17	[사과] [파인애플] --->[감]	3	100,00	1,43	10,00

(1) 연관규칙

[A->B] 연관규칙은 "상품 A 가 구매된 경우는 상품 B 도 구매된다." 라고 해석됩니다.

예시) "사과와 배를 선택한 사람은 감을 선택한다"라고 해석합니다.

(2) 아이템 수

연관성을 정의한 전체 아이템 수를 나타냅니다.

(3) 신뢰도(%)

신뢰도는 항목 A 를 포함하는 거래에서 항목 B 가 포함될 확률은 어느 정도인가를 나타내며 연관성의 정도를 파악할 수 있습니다.

$$\left(\frac{\text{A와 B를 동시에포함하는거래의비율} = \Pr(A \cap B)}{A\text{를 포함하는거래의비율} = \Pr(A)} \right)$$

예시) 사과와 배를 선택한 사람 중 감을 선택할 확률은 100% 입니다..

(4) 향상도(lift)

항목 A 를 구매한 경우 그 거래가 항목 B 를 포함하는 경우와 항목 B 가 임의로 구매되는 경우의 비를 나타냅니다.

$$\left(\frac{\Pr(A \cap B)}{\Pr(A) * \Pr(B)} \right)$$

예시) 사과와 배의 동시 선택과 감의 선택은 양의 연관관계가 있습니다.

(5) 지지율(%)

지지도와 동일한 의미입니다. 즉, 지지도는 A 와 B 를 포함하는 거래의 수를 출력한 것이고, 상대지지율은 그 거래의 수를 전체 거래수로 나눈 값입니다.

$$\left(\frac{\text{A와 B를 동시에 포함하는 거래의 수}}{\text{전체 거래수}} = \Pr(A \cap B) \right)$$

예시) 전체 과일 선택 건수 중 사과와 배를 동시에 선택한 사람이 감을 선택하는 비율은 20% 입니다.

고려사항

두 항목의 기본적인 구매율이 어느 정도 수준이 되어야만 의미가 있습니다. 즉, 지지도가 어느 정도 수준에 도달해야만 합니다.

신뢰도가 높을 경우에는 두 항목 A → B 에서 항목 B 의 확률이 커야지 이 연관성규칙에 의미가 있습니다. 원래 B 상품이 구매되는 기본확률보다 커야 A 를 고려해서 B 를 생각하는 것이 의미가 있습니다. 즉, 향상도(Lift)가 1 보다 큰 값을 주어야 유용한 정보를 준다고 볼 수 있습니다.

3.5.2 CART 모델 노드

CART 모델 노드는 CART 모델링 노드의 수행결과입니다. Training 데이터 셋으로 CART 모델링을 수행하고 모델링 결과를 Test 데이터 셋에 적용하여 **새로운 데이터의 예측이나 분류**에 사용될 수 있습니다.

아래는 모델 정보를 나타내고 있습니다.

예제데이터 (GermanCredit.ecl)

독일의 German Bank 의 대출을 신청한 고객 1000 명(우량고객 700 명, 불량고객 300 명)에 대한 데이터를 사용하여 분류 분석을 수행하였습니다. 변수는 "신용 우/불량 여부, 당좌구좌상태, 대출상환기간, 신용거래내역, 대출목적, 대출금액, 예금액, 근무년수, 소득 중 할부 거래 상환액 비율, 성별 및 혼인관계, 보증인 여부, 거주 기간, 부동산, 연령, 타 여신 개설계획, 주거 형태, 현재 당행 신용계좌수, 직업, 부양가족수, 전화소유유무, 외국인 근로자 여부"로 이루어졌습니다. 이 분석을 통해 신용 불량자에 대한 규칙을 발견할 수 있습니다.

	1	2	3	4	5	6	7	8	9	10	11	12	13
	OBS#	CHK_ACCT	DURATION	HISTORY	NEW_CAR	USED_CAR	FURNITURE	RADIO/TV	EDUCATION	RETRAINING	AMOUNT	SAV_ACCT	EMPLOYMENT
1	0,15385	0,84615	6	4	0	0	0	1	0	0	1,169	4	4
2	0,62245	0,37755	48	2	0	0	0	1	0	0	5,951	0	2
3	0,13129	0,86871	12	4	0	0	0	0	1	0	2,096	0	3
4	0,62245	0,37755	42	2	0	0	1	0	0	0	7,882	0	3
5	0,62245	0,37755	24	3	1	0	0	0	0	0	4,870	0	2
6	0,13129	0,86871	36	2	0	0	0	0	1	0	9,055	4	2
7	0,13129	0,86871	24	2	0	0	1	0	0	0	2,835	2	4
8	0,62245	0,37755	36	2	0	1	0	0	0	0	6,948	0	2
9	0,13129	0,86871	12	2	0	0	0	1	0	0	3,059	3	3
10	0,62245	0,37755	30	4	1	0	0	0	0	0	5,234	0	0
11	0,34715	0,65285	12	2	1	0	0	0	0	0	1,295	0	1
12	0,62245	0,37755	48	2	0	0	0	0	0	1	4,308	0	1
13	0,34715	0,65285	12	2	0	0	0	1	0	0	1,567	0	2
14	0,62245	0,37755	24	4	1	0	0	0	0	0	1,199	0	4
15	0,34715	0,65285	15	2	1	0	0	0	0	0	1,403	0	2
16	0,62245	0,37755	24	2	0	0	0	1	0	0	1,282	1	2
17	0,13129	0,86871	24	4	0	0	0	1	0	0	2,424	4	4
18	0,29268	0,70732	30	0	0	0	0	0	0	1	8,072	4	1
19	0,62245	0,37755	24	2	0	1	0	0	0	0	12,579	0	4
20	0,13129	0,86871	24	2	0	0	0	1	0	0	3,430	2	4

옵션 정보

CART 수행 시 옵션을 보여줍니다.

모델 정보

CART 수행으로 형성된 모델의 정보를 보여줍니다.

오분류 정보

분류 분석의 경우에만 해당됩니다. 원래 클래스와 예측 클래스에 대한 교차표를 통하여 예측이 잘못된 관측치 수와 백분율을 알 수 있습니다. 불량고객을 불량고객으로 예측할 확률은 59.33%, 우량고객을 우량고객으로 예측할 확률은 85.71%, 불량고객을 우량고객으로 예측할 확률은 40.67%, 우량고객을 불량고객으로 예측할 확률은 14.29% 입니다. 오분류 수 및 오분류율이 낮을수록 좋은 모형이라 판단할 수 있습니다.

● 모델링용 데이터 (Training Set)

	[예측] 1	[예측] 2	[예측] 3
1	57 (85,07 %)	4 (5,97 %)	6 (8,96 %)
2	2 (3,08 %)	61 (93,85 %)	2 (3,08 %)
3	0 (0,00 %)	6 (13,33 %)	39 (86,67 %)

오분류 수: 20
오분류율: 11.30%

ANOVA 테이블

예측 분석의 경우에만 해당됩니다. 독립변수가 종속변수를 설명하는데 있어서 통계적으로 유의한지 검정하기 위해 F 검정을 합니다. p-값이 유의수준보다 커서 유의하지 않다면 회귀계수에 대한 해석은 의미가 없습니다.

▶ 아노바 테이블

	DF	SS	MS	F	p
모형(회귀)	8	2601037457,21521	325129682,1519	0,12352	0,99832
잔차	8521	31204042,80088	3662,01652		
합계	8529	2632241500,01608			

표준에러정보

예측분석의 경우에만 해당됩니다.

▶ 표준에러 정보

R - square	0,98815
RMSE	60,48266
MAE	6,287
MAPE	310,59524

(1) R-square(결정계수)

도출된 회귀식이 측정값들을 대표할만한가를 확인하는 것으로, 즉, 추정된 회귀선이 관측치들을 얼마나 잘 설명하는지를 검정하는 값입니다. 1 에 가까울수록 회귀식이 데이터를 잘 설명하고 있음을 의미합니다.

(2) RMSE (Root-Mean Square Error)

$$RMSE = \sqrt{MSE} = \sqrt{SSE/(n-p-1)}$$

(3) MAE (Mean Absolute Error)

$$MAE = \frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{n}$$

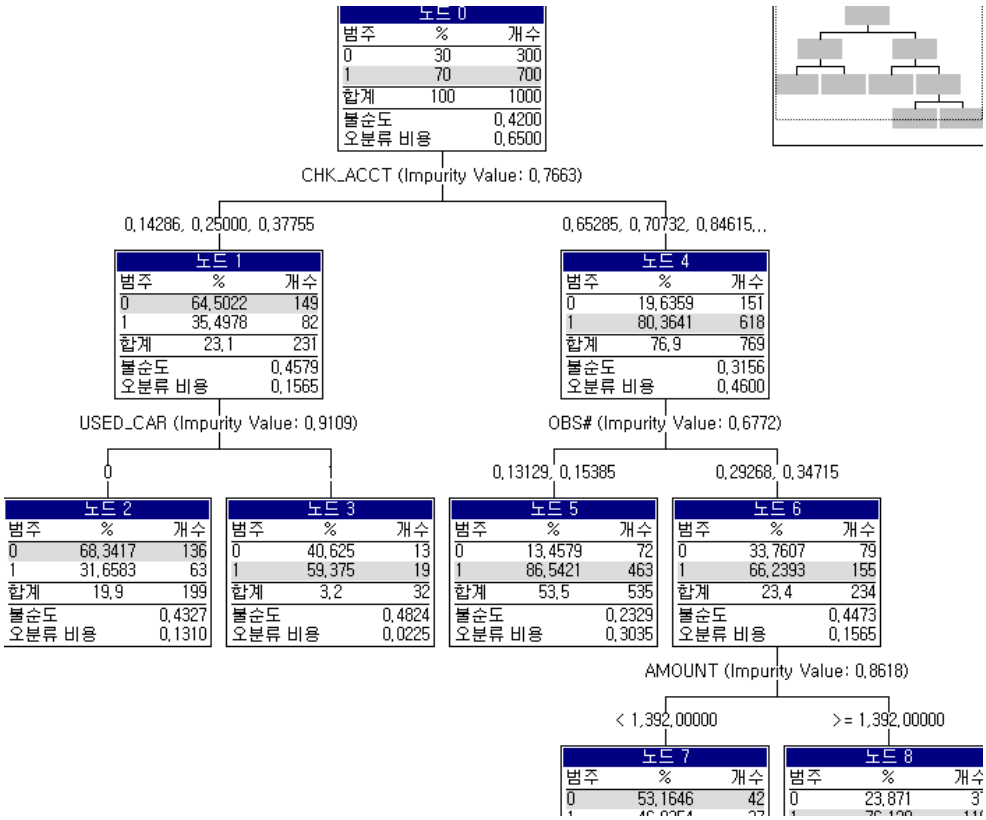
(4) MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{\sum_{i=1}^{n} | (y_i - \hat{y}_i)/y_i |}{n} \times 100$$

(2) ~ (4)번 모두 종속 변수의 원래값과 예측값의 오차를 나타낸 것으로 값이 작을 수록 좋은 모형이라 할 수 있습니다.

Tree

분석과정을 도식화하여 보여줍니다. 아래는 분류 분석 결과입니다. 아래의 트리구조를 통해 우량고객과 불량고객을 분류할 수 있는 몇 가지 룰을 찾을 수 있습니다. 예를 들어 첫 번째 가지는 CHK_ACCT 값이 0.14286, 0.25, 0.37755 중 하나이면서 USED_CAR(0)가 없는 그룹은 우량고객(0)이 136 명, 불량고객(1)이 63 명인 그룹으로 이 룰에 속하는 고객을 우량고객 그룹으로 분류할 수 있습니다(불순도가 43%인 우량고객 그룹). 높은 정확도를 가지는 룰을 얻기 위해서는 불순도 및 오분류 비용이 적은 룰을 만들어야 합니다.



3.5.3 HIERARCHICAL 모델 노드

HIERARCHICAL 모델 노드는 HIERARCHICAL 노드의 분석 결과로서 각 데이터가 할당된 군집에 대한 정보를 가지고 있습니다.

예제데이터 (Mileage.csv)

미국에서 판매되는 자동차의 특성에 관한 자료로, 제조사/차량명(MAKE), 차량크기(VOL), 마력(HP), 연비(MPG), 최고속도(sp), 무게(wt)에 대한 자료를 활용하여 분석을 수행하였습니다.

	1	2	3	4	5	6
	MAKE	VOL	HP	MPG	sp	wt
1	GM/GeoMetr	89	49	65,40000	96	17,50000
2	GM/GeoMetr	92	55	56,00000	97	20,00000
3	GM/GeoMetr	92	55	55,90000	97	20,00000
4	SuzukiSwift	92	70	49,00000	105	20,00000
5	DaihatsuCha	92	53	46,50000	96	20,00000
6	GM/GeoSprii	89	70	46,20000	105	20,00000
7	GM/GeoSprii	92	55	45,40000	97	20,00000
8	HondaCivicC	50	62	59,20000	98	22,50000
9	HondaCivicC	50	62	53,30000	98	22,50000
10	DaihatsuCha	94	80	43,40000	107	22,50000
11	SubaruJustw	89	73	41,10000	103	22,50000
12	HondaCivicC	50	92	40,90000	113	22,50000
13	HondaCivic	99	92	40,90000	113	22,50000
14	SubaruJustw	89	73	40,40000	103	22,50000
15	SubaruJustw	89	66	39,60000	100	22,50000
16	SubaruJustw	89	73	39,30000	103	22,50000
17	ToyotaTerce	91	78	38,90000	106	22,50000
18	HondaCivicC	50	92	38,80000	113	22,50000
19	ToyotaTerce	91	78	38,20000	106	22,50000
20	FordEscort	103	90	42,20000	109	25,00000
21	HondaCivic	99	92	40,90000	110	25,00000

옵션 정보

군집분석 수행 시 사용된 옵션에 대해 파악할 수 있습니다.

거리 계산법: Euclidean
연결방법: 단일 연결법 (Single Linkage)
변수 보정법: 없음
최종 군집수: 5 개

군집 정보

각 군집에 할당된 데이터 개수 및 군집 내 데이터간 거리정보를 제공합니다.

	개수	제곱합	평균 거리	최소 거리	최대 거리
군집1	1	0	0	0	0
군집2	74	119819,91784	33,60946	4,82228	82,77879
군집3	3	2064,42000	24,10785	10,19199	34,60733
군집4	1	0	0	0	0
군집5	3	419,59333	10,85085	4,58609	14,58505

군집 내 원소 개수

각 군집에 속하는 원소의 빈도수 및 백분율을 제공합니다.

Value	빈도수	Graph	백분율
군집1	1		1,22%
군집2	74		90,24%
군집3	3		3,66%
군집4	1		1,22%
군집5	3		3,66%
Total Count	82		

군집의 중심

각 군집에서 변수들의 군집 중심값을 제공합니다.

변수	군집1	군집2	군집3	군집4	군집5	전체 중심
MAKE	BuickReatta	ToyotaCamry	JaguarXJSConvert	BMW750IL	LexusLS400	ToyotaCamry
VOL	50	100,66216	50	119	111,33333	98,80488
HP	165	102,17568	288,33333	295	239,66667	117,13415
MPG	23,60000	35,37027	19,50000	16,70000	17,96667	33,78171
sp	122	108,77027	157,33333	157	139,33333	112,41463
wt	40	29,45946	43,33333	45	46,66667	30,91463

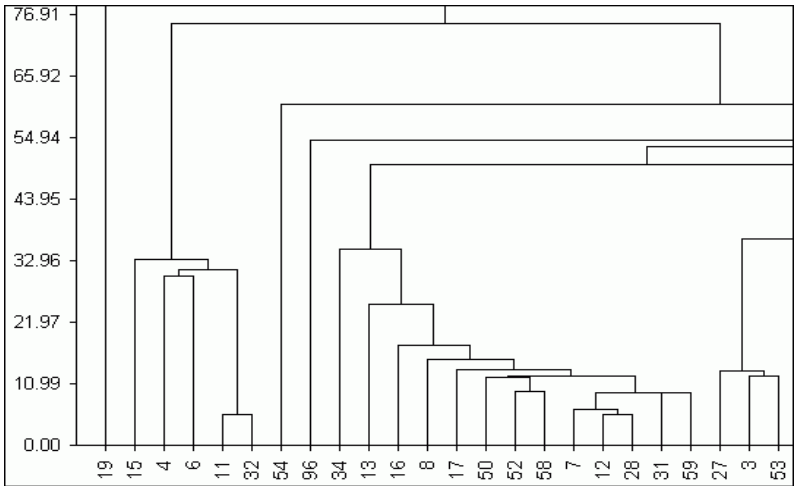
군집간 거리

군집 간의 거리정보를 제공합니다.

	군집2	군집3	군집4	군집5
군집1	83,30446	128,40908	151,52429	98,56074
군집2		200,06365	201,08482	143,35950
군집3			69,40586	80,42953
군집4				58,63679

덴드로그램 결과

덴드로그램은 군집화 과정을 요약하는 나무형태의 도표입니다. 도표의 하단에는 레코드가 표시됩니다. 유사한 레코트들은 수직선에 의해 연결이 되어 있으며, 수직선의 길이는 레코드들 사이의 거리를 반영합니다. 이를 통해 각 관측치들이 묶이는 과정을 한눈에 볼 수 있습니다. 연결 방법에 따라 다른 형태의 덴드로그램이 도출됩니다.



3.5.4 KMEANS 모델 노드

KMEANS 모델 노드는 KMEANS 노드의 분석결과로 각 데이터가 할당된 군집에 대한 정보를 가지고 있습니다.

예제데이터 (Mileage.csv)

미국에서 판매되는 자동차의 특성에 관한 자료로, 제조사/차량명(MAKE), 차량크기(VOL), 마력(HP), 연비(MPG), 최고속도(sp), 무게(wt)에 대한 자료를 활용하여 분석을 수행하였습니다.

	1	2	3	4	5	6
	MAKE	VOL	HP	MPG	sp	wt
1	GM/GeoMetr	89	49	65,40000	96	17,50000
2	GM/GeoMetr	92	55	56,00000	97	20,00000
3	GM/GeoMetr	92	55	55,90000	97	20,00000
4	SuzukiSwift	92	70	49,00000	105	20,00000
5	DaihatsuCha	92	53	46,50000	96	20,00000
6	GM/GeoSpri	89	70	46,20000	105	20,00000
7	GM/GeoSpri	92	55	45,40000	97	20,00000
8	HondaCivicC	50	62	59,20000	98	22,50000
9	HondaCivicC	50	62	53,30000	98	22,50000
10	DaihatsuCha	94	80	43,40000	107	22,50000
11	SubaruJustw	89	73	41,10000	103	22,50000
12	HondaCivicC	50	92	40,90000	113	22,50000
13	HondaCivic	99	92	40,90000	113	22,50000
14	SubaruJustw	89	73	40,40000	103	22,50000
15	SubaruJustw	89	66	39,60000	100	22,50000
16	SubaruJustw	89	73	39,30000	103	22,50000
17	ToyotaTerce	91	78	38,90000	106	22,50000
18	HondaCivicC	50	92	38,80000	113	22,50000
19	ToyotaTerce	91	78	38,20000	106	22,50000
20	FordEscort	103	90	42,20000	109	25,00000
21	HondaCivic	99	92	40,90000	110	25,00000

군집분석정보

몇 개의 군집으로 나뉘었는지에 대한 **거리 계산법**, **군집수**, **각 중심으로부터의 총 거리**, 수렴되기까지의 **전체 시행수**, **전처리 방법을** 나타냅니다.

거리 계산법: Euclidean

군집수 : 3

각 중심으로부터의 총거리 : 19.47

시행수 : 13

전처리 방법 : 정규화

군집 내 원소 개수

생성된 각 군집에 할당된 관측치의 빈도 및 백분율을 나타냅니다.

Value	빈도수	Graph	백분율
군집1	20		24.39%
군집2	40		48.78%
군집3	22		26.83%
Total Count	82		

군집중심정보

군집분석 결과로 생성된 각 군집의 중심좌표로 각 변수들의 값들이 나타납니다.

변수	군집1	군집2	군집3
VOL	81.20000	101.60000	109.72727
HP	69	99.80000	192.40909
MPG	46.14000	33.80500	22.50455
sp	102.15000	108.35000	129.13636
wt	21.87500	29.56250	41.59091

군집간거리

생성된 군집의 각 군집간 거리를 나타냅니다. 이는 각 군집중심 사이의 거리를 측정한 것입니다.

거리	군집2	군집3
군집1	40.18072	133.11358
군집2		96.67964

3.5.5 KNN 모델 노드


KNN

KNN 모델 노드는 KNN 모델링의 결과로 생성되며 모델링 시 입력된 데이터를 저장하고 있습니다. 새로운 데이터가 입력되면 저장되어 있는 데이터와의 거리를 분류하게 됩니다. 저장된 데이터의 클래스가 분류 기준이 됩니다.

예제(학습)데이터 (고객이탈_model.txt)

어느 회사의 고객자료를 바탕으로 고객 이탈 여부를 판단하고자 고객이탈모형을 구축하고자 KNN 모델 노드를 이용하여 분석을 수행하였습니다. 단 KNN 은 독립변수가 모두

연속형이어야 하므로 이산형 변수인 A2, A3, A4, A9 과 테스트 데이터에 없는 A7 을 제거합니다.

	1	2	3	4	5	6	7	8	9	10
	A1	A2	A3	A4	A5	A6	A7	A8	A9	이탈여부
1	24	B	M	B	181,39500	42,00000	213	921,00000	A	0
2	12	B	M	A	218,38000	191,40000	554	1,082,10000	A	1
3	23	B	M	A	168,69900	37,80000	33	909,00000	A	1
4	22	B	M	G	166,95200	128,40000	215	990,00000	A	0
5	56	B	M	H	137,70700	102,00000	421	768,60000	A	1
6	30	B	M	D	131,15100	36,50000	515	947,00000	A	1
7	43	B	M	G	210,27900	93,50000	355	1,067,60000	A	1
8	19	B	M	G	181,52000	70,80000	332	881,40000	A	0
9	19	B	M	A	177,08500	0,00000	144	750,60000	A	1
10	24	B	M	B	80,23190	81,60000	161	913,20000	A	1
11	29	B	M	H	199,45000	78,00000	478	1,069,50000	A	1
12	33	B	MH	H	205,54300	22,80000	406	1,525,50000	A	1
13	38	B	M	G	182,79500	24,00000	485	983,10000	A	1
14	33	B	M	A	156,26800	28,80000	202	913,20000	A	1
15	24	B	M	K	100,79200	77,50000	527	890,80000	A	1
16	43	B	M	B	99,35700	1,80000	531	867,00000	A	0
17	18	B	MH	F	96,22250	0,00000	679	1,455,60000	A	1

예제(테스트)데이터 (고객이탈_testset1.txt)

학습데이터와 마찬가지로 이산형 변수인 A2, A3, A4, A8(=A9)을 제거합니다.

	1	2	3	4	5	6	7	8
	A1	A2	A3	A4	A5	A6	A7	A8
1	24	B	M	B	181,39500	42,00000	921,00000	A
2	12	B	M	A	218,38000	191,40000	1,082,10000	A
3	23	B	M	A	168,69900	37,80000	909,00000	A
4	22	B	M	G	166,95200	128,40000	990,00000	A
5	56	B	M	H	137,70700	102,00000	768,60000	A
6	30	B	M	D	131,15100	36,50000	947,00000	A
7	43	B	M	G	210,27900	93,50000	1,067,60000	A
8	19	B	M	G	181,52000	70,80000	881,40000	A
9	19	B	M	A	177,08500	0,00000	750,60000	A
10	24	B	M	B	80,23190	81,60000	913,20000	A
11	29	B	M	H	199,45000	78,00000	1,069,50000	A
12	33	B	MH	H	205,54300	22,80000	1,525,50000	A
13	38	B	M	G	182,79500	24,00000	983,10000	A
14	33	B	M	A	156,26800	28,80000	913,20000	A
15	24	B	M	K	100,79200	77,50000	890,80000	A

KNN 실행정보

KNN 수행 시 **클래스** 및 **K 의 개수, 전처리 방법**에 대한 정보를 알려줍니다.

변수정렬표

전처리를 통해 변화된 표본데이터가 정렬되어 출력됩니다.

순번	A1	A5	A6	A8	이탈여부
1	24	181.39500	42	921	0
2	21	189.53000	0	1080.30000	0
3	24	259.37500	34.20000	1378.80000	0
4	22	166.95200	128.40000	990	0
5	23	169.88800	106.20000	864.90000	0
6	15	186.49300	76.20000	1157.10000	0
7	14	182.10900	18.50000	1024.10000	0
8	19	181.52000	70.80000	881.40000	0
9	23	159.75900	103.20000	790.50000	0
10	13	179.75800	86	876.20000	0
11	17	154.74100	0	930.30000	0
12	44	220.42200	40	1147	0

요약정보

KNN 노드의 속성창에서 Leave One Out 옵션을 예로 선택할 경우 학습 데이터의 오분류표(Leave One Out)를 확인할 수 있습니다. 이탈고객을 이탈로 예측할 확률은 49.12%, 이탈하지 않을 고객을 이탈하지 않은 고객으로 예측할 확률이 67.76%, 이탈고객을 이탈하지 않은 고객으로 예측할 확률이 50.88%, 이탈하지 않을 고객을 이탈할 고객으로 예측할 확률은 32.24%입니다. 또한 전체 오분류수는 3467, 오분류율은 40.64%입니다.

	[예측] 0	[예측] 1
0	1889 (49.12 %)	1957 (50.88 %)
1	1510 (32.24 %)	3174 (67.76 %)

오분류 수: 3467
오분류율: 40.64%

3.5.6 LDA 모델 노드

LDA 모델 노드는 LDA 분류분석의 결과로 생성됩니다 Training 데이터 셋으로 LDA 분석을 수행하고 분석 결과인 LDA 모델 노드를 Test 데이터 셋(새로운 데이터 셋)에 적용하여 **새로운 데이터의 분류**에 사용될 수 있습니다.

예제데이터 (고객이탈_model.txt)

어느 회사의 고객자료를 바탕으로 고객 이탈 여부를 판단하고자 고객이탈모형을 구축하고자 LDA 모델 노드를 이용하여 분석을 수행하였습니다.

	1	2	3	4	5	6	7	8	9	10
	A1	A2	A3	A4	A5	A6	A7	A8	A9	이탈여부
1	24	B	M	B	181,39500	42,00000	213	921,00000	A	0
2	12	B	M	A	218,38000	191,40000	554	1,082,10000	A	1
3	23	B	M	A	168,69900	37,80000	33	909,00000	A	1
4	22	B	M	G	166,95200	128,40000	215	990,00000	A	0
5	56	B	M	H	137,70700	102,00000	421	768,60000	A	1
6	30	B	M	D	131,15100	36,50000	515	947,00000	A	1
7	43	B	M	G	210,27900	93,50000	355	1,067,60000	A	1
8	19	B	M	G	181,52000	70,80000	332	881,40000	A	0
9	19	B	M	A	177,08500	0,00000	144	750,60000	A	1
10	24	B	M	B	80,23190	81,60000	161	913,20000	A	1
11	29	B	M	H	199,45000	78,00000	478	1,069,50000	A	1
12	33	B	MH	H	205,54300	22,80000	406	1,525,50000	A	1
13	38	B	M	G	182,79500	24,00000	485	983,10000	A	1
14	33	B	M	A	156,26800	28,80000	202	913,20000	A	1
15	24	B	M	K	100,79200	77,50000	527	890,80000	A	1
16	43	B	M	B	99,35700	1,80000	531	867,00000	A	0
17	18	B	MH	F	96,22250	0,00000	679	1,455,60000	A	1

클래스별 독립변수평균

분류된 각 클래스의 변수평균값을 나타냅니다.

계수	0	1
A1	28,67447	31,85696
A2 (A)	-0,12376	-0,24402
A2 (B)	-0,28549	-0,25833
A2 (C)	-0,30499	-0,33604
A2 (D)	-0,20333	-0,15777
A3 (H)	0,06448	0,06191
A3 (L)	0,04992	0,01430
A3 (M)	0,45164	0,60675
A3 (MH)	0,21451	0,22395
A3 (ML)	0,21321	0,09308
A4 (A)	-0,00364	-0,03309
A4 (B)	0,25663	-0,15307

클래스의 분산-공분산행렬

그룹별 분산-공분산이 같기 때문에 한 개의 분산-공분산 행렬이 출력됩니다.

	A1	A2 (A)	A2 (B)	A2 (C)	A2 (D)	A3 (H)	A3 (L)	A3 (M)	A3 (MH)	A3 (ML)	A4 (A)	A4 (B)	A4 (C)	A4 (D)	A4 (E)	A (F
A1	161,56222	-0,08495	-0,12999	-0,07786	0,03796	0,00894	-0,00292	-0,13532	0,07840	0,04929	-0,02881	-0,18096	0,03648	-0,02220	0,01061	0,02
A2 (A)	-0,08495	0,55502	0,34168	0,33018	0,35974	-0,05201	0,02147	0,07768	-0,16216	0,11185	-0,02668	-0,03187	0,02843	0,03486	-0,01659	0,01
A2 (B)	-0,12999	0,34168	0,44043	0,30529	0,34367	-0,04623	0,00869	0,12266	-0,12775	0,04183	-0,01924	-0,00756	-0,00290	0,02241	-0,01126	0,01
A2 (C)	-0,07786	0,33018	0,30529	0,35839	0,33515	-0,04314	0,02287	0,04899	-0,13284	0,10326	-0,03276	-0,03681	0,01223	0,00765	0,00249	0,00
A2 (D)	0,03796	0,35974	0,34367	0,33515	0,57380	-0,05216	0,00582	0,18088	-0,16461	0,02949	-0,04225	-0,03969	0,01188	-0,00822	0,00194	0,01
A3 (H)	0,00894	-0,05201	-0,04623	-0,04314	-0,05216	0,06004	-0,00147	-0,03330	-0,01338	-0,00890	0,00768	0,00724	-0,00194	-0,00036	0,00008	0,00
A3 (L)	-0,00292	0,02147	0,00869	0,02287	0,00582	-0,00147	0,03007	-0,01447	-0,00612	-0,00506	0,00011	-0,00481	0,00325	0,00056	0,00728	0,00
A3	-0,13532	0,07768	0,12266	0,04899	0,18088	-0,03330	-0,01447	0,24368	-0,11786	-0,07398	-0,01848	0,00460	-0,02478	-	-	-

사전확률

점유도 옵션에서 지정한 클래스 별 확률값을 보여줍니다.

클래스	확률
0	0.50000
1	0.50000

그룹별 선형판별함수

각 class 에 대한 선형함수의 상수와 계수를 나타냅니다.

계수	0	1
상수	-115.01354	-119.84451
A1	0.18370	0.20162
A2 (A)	12.10622	10.31288
A2 (B)	-6.39559	-7.79261
A2 (C)	11.64822	14.21276
A2 (D)	-0.90353	-0.35487
A3 (H)	9.56061	9.92711
A3 (L)	76.18010	76.24068
A3 (M)	65.86759	66.18165
A3 (MH)	45.59824	45.94307

Confusion matrix for number & percentage

원래 클래스와 예측 클래스에 대한 교차표를 통하여 예측이 잘못된 관측치 수와 백분율을 알 수 있습니다. 이탈고객을 이탈로 예측할 확률은 77.41%, 이탈하지 않을 고객을 이탈하지 않은 고객으로 예측할 확률이 93.30%, 이탈고객을 이탈하지 않은 고객으로 예측할 확률이 22.59%, 이탈하지 않을 고객을 이탈할 고객으로 예측할 확률은 6.70%입니다. 또한 전체 오분류수는 1183, 오분류율은 13.87%입니다.

	[예측] 0	[예측] 1
0	2977 (77.41 %)	869 (22.59 %)
1	314 (6.70 %)	4370 (93.30 %)

오분류 수: 1183
오분류율: 13.87%

3.5.7 LOGISTIC 모델 노드

LOGISTIC 모델 노드는 로지스틱 분석의 결과로 생성됩니다 Training 데이터 셋으로 로지스틱 분석을 수행한 후 분석 결과인 로지스틱 모델 노드를 Test 데이터 셋(새로운 데이터 셋)에 적용하여 **새로운 데이터의 예측/분류**에 사용될 수 있습니다.

예제데이터 (고객이탈_model.txt)

어느 회사의 고객자료를 바탕으로 고객 이탈 여부를 판단하고자 고객이탈모형을 구축하고자 LOGISTIC 모델 노드를 이용하여 분석을 수행하였습니다.

	1	2	3	4	5	6	7	8	9	10
	A1	A2	A3	A4	A5	A6	A7	A8	A9	이탈여부
1	24	B	M	B	181,39500	42,00000	213	921,00000	A	0
2	12	B	M	A	218,38000	191,40000	554	1,082,10000	A	1
3	23	B	M	A	168,69900	37,80000	33	909,00000	A	1
4	22	B	M	G	166,95200	128,40000	215	990,00000	A	0
5	56	B	M	H	137,70700	102,00000	421	768,60000	A	1
6	30	B	M	D	131,15100	36,50000	515	947,00000	A	1
7	43	B	M	G	210,27900	93,50000	355	1,067,60000	A	1
8	19	B	M	G	181,52000	70,80000	332	881,40000	A	0
9	19	B	M	A	177,08500	0,00000	144	750,60000	A	1
10	24	B	M	B	80,23190	81,60000	161	913,20000	A	1
11	29	B	M	H	199,45000	78,00000	478	1,069,50000	A	1
12	33	B	MH	H	205,54300	22,80000	406	1,525,50000	A	1
13	38	B	M	G	182,79500	24,00000	485	983,10000	A	1
14	33	B	M	A	156,26800	28,80000	202	913,20000	A	1
15	24	B	M	K	100,79200	77,50000	527	890,80000	A	1
16	43	B	M	B	99,35700	1,80000	531	867,00000	A	0
17	18	B	MH	F	96,22250	0,00000	679	1,455,60000	A	1

클래스 수와 총 시행 수

종속변수의 클래스 수와 알고리즘이 실행된 총 시행 수를 알 수 있습니다.

Y Value

종속변수 Y 의 클래스 별 빈도 및 비율을 보여줍니다.

Value	빈도수	Graph	백분율
1	6		50,00%
2	6		50,00%
Total Count	12		

로지스틱 테이블

Parameter Estimate 는 각 변수의 계수값을 나타내며, **Standard Error** 는 각 계수(Beta)의 standard error 입니다. 또한 각 변수 유의성 검정의 **z-value, p-값을** 보여줍니다. p-값이

유의수준보다 작을 때 유의한 변수라고 판단합니다. **Odds Ratio** 는 입력변수가 분류결정에 미치는 영향의 정도를 나타내는 값입니다. 오즈비가 1 보다 작다는 것은 입력변수 xi 가 (-)의 영향을 준다는 것이고, 오즈비가 1 보다 크다는 것은 (+)영향을 줌을 의미합니다.

Predictor	Parameter Estimate	Standard Error	Z	p	Odds 비
FEILD1	-0,00512	0,00247	-2,07505	0,03871	0,99489
FEILD2	0,05182	0,06825	0,75926	0,44774	1,05318
Constant	0,72822	0,16906	4,3074	0,00002	

※ 값이 1.#IO, 1#QO 또는 1.#INF 와 같이 나타난 경우 값이 너무 크거나 작은 경우입니다.

오분류 정보

종속변수의 원래값과 Logistic Regression 을 통해 얻어진 예측값 간의 빈도, 퍼센트 교차표를 보여줍니다. 오분류 정보를 통해 각 범주에 대한 분류정확도를 판단할 수 있습니다. 이탈고객을 이탈로 예측할 확률은 79.04%, 이탈하지 않을 고객을 이탈하지 않은 고객으로 예측할 확률이 93.04%, 이탈고객을 이탈하지 않은 고객으로 예측할 확률이 20.96%, 이탈하지 않을 고객을 이탈할 고객으로 예측할 확률은 6.96%입니다. 또한 전체 오분류수는 1132, 오분류율은 13.27%입니다.

	[예측] 0	[예측] 1
0	3040 (79,04 %)	806 (20,96 %)
1	326 (6,96 %)	4358 (93,04 %)

오분류 수: 1132
오분류율: 13.27%

분리 기준점: 0.500000

3.5.8 MANUAL CART 모델 노드

MANUAL CART 모델 노드는 Manual CART 노드 분석수행의 결과로 생성됩니다.

예제데이터 (GermanCredit.ecl)

독일의 German Bank 의 대출을 신청한 고객 1000 명(우량고객 700 명, 불량고객 300 명)에 대한 데이터를 사용하여 분류 분석을 수행하였습니다. 변수는 "신용 우/불량 여부, 당좌구좌상태, 대출상환기간, 신용거래내역, 대출목적, 대출금액, 예금액, 근무년수, 소득 중

할부 거래 상환액 비율, 성별 및 혼인관계, 보증인 여부, 거주 기간, 부동산, 연령, 타 여신 개설계획, 주거 형태, 현재 당행 신용계좌수, 직업, 부양가족수, 전화소유유무, 외국인 근로자 여부"로 이루어졌습니다. 이 분석을 통해 신용 불량자에 대한 규칙을 발견할 수 있습니다.

	1	2	3	4	5	6	7	8	9	10	11	12	13
	OBS#	CHK_ACCT	DURATION	HISTORY	NEW_CAR	USED_CAR	FURNITURE	RADIO/TV	EDUCATION	RETRAINING	AMOUNT	SAV_ACCT	EMPLOYMEN
1	0,15385	0,84615	6	4	0	0	0	1	0	0	1,169	4	4
2	0,62245	0,37755	48	2	0	0	0	1	0	0	5,951	0	2
3	0,13129	0,86871	12	4	0	0	0	0	1	0	2,096	0	3
4	0,62245	0,37755	42	2	0	0	1	0	0	0	7,882	0	3
5	0,62245	0,37755	24	3	1	0	0	0	0	0	4,870	0	2
6	0,13129	0,86871	36	2	0	0	0	0	1	0	9,055	4	2
7	0,13129	0,86871	24	2	0	0	1	0	0	0	2,835	2	4
8	0,62245	0,37755	36	2	0	1	0	0	0	0	6,948	0	2
9	0,13129	0,86871	12	2	0	0	0	1	0	0	3,059	3	3
10	0,62245	0,37755	30	4	1	0	0	0	0	0	5,234	0	0
11	0,34715	0,65285	12	2	1	0	0	0	0	0	1,295	0	1
12	0,62245	0,37755	48	2	0	0	0	0	0	1	4,308	0	1
13	0,34715	0,65285	12	2	0	0	0	1	0	0	1,567	0	2
14	0,62245	0,37755	24	4	1	0	0	0	0	0	1,199	0	4
15	0,34715	0,65285	15	2	1	0	0	0	0	0	1,403	0	2
16	0,62245	0,37755	24	2	0	0	0	1	0	0	1,282	1	2
17	0,13129	0,86871	24	4	0	0	0	1	0	0	2,424	4	4
18	0,29268	0,70732	30	0	0	0	0	0	0	1	8,072	4	1
19	0,62245	0,37755	24	2	0	1	0	0	0	0	12,579	0	4
20	0,13129	0,86871	24	2	0	0	0	1	0	0	3,430	2	4

오분류표

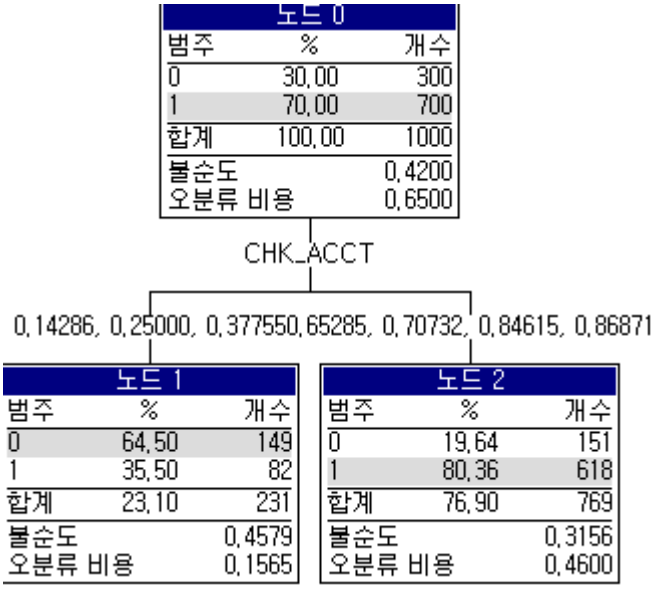
● 모델링용 데이터 (Training Set)

	[예측] 0	[예측] 1
0	2150 (55,90 %)	1696 (44,10 %)
1	125 (2,67 %)	4559 (97,33 %)

오분류 수: 1821
오분류율: 21.35%

Decision Tree

아래 트리구조를 통해 얻을 수 있는 룰은 총 2 가지 이며, CHK_ACCT 의 변수값으로 그룹이 결정됩니다. CHK_ACCT 라는 변수가 0.14286, 0.25, 0.37755, 0.65285 일 경우 신용평가가 양호하게 평가된 그룹으로 나누어지며 하나의 룰로 사용할 수 있습니다. CHK_ACCT 가 0.70732, 0.84615, 0.86871 일 때 신용평가가 불량으로 나타난 그룹으로 나타나며 이 경우 역시 하나의 룰로 사용될 수 있습니다.



3.5.9 MLP 모델 노드

MLP 모델 노드는 MLP 분류분석의 결과로 생성됩니다 Training 데이터 셋으로 MLP 분석을 수행하고 분석 결과인 MLP 모델 노드를 Test 데이터 셋(새로운 데이터 셋)에 적용하여 **새로운 데이터의 분류/예측**에 사용될 수 있습니다.

옵션정보

MLP 를 이용해서 모델링 시 어떤 옵션을 사용했는지를 보여줍니다. 이것을 통해 옵션이 바뀌었을 때 결과가 어떻게 바뀌었는지 판단할 수 있습니다.

Weight 정보

Input Layer 에서 Hidden Layer #1 으로 들어가는 노드의 가중치와 Hidden Layer #1 에서 Output Layer 로 들어가는 노드의 가중치 정보가 나타나있습니다.

ANOVA 테이블

회귀분석의 경우에만 해당됩니다. 독립변수가 종속변수 y 를 설명하는데 있어서 통계적으로 유의한지 검정하기 위해 F 검정을 합니다. p-값이 유의수준보다 커서 유의하지 않다면 회귀계수에 대한 해석은 의미가 없습니다.

▶ 아노바 테이블

	DF	SS	MS	F	p
모형(회귀)	4	476,91747	119,22937	4339,50796	0
잔차	8525	234,2271	0,02748		
합계	8529	711,14457			

오분류정보

분류의 경우에만 해당됩니다. 원래 클래스와 예측 클래스에 대한 교차표를 통하여 예측이 잘못된 관측치 수와 백분율을 알 수 있습니다.

● 모델링용 데이터 (Training Set)

	[예측] 1	[예측] 2	[예측] 3
1	57 (85,07 %)	4 (5,97 %)	6 (8,96 %)
2	2 (3,08 %)	61 (93,85 %)	2 (3,08 %)
3	0 (0,00 %)	6 (13,33 %)	39 (86,67 %)

오분류 수: 20
오분류율: 11,30%

표준에러정보

회귀분석의 경우에만 해당되며 모델의 적합력 및 예측력을 나타냅니다.

▶ 표준에러 정보

R - square	0,67063
RMSE	0,16576
MAE	0,13834
MAPE	139,36989

(1) R-square(결정계수)

도출된 회귀식이 측정값들을 대표할만한가를 확인하는 것으로, 즉, 추정된 회귀선이 관측치들을 얼마나 잘 설명하는지를 검정하는 값입니다. 1 에 가까울수록 회귀식이 데이터를 잘 설명하고 있음을 의미합니다.

(2) RMSE (Root-Mean Square Error)

$$RMSE = \sqrt{MSE} = \sqrt{SSE/(n-p-1)}$$

(3) MAE (Mean Absolute Error)

$$MAE = \frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{n}$$

(4) MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{\sum_{i=1}^{n} | (y_i - \hat{y}_i) / y_i |}{n} \times 100$$

(2) ~ (4)번 모두 종속 변수의 원래값과 예측값의 오차를 나타낸 것으로 값이 작을수록 좋은 모형이라 할 수 있습니다.

3.5.10 MLR 모델 노드

MLR 모델 노드는 MLR 분석 결과로 생성됩니다. Training 데이터 셋으로 MLR 분석을 수행한 후 분석 결과인 MLR 모델 노드를 Test 데이터 셋(새로운 데이터 셋)에 적용하여 **새로운 데이터의 예측**에 사용될 수 있습니다.

예제데이터 (밀가루.ecl)

식품 연구실에서 밀가루의 밀단백질함유율(X1)과 끈적거림의 정도(X2)가 수분흡수율(Y)에 미치는 영향을 조사 자료입니다. MLR 모델을 이용하여 밀단백질함유율(X1)과 끈적거림의 정도(X2)가 수분흡수율에 미치는 영향에 대한 수식(모델)을 만들 수 있습니다.

	1	2	3
	x1	x2	y
1	8,50000	2	30,90000
2	9,80000	22	40,90000
3	12,50000	31	47,20000
4	11,30000	30	46,80000
5	12,90000	28	45,90000
6	11,60000	35	49,20000
7	11,50000	45	49,60000
8	8,90000	3	32,70000
9	10,80000	20	42,90000
10	10,90000	28	44,00000
11	13,00000	27	46,20000
12	13,10000	28	48,80000

회귀식

독립변수들에 의해 생성된 종속변수 y 를 예측하기 위한 회귀식으로 MLR 수행의 최종결과입니다.

회귀계수 추정 $B=(X'X)^{-1}X'Y$

종속변수 예측 $\hat{Y}=XB$

회귀테이블

각 독립변수의 유의성 및 회귀 계수값을 파악하기 위한 테이블입니다.

y = 19.43976 + 1.44228 x1 + 0.33563 x2

● 회귀테이블

Variable	Parameter Estimate	Standard Error	t - value	p - value
상수	19,43976	2,18829	8,88355	0,00000
x1	1,44228	0,20764	6,94619	0,00000
x2	0,33563	0,01814	18,50673	0,00000

※ 값이 1.#IO, 1#QO 또는 1.#INF와 같이 나타난 경우 값이 너무 크거나 작은 경우 입니다.
※ 확률(p) 값이 음수가 나오는 경우, 입력 파라미터 값이 의미없음을 나타냅니다.

(1) Parameter Estimate

각 독립변수의 모수 추정치(회귀계수)입니다.

(2) Standard Error

모수 추정치(회귀계수)의 표준오차입니다.

(3) t-value

'특정변수가 종속변수에 미치는 영향력이 없습니다.' 라는 귀무 가설을 검증하기 위한 각 변수의 t 값을 의미합니다.

(4) p-value

각 추정치를 위한 확률로 p-값이 유의수준보다 크면 귀무가설 '특정변수가 종속변수에 미치는 영향력이 없습니다.' 을 기각할 수 없게 됩니다. 유의수준보다 작으면 귀무가설을 기각하게 됩니다.

ANOVA 테이블

독립변수가 종속변수 y 를 설명하는데 있어서 통계적으로 유의한지 검정하기 위해 F 검정을 합니다. p-값이 유의수준보다 커서 유의하지 않다면 회귀계수에 대한 해석은 의미가 없습니다.

	DF	SS	MS	F	p
모형(회귀)	2	812,37959	406,18980	339,31092	0
잔차	25	29,92755	1,19710		
합계	27	842,30714			

(1) DF(Degree of Freedom)

자유도 입니다.

(2) SS(Sum of Squares)

제곱합 입니다.

(3) MS(Mean Sqare)

제곱합을 자유도로 나눈 값으로 평균제곱입니다.

(4) F

'모든 설명변수들과 반응변수가 서로 관련성이 없다.'라는 귀무가설을 검증하기 위한 F 값을 의미합니다.

(5) p

p--값이 유의수준보다 크면 귀무가설(모든 설명변수들과 반응변수가 서로 관련성이 없다.)을 기각할 수 없게 됩니다. 즉, 적어도 하나의 설명변수는 종속변수에 통계적으로 유의한 영향을 준다고 해석할 수 있습니다. 반대로 p-값이 유의수준보다 작으면 귀무가설을 기각하게 되고 회귀식이 유효하지 않다는 의미입니다.

표준에러정보

Adjust R2	0,96163
R - square	0,96447
RMSE	1,09412
MAE	0,87393
MAPE	1,96048

(1) R-square(결정계수)

도출된 회귀식이 측정값들을 대표할만한가를 확인하는 것으로, 즉, 추정된 회귀선이 관측치들을 얼마나 잘 설명하는지를 검정하는 값입니다. 1 에 가까울수록 회귀식이 데이터를 잘 설명하고 있음을 의미합니다.

(2) Adjust R2

투입되는 독립변수들이 많을수록 종속변수에 대한 설명력인 R-square 는 자연히 증가합니다. 그러므로 실제 적합도를 판정하는 데에는 자유도를 고려하여 조정된 결정계수 R-square(adj)를 많이 활용합니다. 1 에 가까울수록 회귀식이 데이터를 잘 설명하고 있음을 의미합니다.

(3) RMSE (Root-Mean Square Error)

$$RMSE = \sqrt{MSE} = \sqrt{SSE/(n-p-1)}$$

(4) MAE (Mean Absolute Error)

$$MAE = \frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{n}$$

(5) MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{\sum_{i=1}^{n} | (y_i - \hat{y}_i)/y_i |}{n} \times 100$$

(3) ~ (5)번 모두 종속 변수의 원래값과 예측값의 오차를 나타낸 것으로 값이 작을수록 좋은 모형이라 할 수 있습니다.

Trace and Statistical Table

STEP	x1	x2	RMES	R2	R2 - Adj
Step1	0	1	1,83646	0,89590	0,89189
Step2	1	1	1,09412	0,96447	0,96163

(1) 이 결과는 기법을 stepwise 로 선택했을 경우에만 출력됩니다. 각 단계별 모형에 대해 설명하고 있습니다. Trace 값은 각 단계에서 어떤 변수가 입력되었는지를 나타냅니다.

(2) Step1 에서 x2 만 Trace 값이 1 임을 알 수 있습니다. 이는 Step1 에서 x2 변수가 모형에 추가되었음을 의미합니다.

(3) Step2 에서는 x1 변수가 모형에 추가되었음을 알 수 있습니다. 이 x1 변수가 모형에 추가되었을 때 R2(R-Square)값은 Step1 보다 좋아졌음을 알 수 있습니다.

(4) Stepwise 를 선택하면 앞에서의 회귀테이블에 마지막 단계에 포함되었던 변수들만이 출력되어 위와 같은 경우는 변수 x1, x2 에 대한 회귀 테이블이 생성됩니다.

3.5.11 PCA 모델 노드

PCA 모델 노드는 PCA 분석 시 생성되는 노드로 원래 데이터가 주성분으로 축약되는 정보를 보여줍니다.

예제데이터 (조업편차분석.txt)

어떤 제조 공정 중에 발생한 데이터를 이용하여 어떤 원인에 의해 같은 공정과정임에도 불구하고 조업 편차가 발생시키는 원인에 대한 분석을 하고자 합니다. 공정과정 중에 발생하는 변수는 총 A1 ~ A54 로 54 개의 변수가 발생하고, 데이터 개수는 7,596 개입니다. 데이터가 다 변량이고 변수간 다수의 상관관계가 존재하기 때문에 모든 변수를 고려하여 분석하기에는 어려움이 따르므로, 주성분 분석을 이용하여 변수 차원 축소를 통해 군집화를 하고 이를 통해 어떻게 그룹화가 이루어 졌으며, 어떤 원인에 의해 편차가 이루어 졌는지 알아 보고자 합니다.

	1	2	3	4	5	6	7	8	9	10	11	12
	No.	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
1	1	110,06100	1,95173	115,15100	27,27380	72,20820	109,96700	0,55476	66,99480	75,06930	38,00220	25,4
2	2	109,95900	1,97996	115,48600	27,05320	72,20820	109,96700	0,55524	66,99330	75,05090	37,96510	25,7
3	3	110,01400	2,02079	115,31600	27,05140	72,20820	110,07900	0,55242	67,05090	75,06930	37,98350	25,6
4	4	110,01400	2,00120	115,33800	27,31790	72,20820	109,96700	0,55388	66,93840	75,05210	37,98350	25,7
5	5	109,99800	2,02360	115,19200	27,04110	72,20820	109,96700	0,55435	66,99420	75,14410	37,98350	25,7
6	6	109,95900	2,02063	115,62000	27,29520	72,20820	109,96700	0,55059	67,08870	75,05210	37,98350	25,4
7	7	110,06100	1,99856	115,26400	27,22800	72,22000	109,96700	0,55992	66,99470	75,05090	37,98350	25,6
8	8	109,95900	2,02636	115,48900	27,15290	72,27770	110,07900	0,55333	66,99420	74,95700	37,98350	25,6
9	9	109,99900	2,01522	115,46700	27,35910	72,25440	109,96700	0,55808	66,96790	74,95670	38,00220	25,4
10	10	109,99100	2,00764	115,39400	27,08530	72,23120	109,94900	0,55435	67,09010	75,14410	37,98380	25,6
11	11	110,04600	1,98164	115,31500	27,11020	72,23140	109,96700	0,55759	66,86390	75,04900	38,05830	25,6
12	12	109,99000	1,99189	114,94700	27,02340	72,23140	110,06000	0,55384	67,02280	74,93680	38,05830	25,4
13	13	110,01400	1,97797	114,98400	27,05420	72,23120	109,96700	0,55715	66,94860	74,71240	38,17050	25,9
14	14	109,93600	1,97222	115,30400	27,34290	72,21980	110,07900	0,55242	67,04990	74,20900	38,17050	25,7
15	15	110,03000	1,95748	115,33600	27,14210	72,21980	109,96700	0,55855	66,94800	73,79750	38,17050	25,6
16	16	110,04400	2,13262	115,15300	27,13500	72,23120	110,07900	0,55711	66,99650	73,79750	38,05830	25,8

PCA 수행정보

전체변수를 몇 개의 주성분으로 축약했는지 표시해주고 T^2 의 Limit 를 알 수 있습니다.

로드값

각 변수를 이용해 주성분을 생성하는 선형결합을 표시하고 있는 값으로 선형 결합 시 계수값을 의미합니다. 즉, 주성분 1 은 다음과 같이 생성됩니다.

주성분 1 = (-0.24360*A1)+(0.00132*A2)+(-0.24335*A3)+…

변수	주성분1	주성분2	주성분3	주성분4	주성분5
A1	-0,24360	0,09108	-0,15877	0,12632	-0,02986
A2	0,00132	0,00041	-0,02010	-0,01280	-0,05525
A3	-0,24335	0,05163	-0,05123	0,17902	0,07103
A4	0,02542	0,22263	0,05792	0,07172	0,16654
A5	-0,04617	0,09046	0,17638	0,21294	0,11003
A6	-0,24455	0,09214	-0,16095	0,12395	-0,03353
A7	-0,23241	-0,02547	-0,14758	0,14832	-0,05053
A8	0,11767	0,19428	-0,17512	0,02119	0,20550
A9	-0,01627	0,10795	-0,30890	-0,12819	-0,01167
A10	0,00410	-0,17923	0,04477	-0,03155	-0,34649
A11	-0,07708	0,15456	-0,07202	-0,11157	-0,05891
A12	-0,08748	0,25798	0,06999	-0,05389	0,09196
A13	0,00390	-0,00343	0,01250	0,00071	-0,00254

분산설명력

각 주성분이 전체데이터에 대한 정보를 어느 정도 가지고 있는지를 표시하고 있습니다. 즉, 각 주성분이 데이터 변동성을 설명하는데 얼마만큼의 기여를 했는지를 보여줍니다.(관심 주성분의 갯수는 PCA 노드의 속성창 중 주성분수에 해당함.)

● 분산 설명력 (관심 주성분)

	Eigen Value of Cov(X)	X기여율(%)	X누적기여율(%)
주성분1	10,87868	20,14571	20,14571
주성분2	7,49760	13,88444	34,03015
주성분3	5,21985	9,66640	43,69655
주성분4	4,42837	8,20068	51,89723
주성분5	3,20759	5,93997	57,83720

● 분산 설명력 (기타 주성분)

	Eigen Value of Cov(X)	X기여율(%)	X누적기여율(%)
주성분6	2,51237	4,65253	62,48973
주성분7	1,80294	3,33878	65,82851
주성분8	1,57459	2,91591	68,74442
주성분9	1,50619	2,78924	71,53365
주성분10	1,22665	2,27157	73,80523

.

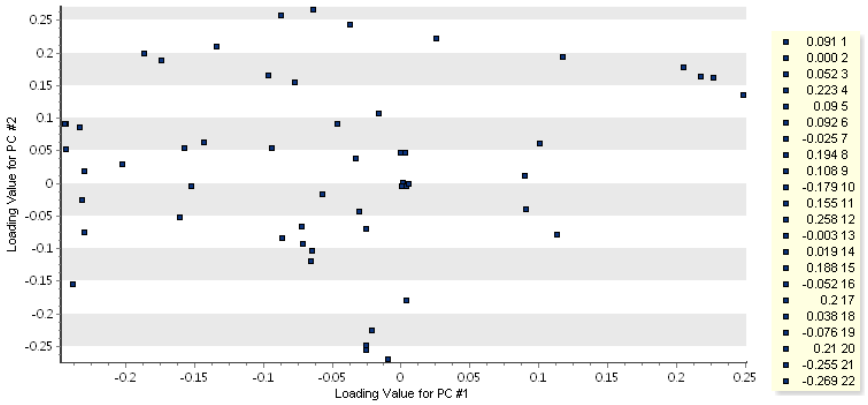
주성분별 Control Limit

주성분별 Control Limit 을 통해 관심 주성분의 95% Limit 및 99% Limit 를 제공합니다.

주성분 #	95% Limit	99% Limit
주성분1	6,46555	8,49796
주성분2	5,36758	7,05484
주성분3	4,47864	5,88648
주성분4	4,12514	5,42186
주성분5	3,51080	4,61440

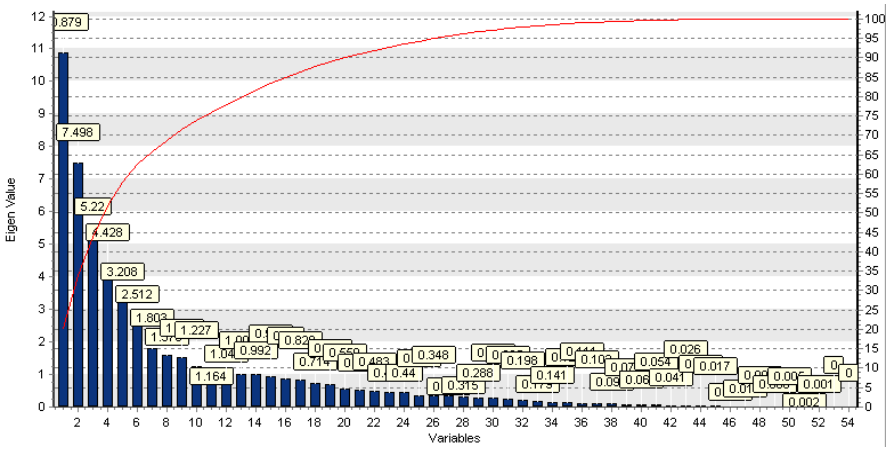
Loading Plot

Loading plot 은 선택된 두 개의 원 데이터 혹은 주성분의 loading 값에 대한 정보를 제공합니다.



Scree Plot

Scree plot 은 각 주성분에 대한 Eigenvalue 값을 표시한 것입니다. Scree plot 은 주성분의 수를 결정하는 하나의 방법으로 Eigenvalue 값의 변화가 급격히 이루어지는 수에서 주성분의 수를 결정할 수 있습니다.



3.5.12 PCR 모델 노드

PCR 모델 노드는 PCR 분석 결과로 생성됩니다. Training 데이터 셋으로 PCR 분석을 수행한 후 분석 결과인 PCR 모델 노드를 Test 데이터 셋(새로운 데이터 셋)에 적용하여 **새로운 데이터의 예측**에 사용될 수 있습니다.

예제데이터 (조업편차분석.txt)

공정 과정에서 발생하는 공정 변수들 중에 한 변수(A1)를 종속변수로 하여 나머지 변수들을 독립변수로 이용하여 종속변수를 예측해 보고자 한다.

	1	2	3	4	5	6	7	8
	No.	A1	A2	A3	A4	A5	A6	A7
1	1	110,06100	1,95173	115,15100	27,27380	72,20820	109,96700	0,55476
2	2	109,95900	1,97996	115,48600	27,05320	72,20820	109,96700	0,55524
3	3	110,01400	2,02079	115,31600	27,05140	72,20820	110,07900	0,55242
4	4	110,01400	2,00120	115,33800	27,31790	72,20820	109,96700	0,55388
5	5	109,99800	2,02360	115,19200	27,04110	72,20820	109,96700	0,55435
6	6	109,95900	2,02063	115,62000	27,29520	72,20820	109,96700	0,55059
7	7	110,06100	1,99856	115,26400	27,22800	72,22000	109,96700	0,55992
8	8	109,95900	2,02636	115,48900	27,15290	72,27770	110,07900	0,55333
9	9	109,99900	2,01522	115,46700	27,35910	72,25440	109,96700	0,55808
10	10	109,99100	2,00764	115,39400	27,08530	72,23120	109,94900	0,55435
11	11	110,04600	1,98164	115,31500	27,11020	72,23140	109,96700	0,55759
12	12	109,99000	1,99189	114,94700	27,02340	72,23140	110,06000	0,55384
13	13	110,01400	1,97797	114,98400	27,05420	72,23120	109,96700	0,55715
14	14	109,93600	1,97222	115,30400	27,34290	72,21980	110,07900	0,55242
15	15	110,03000	1,95748	115,33600	27,14210	72,21980	109,96700	0,55855
16	16	110,04400	2,13262	115,15300	27,13500	72,23120	110,07900	0,55711
17	17	109,92900	1,96840	116,10400	27,04340	72,21980	109,96700	0,55341
18	18	110,09900	1,94347	114,64800	27,22280	72,21980	109,96700	0,55710
19	19	109,90600	1,96301	116,02800	27,02510	72,23140	109,96700	0,55246
20	20	110,02900	1,99941	115,02400	27,22800	72,23140	109,96800	0,55663

분산설명력

각 주성분이 전체변수에 대한 정보를 어느 정도 가지고 있는지를 표시하고 있습니다.

주성분 #	Eigen Value of Cov(X)	X 기여율	X 누적 기여율
주성분1	10,27564	19,38800	19,38800
주성분2	7,42931	14,01756	33,40556
주성분3	5,08751	9,59908	43,00465

X-loadings

각 변수를 이용해 주성분을 생성하는 선형결합을 표시하고 있는 값으로 선형 결합 시 계수값을 의미합니다. 즉, 주성분 1 은 다음과 같이 생성됩니다.

주성분 1 = (-0.00143*A2)+(0.23957*A3)+(-0.03681*A4)+…

변수	주성분1	주성분2	주성분3
A2	-0,00143	0,00006	-0,02236
A3	0,23957	0,06468	-0,04607
A4	-0,03681	0,22270	0,05653
A5	0,04412	0,09681	0,19738
A6	0,23420	0,10138	-0,16262
A7	0,22797	-0,01757	-0,13887
A8	-0,13783	0,17915	-0,17459
A9	0,00195	0,10036	-0,33283
A10	0,00679	-0,17867	0,05286
A11	0,06843	0,15912	-0,09702
A12	0,08027	0,26967	0,04169
A13	-0,00341	-0,00334	0,01319
A14	0,24035	0,04159	0,03813
A15	0,17611	0,20899	0,04425

Regression coefficients

표준화 데이터 및 주성분으로 표현된 회귀모델과 주성분의 개수가 한 개부터 선택된 주성분 개수까지 늘어날 때의 표준화 데이터에 대한 회귀계수를 보여줍니다.

A1 = 0.0032 A2 + 0.0697 A3 + 0.0047 A4 - 0.0115 A5 + 0.0907 A6 + 0.0735 A7 + 0.0137 A8 + 0.0636 A9 - 0.0247 A10 + 0.0474 A11 + 0.0391 A12 - 0.0032 A13 + 0.0541 A14 + 0.0549 A15 + 0.0201 A16 + 0.0742 A17 + 0.0410 A18 + 0.0490 A19 + 0.0362 A20 + 0.0133 A21 - 0.0061 A22 + 0.0351 A23 + 0.0281 A24 - 0.0481 A25 - 0.0190 A26 + 0.0458 A27 + 0.0295 A28 + 0.0485 A29 - 0.0158 A30 - 0.0117 A31 + 0.0698 A32 + 0.0133 A33 + 0.0402 A34 + 0.0443 A35 + 0.0604 A36 + 0.0547 A37 - 0.0248 A38 - 0.0226 A39 - 0.0086 A40 - 0.0010 A41 - 0.0279 A42 - 0.0153 A43 + 0.0482 A44 - 0.0199 A45 - 0.0475 A46 - 0.0226 A47 - 0.0238 A48 - 0.0009 A49 + 0.0863 A50 + 0.0015 A51 + 0.0569 A52 - 0.0214 A53 - 0.0026 A54

A1 = 0.2331 PC1 + 0.1001 PC2 - 0.1596 PC3

● Regression coefficients

변수	주성분1	주성분2	주성분3
A2	-0,00033	-0,00033	0,00324
A3	0,05584	0,06231	0,06967
A4	-0,00858	0,01372	0,00469

표준에러정보

Adjust R2	0.76063
R - square	0.76230
RMSE	0.85279
MAE	0.67251
MAPE	0.61714

(1) R-square(결정계수)

도출된 회귀식이 측정값들을 대표할만한가를 확인하는 것으로, 즉, 추정된 회귀선이 관측치들을 얼마나 잘 설명하는지를 검정하는 값입니다. 1 에 가까울수록 회귀식이 데이터를 잘 설명하고 있음을 의미합니다.

(2) Adjust R2

투입되는 독립변수들이 많을수록 종속변수에 대한 설명력인 R-square 는 자연히 증가합니다. 그러므로 실제 적합도를 판정하는 데에는 자유도를 고려하여 조정된 결정계수 R-square(adj)를 많이 활용합니다. 1 에 가까울수록 회귀식이 데이터를 잘 설명하고 있음을 의미합니다.

(3) RMSE (Root-Mean Square Error)

$$RMSE = \sqrt{MSE} = \sqrt{SSE/(n-p-1)}$$

(4) MAE (Mean Absolute Error)

$$MAE = \frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{n}$$

(5) MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{\sum_{i=1}^{n} | (y_i - \hat{y}_i)/y_i |}{n} \times 100$$

(3) ~ (5)번 모두 종속 변수의 원래값과 예측값의 오차를 나타낸 것으로 값이 작을수록 좋은 모형이라 할 수 있습니다.

3.5.13 PLS 모델 노드

PLS

PLS 모델 노드는 PLS 분석 결과로 생성됩니다. Training 데이터 셋으로 PLS 분석을 수행한 후 분석 결과인 PLS 모델 노드를 Test 데이터 셋(새로운 데이터 셋)에 적용하여 **새로운 데이터의 예측**에 사용될 수 있습니다.

예제데이터 (조업편차분석.txt)

공정과정 중에 발생하는 A1 ~ A54, 총 54 개의 변수 중 A1 변수를 종속변수로 하고 나머지 변수(A2 ~ A54)를 독립변수로 하여 A1 을 예측하는 분석을 수행합니다.

	1	2	3	4	5	6	7	8	9	10	11	12
	No.	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
1	1	110,06100	1,95173	115,15100	27,27380	72,20820	109,96700	0,55476	66,99480	75,06930	38,00220	25,4
2	2	109,95900	1,97996	115,48600	27,05320	72,20820	109,96700	0,55524	66,99330	75,05090	37,96510	25,7
3	3	110,01400	2,02079	115,31600	27,05140	72,20820	110,07900	0,55242	67,05090	75,06930	37,98350	25,6
4	4	110,01400	2,00120	115,33800	27,31790	72,20820	109,96700	0,55388	66,93840	75,05210	37,98350	25,7
5	5	109,99800	2,02360	115,19200	27,04110	72,20820	109,96700	0,55435	66,99420	75,14410	37,98350	25,7
6	6	109,95900	2,02063	115,62000	27,29520	72,20820	109,96700	0,55059	67,08870	75,05210	37,98350	25,4
7	7	110,06100	1,99856	115,26400	27,22800	72,22000	109,96700	0,55992	66,99470	75,05090	37,98350	25,6
8	8	109,95900	2,02636	115,48900	27,15290	72,27770	110,07900	0,55333	66,99420	74,95700	37,98350	25,6
9	9	109,99900	2,01522	115,46700	27,35910	72,25440	109,96700	0,55808	66,96790	74,95670	38,00220	25,4
10	10	109,99100	2,00764	115,39400	27,08530	72,23120	109,94900	0,55435	67,09010	75,14410	37,98380	25,6
11	11	110,04600	1,98164	115,31500	27,11020	72,23140	109,96700	0,55759	66,86390	75,04900	38,05830	25,6
12	12	109,99000	1,99189	114,94700	27,02340	72,23140	110,06000	0,55384	67,02280	74,93680	38,05830	25,4
13	13	110,01400	1,97797	114,98400	27,05420	72,23120	109,96700	0,55715	66,94860	74,71240	38,17050	25,9
14	14	109,93600	1,97222	115,30400	27,34290	72,21980	110,07900	0,55242	67,04990	74,20900	38,17050	25,7
15	15	110,03000	1,95748	115,33600	27,14210	72,21980	109,96700	0,55855	66,94800	73,79750	38,17050	25,6
	16	110,04400	2,13262	115,15300	27,13500	72,23120	110,07900	0,55711	66,99650	73,79750	38,05830	25,8

독립변수로딩

PLS 기본설명에서 P 에 대한 값들 입니다. 알고리즘의 각 iteration 에서 X 를 업데이트할 때 사용됩니다.

	잠재변수1	잠재변수2	잠재변수3	잠재변수4
A2	0,00024	0,00238	-0,02664	-0,03762
A3	0,27630	0,04942	0,07208	0,01902
A4	0,00757	0,07909	-0,23272	0,12696
A5	0,05397	-0,00264	-0,00456	0,11767
A6	0,29658	0,13423	0,06297	0,05049
A7	0,26450	0,09502	0,19407	-0,00457
A8	-0,06243	0,24440	-0,21104	0,04243
A9	0,06771	0,19069	-0,14092	-0,05305
A10	-0,03946	-0,08924	0,21377	-0,06352
A11	0,10887	0,06596	-0,08627	0,18075
A12	0,11819	-0,01029	-0,28940	0,16308
A13	-0,00603	-0,00647	0,00658	0,00635
A14	0,22064	-0,18366	-0,08281	-0,04296
A15	0,18457	-0,13339	-0,25931	0,11478

독립변수가중치

잠재변수의 스코어를 구하기 위한 X 행렬에 대한 weight 입니다.

	잠재변수1	잠재변수2	잠재변수3	잠재변수4
A2	-0,00383	-0,01686	-0,06130	-0,02097
A3	0,31954	0,10644	0,15644	0,04626
A4	0,01148	0,01413	-0,23856	0,14714
A5	0,05922	0,00760	0,03323	0,04505
A6	0,37134	0,23120	0,24374	0,15951
A7	0,33701	0,23028	0,38106	0,07988
A8	-0,02058	0,18894	-0,28254	0,05979
A9	0,10639	0,14195	-0,23755	-0,01446
A10	-0,05033	-0,03455	0,21094	-0,14427
A11	0,13297	0,07107	-0,01338	0,13805
A12	0,09824	-0,11309	-0,31720	0,16013
A13	-0,00750	-0,00450	0,00902	-0,00163
A14	0,17352	-0,25186	-0,13164	0,00035
A15	0,13829	-0,23900	-0,27113	0,15804

종속변수로딩

PLS 기본설명에서 Q 에 대한 값들입니다. 알고리즘의 각 iteration 에서 Y 를 업데이트할 때 사용됩니다.

	잠재변수1	잠재변수2	잠재변수3	잠재변수4
A1	1	1	1	1

Inner Relation Coefficients

잠재변수들 간의 회귀계수입니다.

잠재변수1	잠재변수2	잠재변수3	잠재변수4
0,29507	0,13436	0,06783	0,05544

Coefficient Beta

표준화 데이터에 대한 회귀계수를 나타냅니다. 잠재변수수가 늘어남에 따라 y1 과 y2 의 식이 변화되는 정보가 나옵니다.

A1 = - 0.0119 A2 + 0.1475 A3 - 0.0103 A4 + 0.0276 A5 + 0.2032 A6 + 0.2012 A7 + 0.0009 A8 + 0.0365 A9 - 0.0094 A10 + 0.0646 A11 - 0.0111 A12 - 0.0025 A13 + 0.0001 A14 - 0.0155 A15 + 0.0331 A16 + 0.0552 A17 + 0.0789 A18 + 0.0022 A19 - 0.0048 A20 - 0.0157 A21 - 0.0526 A22 + 0.0265 A23 - 0.0189 A24 - 0.0173 A25 + 0.0325 A26 + 0.0273 A27 - 0.0098 A28 + 0.0330 A29 - 0.0044 A30 + 0.0070 A31 + 0.0353 A32 - 0.0168 A33 + 0.0060 A34 + 0.0771 A35 + 0.0144 A36 + 0.0135 A37 + 0.0017 A38 + 0.0189 A39 + 0.0153 A40 + 0.0105 A41 + 0.0053 A42 - 0.0062 A43 + 0.0519 A44 + 0.0053 A45 - 0.0104 A46 - 0.0011 A47 + 0.0124 A48 + 0.0269 A49 + 0.1878 A50 - 0.0042 A51 + 0.0826 A52 + 0.0065 A53 - 0.0002 A54

	잠재변수1 - A1	잠재변수2 - A1	잠재변수3 - A1	잠재변수4 - A1
A2	-0,00113	-0,00353	-0,00823	-0,01192
A3	0,09429	0,11977	0,13611	0,14750
A4	0,00339	0,00569	-0,00998	-0,01027
A5	0,01747	0,02057	0,02352	0,02761
A6	0,10957	0,15363	0,18007	0,20319
A7	0,09944	0,14218	0,17763	0,20120
A8	-0,00607	0,01859	0,00498	0,00095
A9	0,03139	0,05419	0,04320	0,03647
A10	-0,01485	-0,02125	-0,00839	-0,00944
A11	0,03924	0,05344	0,05573	0,06461

분산설명력

각 잠재변수가 전체데이터에 대한 정보를 어느 정도 가지고 있는지를 표시하고 있습니다.

잠재변수	X 기여율	X 누적 기여율	Y 기여율	Y 누적 기여율
1	18,17533	18,17533	83,86942	83,86942
2	10,39598	28,57131	9,94633	93,81576
3	9,17679	37,74809	2,23753	96,05328
4	7,41191	45,16000	1,20745	97,26073

표준에러정보

	A1
Adjust R2	0,97259
R - square	0,97261
RMSE	0,28858
MAE	0,20061
MAPE	0,18404

종속변수가 여러 개이므로 각 종속변수에 대한 에러정보가 출력됩니다.

(1) R-square(결정계수)

도출된 회귀식이 측정값들을 대표할만한가를 확인하는 것으로, 즉, 추정된 회귀선이 관측치들을 얼마나 잘 설명하는지를 검정하는 값입니다. 1 에 가까울수록 회귀식이 데이터를 잘 설명하고 있음을 의미합니다.

(2) Adjust R2

투입되는 독립변수들이 많을수록 종속변수에 대한 설명력인 R-square 는 자연히 증가합니다. 그러므로 실제 적합도를 판정하는 데에는 자유도를 고려하여 조정된 결정계수 R-square(adj)를 많이 활용합니다. 1 에 가까울수록 회귀식이 데이터를 잘 설명하고 있음을 의미합니다.

(3) RMSE (Root-Mean Square Error)

$$RMSE = \sqrt{MSE} = \sqrt{SSE/(n-p-1)}$$

(4) MAE (Mean Absolute Error)

$$MAE = \frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{n}$$

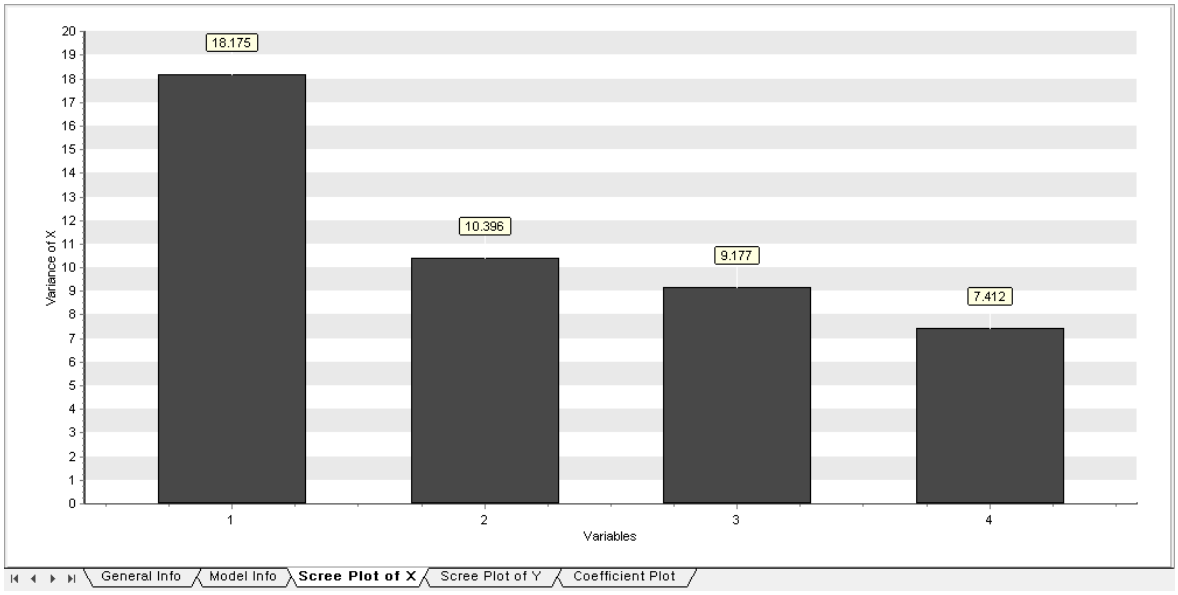
(5) MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{\sum_{i=1}^{n} | (y_i - \hat{y}_i)/y_i |}{n} \times 100$$

(3) ~ (5)번 모두 종속 변수의 원래값과 예측값의 오차를 나타낸 것으로 값이 작을수록 좋은 모형이라 할 수 있습니다.

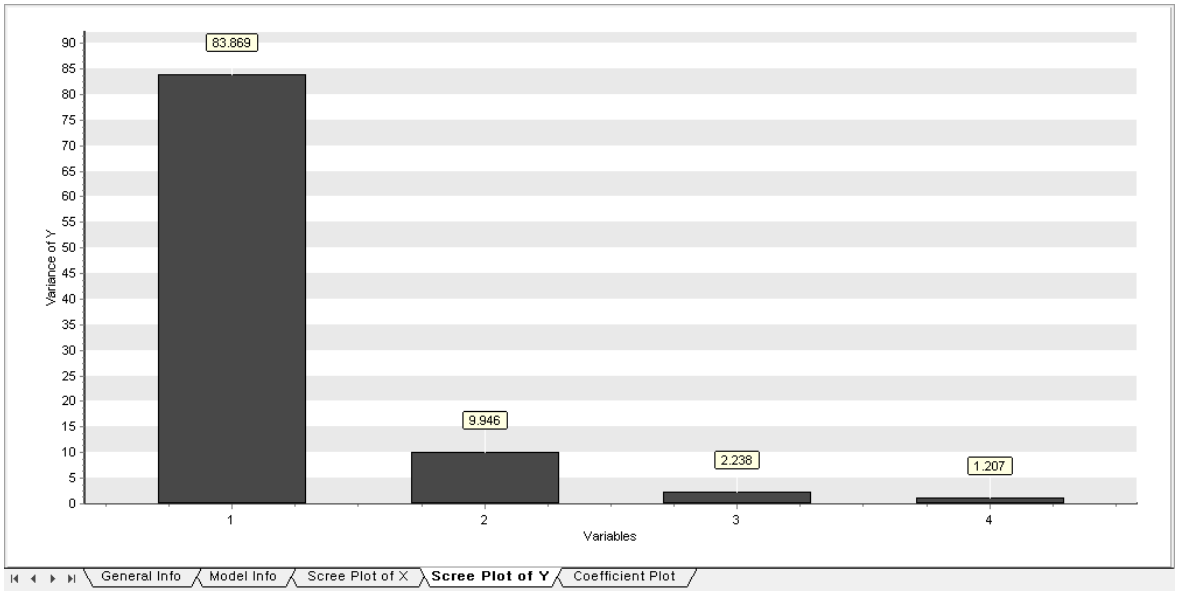
Scree Plot of X

각 잠재변수가 X 에 대한 설명력을 얼마나 가지고 있는지를 plot 으로 나타내고 있습니다.



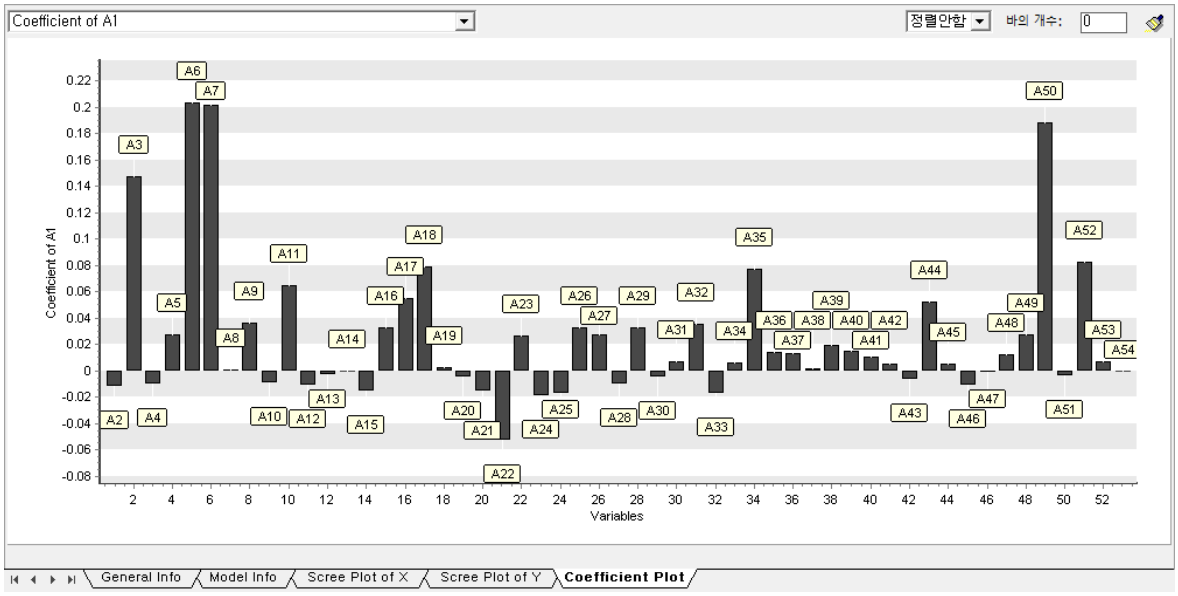
Scree Plot of Y

각 잠재변수가 Y 에 대한 설명력을 얼마나 가지고 있는지를 plot 으로 나타내고 있습니다.



Coefficient Plot

각 종속변수에 대해 어떤 변수가 가장 영향을 많이 미치는지를 파악할 수 있습니다.



3.5.14 QDA 모델 노드

QDA 모델 노드는 QDA 분류분석의 결과로 생성됩니다 Training 데이터 셋으로 QDA 분석을 수행하고 분석 결과인 QDA 모델 노드를 Test 데이터 셋(새로운 데이터 셋)에 적용하여 **새로운 데이터의 분류**에 사용될 수 있습니다.

예제데이터 (고객이탈_model.txt)

어느 회사의 고객자료를 바탕으로 고객 이탈 여부를 판단하고자 고객이탈모형을 구축하고자 QDA 모델 노드를 이용하여 분석을 수행하였습니다.

	1	2	3	4	5	6	7	8	9	10
	A1	A2	A3	A4	A5	A6	A7	A8	A9	이탈여부
1	24	B	M	B	181,39500	42,00000	213	921,00000	A	0
2	12	B	M	A	218,38000	191,40000	554	1,082,10000	A	1
3	23	B	M	A	168,69900	37,80000	33	909,00000	A	1
4	22	B	M	G	166,95200	128,40000	215	990,00000	A	0
5	56	B	M	H	137,70700	102,00000	421	768,60000	A	1
6	30	B	M	D	131,15100	36,50000	515	947,00000	A	1
7	43	B	M	G	210,27900	93,50000	355	1,067,60000	A	1
8	19	B	M	G	181,52000	70,80000	332	881,40000	A	0
9	19	B	M	A	177,08500	0,00000	144	750,60000	A	1
10	24	B	M	B	80,23190	81,60000	161	913,20000	A	1
11	29	B	M	H	199,45000	78,00000	478	1,069,50000	A	1
12	33	B	MH	H	205,54300	22,80000	406	1,525,50000	A	1
13	38	B	M	G	182,79500	24,00000	485	983,10000	A	1
14	33	B	M	A	156,26800	28,80000	202	913,20000	A	1
15	24	B	M	K	100,79200	77,50000	527	890,80000	A	1
16	43	B	M	B	99,35700	1,80000	531	867,00000	A	0
17	18	B	MH	F	96,22250	0,00000	679	1,455,60000	A	1

클래스별 독립변수평균

분류된 각 클래스의 변수평균값을 나타냅니다.

변수	0	1
A1	28,67447	31,85696
A2 (A)	-0,12376	-0,24402
A2 (B)	-0,28549	-0,25833
A2 (C)	-0,30499	-0,33604
A2 (D)	-0,20333	-0,15777
A3 (H)	0,06448	0,06191
A3 (L)	0,04992	0,01430
A3 (M)	0,45164	0,60675
A3 (MH)	0,21451	0,22395
A3 (ML)	0,21321	0,09308

클래스의 분산-공분산 행렬

클래스의 분산-공분산 행렬로 각 그룹의 분산-공분산 행렬이 다르기 때문에 그룹 수만큼의 분산-공분산 행렬이 출력됩니다.

	A1	A2 (A)	A2 (B)	A2 (C)	A2 (D)	A3 (H)	A3 (L)	A3 (M)	A3 (MH)	A3 (ML)	A4 (A)
A1	158,56500	0,37608	0,09612	0,26974	0,37879	-0,06353	0,00221	-0,07192	-0,08542	0,21507	0,0053
A2 (A)	0,37608	0,62811	0,34827	0,34586	0,35844	-0,05834	0,03349	0,01300	-0,17631	0,18114	-0,0017
A2 (B)	0,09612	0,34827	0,40014	0,29652	0,32555	-0,04661	0,01426	0,09490	-0,12678	0,06245	-0,0031
A2 (C)	0,26974	0,34586	0,29652	0,36911	0,32159	-0,04561	0,03786	0,02127	-0,13638	0,12096	-0,0024
A2 (D)	0,37879	0,35844	0,32555	0,32159	0,52250	-0,05217	0,01015	0,15453	-0,15819	0,04440	-0,0023
A3 (H)	-0,06353	-0,05834	-0,04661	-0,04561	-0,05217	0,06242	-0,00218	-0,02809	-0,01280	-0,01271	0,0010
A3 (L)	0,00221	0,03349	0,01426	0,03786	0,01015	-0,00218	0,04952	-0,02151	-0,00967	-0,00961	0,0001
A3 (M)	-0,07192	0,01300	0,09490	0,02127	0,15453	-0,02809	-0,02151	0,24981	-0,09587	-0,09528	-0,0027
A3 (MH)	-0,08542	-0,17631	-0,12678	-0,13638	-0,15819	-0,01280	-0,00967	-0,09587	0,17062	-0,04471	0,0015

사전확률

점유도 옵션에서 지정한 클래스별 확률값을 보여줍니다.

클래스	확률
0	0,50000
1	0,50000

Confusion matrix for number & percentage

원래 클래스와 예측 클래스에 대한 교차표를 통하여 예측이 잘못된 관측치 수와 백분율을 알 수 있습니다. 이탈고객을 이탈로 예측할 확률은 57.57%, 이탈하지 않을 고객을 이탈하지 않은 고객으로 예측할 확률이 97.25%, 이탈고객을 이탈하지 않은 고객으로 예측할 확률이 42.43%, 이탈하지 않을 고객을 이탈할 고객으로 예측할 확률은 2.75%입니다. 또한 전체 오분류수는 1761, 오분류율은 20.64%입니다.

	[예측] 0	[예측] 1
0	2214 (57.57 %)	1632 (42.43 %)
1	129 (2.75 %)	4555 (97.25 %)

오분류 수: 1761
오분류율: 20.64%

3.5.15 RBF 모델 노드

RBF 모델 노드는 RBF 분류분석의 결과로 생성됩니다 Training 데이터 셋으로 RBF 분석을 수행하고 분석 결과인 RBF 모델 노드를 Test 데이터 셋(새로운 데이터 셋)에 적용하여 **새로운 데이터의 분류**에 사용될 수 있습니다.

예제데이터 (RBF.csv)

독일의 German Bank 의 대출을 신청한 고객 1000 명(우량고객 700 명, 불량고객 300 명)에 대한 데이터를 사용하여 분류 분석을 수행하였습니다. 변수는 "신용 우/불량 여부, 당좌구좌상태, 대출상환기간, 신용거래내역, 대출목적, 대출금액, 예금액, 근무년수, 소득 중 할부 거래 상환액 비율, 성별 및 혼인관계, 보증인 여부, 거주 기간, 부동산, 연령, 타 여신 개설계획, 주거 형태, 현재 당행 신용계좌수, 직업, 부양가족수, 전화소유유무, 외국인 근로자 여부"로 이루어졌습니다.

	1	2	3	4	5	6	7	8	9	10	11	12	13
	OBS#	CHK_ACCT	DURATION	HISTORY	NEW_CAR	USED_CAR	FURNITURE	RADIO/TV	EDUCATION	RETRAINING	AMOUNT	SAV_ACCT	EMPLOYMEN
1	0,15385	0,84615	6	4	0	0	0	1	0	0	1,169	4	4
2	0,62245	0,37755	48	2	0	0	0	1	0	0	5,951	0	2
3	0,13129	0,86871	12	4	0	0	0	0	1	0	2,096	0	3
4	0,62245	0,37755	42	2	0	0	1	0	0	0	7,882	0	3
5	0,62245	0,37755	24	3	1	0	0	0	0	0	4,870	0	2
6	0,13129	0,86871	36	2	0	0	0	0	1	0	9,055	4	2
7	0,13129	0,86871	24	2	0	0	1	0	0	0	2,835	2	4
8	0,62245	0,37755	36	2	0	1	0	0	0	0	6,948	0	2
9	0,13129	0,86871	12	2	0	0	0	1	0	0	3,059	3	3
10	0,62245	0,37755	30	4	1	0	0	0	0	0	5,234	0	0
11	0,34715	0,65285	12	2	1	0	0	0	0	0	1,295	0	1
12	0,62245	0,37755	48	2	0	0	0	0	0	1	4,308	0	1
13	0,34715	0,65285	12	2	0	0	0	1	0	0	1,567	0	2
14	0,62245	0,37755	24	4	1	0	0	0	0	0	1,199	0	4
15	0,34715	0,65285	15	2	1	0	0	0	0	0	1,403	0	2
16	0,62245	0,37755	24	2	0	0	0	1	0	0	1,282	1	2
17	0,13129	0,86871	24	4	0	0	0	1	0	0	2,424	4	4
18	0,29268	0,70732	30	0	0	0	0	0	0	1	8,072	4	1
19	0,62245	0,37755	24	2	0	1	0	0	0	0	12,579	0	4
20	0,13129	0,86871	24	2	0	0	0	1	0	0	3,430	2	4

옵션정보

RBF 이용해서 모델링 시 어떤 옵션을 사용했는지를 보여줍니다. 이것을 통해 옵션이 바뀌었을 때 결과가 어떻게 바뀌었는지 판단할 수 있습니다.

중심

Radial Basis Function 의 중심값입니다. 사용자가 모델 옵션에서 중심수를 선택하게 되면 그에 해당하는 k-means 알고리즘이 실행되며 최종 RBF 중심의 좌표를 확인 할 수 있습니다.

변수	중심1	중심2	중심3	중심4	중심5
CHK_ACCT	1.65789	1.60000	1.95455	1.40000	1.63889
DURATION	13.86842	31.11429	16.95455	35.05000	28.13889
HISTORY	2.63158	3	2.72727	2.30000	2.66667
NEW_CAR	0.36842	0.08571	0.22727	0.20000	0.13889
USED_CAR	0	0.28571	0.04545	0.35000	0.25000

오분류정보

분류의 경우에만 해당됩니다. 원래 클래스와 예측 클래스에 대한 교차표를 통하여 예측이 잘못된 관측치 수와 백분율을 알 수 있습니다. 불량고객을 불량고객으로 예측할 확률은 3%, 우량고객을 우량고객으로 예측할 확률은 99%, 불량고객을 우량고객으로 예측할 확률은 97%, 우량고객을 불량고객으로 예측할 확률은 1% 입니다. 오분류 수 및 오분류율이 낮을수록 좋은 모형이라 판단할 수 있습니다.

	[예측] 0	[예측] 1
0	9 (3.00 %)	291 (97.00 %)
1	7 (1.00 %)	693 (99.00 %)

오분류 수: 298
오분류율: 29.80%

ANOVA 테이블

회귀분석의 경우에만 해당됩니다. 독립변수가 종속변수 y 를 설명하는데 있어서 통계적으로 유의한지 검정하기 위해 F 검정을 합니다. P-값이 유의수준보다 커서 유의하지 않다면 회귀계수에 대한 해석은 의미가 없습니다.

▶ ANOVA Table

	DF	SS	MS	F	p
모형(회귀)	2	812.37959	406.18980	339.31092	0.00000
잔차	25	29.92755	1.19710		
합계	27	842.30714			

표준에러정보

회귀분석의 경우에만 해당됩니다.

▶ 표준에러 정보

Adjust R2	0.96163
R - square	0.96447
RMSE	1.09412
MAE	0.87393
MAPE	1.96048

(1) Adjust R2

투입되는 독립변수들이 많을수록 종속변수에 대한 설명력인 R-square 는 자연히 증가합니다. 그러므로 실제 적합도를 판정하는 데에는 자유도를 고려하여 조정된 결정계수 R-square(adj)를 많이 활용합니다. 1 에 가까울수록 회귀식이 데이터를 잘 설명하고 있음을 의미합니다.

(2) R-square(결정계수)

도출된 회귀식이 측정값들을 대표할만한가를 확인하는 것으로, 즉, 추정된 회귀선이 관측치들을 얼마나 잘 설명하는지를 검정하는 값입니다. 1 에 가까울수록 회귀식이 데이터를 잘 설명하고 있음을 의미합니다.

(3) Root-MSE (Root-Mean Square Error)

$$RMSE = \sqrt{MSE} = \sqrt{SSE/(n-p-1)}$$

(4) MAE (Mean Absolute Error)

$$MAE = \frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{n}$$

(5) MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{\sum_{i=1}^{n} | (y_i - \hat{y}_i)/y_i |}{n} \times 100$$

(3) ~ (5)번 모두 종속 변수의 원래값과 예측값의 오차를 나타낸 것으로 값이 작을수록 좋은 모형이라 할 수 있습니다.

3.5.16 순차 연관성 분석 모델 노드

순차 연관성 분석 모델 노드는 어떤 변수들이 순차적인 관련성을 가지고 있을 것이라 가정하고 그 변수들의 순차 연관성을 관찰하고자 하는 노드입니다.

예제데이터 (순차연관성 분석.csv)

어느 과일가게에 고객이 구입한 과일의 정보가 아래와 같을 때, 구입한 순서를 고려한 과일 선호에 따른 규칙 탐사를 순차 연관성 분석 기법을 활용하여 수행하였습니다. 고객별 시간에 따른 구매품목(과일)과 이력을 나열한 데이터를 사용하여 입력 형태의 세 변수로 구성하였습니다. (1 개 CID(고객), 1 개 TID(순서), 5 개 ITEMID(품목))

고객 No.	구입한 과일
1	귤, 귤
2	감, 바나나, 감
3	귤, 귤, 대추
4	감, 감, 바나나, 사과
5	사과, 감, 감
6	대추, 귤, 귤

	1	2	3
	고객No.	시간	과일
1	1	1	귤
2	1	2	귤
3	2	3	감
4	2	4	바나나
5	2	5	감
6	3	6	귤
7	3	7	귤
8	3	8	대추
9	4	9	감
10	4	10	감
11	4	11	바나나
12	4	12	사과
13	5	13	사과
14	5	14	감
15	5	15	감
16	6	16	대추
17	6	17	귤
18	6	18	귤

연관규칙 정보

전체 고객 중에서 감을 구입한 후 바나나를 구입할 비율은 지지도가 33.33%이고, 감을 구입한 고객 중에서 이후에 바나나를 구입할 비율인 신뢰도는 66.67%입니다. 순차 연관 규칙의 두 품목은 서로 구입시점이 다르므로 인과 관계 적인 해석이 가능합니다.

● 연관규칙 정보

※ 총 3개의 연관 규칙을 발견하였습니다.

순번	연관규칙	연결 수	신뢰도(%)	향상도	지지율(%)
1	[감]----->[감]	2	100,00	2,00	50,00
2	[감]----->[바나나]	2	66,67	2,00	33,33
3	[귤]----->[귤]	2	100,00	2,00	50,00

(1) 연관규칙

[A->B] 연관규칙은 "상품 A 가 구매된 경우는 상품 B 도 구매된다." 라고 해석됩니다.

예시) "감을 선택한 사람은 이후 감을 또 선택한다"라고 해석합니다.

(2) 신뢰도(%)

신뢰도는 항목 A 를 포함하는 거래에서 항목 B 가 포함될 확률은 어느 정도인가를 나타내며 연관성의 정도를 파악할 수 있습니다.

예시) 감을 선택한 사람 중 이후에 감을 선택할 확률이 100% 입니다.

$$\left(\frac{\text{A와 B를 동시에 포함하는 거래의 비율} = \Pr(A \cap B)}{\text{A를 포함하는 거래의 비율} = \Pr(A)} \right)$$

(3) 향상도(lift)

항목 A 를 구매한 경우 그 거래가 항목 B 를 포함하는 경우와 항목 B 가 임의로 구매되는 경우의 비를 나타냅니다.

$$\left(\frac{\Pr(A \cap B)}{\Pr(A) * \Pr(B)} \right)$$

예시) 감의 선택과 이후의 감 선택은 양의 연관관계가 있음

(4) 지지율(%)

지지도와 동일한 의미입니다. 즉, 지지도는 A 와 B 를 포함하는 거래의 수를 출력한 것이고, 상대지지율은 그 거래의 수를 전체 거래수로 나눈 값입니다.

$$\left(\frac{\text{A와 B를 동시에 포함하는 거래의 수}}{\text{전체 거래수}} = \Pr(A \cap B) \right)$$

예시) 전체 과일 선택 건수 중 감을 선택한 사람이 이후에 감을 선택하는 비율은 18.52% 입니다.

3.5.17 SOM 모델 노드

SOM 모델 노드는 SOM 분석의 결과로 생성됩니다 Training 데이터 셋으로 SOM 분석을 수행하고 분석 결과인 SOM 모델 노드를 Test 데이터 셋(새로운 데이터 셋)에 적용하여 **새로운 데이터의 Clustering** 에 사용될 수 있습니다.

예제데이터 (축구선수 능력치 데이터.ecl)

유명 축구 게임에서 전세계 축구 선수 4,585 명의 신상 및 능력치를 데이터화 된 자료를 이용하여 축구 선수의 군집 및 같은 군집 내에서 저평가된(몸값이 낮은) 선수를 발굴하고 SOM 모델 분석을 수행하였습니다.

	1	2	3	4	5	6	7	8	9
	Nationallity	Position	Player	Cost	Age	Height	Foot	Attack	Defe
1	프랑스	CF	앙리	1,485	27	188	R	98	
2	체코	OMF	네드베드	1,418	32	177	R	85	
3	네덜란드	CF	반 니스텔루이	1,384	29	188	R	94	
4	독일	OMF	발락	1,384	28	189	R	86	
5	이탈리아	CB	말디니	1,315	37	186	R	72	
6	이탈리아	ST	토티	1,281	28	180	R	90	
7	이탈리아	CB	네스타	1,248	29	187	R	63	
8	프랑스	CMF	비에이라	1,248	29	191	R	73	
9	우크라이나	CF	쉐브첸코	1,193	28	183	R	97	
10	포루투갈	SMF	피구	1,180	32	180	R	84	
11	스페인	GK	카시아스	1,161	24	185	L	30	
12	이탈리아	CF	비에리	1,145	32	185	L	94	
13	브라질	ST	호나우딩요	1,142	25	181	R	93	
14	브라질	OMF	카카	1,129	23	186	R	85	
15	코트디부아르	CF	드록바	1,129	31	188	R	87	
16	영국	CMF	제라드	1,129	25	186	R	83	
17	브라질	CMF	에메우손	1,111	29	184	R	79	
18	브라질	SMF	제 로베르토	1,111	31	172	L	85	
19	프랑스	OMF	피레	1,111	32	188	R	85	
20	브라질	CF	아드리아노	1,078	23	189	L	90	

SOM 모형 정보

SOM 모형 정보를 통해 SOM 에 토폴로지와 Training 에 대한 정보를 얻을 수 있습니다.

토폴로지(Topology) 정보

토폴로지의 종류: Hexagonal 혹은 Grid 를 선택할 수 있습니다.

가로, 세로 방향의 격자 수: 선택한 토폴로지에서 가로, 세로에 몇 개의 격자를 두어서 최종 Topology 를 구성할지를 정하는 것입니다.

Training 정보

반복 수: 몇 번의 반복을 통해 Training 을 완료되었는지를 보여줍니다.

학습률의 변화: Training 을 시작할 때와 끝낼 때의 학습률이 어떻게 변화하였는지를 보여줍니다.

이웃 반경의 변화: Training 을 시작할 때와 끝낼 때의 이웃 반경이 어떻게 변화하였는지를 보여줍니다.

Chart 및 Statistics

SOM 모델링 노드에 대해서 설명할 때 언급하였듯이 모델 노드에서도 동일한 Chart 를 제공합니다.

SOM Weight Position Chart

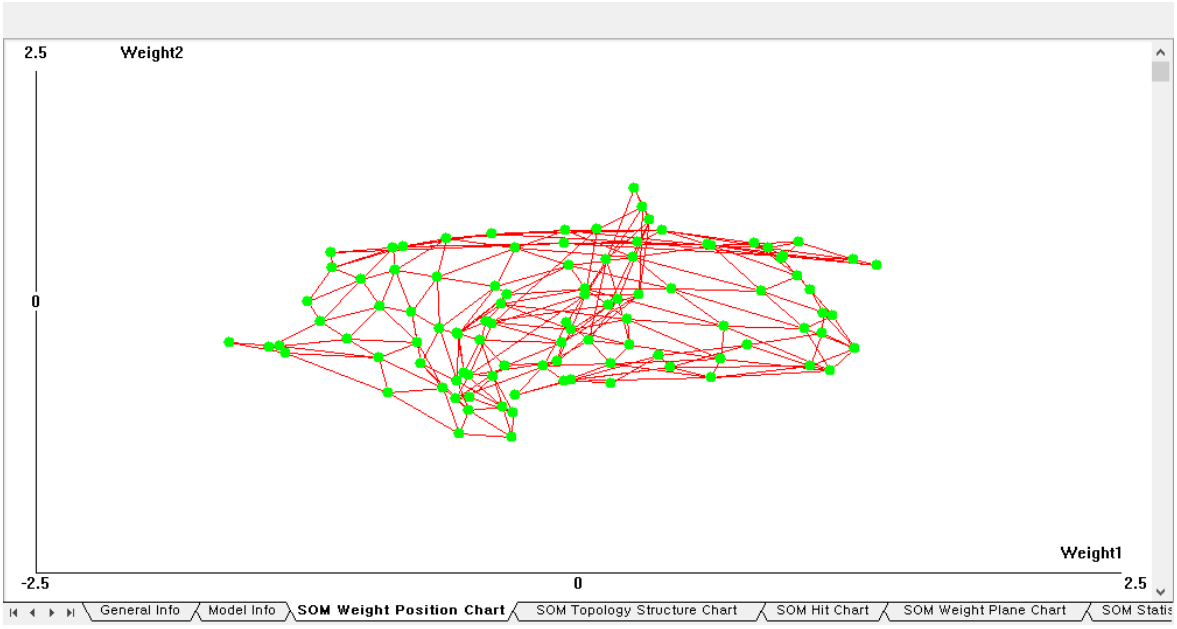
SOM Topology Structure Chart

SOM Hit Chart

SOM Weight Plane Chart

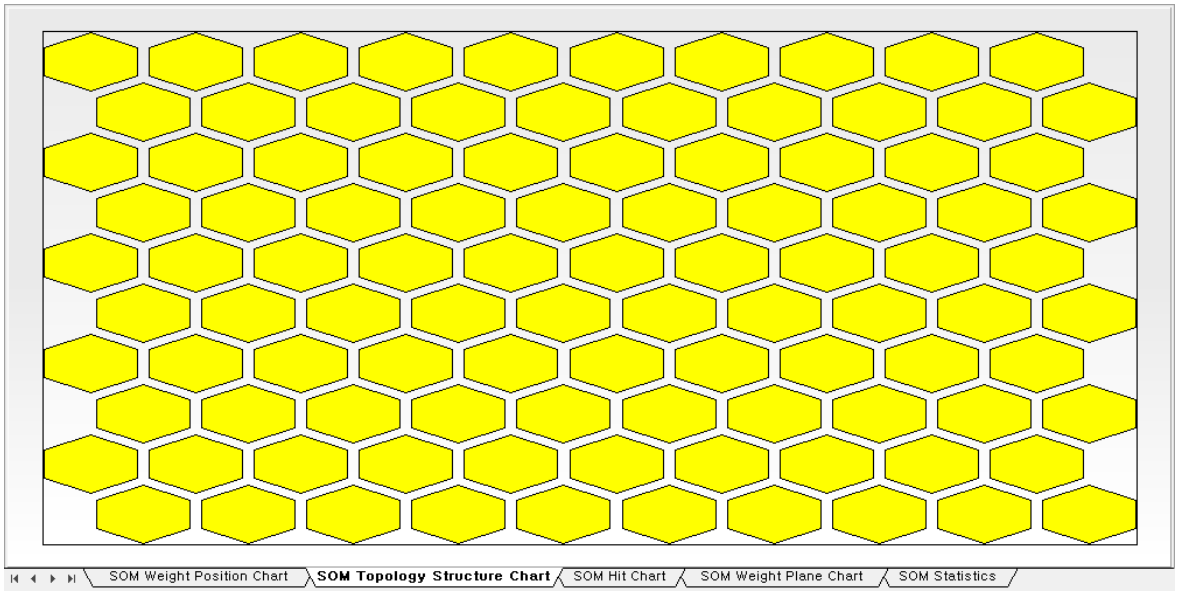
▪ SOM Weight Position Chart

본 Chart 를 통해서 Weight 가 공간상에 어떻게 배치되어 있는지와 각 Weight 간 어떠한 연결관계가 있는지를 알 수 있습니다.



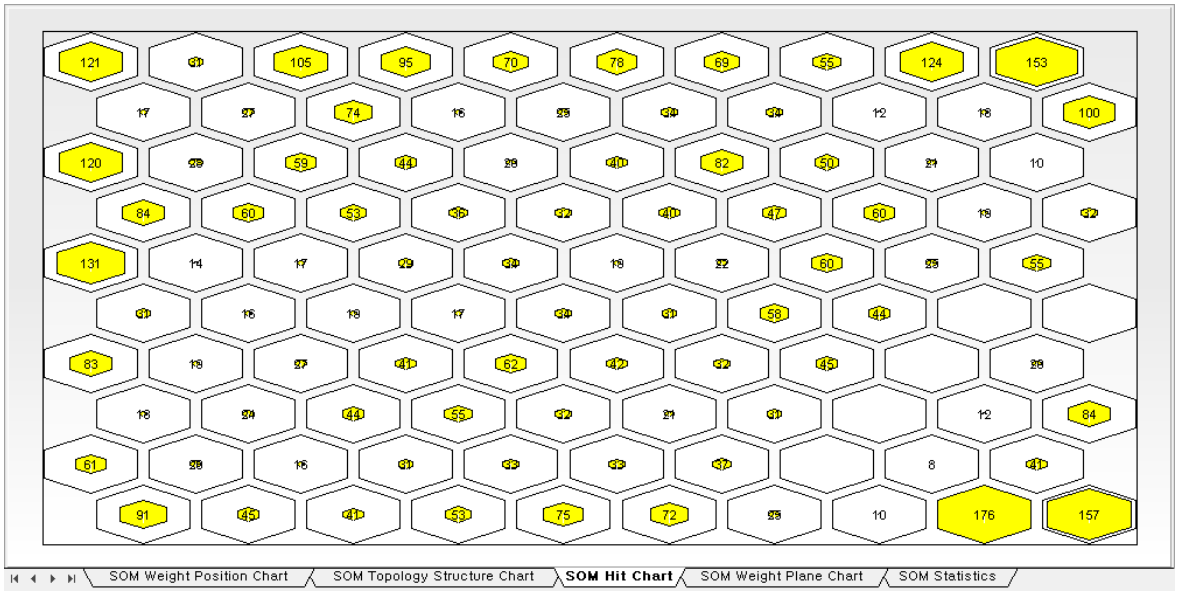
▪ **SOM Topology Structure Chart**

본 Chart 를 통해서 Topology 의 Structure 를 시각적으로 볼 수 있습니다. 아래의 예시는 Hexagonal Structure 를 시각화한 것입니다.



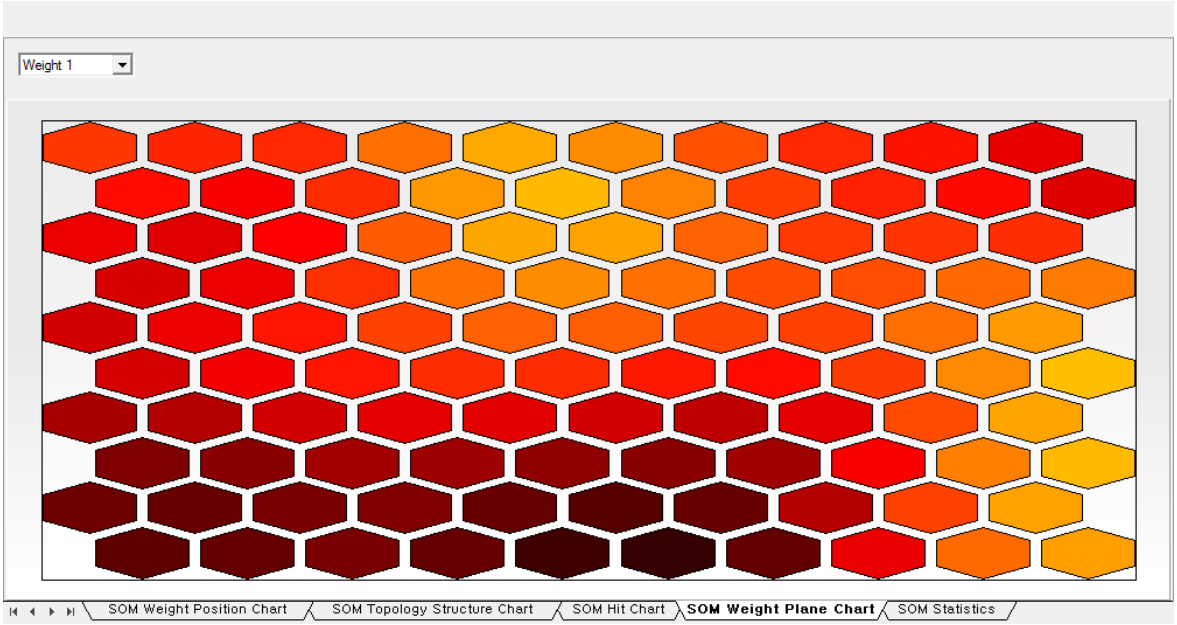
▪ **SOM Hit Chart**

본 Chart 를 통해서 각 Group 당 가장 가까운 데이터가 몇 개씩 있는지를 시각적으로 확인할 수 있습니다.



▪ **SOM Weight Plane Chart**

본 Chart 를 통해서 각 Weight 의 성분의 값을 시각적으로 알 수 있습니다. Weight 1 이라는 것은 Weight Vector 의 첫 번째 성분을 나타냅니다. Topology 상에서 각 Weight 의 성분(여기서는 첫번째 성분)의 값이 클수록 밝은 색이 되고 작을수록 짙은 색이 됩니다. ECMiner™ SOM 에서는 빨강색을 기준으로 Weight 의 성분 값이 0 보다 크면 노란색에 가까워지고, 0 보다 작으면 검은색에 가까워지도록 하였습니다.



▪ **SOM Statistics**

본 Statistics 를 통하여 Topology 상에 위치한 Weight Vector 의 성분 값을 구체적으로 알 수 있고 화면표시에서 나타나는 SOM_YHAT 이 나타내는 숫자에 해당하는 가로, 세로 격자 순서를 알 수 있습니다.

	1	2	3	4	5	6	7	8	9	10	11	
	세로 격자 순서	가로 격자 순서	Group Numbe	Weight1	Weight2	Weight3	Weight4	Weight5	Weight6	Weight7	Weight8	W
1	1	1	1	0,327905	-0,634704	0,91997	-0,642454	0,079498	0,747604	0,88671	1,365831	
2	1	2	2	0,217939	-0,622986	0,863317	-0,727052	-0,026231	0,560796	0,658245	1,179924	
3	1	3	3	0,244718	-0,626727	0,749674	-0,614102	-0,257809	0,322861	0,365286	0,763721	
4	1	4	4	0,657363	-0,687467	0,562843	-0,105627	-0,428812	0,253078	0,027108	0,155779	
5	1	5	5	0,999175	-0,699603	0,42922	0,401842	-0,356041	0,446084	-0,156141	-0,208089	
6	1	6	6	0,8242	-0,638509	0,476523	0,478725	-0,072834	0,715425	-0,035742	-0,094914	
7	1	7	7	0,478259	-0,466853	0,645471	0,217163	0,3734	0,837038	0,276125	0,329515	
8	1	8	8	0,241775	-0,149932	0,795897	-0,18843	0,700178	0,730627	0,549686	0,642798	
9	1	9	9	0,111396	0,445023	0,814863	-0,576172	0,770161	0,440623	0,640227	0,603598	
10	1	10	10	-0,132796	1,062177	0,771284	-0,930434	0,677138	0,170086	0,495218	0,300125	
11	2	1	11	0,069854	-0,683937	0,819564	-0,817509	-0,095698	0,615802	0,683023	1,23724	
12	2	2	12	-0,021141	-0,720363	0,722217	-0,757593	-0,309132	0,432505	0,441571	0,954965	
13	2	3	13	0,263244	-0,784021	0,57338	-0,350063	-0,49491	0,339862	0,195022	0,484964	
14	2	4	14	0,899578	-0,766051	0,373317	0,280549	-0,492194	0,403032	-0,076526	-0,036001	
15	2	5	15	1,097042	-0,687129	0,316159	0,587476	-0,267559	0,571207	-0,123923	-0,207018	
16	2	6	16	0,766867	-0,50324	0,395001	0,503337	0,136142	0,707191	0,074768	0,012945	
17	2	7	17	0,364988	-0,231756	0,521502	0,222341	0,58031	0,706224	0,336792	0,362268	
18	2	8	18	0,201268	0,155549	0,540954	-0,05337	0,744875	0,498605	0,426953	0,422433	
19	2	9	19	0,048289	0,70684	0,503387	-0,375861	0,687471	0,22506	0,316614	0,170127	
20	2	10	20	-0,186141	1,106997	0,529201	-0,737061	0,546156	0,066263	0,123955	-0,128902	
21	3	1	21	-0,110838	-0,640637	0,790196	-0,98708	-0,311946	0,5698	0,780287	1,256771	

SOM Weight Position Chart | SOM Topology Structure Chart | SOM Hit Chart | SOM Weight Plane Chart | **SOM Statistics**

3.5.18 Factor Analysis 모델 노드

Factor Analysis 모델 노드는 Factor Analysis 의 결과로 생성됩니다. 본 모델 노드를 통해서 Factor Analysis 의 결과 어떠한 공통 요인이 추출되었는지를 확인하고 해석할 수 있습니다.

일반정보

Factor Analysis 를 이용해서 모델링 시 사용한 분석방법에 대한 요약정보를 보여줍니다 (요인수, 추출 방법, 회전 방법, 최적화 결과). 알고리즘 반복회수에 따른 함수값 출력을 통해 더 이상 작아지지 않는 시점의 회수를 선택할 수 있습니다.

예제데이터 (Factor Analysis.csv)

현대인들의 행복지수에 영향을 주는 factor 에 대해 요인분석을 수행하였습니다. Maximum Likelihood 방법으로 비회전 옵션을 적용하여 분석하였습니다.

	1	2	3	4	5	6	7
	키	몸무게	나이	연소득	주중근무시간	연차사용일수	행복지수
1	164	50,60000	46	3,146	24	7	44,64000
2	155	44,50000	48	2,584	43	18	55,36000
3	153	44,70000	65	1,062	42	16	52,45000
4	179	62,10000	51	3,104	38	25	63,60000
5	188	69,20000	36	4,734	54	2	54,89000
6	164	58,60000	35	5,478	57	5	59,09000
7	165	63,50000	31	6,700	64	21	76,28000
8	150	55,00000	46	8,151	52	30	88,11000
9	173	71,70000	56	7,181	61	17	82,30000
10	186	74,40000	38	7,388	17	26	72,58000
11	171	72,90000	25	5,197	23	17	55,25000
12	179	81,10000	33	5,336	58	4	59,52000
13	166	53,40000	42	5,684	53	26	75,12000
14	187	80,30000	34	9,968	44	29	91,65000
15	181	82,90000	38	9,045	17	20	75,33000

Factor Loading(비회전)

회전되지 않은 Factor Loading 을 확인할 수 있습니다. Factor Loading 은 각 요인의 Loading 값으로 해당 변수와 요인의 설명 정도를 확인할 수 있습니다. Loading 값의 크기를 보고 주요 요인으로 선택되는 변수 확인합니다. 공통성은 추출된 요인이 변수들의 속성을 얼마나 잘 반영하고 있는가에 대한 설명력을 의미합니다.

아래 표에 대한 전반적인 해석은 다음과 같습니다. 일반적으로 Factor Loading 이±0.3 이상이면 유의한 값으로 보고 ±0.5 이상이면 높은 유의성이 있다고 해석합니다. 따라서 Factor1 은 키에서 loading 값이 0.85 이고, 몸무게는 0.96 이므로 Factor1 은 키와 몸무게의 속성을 가장 많이 포함하고 있는 요인이라고 해석할 수 있습니다. 반면 Factor2 에서 연소득의 loading 값이 0.99 이고 나머지 값들은 모두 0.3 이하이므로 Factor2 는 연소득의 속성을 가장 많이 포함하고 있는 요인이라고 판단할 수 있습니다.

2. Factor Loading (비회전)

변수	Factor1	Factor2	Factor3	Factor4	공통성
키	0,86591	0,12588	0,06940	-0,01487	0,77069
몸무게	0,96397	0,18462	0,00428	0,00934	0,96343
나이	-0,10140	-0,06320	-0,09602	0,26428	0,09334
연소득	-0,02353	0,99738	0,01740	0,00074	0,99563
주중근무시간	0,00019	0,00265	-0,99967	-0,00003	0,99934
연차사용일수	0,09562	0,11502	0,00823	-0,72948	0,55458
분산	1,69902	1,06194	1,01376	0,60229	4,37701
%분산	0,28317	0,17699	0,16896	0,10038	0,72950

요인별 점수 계수

각 요인당 점수를 구하기 위한 계수가 어떠한지를 확인할 수 있습니다. 이는 각 요인에 대한 독립변수들의 가중치를 의미합니다.

4. 요인별 점수 계수

변수	Factor1	Factor2	Factor3	Factor4
키	0.12222	-0.00355	-0.01162	-0.04180
몸무게	0.88904	-0.00667	-0.05314	0.01565
나이	0.01084	0.00801	0.00941	0.12712
연소득	-0.08544	-0.00270	1.00902	0.01032
주중근무시간	0.00004	-0.99898	0.00850	0.04392
연차사용일수	-0.07465	-0.04514	-0.05212	-0.71264

3.5.19 RBF DDA 모델 노드

RBF DDA 모델 노드는 RBF DDA 분류분석의 결과로 생성됩니다 Training 데이터 셋으로 RBF DDA 분석을 수행하고 분석 결과인 RBF DDA 모델 노드를 Test 데이터 셋(새로운 데이터 셋)에 적용하여 **새로운 데이터의 분류**에 사용될 수 있습니다.

예제데이터 (RBF.csv)

독일의 German Bank 의 대출을 신청한 고객 1000 명(우량고객 700 명, 불량고객 300 명)에 대한 데이터를 사용하여 분류 분석을 수행하였습니다. 변수는 "신용 우/불량 여부, 당좌구좌상태, 대출상환기간, 신용거래내역, 대출목적, 대출금액, 예금액, 근무년수, 소득 중 할부 거래 상환액 비율, 성별 및 혼인관계, 보증인 여부, 거주 기간, 부동산, 연령, 타 여신 개설계획, 주거 형태, 현재 당행 신용계좌수, 직업, 부양가족수, 전화소유유무, 외국인 근로자 여부"로 이루어졌습니다.

	1	2	3	4	5	6	7	8	9	10	11	12	13
	OBS#	CHK_ACCT	DURATION	HISTORY	NEW_CAR	USED_CAR	FURNITURE	RADIO/TV	EDUCATION	RETRAINING	AMOUNT	SAV_ACCT	EMPLOYMENT
1	0.15385	0.84615	6	4	0	0	0	1	0	0	1,169	4	4
2	0.62245	0.37755	48	2	0	0	0	1	0	0	5,951	0	2
3	0.13129	0.86871	12	4	0	0	0	0	1	0	2,096	0	3
4	0.62245	0.37755	42	2	0	0	1	0	0	0	7,882	0	3
5	0.62245	0.37755	24	3	1	0	0	0	0	0	4,870	0	2
6	0.13129	0.86871	36	2	0	0	0	0	1	0	9,055	4	2
7	0.13129	0.86871	24	2	0	0	1	0	0	0	2,835	2	4
8	0.62245	0.37755	36	2	0	1	0	0	0	0	6,948	0	2
9	0.13129	0.86871	12	2	0	0	0	1	0	0	3,059	3	3
10	0.62245	0.37755	30	4	1	0	0	0	0	0	5,234	0	0
11	0.34715	0.65285	12	2	1	0	0	0	0	0	1,295	0	1
12	0.62245	0.37755	48	2	0	0	0	0	0	1	4,308	0	1
13	0.34715	0.65285	12	2	0	0	0	1	0	0	1,567	0	2
14	0.62245	0.37755	24	4	1	0	0	0	0	0	1,199	0	4
15	0.34715	0.65285	15	2	1	0	0	0	0	0	1,403	0	2
16	0.62245	0.37755	24	2	0	0	0	1	0	0	1,282	1	2
17	0.13129	0.86871	24	4	0	0	0	1	0	0	2,424	4	4
18	0.29268	0.70732	30	0	0	0	0	0	0	1	8,072	4	1
19	0.62245	0.37755	24	2	0	1	0	0	0	0	12,579	0	4
20	0.13129	0.86871	24	2	0	0	0	1	0	0	3,430	2	4

옵션정보

RBF DDA 를 이용해서 모델링 시 어떤 옵션을 사용했는지를 보여줍니다. 이것을 통해 옵션이 바뀌었을 때 결과가 어떻게 바뀌었는지 판단할 수 있습니다.

Weight, Sigma, 중심

RBF DDA 의 Training 의 결과로 얻어진 Weight, Sigma, 중심을 보여줍니다.

Class: 1

변수	Weights	Sigma	FIELD1	FIELD2	FIELD3	FIELD4
중심1	1	2.53390	-0.45880	2.24707	-0.50318	-0.55243
중심2	2	3.63415	-1.25394	0.25682	0.42008	1.80838
중심3	2	3.30909	-0.45880	0.75439	1.34334	1.80838
중심4	1	2.17008	-1.25394	0.25682	1.34334	1.80838
중심5	1	2.06282	-1.25394	0.25682	-0.50318	-0.55243
중심6	1	3.68657	-0.45880	0.25682	-0.50318	-0.55243
중심7	1	4.41723	-1.25394	3.24219	0.42008	-0.55243

오분류정보

원래 클래스와 예측 클래스에 대한 교차표를 통하여 예측이 잘못된 관측치 수와 백분율을 알 수 있습니다. 불량고객을 불량고객으로 예측할 확률은 7%, 우량고객을 우량고객으로 예측할 확률은 100%, 불량고객을 우량고객으로 예측할 확률은 93%, 우량고객을 불량고객으로 예측할 확률은 0% 입니다. 오분류 수 및 오분류율이 낮을수록 좋은 모형이라 판단할 수 있습니다.

	[예측] 0	[예측] 1
0	21 (7.00 %)	279 (93.00 %)
1	0 (0.00 %)	700 (100.00 %)

오분류 수: 279
오분류율: 27.90%

3.5.20 ScoreCard 모델 노드


ScoreCard

ScoreCard 모델 노드는 ScoreCard 분석의 결과로 생성됩니다 Training 데이터 셋으로 ScoreCard 분석을 수행하고 분석 결과인 ScoreCard 모델 노드를 Test 데이터 셋(새로운 데이터 셋)에 적용하여 **새로운 데이터의 분류**에 사용될 수 있을 뿐 아니라 새로운 데이터에 대해서 Score 를 구하는 데에도 사용할 수 있습니다.

예제데이터 (고객이탈_model.txt)

어느 회사의 고객자료를 바탕으로 고객 이탈 여부를 판단하고자 고객이탈모형을 구축하고자 ScoreCard 모델 노드를 이용하여 분석을 수행하였습니다. 단 ScoreCard 은 독립변수가 모두 이산형이어야 하므로 연속형 변수인 A1, A5, A6, A8 을 제거합니다.(단, 구간화 노드를 이용하여 연속형 변수를 이산형으로 변환한 후 분석 가능합니다. 본 예시에서는 이산형 변수만으로 분석 하였습니다.)

	1	2	3	4	5	6	7	8	9	10
	A1	A2	A3	A4	A5	A6	A7	A8	A9	이탈여부
1	24	B	M	B	181,39500	42,00000	213	921,00000	A	0
2	12	B	M	A	218,38000	191,40000	554	1,082,10000	A	1
3	23	B	M	A	168,69900	37,80000	33	909,00000	A	1
4	22	B	M	G	166,95200	128,40000	215	990,00000	A	0
5	56	B	M	H	137,70700	102,00000	421	768,60000	A	1
6	30	B	M	D	131,15100	36,50000	515	947,00000	A	1
7	43	B	M	G	210,27900	93,50000	355	1,067,60000	A	1
8	19	B	M	G	181,52000	70,80000	332	881,40000	A	0
9	19	B	M	A	177,08500	0,00000	144	750,60000	A	1
10	24	B	M	B	80,23190	81,60000	161	913,20000	A	1
11	29	B	M	H	199,45000	78,00000	478	1,069,50000	A	1
12	33	B	MH	H	205,54300	22,80000	406	1,525,50000	A	1
13	38	B	M	G	182,79500	24,00000	485	983,10000	A	1
14	33	B	M	A	156,26800	28,80000	202	913,20000	A	1
15	24	B	M	K	100,79200	77,50000	527	890,80000	A	1
16	43	B	M	B	99,35700	1,80000	531	867,00000	A	0
17	18	B	MH	F	96,22250	0,00000	679	1,455,60000	A	1

클래스 수와 총 시행 수

종속변수의 클래스 수와 알고리즘이 실행된 총 시행 수를 알 수 있습니다.

Y Value

종속변수 Y 의 클래스 별 빈도 및 비율을 보여줍니다.

Value	빈도수	Graph	백분율
0	3846		45.09%
1	4684		54.91%
Total Count	8530		

로지스틱 테이블

Parameter Estimate 는 각 변수의 계수값을 나타내며, **Standard Error** 는 각 계수(Beta)의 standard error 입니다. 각 변수 유의성 검정의 **z-value, p-value** 를 보여줍니다. p-값이 유의수준보다 작을 때 유의한 변수라고 판단합니다. **Odds Ratio** 는 입력변수가 분류결정에 미치는 영향의 정도를 나타내는 값입니다. 오즈비가 1 보다 작다는 것은 입력변수 xi 가 (-)의 영향을 준다는 것이고, 오즈비가 1 보다 크다는 것은 (+)영향을 줌을 의미합니다.

Predictor	Dummy	Parameter Estimate	Standard Error	Z	p	Odds 비
A2	A	-1.39756	0.16718	-8.35961	0.00000	0.24720
A2	B	-0.89345	0.13656	-6.54273	0.00000	0.40924
A2	C	1.26250	0.22055	5.72435	0.00000	3.53426
A2	D	0.41187	0.11032	3.73334	0.00020	1.50964
A3	H	7.50626	135.04214	0.05558	0.95567	1819.38989
A3	L	8.11949	135.04205	0.06013	0.95206	3359.30644
A3	M	7.78500	135.04204	0.05765	0.95403	2404.26210
A3	MH	7.80809	135.04209	0.05782	0.95389	2460.43879
A3	ML	6.63574	135.04202	0.04914	0.96081	761.84233
A4	A	0.24965	0.26590	0.93889	0.34791	1.28358

Score 테이블

가변수	Dummy	Parameter Estimate	Min Parameter Estimate	양수화	스코어최대값	스코어환산	Max	Min
A2	A	-1.39756		0		2.50000		
A2	B	-0.89345		0.50411		3.81751		
A2	C	1.26250		2.66006		9.45222		
A2	D	0.41187		1.80943		7.22905		
A2	E	0	-1.39756	1.39756	2.66006	6.15260	9.45222	2.50000
A3	H	7.50626		7.50626		22.11806		
A3	L	8.11949		8.11949		23.72079		
A3	M	7.78500		7.78500		22.84657		
A3	MH	7.80809		7.80809		22.90694		
A3	ML	6.63574		6.63574		19.84291		
A3	N	0	0	0	8.11949	2.50000	23.72079	2.50000
A4	A	0.24965		6.57581		19.68629		
A4	B	-5.94878		0.37738		3.48630		

Score 테이블을 통해서 Logistic 회귀분석의 결과를 더 쉽게 해석할 수 있습니다. 예를 들어 위의 A2 은 A2_A, A2_B, A2_C, A2_D, A2_E 으로 새로운 파생변수를 만들어서 Score

Card 에서 활용하는데 이는 A2 변수의 범주가 5 개이기 때문입니다. A2_A, A2_B, A2_C, A2_D, A2_E 중 스코어가 가장 큰 것은 A2_C 로 이는 이중에 범주 3 이 가장 큰 영향을 미친다는 의미입니다. 이런 식으로 다른 변수들 또한 해석할 수 있고 이를 통해 어떠한 독립변수가 분류에 큰 영향을 주는지 알 수 있습니다.

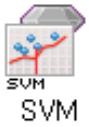
오분류 정보

종속변수의 원래값과 Logistic Regression 을 통해 얻어진 예측값 간의 빈도, 퍼센트 교차표를 보여줍니다. 오분류 정보를 통해 각 범주에 대한 분류정확도를 판단할 수 있습니다. 오분류 정보를 통해 각 범주에 대한 분류정확도를 판단할 수 있습니다. 이탈고객을 이탈로 예측할 확률은 78.00%, 이탈하지 않을 고객을 이탈하지 않은 고객으로 예측할 확률이 93.10%, 이탈고객을 이탈하지 않은 고객으로 예측할 확률이 22.00%, 이탈하지 않을 고객을 이탈할 고객으로 예측할 확률은 6.90%입니다. 또한 전체 오분류수는 1169, 오분류율은 13.70%입니다.

	[예측] 0	[예측] 1
0	3000 (78.00 %)	846 (22.00 %)
1	323 (6.90 %)	4361 (93.10 %)

오분류 수: 1169
오분류율: 13.70%

3.5.21 SVM 모델 노드



SVM 모델 노드는 Support Vector Machine 분석의 결과로 생성됩니다 Training 데이터 셋으로 SVM 분석을 수행하고 분석 결과인 SVM 모델 노드를 Test 데이터 셋(새로운 데이터 셋)에 적용하여 **새로운 데이터의 분류**에 사용될 수 있습니다.

예제데이터 (GermanCredit.ecl)

독일의 German Bank 의 대출을 신청한 1000 명에 대한 자료로, 이중 700 명은 우량 고객, 300 명은 불량 고객입니다. 이 데이터를 이용하여 신용 불량 고객 예측 모델을 개발을 위하여 SVM 모델 분석을 수행하였습니다.

	1	2	3	4	5	6	7	8
	OBS#	CHK_ACCT	DURATION	HISTORY	NEW_CAR	USED_CAR	FURNITURE	RADIO
1	1	0	6	4	0	0	0	
2	2	1	48	2	0	0	0	
3	3	3	12	4	0	0	0	
4	4	0	42	2	0	0	1	
5	5	0	24	3	1	0	0	
6	6	3	36	2	0	0	0	
7	7	3	24	2	0	0	1	
8	8	1	36	2	0	1	0	
9	9	3	12	2	0	0	0	
10	10	1	30	4	1	0	0	
11	11	1	12	2	1	0	0	
12	12	0	48	2	0	0	0	
13	13	1	12	2	0	0	0	
14	14	0	24	4	1	0	0	
15	15	0	15	2	1	0	0	
16	16	0	24	2	0	0	0	
17	17	3	24	4	0	0	0	
18	18	0	30	0	0	0	0	

옵션정보

SVM 를 이용해서 모델링 시 어떤 옵션을 사용했는지를 보여줍니다. 이것을 통해 옵션이 바뀌었을 때 결과가 어떻게 바뀌었는지 판단할 수 있습니다.

SVM Type : C-SVM
C : 1.000000
Kernel Type : Radial Basis Function
gamma : 0.014493

오분류 정보

원래 클래스와 예측 클래스에 대한 교차표를 통하여 예측이 잘못된 관측치 수와 백분율을 알 수 있습니다. 불량고객을 불량고객으로 예측할 확률은 62.67%, 우량고객을 우량고객으로 예측할 확률은 97.29%, 불량고객을 우량고객으로 예측할 확률은 37.33%, 우량고객을 불량고객으로 예측할 확률은 2.71% 입니다. 또한 전체 오분류수는 131, 오분류율은 13.10%입니다.

	[예측] 0	[예측] 1
0	188 (62.67 %)	112 (37.33 %)
1	19 (2.71 %)	681 (97.29 %)

오분류 수: 131
오분류율: 13.10%

3.5.22 SVR 모델 노드

SVR 모델 노드는 Support Vector Regression 분석의 결과로 생성됩니다 Training 데이터 셋으로 SVR 분석을 수행하고 분석 결과인 SVR 모델 노드를 Test 데이터 셋(새로운 데이터 셋)에 적용하여 **새로운 데이터의 예측**에 사용될 수 있습니다.

예제데이터 (SVR_BostonHousing.csv)

보스턴의 506 개 주택에 대한 데이터(1993)를 통해 주택가격을 예측하는 분석을 수행하였습니다. [자치시(town) 별 1 인당 범죄율, 25,000 평방피트를 초과하는 거주지역의 비율, 비소매상업지역이 점유하고 있는 토지의 비율, 찰스강에 대한 더미변수(강의 경계에 위치한 경우는 1, 아니면 0), 10ppm 당 농축 일산화질소, 주택 1 가구당 평균 방의 개수, 1940 년 이전에 건축된 소유주택의 비율, 5 개의 보스턴 직업센터까지의 접근성 지수, 방사형 도로까지의 접근성 지수, 10,000 달러 당 재산세율, 자치시(town)별 학생/교사 비율, 1000(Bk-0.63)^2(여기서 Bk 는 자치시별 흑인의 비율을 말함), 모집단의 하위계층의 비율(%)을 이용하여 본인 소유의 주택가격(중앙값) (단위: $1,000)을 예측합니다.]

	1	2	3	4	5	6	7	8	9	10
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX
1	0,00632	18,00000	2,31000	0	0,53800	6,57500	65,20000	4,09000	1	
2	0,02731	0,00000	7,07000	0	0,46900	6,42100	78,90000	4,96710	2	
3	0,02729	0,00000	7,07000	0	0,46900	7,18500	61,10000	4,96710	2	
4	0,03237	0,00000	2,18000	0	0,45800	6,99800	45,80000	6,06220	3	
5	0,06905	0,00000	2,18000	0	0,45800	7,14700	54,20000	6,06220	3	
6	0,02985	0,00000	2,18000	0	0,45800	6,43000	58,70000	6,06220	3	
7	0,08829	12,50000	7,87000	0	0,52400	6,01200	66,60000	5,56050	5	
8	0,14455	12,50000	7,87000	0	0,52400	6,17200	96,10000	5,95050	5	
9	0,21124	12,50000	7,87000	0	0,52400	5,63100	100,00000	6,08210	5	
10	0,17004	12,50000	7,87000	0	0,52400	6,00400	85,90000	6,59210	5	
11	0,22489	12,50000	7,87000	0	0,52400	6,37700	94,30000	6,34670	5	
12	0,11747	12,50000	7,87000	0	0,52400	6,00900	82,90000	6,22670	5	
13	0,09378	12,50000	7,87000	0	0,52400	5,88900	39,00000	5,45090	5	
14	0,62976	0,00000	8,14000	0	0,53800	5,94900	61,80000	4,70750	4	
15	0,63796	0,00000	8,14000	0	0,53800	6,09600	84,50000	4,46190	4	
16	0,62739	0,00000	8,14000	0	0,53800	5,83400	56,50000	4,49860	4	
17	1,05393	0,00000	8,14000	0	0,53800	5,93500	29,30000	4,49860	4	
18	0,78420	0,00000	8,14000	0	0,53800	5,99000	81,70000	4,25790	4	

옵션정보

SVR 을 이용해서 모델링 시 어떤 옵션을 사용했는지를 보여줍니다. 이것을 통해 옵션이 바뀌었을 때 결과가 어떻게 바뀌었는지 판단할 수 있습니다.

ANOVA 테이블

독립변수가 종속변수 y 를 설명하는데 있어서 통계적으로 유의한지 검정하기 위해 F 검정을 합니다. P-값이 유의수준보다 커서 유의하지 않다면 회귀계수에 대한 해석은 의미가 없습니다

▶ ANOVA Table

	DF	SS	MS	F	p
모형(회귀)	13	29580,92501	2275,45577	85,22974	0
잔차	492	13135,37041	26,69791		
합계	505	42716,29542			

표준에러정보

▶ 표준에러 정보

Adjust R2	0,68437
R - square	0,69250
RMSE	5,16700
MAE	2,93753
MAPE	13,98262

(1) R-square(결정계수)

도출된 회귀식이 측정값들을 대표할만한가를 확인하는 것으로, 즉, 추정된 회귀선이 관측치들을 얼마나 잘 설명하는지를 검정하는 값입니다. 1 에 가까울수록 회귀식이 데이터를 잘 설명하고 있음을 의미합니다.

(2) R-square(adj)

투입되는 독립변수들이 많을수록 종속변수에 대한 설명력인 R-square 는 자연히 증가합니다. 그러므로 실제 적합도를 판정하는 데에는 자유도를 고려하여 조정된 결정계수 R-square(adj)를 많이 활용합니다. 1 에 가까울수록 회귀식이 데이터를 잘 설명하고 있음을 의미합니다.

(3) Root-MSE (Root-Mean Square Error)

$$RMSE = \sqrt{MSE} = \sqrt{SSE/(n-p-1)}$$

(4) MAE (Mean Absolute Error)

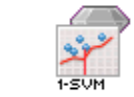
$$MAE = \frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{n}$$

(5) MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{\sum_{i=1}^{n} | (y_i - \hat{y}_i)/y_i |}{n} \times 100$$

(3) ~ (5)번 모두 종속 변수의 원래값과 예측값의 오차를 나타낸 것으로 값이 작을수록 좋은 모형이라 할 수 있습니다.

3.5.23 One class SVM 모델 노드



One class SVM 모델 노드는 One class SVM 분석 결과로 생성됩니다. Training 데이터 셋으로 이상치를 탐지하는 Support vector 를 찾고, 이를 이용하여 새로운 데이터 셋에 대해 이상치를 판별할 수 있습니다.

예제(학습)데이터 (Traning Data.ecl)

대장균이 Cytoplasm(세포질)에 존재하는 데이터를 Training Data 로 학습을 위한 데이터입니다.

	1	2	3	4	5	6	7	8	9	10
	A1	A2	A3	A4	A5	A6	A7	A8	A9	이탈여부
1	24	B	M	B	181,39500	42,00000	213	921,00000	A	0
2	12	B	M	A	218,38000	191,40000	554	1,082,10000	A	1
3	23	B	M	A	168,69900	37,80000	33	909,00000	A	1
4	22	B	M	G	166,95200	128,40000	215	990,00000	A	0
5	56	B	M	H	137,70700	102,00000	421	768,60000	A	1
6	30	B	M	D	131,15100	36,50000	515	947,00000	A	1
7	43	B	M	G	210,27900	93,50000	355	1,067,60000	A	1
8	19	B	M	G	181,52000	70,80000	332	881,40000	A	0
9	19	B	M	A	177,08500	0,00000	144	750,60000	A	1
10	24	B	M	B	80,23190	81,60000	161	913,20000	A	1
11	29	B	M	H	199,45000	78,00000	478	1,069,50000	A	1
12	33	B	MH	H	205,54300	22,80000	406	1,525,50000	A	1
13	38	B	M	G	182,79500	24,00000	485	983,10000	A	1
14	33	B	M	A	156,26800	28,80000	202	913,20000	A	1
15	24	B	M	K	100,79200	77,50000	527	890,80000	A	1
16	43	B	M	B	99,35700	1,80000	531	867,00000	A	0
17	18	B	MH	F	96,22250	0,00000	679	1,455,60000	A	1

예제(테스트)데이터 (Test Data.ecl)

대장균이 Cytoplasm 에 존재하는 데이터와 다른 곳에 존재하는 데이터가 같이 있는 데이터를 Test Data 로 사용하여, 학습 데이터를 통해 학습된 One class SVM 모델 노드를 이용하여 특정 집단에 속하는 데이터와 속하지 않는 이상치 데이터를 구분하는 분석을 수행합니다.

	1	2	3	4	5	6	7	8
	A1	A2	A3	A4	A5	A6	A7	A8
1	24	B	M	B	181,39500	42,00000	921,00000	A
2	12	B	M	A	218,38000	191,40000	1,082,10000	A
3	23	B	M	A	168,69900	37,80000	909,00000	A
4	22	B	M	G	166,95200	128,40000	990,00000	A
5	56	B	M	H	137,70700	102,00000	768,60000	A
6	30	B	M	D	131,15100	36,50000	947,00000	A
7	43	B	M	G	210,27900	93,50000	1,067,60000	A
8	19	B	M	G	181,52000	70,80000	881,40000	A
9	19	B	M	A	177,08500	0,00000	750,60000	A
10	24	B	M	B	80,23190	81,60000	913,20000	A
11	29	B	M	H	199,45000	78,00000	1,069,50000	A
12	33	B	MH	H	205,54300	22,80000	1,525,50000	A
13	38	B	M	G	182,79500	24,00000	983,10000	A
14	33	B	M	A	156,26800	28,80000	913,20000	A
15	24	B	M	K	100,79200	77,50000	890,80000	A

옵션정보

One class SVM 를 이용해서 모델링 시 어떤 옵션을 사용했는지를 보여줍니다. 이것을 통해 옵션이 바뀌었을 때 결과가 어떻게 바뀌었는지 판단할 수 있습니다.

SVM Type : One-Class SVM
nu : 0.500000
Kernel Type : Radial Basis Function
gamma : 0.010000

Support Vector List

모델링 결과 선정된 support vector 의 개수와 리스트를 보여줍니다. 이는 모델링 수행 결과 alpha 값이 0 보다 큰, 즉 support vector 로 선정된 데이터 값과 alpha 값을 보여줍니다.

Support Vector 갯수 : 37
rho : 32.246525

FIELD1	FIELD2	FIELD3	FIELD4	FIELD5	Alpha
0,49000	0,29000	0,56000	0,24000	0,35000	1
0,56000	0,40000	0,49000	0,37000	0,46000	0,18180
0,23000	0,32000	0,55000	0,25000	0,35000	1
0,29000	0,28000	0,44000	0,23000	0,34000	1
0,20000	0,44000	0,46000	0,51000	0,57000	1
0,42000	0,24000	0,57000	0,27000	0,37000	1
0,22000	0,43000	0,48000	0,16000	0,28000	1
0,40000	0,45000	0,38000	0,22000	0	1
0,51000	0,54000	0,41000	0,34000	0,43000	1
0,36000	0,39000	0,48000	0,22000	0,23000	1
0,56000	0,51000	0,34000	0,37000	0,46000	1

분석결과정보

화면표시 노드에서 One class SVM 결과인 decision value 와 그에 따른 이상치 판별 결과를 확인할 수 있습니다.

	1	2	3	4	5	6	7
	FIELD1	FIELD2	FIELD3	FIELD4	FIELD5	OneClassSVM10_DecisionValue	OneClassSVM10_YHA
1	0,07000	0,40000	0,54000	0,35000	0,44000	-1,17942	1
2	0,59000	0,49000	0,52000	0,45000	0,36000	-1,19492	1
3	0,67000	0,39000	0,36000	0,38000	0,46000	-1,44222	1
4	0,21000	0,34000	0,51000	0,28000	0,39000	0,21324	0
5	0,42000	0,40000	0,56000	0,18000	0,30000	-0,45184	1
6	0,25000	0,48000	0,44000	0,17000	0,29000	-0,51290	1
7	0,51000	0,50000	0,46000	0,32000	0,35000	0,16247	0
8	0,25000	0,40000	0,46000	0,44000	0,52000	-0,14243	1
9	0,44000	0,27000	0,55000	0,52000	0,58000	-2,74568	1
10	0,41000	0,57000	0,39000	0,21000	0,32000	-0,95571	1
11	0,31000	0,23000	0,73000	0,05000	0,14000	-7,49022	1
12	0,30000	0,16000	0,56000	0,11000	0,23000	-4,36672	1
13	0,29000	0,37000	0,48000	0,44000	0,52000	-0,08791	1
14	0,21000	0,51000	0,50000	0,32000	0,41000	-0,05942	1
15	0,43000	0,39000	0,47000	0,31000	0,41000	1,02078	0

Decision value 이상치 판별 결과

3.5.24 LOF 모델 노드


LOF

LOF 모델 노드는 Local Outlier Factor 분석의 결과로 생성됩니다. Training 데이터 셋으로 LOF 분석을 수행하고 LOF 모델 노드를 새로운 데이터 셋에 적용하여 새로운 데이터에 대한 이상치 판별을 진행할 수 있습니다.

예제데이터 (LOF_training.ecl, LOF_test.ecl)

Protein Localization Sites (단백질(대장균) 존재 위치)

대장균이 Cytoplasm(세포질)에 존재하는 데이터를 Training Data 로 학습한 후, 대장균이 Cytoplasm 에 존재하는 데이터와 다른 곳에 존재하는 데이터가 같이 있는 데이터를 Test Data 로 사용하여 이상치로 판단할 수 있는지를 분석했습니다. - 출처: UCI Machine

	1	2	3	4	5
	mcg	gvh	aac	alm1	alm2
1	0,49000	0,29000	0,56000	0,24000	0,35000
2	0,56000	0,40000	0,49000	0,37000	0,46000
3	0,23000	0,32000	0,55000	0,25000	0,35000
4	0,29000	0,28000	0,44000	0,23000	0,34000
5	0,20000	0,44000	0,46000	0,51000	0,57000
6	0,42000	0,24000	0,57000	0,27000	0,37000
7	0,39000	0,32000	0,46000	0,24000	0,35000
8	0,22000	0,43000	0,48000	0,16000	0,28000

옵션정보

LOF 를 이용해서 모델링 시 어떤 옵션을 사용했는지를 보여줍니다. 이것을 통해 옵션이 바뀌었을 때 결과가 어떻게 바뀌었는지 판단할 수 있습니다.

통계정보

통계정보에는 변수 정렬표를 나열합니다. 변수 정렬표란 training 데이터에서 모델링 결과값인 LOF score 가 높은 순서로 나열합니다. LOF 값이 클수록 군집에서 떨어져 있음을 의미하며 이를 참고하여 Score 값이 큰 데이터를 제거한 후, 모델링 데이터를 정제할 수 있습니다.

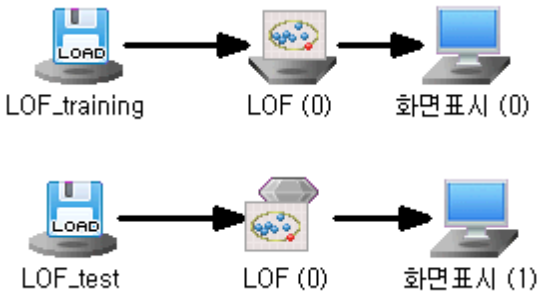
2. 통계정보

● 변수정렬표

순번	LOF Score
59	2.1608
9	2.1191
46	2.1189
44	1.9179
5	1.7894
40	1.7667
39	1.7376
48	1.5795
49	1.5795
47	1.5795
71	1.4858
15	1.4436
27	1.3885
52	1.3881
56	1.3658
68	1.3552

LOF 모델링을 통한 이상치 판별

LOF 모델링을 통한 이상치 판별은 다음 그림과 같이 진행됩니다.



1) 데이터를 training 용과 test 용으로 나누고 training 데이터로 "이상치 판별 여부" 옵션을 '예'로 설정하여 모델링 합니다.

2) Test 용 데이터를 생성된 모델에 적용합니다. (이때, 생성된 모델 노드에서 변수매칭방법을 '이름'으로 설정합니다.)

3) 화면표시(1)을 통해 나온 결과에서 LOF1_Outlier 가 1 이면 이상치, 0 이면 정상범위 안에 있는 데이터로 해석할 수 있습니다.

	1	2	3	4	5	6	7	8
	FIELD1	FIELD2	FIELD3	FIELD4	FIELD5	FIELD6	LOF1_Score	LOF1_Outlier
1	0,07000	0,40000	0,54000	0,35000	0,44000	0,00000	1,09424	0,00000
2	0,59000	0,49000	0,52000	0,45000	0,36000	0,00000	1,53947	0,00000
3	0,67000	0,39000	0,36000	0,38000	0,46000	0,00000	1,67055	0,00000
4	0,21000	0,34000	0,51000	0,28000	0,39000	0,00000	0,95665	0,00000
5	0,42000	0,40000	0,56000	0,18000	0,30000	0,00000	0,83815	0,00000
6	0,25000	0,48000	0,44000	0,17000	0,29000	0,00000	1,06212	0,00000
7	0,51000	0,50000	0,46000	0,32000	0,35000	0,00000	0,97319	0,00000
8	0,25000	0,40000	0,46000	0,44000	0,52000	0,00000	1,04637	0,00000
9	0,44000	0,27000	0,55000	0,52000	0,58000	0,00000	0,99424	0,00000
10	0,41000	0,57000	0,39000	0,21000	0,32000	0,00000	1,60920	0,00000
11	0,31000	0,23000	0,73000	0,05000	0,14000	0,00000	1,54296	0,00000
12	0,30000	0,16000	0,56000	0,11000	0,23000	0,00000	1,15020	0,00000
13	0,29000	0,37000	0,48000	0,44000	0,52000	0,00000	1,04637	0,00000

3.6 출력노드

출력 노드는 스트림을 통해 실행된 데이터를 테이블 또는 데이터베이스, 파일로 출력하기 위한 노드입니다. 이들 **출력 노드**는 **ECMiner™**스트림 상에서 실행된 데이터를 일반적인 형식의 데이터(파일 또는 데이터베이스)로 전환하여, **ECMiner™**뿐만 아니라 다른 솔루션에서 사용이 가능할 수 있도록 해줍니다.

ECMiner™에서는 다음과 같은 데이터 출력 노드를 지원합니다.

- **ODBC 출력**

ODBC 를 이용하여 데이터를 데이터베이스로 저장합니다.

- **OLEDB 출력**

  OLE DB 기술을 이용하여 데이터베이스로 데이터를 저장합니다. ODBC 보다 좋은 성능으로 데이터베이스를 조작할 수 있습니다.

- **피봇**

  입력된 정보에 따라 피봇팅을 수행합니다.

- **오라클 출력**

  오라클 데이터베이스 데이터를 저장합니다.

- **원인/결과 연관**

  원인 변수 값들과 결과 변수 값 간의 연관성을 분석합니다.

- **통계 분석**

  각 변수에 대한 통계분석을 수행합니다.

- **파일 출력**

  파일에 데이터를 저장합니다.

- **화면표시**

  결과를 화면에 표시합니다.

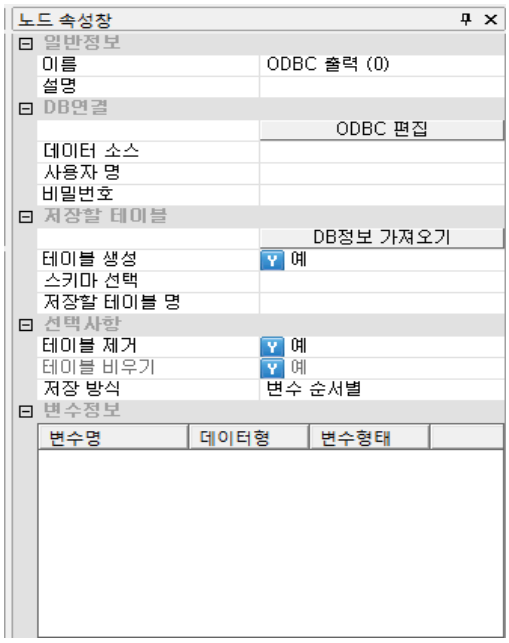
- **분리저장**

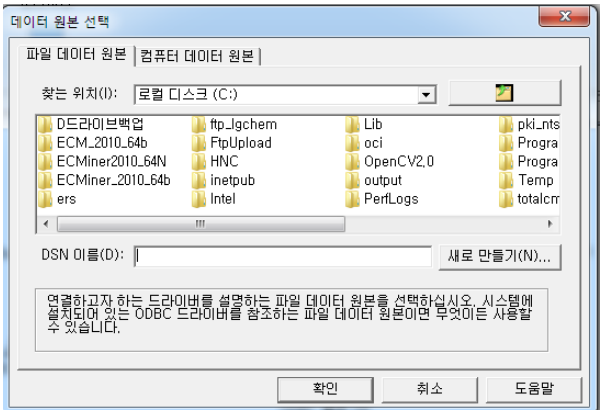
  데이터를 분리하여 저장합니다.

3.6.1 ODBC 출력노드

ODBC 출력 노드는 ODBC(Open Database Connectivity) 기술을 이용하여 데이터를 데이터 베이스로 저장하는 노드입니다.

사용법





- 데이터베이스에 접속하기 위하여 **데이터 소스, 사용자명, 비밀번호**를 입력합니다.
- 연결하고자 하는 데이터베이스가 없다면 **ODBC 편집** 버튼을 눌러 새로운 DSN 을 추가합니다.
- **DB 정보 가져오기** 버튼을 눌러 지정된 데이터베이스의 정보를 가져옵니다.
- 스키마를 선택하고, 데이터를 저장할 테이블명을 입력합니다.
- 만약 테이블 생성 옵션을 **'예'**로 선택한 경우, **테이블 제거** 및 **테이블 비우기**의 옵션을 확인 또는 변경합니다.
- **실행** 버튼을 누르면, 데이터가 DB 로 저장됩니다.

속성

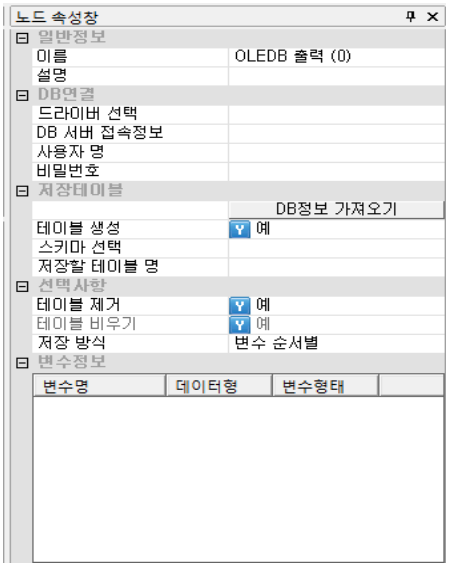
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
DB 연결	ODBC 편집	ODBC 연결을 위한 DSN(Data Source Name)을 새로 만들기 위하여 사용합니다.	버튼	
	데이터 소스	현재 PC 내에 등록되어 있는 ODBC DSN 의 목록을 나타내며 이 중 사용하고자 하는 DSN 을	필수	

		선택하면 됩니다. 만약 존재하지 않는다면 "ODBC 편집"을 이용하여 추가할 수 있습니다.		
	사용자명	데이터베이스에 접속하기 위한 사용자명을 입력합니다.	필수	
	비밀번호	데이터베이스에 접속하기 위한 사용자의 비밀번호를 입력합니다.	필수	
저장할 테이블	DB 정보 가져오기	입력된 데이터베이스 접속정보를 이용하여 데이터베이스의 스키마, 테이블, 뷰 등의 정보를 읽어옵니다. 가져온 정보는 아래쪽에 표시됩니다.	버튼	
	테이블 생성	테이블을 새로 생성할 지의 여부를 나타냅니다. **'예'**로 선택될 경우에는 새로운 테이블을 생성하고, **'아니오'**로 선택되면 기존 테이블을 이용하여 저장됩니다.		예, 아니오
	스키마 선택	테이블 속성에 목록 될 데이터베이스 스키마를 선택합니다. 테이블과 뷰를 지원합니다.		All, Table, View
	저장할 테이블 명	DB 로 저장할 테이블 이름을 지정합니다.		
선택사항	테이블 제거	지정된 테이블이 존재할 경우 테이블을 제거하도록 설정합니다. 단 **테이블 생성** 옵션이 **'예'**일 경우에만 활성화 됩니다.	버튼	
	테이블 비우기	지정된 테이블 내의 모든 데이터를 삭제합니다. 만약 **'아니오'**로 선택될 경우는 기존 데이터에 이어서 저장하게 됩니다. **테이블 제거**와 마찬가지로 **테이블 생성** 옵션이 **'예'**일 경우에만 활성화 됩니다.		
	저장 방식	**변수 순서별** 혹은 **변수명 일치**를 선택할 수 있습니다. 이는 기존에 있던 테이블에 데이터를 추가적으로 입력할 때 Table 의 변수 순서와 추가하려는 데이터 Table 의 변수 순서가 일치하면 같은 변수로 볼지 혹은 변수명이 같을 때 같은 변수로 볼지를 선택하는 옵션입니다.		변수 순서별, 변수명 일치
변수정보	변수정보	DB 로 저장할 변수의 필드명 및 타입이 나타납니다.		

3.6.2 OLEDB 출력 노드

OLEDB 출력 노드는 OLEDB 기술을 이용하여 데이터베이스에 데이터를 저장하는 노드입니다. 테이블을 직접 선택하거나 새로이 테이블 명을 입력하여 해당하는 데이터를 저장할 수 있습니다. ODBC 및 OLEDB 를 모두 지원하는 데이터베이스라면 OLEDB 를 사용할 것을 권장합니다.

사용법



- 접속하려는 데이터베이스에 맞는 **드라이버**를 지정합니다.
- 데이터베이스에 접속하기 위하여 **DB 서버 접속정보, 사용자명, 비밀번호**를 입력합니다.
- **DB 정보 가져오기** 버튼을 눌러 지정된 데이터베이스의 정보를 가져옵니다.
- 저장하기 전 **저장할 변수** 부분에 나타난 데이터를 확인합니다.
- 기존에 존재하는 테이블에 저장할 경우 데이터 형이 맞지 않으면 에러가 발생합니다.
- 데이터를 저장할 테이블을 선택하거나 아니면 테이블 명을 입력합니다.

저장 테이블의 **테이블 생성** 옵션에서 **'예'**를 선택하면 기존 테이블 삭제 여부를 결정할 수 있고 **'아니오'**를 선택하면 기존 테이블 내용 지우기 여부를 결정할 수 있습니다.
이미 존재하는 테이블에 저장할 경우 **테이블 생성** 옵션에서 **'예'**를 선택한 후 테이블 삭제여부를 **'예'**로 선택하지 않으면 에러가 발생합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 입력합니다.	선택	
DB 연결	드라이버 선택	접속할 데이터베이스 종류에 맞는 OLE DB 드라이버를 선택합니다. ECMiner™은 이 드라이버를 통하여 데이터베이스에	필수	

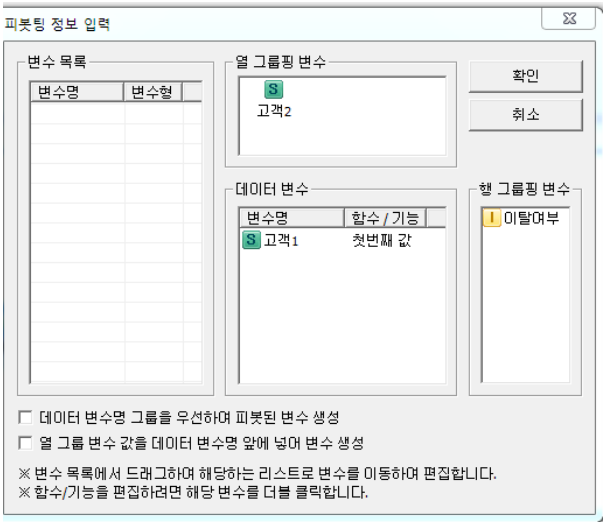
		접근합니다.		
	DB 서버 접속정보	데이터베이스를 인식할 수 있는 정보를 입력합니다. SQL 서버인 경우 데이터베이스 서버의 IP 어드레스를 입력하며, Oracle 인 경우 TNS 명을 입력합니다. 기타 데이터베이스인 경우 데이터베이스를 인식할 수 있는 특정 정보를 입력합니다.	필수	
	사용자명	데이터베이스에 접속하기 위한 사용자명을 입력합니다.	필수	
	비밀번호	데이터베이스에 접속하기 위한 사용자의 비밀번호를 입력합니다.	필수	
저장테이블	DB 정보 가져오기	입력된 데이터베이스 접속정보를 이용하여 데이터베이스의 스키마, 테이블 등의 정보를 읽어옵니다. 가져온 정보는 아래쪽에 목록이 생성 됩니다.	버튼	
	테이블 생성	테이블을 새로 생성할 것인지 아니면 기존 테이블을 그대로 사용할 것인지 여부를 결정합니다.	필수	예, 아니오
	스키마 선택	스키마를 선택합니다.		
	저장할 테이블 명	데이터를 저장할 테이블을 선택하거나 입력합니다.	필수	
선택사항	테이블 제거	**테이블 생성** 속성이 **'예'**일 경우 활성화 됩니다. 이미 존재하는 테이블의 내용 삭제여부를 결정합니다.	필수	예, 아니오
	테이블 비우기	**테이블 생성** 속성이 **'아니오'**일 경우 활성화 됩니다. 이미 존재하는 테이블의 내용 삭제 여부를 결정합니다.	필수	예, 아니오
	저장 방식	**변수 순서별** 혹은 **변수명 일치**를 선택할 수 있습니다. 이는 기존에 있던 테이블에 데이터를 추가적으로 입력할 때 Table 의 변수 순서와 추가하려는 데이터 Table 의 변수 순서가 일치하면 같은 변수로 볼지 혹은 변수명이 같을 때 같은 변수로 볼지를 선택하는 옵션입니다.		변수 순서별, 변수명 일치

변수정보	변수정보	저장될 변수 정보를 나타냅니다.		정수, 실수, 문자, 날짜

3.6.3 피봇 노드

피봇 노드는 행과 열을 그룹핑하여 데이터 변수를 화면상에 테이블로 보여주는 노드입니다.

사용법



▪ 피봇팅 정보 구성을 클릭합니다. 그러면 다음과 같은 창이 뜹니다.

▪ 데이터 변수를 선택하고 열 그룹핑변수와 행 그룹핑 변수를 지정합니다.

▪ 행과 열을 그룹핑하여 이탈 여부를 확인할 수 있습니다.

	1	2	3	4	5	6
	A4	이탈여부_M	이탈여부_MH	이탈여부_ML	이탈여부_L	이탈여부_H
1	A	1	1			1
2	B					
3	C					1

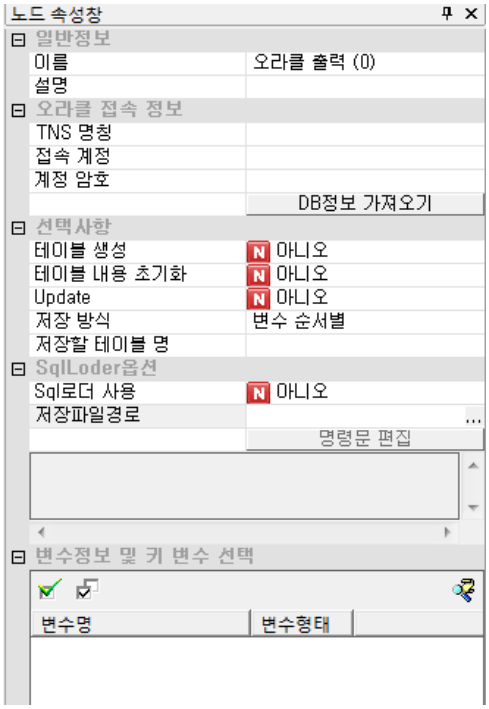
속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
설정사항	피봇팅 정보 구성	행과 열 그룹핑 변수를 지정하고 데이터 변수를 지정합니다.	필수	
	파일로 저장	파일로 저장할 것인지를 선택합니다.	선택	예, 아니오
	파일 경로	파일 경로를 지정하여 원하는 경로에 파일을 저장하도록 합니다.	선택	

3.6.4 오라클 출력노드

오라클 출력

오라클 출력 노드는 오라클 데이터를 데이터 베이스로 저장하는 노드입니다



사용법

- 데이터베이스에 접속하기 위하여 TNS 명칭, 접속 계정, 계정암호를 입력합니다.
- DB 정보 가져오기 버튼을 눌러 지정된 데이터베이스의 정보를 가져옵니다.
- 스키마를 선택하고, 데이터를 저장할 테이블명을 입력합니다.
- 만약 테이블 생성 옵션을 예로 선택한 경우, 테이블 제거 및 테이블 비우기의 옵션을 확인 또는 변경합니다.
- 실행 버튼을 누르면, 데이터가 DB 로 저장됩니다.

속성

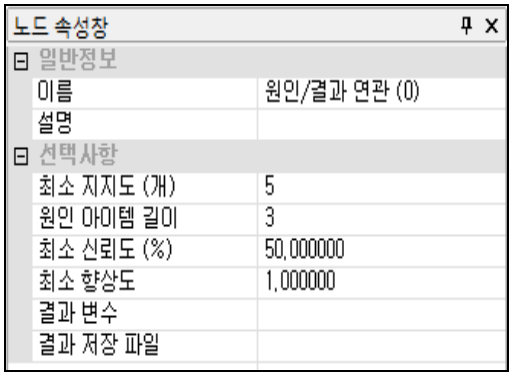
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
오라클 접속 정보	TNS 명칭	접속할 오라클 데이터베이스의 TNS 명을 입력합니다.	버튼	
	접속 계정	데이터베이스 접속 계정명을 입력합니다.	필수	
	계정 암호	데이터베이스 접속 계정의 암호를 입력합니다.	필수	
선택사항	DB 정보 가져오기	입력된 데이터베이스 접속정보를 이용하여 데이터베이스의 스키마, 테이블, 뷰 등의 정보를 읽어옵니다. 가져온 정보는 아래쪽에 표시됩니다.	버튼	
	테이블	테이블을 새로 생성할 지의 여부를 나타냅니다.		예,

	생성	예로 선택될 경우에는 새로운 테이블을 생성하고, 아니오로 선택되면 기존 테이블을 이용하여 저장됩니다.		아니오
	테이블 내용 초기화	테이블 내용을 초기화 합니다.		All, Table, View
	Update	선택된 키 변수와 일치하는 date 을 Update 합니다.		
	저장 방식	변수 저장 방식을 지정합니다.		변수 순서별, 변수명 일치
	저장할 테이블 명	저장할 테이블을 선택합니다.		
SqlLoader 옵션	Sql 로더 사용	Sql Loader 사용여부를 결정합니다.		
	저장파일경로	저장할 파일의 경로를 지정합니다.		
	명령문 편집	쿼리문을 편집합니다.		
변수정보	변수정보	DB 로 저장할 변수의 필드명 및 타입이 나타납니다.		

3.6.5 원인/결과 연관 노드

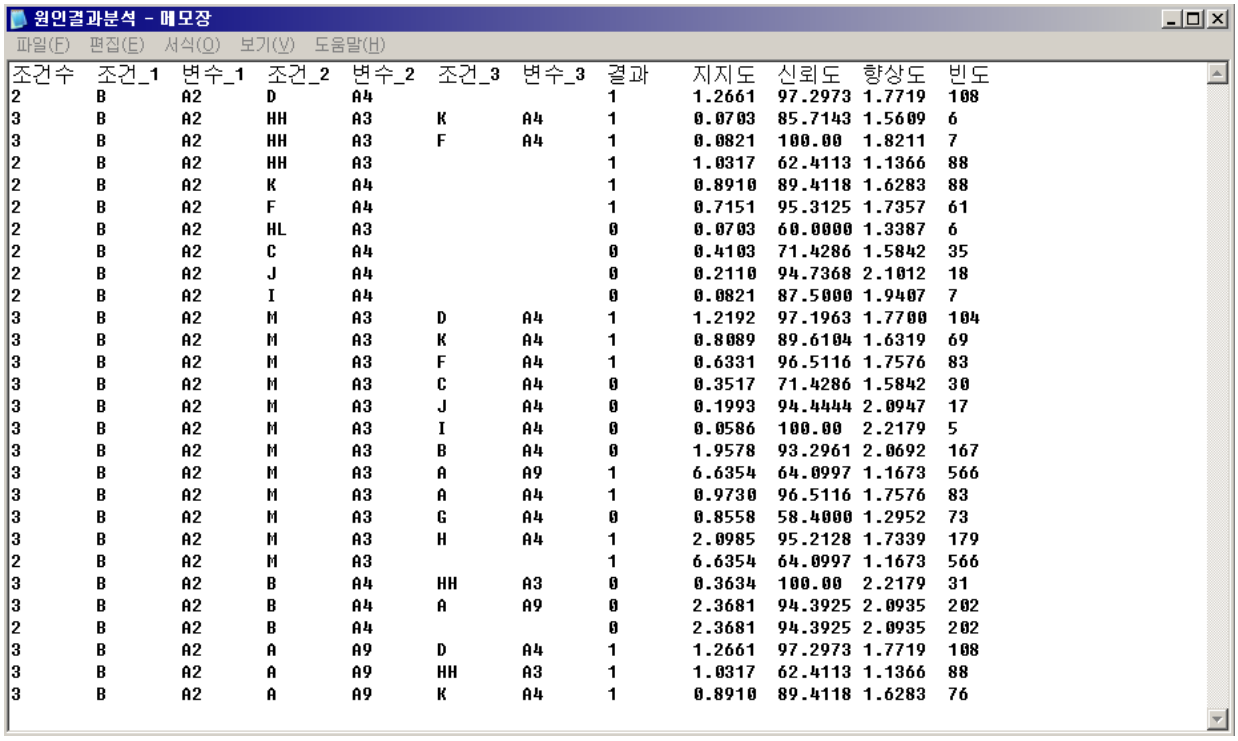
원인/결과 연관 노드는 이산형 데이터 변수를 이용하여 연관성을 파일로 출력하는 노드입니다.

사용법



- 분석에 사용되는 변수는 모두 이산형이어야 하며 결과변수와 결과저장파일을 필수로 선택하여야 합니다.

연관성 분석과 마찬가지로 지지도, 신뢰도, 향상도 등이 출력됩니다.



조건수	조건_1	변수_1	조건_2	변수_2	조건_3	변수_3	결과	지지도	신뢰도	향상도	빈도
2	B	A2	D	A4			1	1.2661	97.2973	1.7719	108
3	B	A2	HH	A3	K	A4	1	0.0703	85.7143	1.5609	6
3	B	A2	HH	A3	F	A4	1	0.0821	100.00	1.8211	7
2	B	A2	HH	A3			1	1.0317	62.4113	1.1366	88
2	B	A2	K	A4			1	0.8910	89.4118	1.6283	88
2	B	A2	F	A4			1	0.7151	95.3125	1.7357	61
2	B	A2	HL	A3			0	0.0703	60.0000	1.3387	6
2	B	A2	C	A4			0	0.4103	71.4286	1.5842	35
2	B	A2	J	A4			0	0.2110	94.7368	2.1012	18
2	B	A2	I	A4			0	0.0821	87.5000	1.9407	7
3	B	A2	M	A3	D	A4	1	1.2192	97.1963	1.7700	104
3	B	A2	M	A3	K	A4	1	0.8089	89.6104	1.6319	69
3	B	A2	M	A3	F	A4	1	0.6331	96.5116	1.7576	83
3	B	A2	M	A3	C	A4	0	0.3517	71.4286	1.5842	30
3	B	A2	M	A3	J	A4	0	0.1993	94.4444	2.0947	17
3	B	A2	M	A3	I	A4	0	0.0586	100.00	2.2179	5
3	B	A2	M	A3	B	A4	0	1.9578	93.2961	2.0692	167
3	B	A2	M	A3	A	A9	1	6.6354	64.0997	1.1673	566
3	B	A2	M	A3	A	A4	1	0.9730	96.5116	1.7576	83
3	B	A2	M	A3	G	A4	0	0.8558	58.4000	1.2952	73
3	B	A2	M	A3	H	A4	1	2.0985	95.2128	1.7339	179
2	B	A2	M	A3			1	6.6354	64.0997	1.1673	566
3	B	A2	B	A4	HH	A3	0	0.3634	100.00	2.2179	31
3	B	A2	B	A4	A	A9	0	2.3681	94.3925	2.0935	202
2	B	A2	B	A4			0	2.3681	94.3925	2.0935	202
3	B	A2	A	A9	D	A4	1	1.2661	97.2973	1.7719	108
3	B	A2	A	A9	HH	A3	1	1.0317	62.4113	1.1366	88
3	B	A2	A	A9	K	A4	1	0.8910	89.4118	1.6283	76

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	

선택사항	최소 지지도(개)	최소 지지도 조건을 입력합니다.	필수	
	원인 아이템 길이	최대 아이템 길이 조건을 입력합니다. 2 이상의 정수를 입력해야 합니다.	필수	
	최소신뢰도(%)	최소 신뢰도 조건을 입력합니다. 0 이상 100 미만의 값을 입력해야 합니다.	필수	
	최소향상도	최소 향상도(Lift) 조건을 입력합니다. 0 이상의 값을 입력해야 합니다	필수	
	결과 변수	결과 변수를 지정합니다.	필수	
	결과 저장 파일	결과 저장 파일을 지정합니다.	필수	

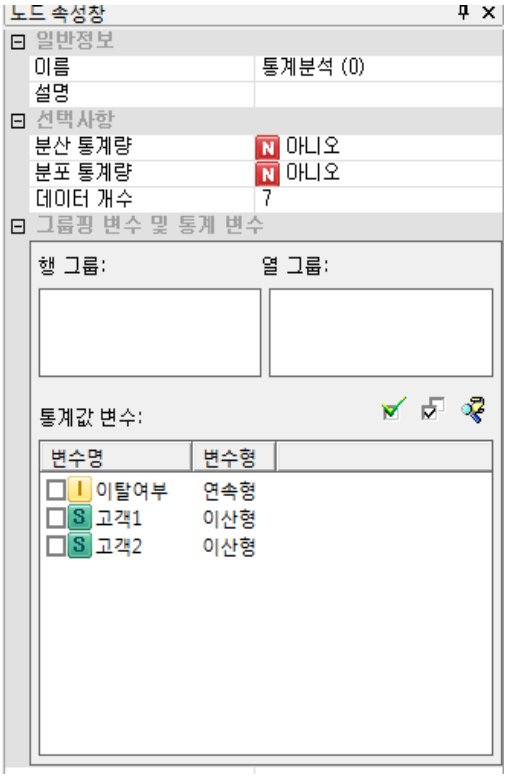
3.6.6 통계 분석 노드

통계 분석 노드는 데이터 통계 분석 결과를 화면에 출력하는 노드입니다.

사용법

통계 분석 노드는 다음과 같은 속성변경 인터페이스를 갖습니다.

그룹핑 변수 및 통계 변수 인터페이스는 통계값을 구할 변수를 선택하고, 행 혹은 열로 그룹핑 할 변수를 드래그하여 선택하게 합니다.



	선택된 변수를 사용하도록 설정합니다.
	선택된 변수와 선택되지 않은 변수들을 반전시킵니다.
	 버튼 클릭 시 왼쪽과 같은 창이 생성되어 사용자가 선택한 변수형 또는 변수명 조건에 일치하는 변수들을 선택합니다

사용법

- 선택사항에서 알고자 하는 통계량을 선택합니다.
- 분석할 변수를 선택하고 그룹핑 변수를 드래그합니다.
- 스트림을 수행합니다. 통계 분석 노드는 별다른 속성 지정 없이 사용 가능합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 입력합니다.	선택	
선택사항	분산통계량	연속형 변수의 분산, 표준편차, 첨도를 추가로 계산합니다..	필수	예, 아니오
	분포통계량	연속형 변수의 왜도, 중앙값, Q0, Q1, Q2, Q3, Q4, IQR 을 추가로 계산합니다.	필수	예, 아니오

	데이터 개수	이산형 변수의 값 중 큰 순서대로 몇 개를 표시할 지 지정합니다.	필수	
그룹핑 변수 및 통계 변수	행 그룹	행 기준으로 그룹핑 할 변수를 드래그 합니다.	기타 UI	사용자 인터페이스 참조
	열 그룹	열 기준으로 그룹핑 할 변수를 드래그 합니다.	기타 UI	사용자 인터페이스 참조
	통계값 변수	통계값을 구할 변수 필드를 선택합니다.	기타 UI	사용자 인터페이스 참조

결과

	NEW_CAR	0	1
HISTORY	통/변	AMOUNT	AMOUNT
0	총합	191,574	20,653
	평균	5,805.272727	2,950.428571
	최소값	426	950
	최대값	18,424	8,358
	범위	17,998	7,408
	결측치 수	0	0
	개수	33	7
1	총합	126,703	37,196
	평균	3,424.405405	3,099.666667
	최소값	339	697
	최대값	14,782	12,169
	범위	14,443	11,472
	결측치 수	0	0
	개수	37	12
2	총합	1,221,451	390,257
	평균	2,979.14878	3,252.141667
	최소값	343	276
	최대값	15,857	14,896
	범위	15,514	14,620
	결측치 수	0	0
	개수	410	120
	총합	322,840	55,789

3.6.7 파일 출력 노드

파일 출력

파일 출력 노드는 데이터 분석 결과를 파일로 저장하는 노드입니다.

사용법

- 데이터를 저장할 파일의 파일경로를 **파일경로** 속성에 지정합니다. 직접 입력할 수도 있으며 파일 대화상자를 통하여 지정할 수도 있습니다.
- 만약 변수명을 저장하지 않을 경우, **변수명 저장**을 **아니오**로 설정합니다.
- 만약 **덮어쓰기**로 저장할 경우, **저장 방법**을 **덮어쓰기**로 설정합니다.
- 만약 **구분자**를 **탭** 이외의 문자로 할 경우, **구분자** 콤보 박스에서 선택하고, 만약 원하는 구분자가 없을 경우 **기타**를 선택하고, **기타 구분자**에서 직접 입력합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
선택사항	변수명 저장	파일의 첫행에 변수명을 저장할 지 여부를 지정합니다. 변수명이 있다면 **'예'**로 설정합니다. **'아니오'**로 설정되면 변수명을 자동으로 생성합니다. 첫행에 변수명이 있는데 **'아니오'**로 설정할 경우 변수의 형태가 모두 문자형으로 설정될 수 있습니다.	선택	예, 아니오
	저장 방법	파일을 덮어쓸지, 이어 쓸지를 결정합니다. **덮어쓰기**를 선택하면, 기존의 파일 내용을 무시하고 새로운 내용을 쓰고, **파일 존재 시 오류**를 선택하면, 기존의 동일한 파일이 존재할 때 저장이	선택	덮어쓰기, 파일 존재 시 오류

		되지 않으며 오류가 발생합니다.		
	구분자	컬럼을 구분하는 구분자를 입력합니다. 기본적으로 (탭), (공백), |(바), ,(콤마), ;(세미콜론), **:(콜론)**을 지원하며 이들 이외의 구분자를 사용할 경우 **(기타)**를 선택합니다.	필수	(탭) (공백) '|' ',' ';' ':' (기타)
	기타 구분자	**구분자** 속성이 **(기타)**일 경우 활성 되며 기타 구분자를 입력합니다.	필수	
	문자열 따옴표로 묶음	파일에서 문자열을 따옴표로 묶을 지의 여부를 선택합니다.	필수	예, 아니오
파일	파일경로	저장할 파일의 경로와 파일명을 입력합니다. 직접 입력할 수도 있으며, (...) 버튼을 누르면 대화상자를 통하여 파일경로를 지정할 수 있습니다.	필수	
	FTP 사용	FTP 서버에 파일을 저장하여 올려놓을 것인지의 여부를 선택	필수	예, 아니오
FTP 연결	FTP 서버	사용할 FTP 서버를 선택합니다.		
	FTP 사용자	FTP 사용자를 입력합니다.		
	FTP 암호	FTP 사용자에 대한 암호를 입력합니다.		
	FTP 파일선택	FTP 파일을 선택합니다.		
	FTP 파일경로	FTP 서버의 어떤 경로에 파일을 저장할지를 입력합니다.		
변수 정보	변수 정보	파일로 저장할 변수의 필드명이 나타납니다.		

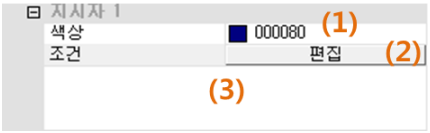
3.6.8 화면 표시 노드

화면 표시 노드는 데이터 분석 결과를 화면상에 테이블로 보여주는 노드입니다.

사용법

화면표시 노드는 별다른 속성 지정 없이 사용 가능합니다. 다만 특정 행을 지시하기 위한 지시자 추가를 위해서는, 아래와 같은 과정을 거치면 됩니다.

- 필요한 지시자 개수만큼 지시자 추가 버튼을 눌러 추가합니다. 다음과 같은 속성이 노드의 속성창에 추가됩니다.



- 지시자의 색상을 변경하려면 (1) 선택 시에 나타나는 색상 다이얼로그를 이용하여 색상을 변경합니다.
- (2)번 버튼을 누르면 수식 편집기가 실행됩니다. 수식 편집기를 이용하여 지시자에 대한 수식을 입력합니다.
- (3)을 통하여 입력된 수식을 확인할 수 있습니다. 그러나, (3)을 이용하여 직접 수식을 편집할 수 없습니다. 꼭 (2)의 편집 버튼을 사용하여야 합니다.

속성

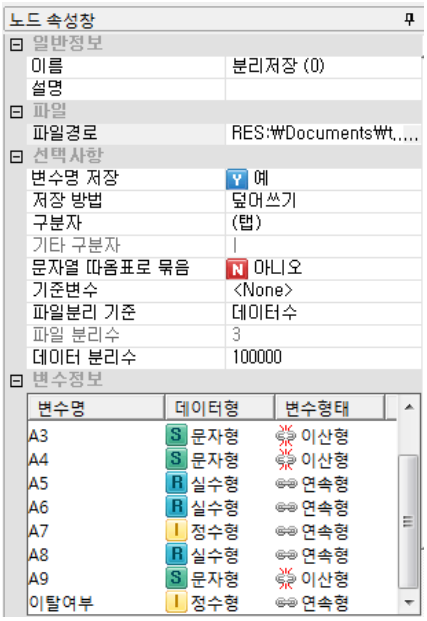
속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
선택사항	새로 만들기	테이블을 새로 만들지 여부를 지정합니다. 기본값인 **'아니오'**일 경우 실행 시에 하나의 출력테이블만 만들어지나, **'예'**로 설정되면, 실행 시마다 새로운 출력 테이블이 나타납니다.	선택(기본값: **아니오**)	예, 아니오
지시자	지시자 추가	특정 행을 다른 행들과 구분 짓기 위한 색깔 지시자를 추가하는 버튼입니다.	버튼	
	마지막 지시자 삭제	추가된 지시자 중, 마지막 지시자를 삭제합니다.	버튼	
변수 정보	변수 정보	테이블로 보여줄 변수의 이름과 타입이 나타납니다.		

3.6.9 분리 저장

분리 저장 노드는 사용자의 목적에 따라 데이터를 저장할 때 분리해서 저장하기 위해서 만든 노드입니다.

대량의 데이터를 처리할 때 대상 데이터를 한번에 처리할 수도 있지만 분리하여 처리하고 분리된 처리 결과를 합치는 방식으로 스트림을 구성하면 효율적인 프로젝트를 구성할 수 있으며, 이를 위하여 분리저장 노드를 사용할 수 있습니다. 참고로, ECMiner™의 경우 레코드 개수 기준으로 3000 여만건 이하의 데이터에서 가장 효율적으로 데이터를 처리합니다.

사용법



- **파일 경로**를 통해서 저장할 파일의 이름과 경로를 지정합니다.
- **변수명 저장**을 통해서 변수명을 저장할 지의 여부를 선택합니다.
- **저장 방법**을 통해서 파일 이름이 기존 파일과 같을 경우 기존 파일에 덮어쓸지 혹은 에러 메시지를 줄지를 선택합니다.
- **구분자, 기타 구분자**를 이용하여 구분자를 설정하고, 파일 분리 기준, 기준 변수 설정, 파일 분리수, 데이터 분리수를 설정하여 분리의 기준 및 세부 옵션을 설정합니다.

속성

속성그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
파일	파일경로	저장할 파일의 경로 및 이름을 지정합니다.	필수	
선택사항	변수명 저장	변수명을 저장할지 하지 않을지를 지정합니다.	선택(기본값: **예**)	예, 아니오
	저장방법	파일을 저장할 방식을 지정합니다. 덮어쓰기를 선택할 경우 같은 이름이 있더라도 덮어쓰고 파일 존재 시 오류를 선택할 경우 같은 이름의 파일이 있으면 오류 메시지를 줍니다.	선택(기본값: 덮어쓰기)	덮어쓰기, 파일 존재 시 오류
	구분자	컬럼을 구분하는 구분자를 입력합니다. 기본적으로 **(탭), (공백), |(바), ,(콤마), ;(세미콜론), :(콜론)**을 지원하며 이들 이외의 구분자를 사용할 경우 **(기타)**를 선택합니다.	필수	(탭) (공백) '|' ',' ';' ':' (기타)
	기타 구분자	**구분자** 속성이 **(기타)**일 경우 활성 되며 기타 구분자를 입력합니다.		
	문자열 따옴표로 묶음	문자열을 따옴표로 묶을지 여부를 선택합니다.	선택(기본: 아니오)	예 아니오
	기준변수	파일 분리의 기준이 되는 변수를 선택합니다.	필수	None 선택 가능
	파일 분리 기준	파일을 분리해서 저장할 기준을 선택합니다. 데이터 수, 파일 수, 기준 변수 범주 별을 선택할 수 있습니다.	필수	
	파일 분리수	파일 분리 기준이 파일 수일 경우 선택할 수 있습니다. 만약 기준 변수가 None 이면 파일 수에 맞게 파일을 분리해서 저장해 줍니다. 만약 기준 변수가 있으면 파일 수가 기준 변수보다 변수를 가르는 우선 순위를 가져서 지정한 파일 수 이하로 분리 저장하게	파일 분리 기준이 파일 수일 때 필수	

		됩니다.		
	데이터 분리수	파일 분리 기준이 데이터 수일 경우 선택할 수 있습니다. 기준 변수가 None 이면 단순하게 그 데이터 수에 해당하도록 파일을 분리해 저장합니다. 기준 변수가 있으면 데이터 수가 기준 변수보다 변수를 가르는 기준에서 우선 순위를 갖습니다. 예를 들어 기준 변수가 1,1,1,2,2,2,2,1,1,1,1 순서로 나타나고 데이터 수를 3 으로 하면 1,1,1 을 한 파일로 하고, 2,2,2,2 를 또 한파일로 합니다. 그리고 1,1,1,1 을 또 한파일로 합니다.	파일 분리 기준이 데이터 수일 때 필수	
변수 정보	변수 정보	테이블로 보여줄 변수의 이름과 타입이 나타납니다.		

<기준변수 설정되어 있을 경우 분리 저장 방법>

- 데이터 분리수로 분리 저장한 경우

예를 들어, 데이터 분리수가 10 으로 설정한 후 기준 변수를 A 라고 하면, 데이터 분리수 옵션이 우선순위가 높으므로, 데이터를 10 개부터 분리합니다. 그 후, A 변수의 10 번째 값과 11 번째 값을 비교하여 같으면 합쳐서 11 개의 데이터 셋을 구성하고, 다를 경우 10 개의 데이터를 분리 저장합니다. 분리 저장되면 다시 10 개 데이터를 순서대로 분리한 후 이와 같은 방법을 동일하게 적용하여 최종적으로 하나의 데이터파일을 여러 개로 분리 저장하게 됩니다.

- 파일 수로 분리 저장한 경우

파일 수를 설정하게 되면, 설정한 파일 분리수를 기준으로 전체 데이터를 나눌 데이터 분리수를 내부적으로 정의합니다. (예 – 100 개의 행으로 구성된 데이터를 6 개의 파일로 구분하려 할 때, 데이터 분리수는 17 개입니다) 데이터 분리수가 정의되면, 위의 데이터 분리수로 분리 저장하는 방법과 동일한 방법에 의해 데이터를 분리 저장합니다.

3.7 모델 평가 노드

모델 평가 노드는 모델링을 통해 생성된 모델의 예측 정확도를 비교, 평가하기 위한 노드입니다. 이들 **모델 평가 노드**를 통해 사용자는 각각의 자료를 가장 잘 예측할 수 있는 모델을 선택할 수 있습니다.

ECMiner™에서는 다음과 같은 모델 평가 노드들을 지원합니다.

- **ROC CHART**

  입력된 모델들의 ROC 곡선을 그립니다.

- **모델 평가**

  입력된 모델들을 비교할 수 있는 보고서를 생성합니다.

- **이익도표(Lift chart)**

  입력된 모델들의 이익도표를 그립니다.

3.7.1 ROC 차트 노드

ROC 차트 노드는 모델링을 통해 생성된 **분류모델**의 성능을 특이도와 민감도를 이용하여 차트를 그립니다. 이는 분류모델의 성능 평가 비교에 좋은 지표가 됩니다.

민감도(sensitivity, True positive rate)란 관심범주를 올바르게 분류할 확률을 말합니다.

민감도 = 관심범주 예측의 정분류 빈도 / 관심범주 빈도

그리고 **특이도(specificity, True negative rate)**는 비관심범주를 올바르게 분류할 확률을 말합니다.

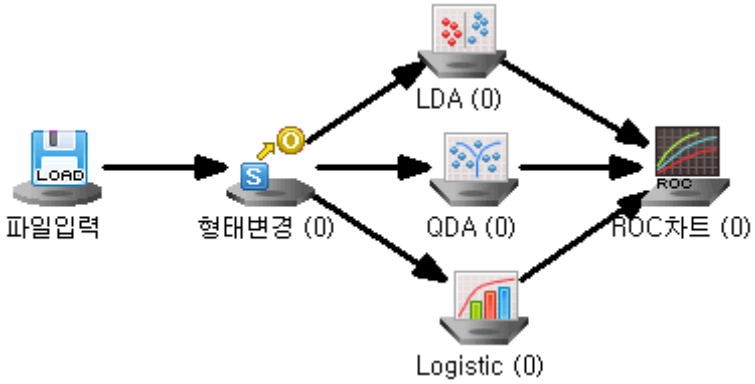
특이도 = 비관심범주 예측의 정분류 빈도 / 비관심범주 빈도

ROC 곡선은 분류 기준값(cutoff)을 0 에서 1 로 증가시키면서 {1-특이도, 민감도} 쌍을 그린 곡선입니다. ROC 곡선이 왼쪽 위 모퉁이에 가까울수록 좋은 성능의 모델임을 의미합니다.

고려사항

- 현재 분류모델 중 LDA, QDA, Logistic 모델 노드만 연결 가능합니다..

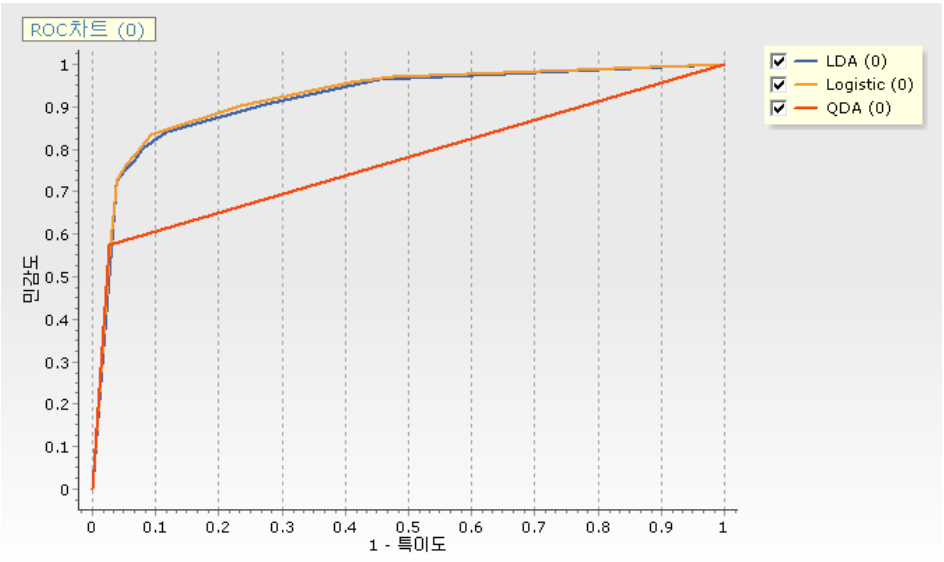
사용법 - Diagram 예시



속성

속성 그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
선택사항	관심등급	관심등급의 개수를 입력합니다.	필수	자연수
	분류 기준값 개수	분류 기준 값(cutoff)의 개수를 많이 선택할수록 여러 포인트에 대해서 민감도, 특이도를 구해 세밀한 ROC 곡선을 생성할 수 있습니다.	필수	

결과



ROC 곡선은 생성된 모델의 성능을 보여줍니다.

위의 경우, QDA **모델**이 특정 특이도에서 민감도가 가장 낮음을 알 수 있습니다. 이것은 이 모델이 해당 특이도에서 오분류율이 가장 높음을 의미해, 세 모델 중 가장 나쁜 모델임을 나타냅니다.

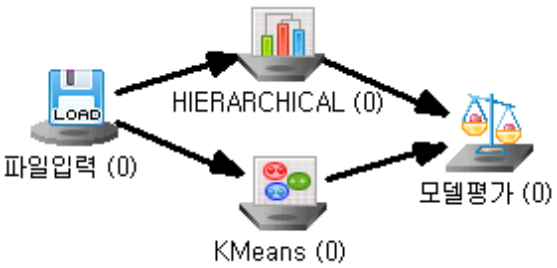
모델의 ROC 곡선이 정확하게 대각선으로 나타났다면 이는 모델링의 효과가 전혀 없음을 나타냅니다. 모든 분류 기준값에서 특이도와 민감도가 동일하여 오분류 행렬의 대각원소의 빈도가 비대각원소의 빈도와 동일하기 때문입니다.

3.7.2 모델평가 노드

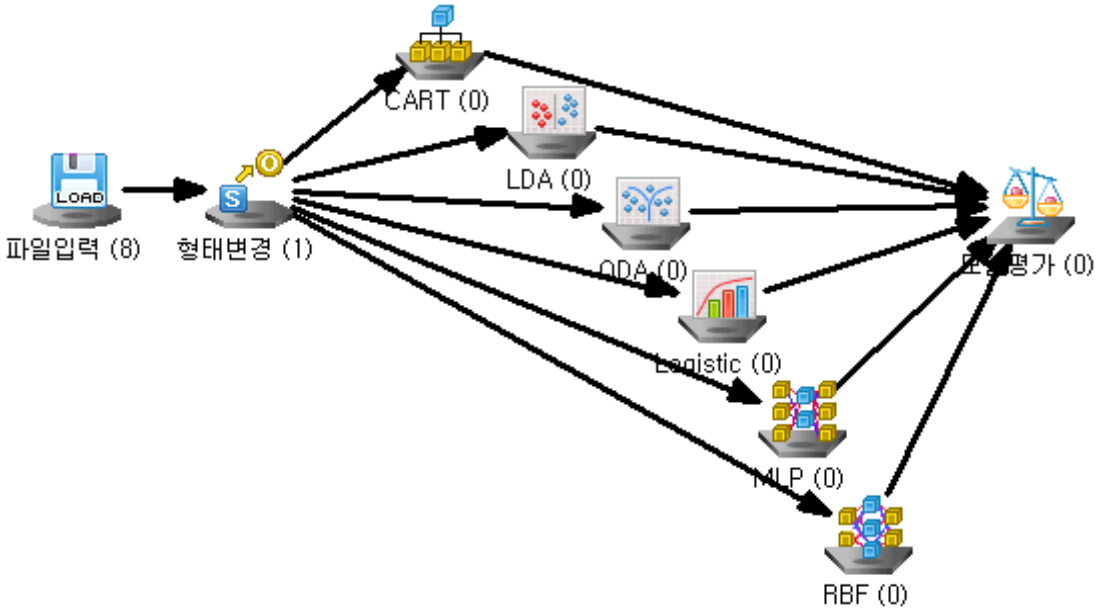
모델평가 노드는 분류 혹은 예측 정확도를 각 모델 별로 한눈에 볼 수 있도록 하여 쉽게 모델의 성능을 평가하게 합니다. 그리고 군집분석의 경우, 그 결과정보를 나열하여 사용자가 한눈에 볼 수 있게 합니다.

사용법 - Diagram 예시

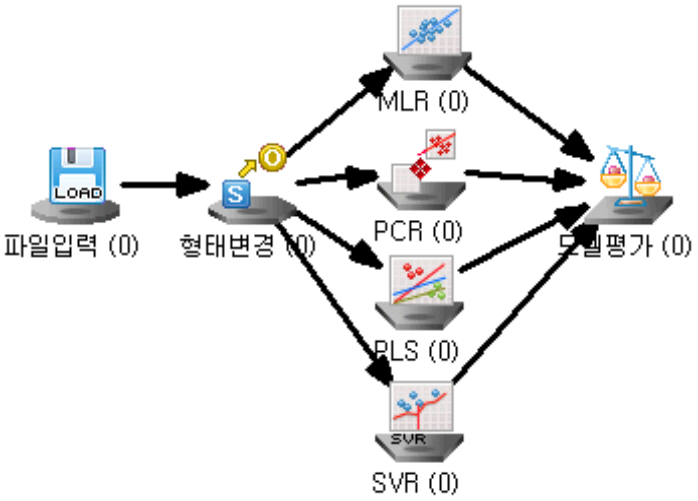
- **군집 분석**



- **분류 분석**



- **회귀 분석**



결과

- **군집 분석**

  아래와 같이 KMEANS, HIERARCHICAL 군집정보를 나열합니다.

HIERARCHICAL (0)

● 군집 내 원소 개수

Value	빈도수	Graph	백분율
군집1	1		1,22%
군집2	74		90,24%
군집3	3		3,66%
군집4	1		1,22%
군집5	3		3,66%
Total Count	82		

● 군집의 중심

변수	군집1	군집2	군집3	군집4	군집5	전체 중심
MAKE	BuickReatta	ToyotaCamry	JaguarXJSConvert	BMW750IL	LexusLS400	ToyotaCamry
VOL	50	100,66216	50	119	111,33333	98,80488
HP	165	102,17568	288,33333	295	239,66667	117,13415
MPG	23,60000	35,37027	19,50000	16,70000	17,96667	33,78171
sp	122	108,77027	157,33333	157	139,33333	112,41463
wt	40	29,45946	43,33333	45	46,66667	30,91463

● 군집간 거리

	군집2	군집3	군집4	군집5
군집1	83,30446	128,40908	151,52429	98,56074
군집2		200,06365	201,08482	143,35950
군집3			69,40586	80,42953
군집4				58,63679

KMeans (0)

● 군집 내 원소 개수

Value	빈도수	Graph	백분율
군집1	9		10.98%
군집2	15		18.29%
군집3	22		26.83%
군집4	18		21.95%
군집5	18		21.95%
Total Count	82		

● 중심

변수	군집1	군집2	군집3	군집4	군집5
MAKE	FordEscort	HondaCivic	ToyotaCamry	SubaruJusty	DaihatsuCharade
VOL	100.44444	91.53333	95	104.50000	103
HP	79.66667	97.53333	101.18182	118.16667	170.66667
MPG	40.17778	36.52000	37.82727	31.38333	25.75556
sp	103.66667	110.06667	108.86364	111.55556	123.94444
wt	25	27.16667	28.75000	32.50000	38.05556

● 군집간거리

거리	군집2	군집3	군집4	군집5
군집1	21.39298	23.21938	41.16460	95.27417
군집2		5.56708	25.51305	76.85703
군집3			21.01353	73.15664
군집4				54.53917

- **분류 분석**

  아래와 같이 각 모델 별로 오분류 표와 오분류 비율을 나열합니다.

CART (0)

● 모델링용 데이터 (Training Set)

	[예측] 0	[예측] 1
0	3064 (79.67 %)	782 (20.33 %)
1	251 (5.36 %)	4433 (94.64 %)

오분류 수: 1033
오분류율: 12.11%

LDA (0)

● 모델링용 데이터 (Training Set)

	[예측] 0	[예측] 1
0	2977 (77.41 %)	869 (22.59 %)
1	314 (6.70 %)	4370 (93.30 %)

오분류 수: 1183
오분류율: 13.87%

Logistic (0)

● 모델링용 데이터 (Training Set)

	[예측] 0	[예측] 1
0	3040 (79.04 %)	806 (20.96 %)
1	326 (6.96 %)	4358 (93.04 %)

오분류 수: 1132
오분류율: 13.27%

- **회귀 분석**

  아래와 같이 각 모델 별로 ANOVA 테이블과 표준에러 정보를 나열합니다.

MLR (0)

● 모델링용 데이터 (Training Set)

▶ ANOVA Table

	DF	SS	MS	F	p
모형(회귀)	17	34.57435	2.03379	5.82932	0.00020
잔차	19	6.62889	0.34889		
합계	36	41.20324			

▶ 표준에러 정보

Adjust R2	0.69517
R - square	0.83912
RMSE	0.59067
MAE	0.35802
MAPE	7.58555

SVR (0)

● 모델링용 데이터 (Training Set)

▶ ANOVA Table

	DF	SS	MS	F	p
모형(회귀)	17	34.60823	2.03578	5.86500	0.00019
잔차	19	6.59502	0.34711		
합계	36	41.20324			

▶ 표준에러 정보

Adjust R2	0.69673
R - square	0.83994
RMSE	0.58916
MAE	0.27829
MAPE	5.59351

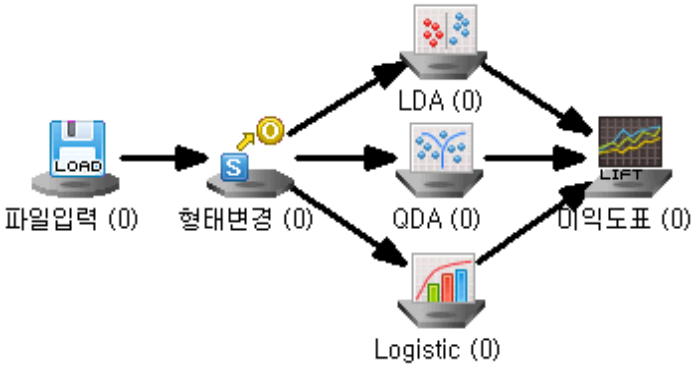
3.7.3 이익도표 노드

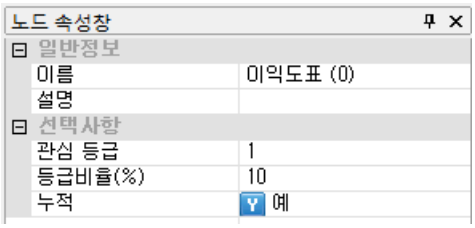
이익도표 노드는 사후 확률값을 이용하여 분류모델을 평가 하는 평가도구입니다. **이익도표 노드**는 **Captured Response, Response, Lift** 의 세 가지 값을 통해 모형 평가를 수행합니다.

고려사항

- 현재 분류모델 중 LDA, QDA, Logistic 모델 노드만 연결 가능합니다.

사용법 - Diagram 예시



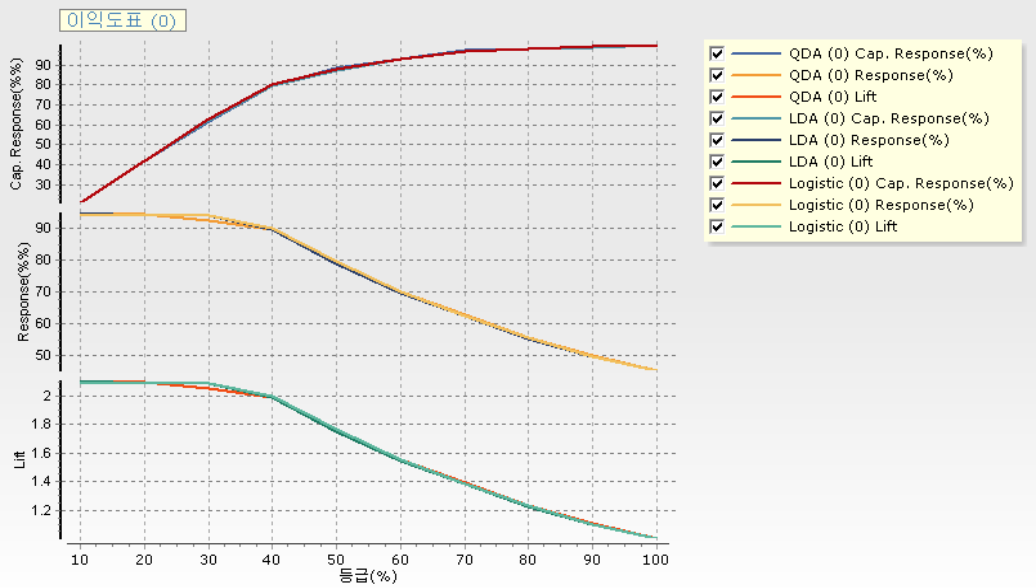


- 관심 등급을 통해 데이터를 몇 개의 관심등급으로 나눌지를 설정합니다.
- 등급 비율을 통해 각 등급당 전체 데이터의 몇 Percent 를 가질지를 설정합니다.
- 누적을 통해서 Captured Response 를 누적으로 보여줄지 아닐지를 선택합니다.

속성

속성 그룹	속성명	설명	기타	비고
일반정보	이름	노드의 이름을 입력합니다.	선택	
	설명	노드에 대한 간단한 주석을 달 수 있습니다.	선택	
선택사항	관심등급	데이터를 몇 개의 관심등급으로 나눌지 설정합니다.	필수	자연수
	등급 비율	각 등급당 데이터의 몇 Percent 를 가질지 설정합니다.	필수	
	누적	누적 Captured Response 를 가질 지의 여부를 선택합니다.	필수	예 아니오

결과



Captured Response 는 특정 범주 내에서 특정 집단이 차지하고 있는 점유율로 해석할 수 있습니다. 이 값이 높다는 것은 해당 집단에 특정 범주가 상대적으로 많이 밀집함을 의미합니다.

Response 는 각 집단 내에서 선택된 범주의 빈도와 집단 내 관찰치의 빈도의 비를 나타냅니다. 목표변수 내에서 **특정 범주**의 점유율을 각 집단에 대해 구한 값이라고 해석할 수 있습니다.

Lift 는 **특정 범주**가 각 집단에서 평균적으로 가지는 빈도와 해당 집단내의 **특정 범주**에 대한 비율입니다.
위와 같은 통계량 각각에 대하여 도식화한 것을 총칭하여 **이익도표**라고 합니다.

제 4 장 예제

ECMiner™를 효율적으로 활용하기 위한 예제를 보여줍니다. 사용자는 이 예제를 통해 **ECMiner™**의 활용법을 익힐 수 있습니다.

다음과 같은 예제 내용들이 있습니다.

- **고객 세분화 분석**

  고객 세분화 과정을 단계별로 나타내어 줍니다.

- **고객 이탈 예측 모델**

  이탈 고객 자료를 통해 향후 잠재 이탈 고객에 대한 분류와 그 확률을 구하는 과정을 나타내어 줍니다.

- **상품 연관성 분석**

  구매한 제품 사이의 연관성을 분석함으로써 동시에 구매되는 관계(즉, 관련성이 있는 상품)를 알아냅니다.

- **조업 편차 분석**

  생산공정에서 설비간의 조업 편차 유무와 그 요인을 찾아내어 일정한 품질을 유지하고자 합니다.

4.1 고객 세분화 분석

분석 목적

카드 발급 시 특성이 비슷한 고객의 세분화로 카드 발급의 효율성을 높이고자 합니다.

분석 방법

전처리 과정을 통하여 데이터의 특성을 파악합니다.

K-Means 클러스터링 방법을 사용하여 고객을 세분화합니다.

데이터

고객 세분화 분석	
예제 프로젝트 파일	'고객세분화.ecm'
데이터 파일	'고객세분화.txt'

데이터의 설명

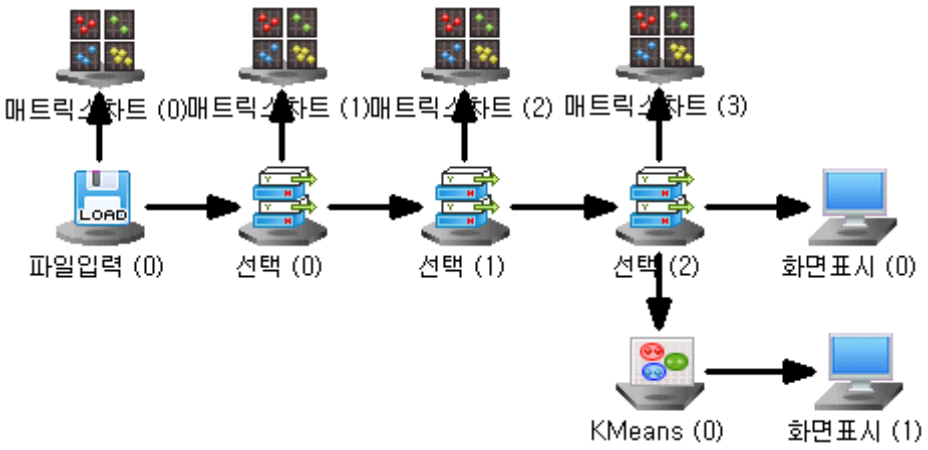
카드 발급용 신청서의 내용으로 의미 있는 변수는 변수변환의 과정을 거쳤습니다. 총 15 개의 변수로서 9 개의 범주형 변수와 6 개의 연속형 범주로 구성되어 있습니다.

변수명	데이터 형태
FIELD1	범주형(b, a) 변수
FIELD2	연속형 변수
FIELD3	연속형 변수
FIELD4	범주형(u, y, l, t) 변수
FIELD5	범주형(g, p, gg) 변수
FIELD6	범주형(c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff) 변수
FIELD7	범주형(v, h, bb, j, n, z, dd, ff, o) 변수
FIELD8	연속형
FIELD9	범주형(t, f) 변수
FIELD10	범주형(t, f) 변수
FIELD11	연속형
FIELD12	범주형(t, f) 변수

FIELD13	범주형(g, p, s) 변수
FIELD14	연속형
FIELD15	연속형

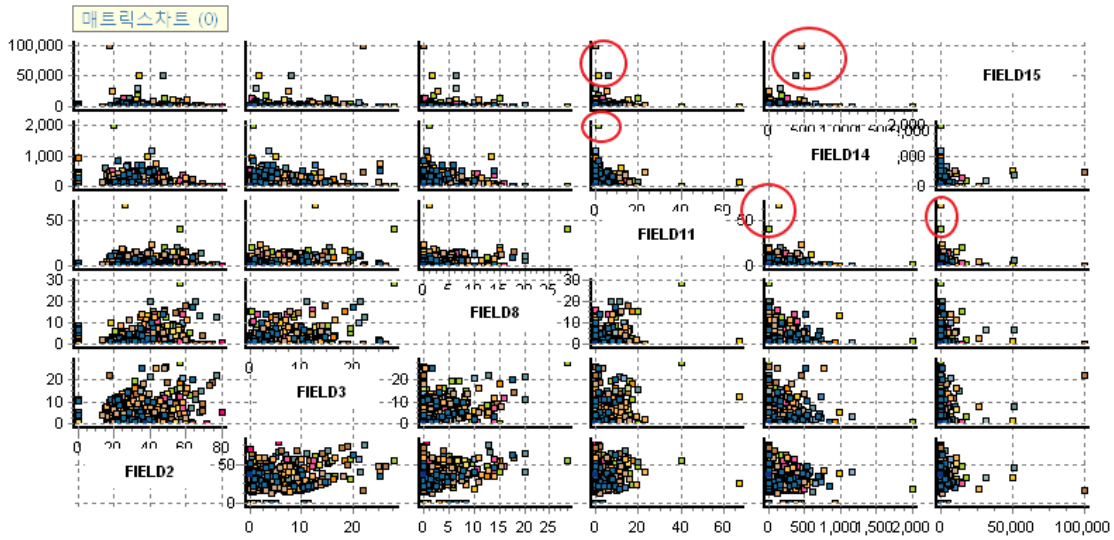
필요 NODE 설명 및 처리 방법

고객 세분화 모형 분석 시 전체 필요 노드



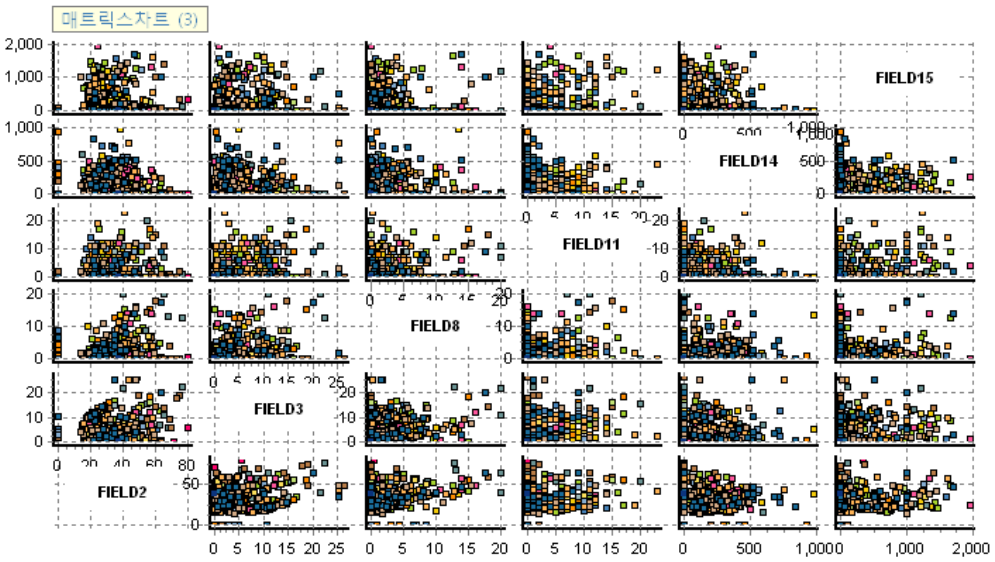
데이터의 전체적인 형태를 알아 보기 위해서 차트 노드의 매트릭스 차트를 이용하여 살펴봅니다.

그림에서 FIELD11, FIELD14, FIELD15 의 일부 점들이 바깥쪽(outlier)으로 치우쳐져 있는 것을 볼 수 있습니다.



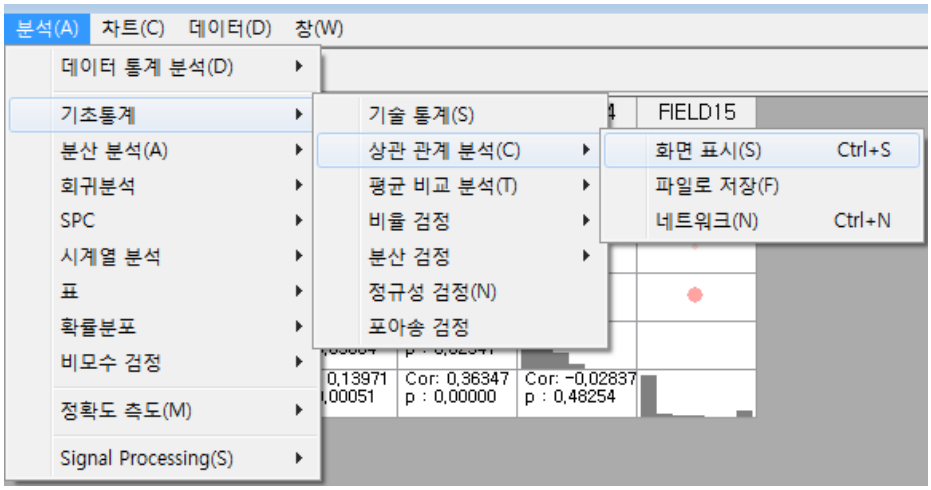
바깥쪽으로 치우쳐진 점들은 모형 설정 시 많은 영향을 주므로 제거하는 것이 좋습니다. 전처리 노드에서 '선택' 노드로 이상치(outlier)들을 제거해 줍니다.

'FIELD11'에서 40 보다 큰 수를, 'FIELD14'에서는 1000 보다 큰 수를, 'FIELD15'에서는 2000 보다 큰 수를 제거하고 매트릭스 차트를 그려본 것입니다.



위 그래프에서 보듯이 이상치(outlier)가 없는 것을 볼 수 있습니다.

이상치(Outlier)가 제거된 데이터의 상관관계를 살펴보기 위하여 출력노드의 '화면표시' 에서 [분석] ➔ [기초통계] ➔ [상관관계분석(C)]➔[화면표시]를 클릭합니다.



	FIELD2	FIELD3	FIELD8	FIELD11	FIELD14	FIELD15
FIELD2						
FIELD3	Cor: 0,18055 p : 0,00001					
FIELD8	Cor: 0,34192 p : 0	Cor: 0,24325 p : 0,00000				
FIELD11	Cor: 0,19244 p : 0,00000	Cor: 0,23752 p : 0,00000	Cor: 0,28023 p : 0,00000			
FIELD14	Cor: -0,08696 p : 0,03107	Cor: -0,18647 p : 0,00000	Cor: -0,01611 p : 0,69004	Cor: -0,09013 p : 0,02541		
FIELD15	Cor: 0,08319 p : 0,03917	Cor: 0,12616 p : 0,00172	Cor: 0,13971 p : 0,00051	Cor: 0,36347 p : 0,00000	Cor: -0,02837 p : 0,48254	

모든 변수들 간의 상관관계가 낮게 나오는 것을 볼 수 있습니다.

데이터의 전처리 후 고객의 세분화를 하기 위하여 모델링 노드에 있는 'KMEANS' 알고리즘을 사용합니다. 이 모델은 종속변수(Y)값이 존재하지 않을 때 비슷한 속성에 있는 그룹끼리 묶는 통계적인 방법입니다. 생성된 KMEANS 모델 노드를 클릭하면 'General Info'에서 기본적인 변수의 정보와 'Model info'에서 세분화 정보들이 나타나는 것을 볼 수 있습니다.

10	11	12	13	14	15	16	17
FIELD10	FIELD11	FIELD12	FIELD13	FIELD14	FIELD15	KM1_YHAT	KM1_DISTANCE
t	1	f	g	202	0	1	1,19785
t	6	f	g	43	560	2	0,39225
f	0	f	g	280	824	2	1,09203
t	5	t	g	100	3	4	1,51847
f	0	f	s	120	0	5	1,72044
f	0	t	g	360	0	4	0,85728
f	0	f	g	80	1,349	2	1,76906
f	0	f	g	180	314	3	2,18408
f	0	t	g	52	1,442	4	1,99213
f	0	t	g	128	0	5	1,42067
f	0	f	g	260	200	4	1,42440
f	0	t	g	0	0	4	1,33301
t	7	t	g	0	0	2	1,45937
t	10	t	g	320	0	1	2,09591
t	3	t	g	396	0	4	1,34081
t	10	f	g	120	245	2	1,03048
f	0	t	g	0	0	4	1,45781

출력노드의 '화면표시'를 이용하여 각 변수들이 어느 집단으로 분류되었는지를 알아볼 수 있으며(KM_YHAT) 중심으로부터의 거리 (KM_DISTANCE)도 알 수 있습니다.

고객세분화는 종속변수(y 변수)가 없는 경우 사용하는 방법입니다. 클러스터링 방법을 사용하며 클러스터링 방법에는 HIERARCHICAL, K-means 방법 등이 있습니다. K-Means 의 경우에는 마지막 결과물 중 'KM-YHAT'값이 그룹을 분류한 값이 됩니다.

4.2 고객 이탈 예측 분석

분석 목적

고객의 이탈 가능성을 예측하는 모델을 만들어 차후 이탈 고객을 분류하고 이탈 확률을 알아보고자 합니다.

분석 방법

- 전처리 과정을 통하여 데이터의 특성을 파악합니다.
- Logistic, LDA, QDA 방법을 사용하여 모델을 생성 후 비교합니다.
- 비교된 모델 중 가장 좋은 모델로 이탈 고객을 예측합니다.

데이터

고객이탈 모델	
예제 프로젝트 파일	'고객이탈.ecm'
데이터 파일	Model data: '고객이탈_model.txt' Test Data: '고객이탈_testset.txt'

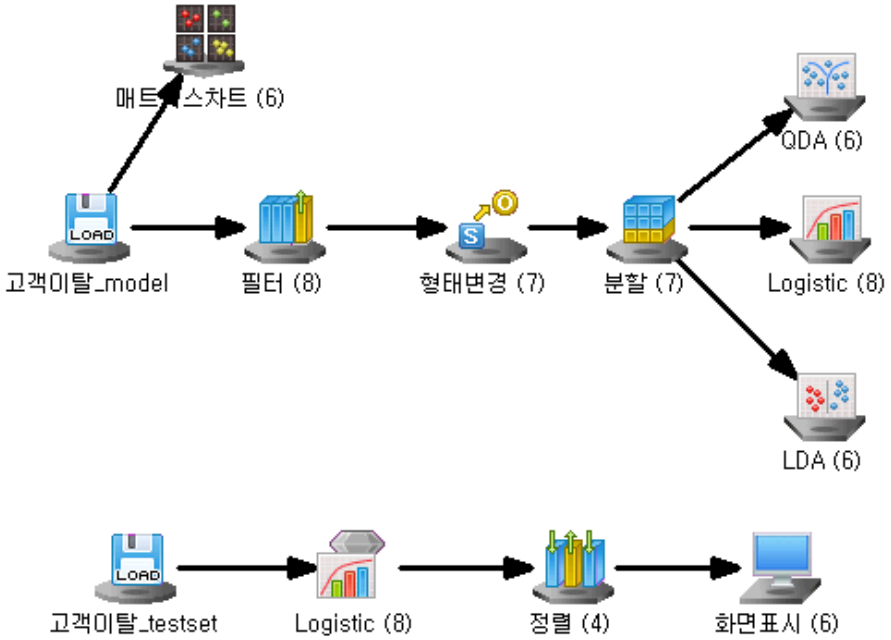
데이터의 설명

데이터 파일은 어느 회사 고객의 특성을 나타낸 자료입니다.

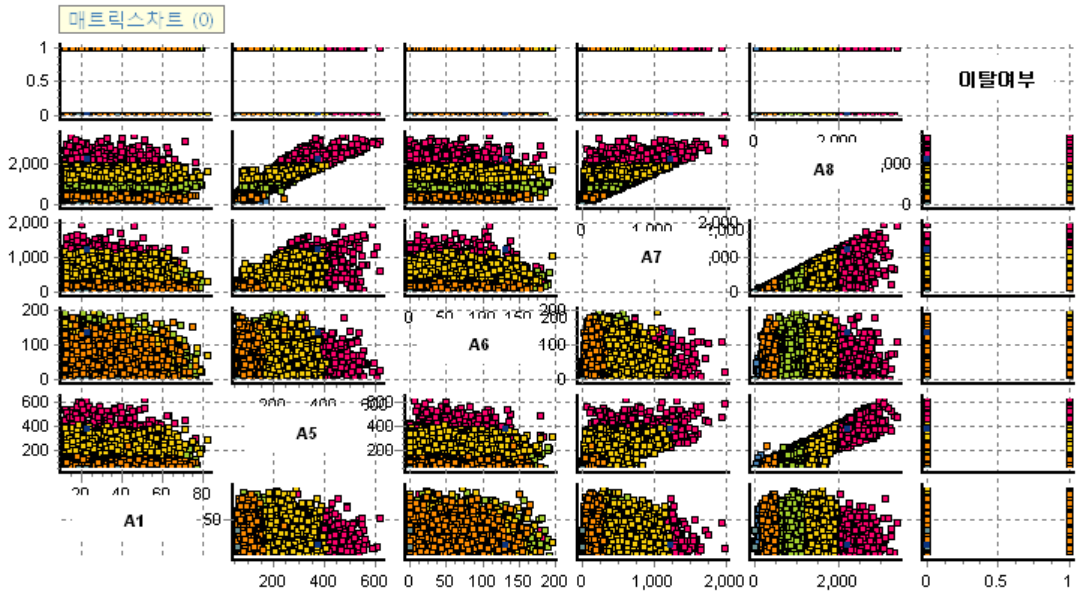
변수명	데이터 형태
A1	연속형 변수
A2	범주형(A, B, C, D, E) 변수
A3	범주형(M, MH, ML, H, L, N) 변수
A4	범주형(G, C, B, H, E, K, A, 기타 4 개) 변수
A5	연속형 변수
A6	연속형 변수
A7	연속형 변수
A8	연속형 변수
A9	범주형(A, B, C, D) 변수
이탈여부	0,1 의 이산형 변수

필요 NODE 설명 및 처리 방법

- 고객 이탈 모형 분석 시 전체 필요 노드

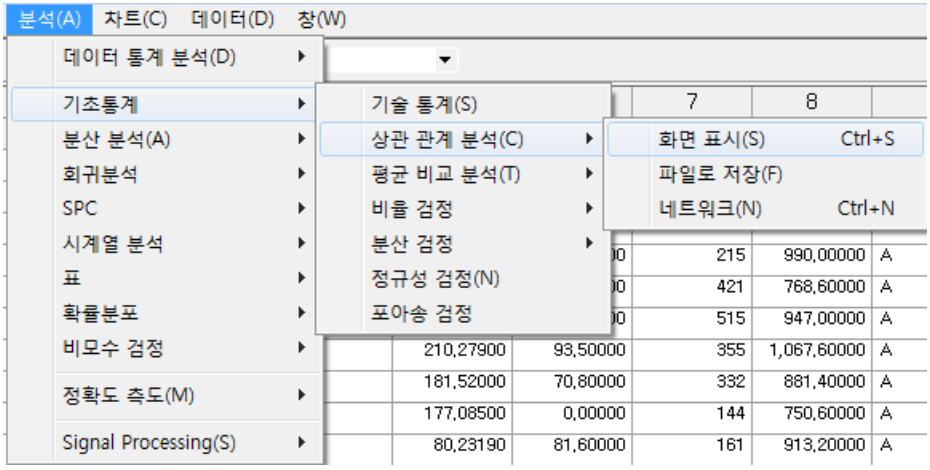


▪ 매트릭스 차트로 전체 데이터의 형태를 파악합니다.

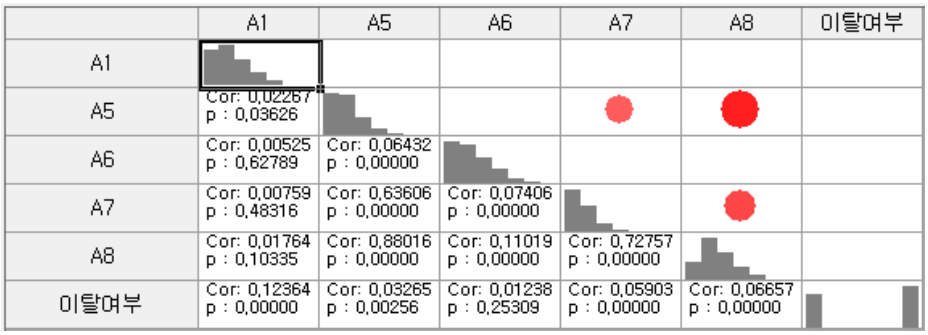


위의 그림은 매트릭스 차트를 나타낸 것입니다. 매트릭스 차트는 데이터간의 상관관계를 그래프로 나타낸 것입니다. 위 그림에서 'A8'변수가 'A5'변수 그리고 'A7'변수와 대략적인 선형관계를 가짐을 알 수 있습니다.

변수들 간의 상관관계를 자세히 알아보기 위하여 '파일입력' 노드를 더블 클릭하면 아래와 같은 창이 나오게 됩니다.

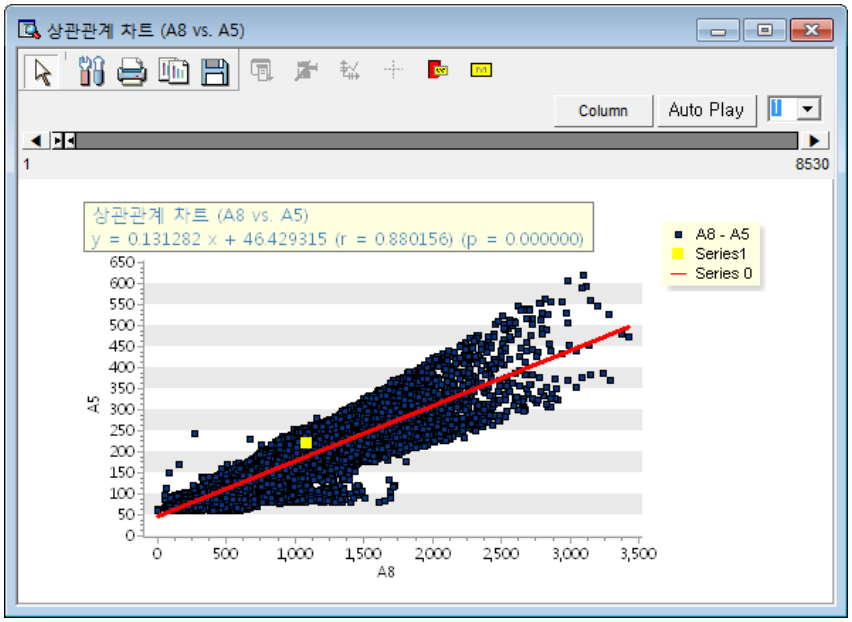


위와 같은 창에서 변수들 간의 간단한 정보를 볼 수 있습니다. 먼저 변수들 간의 상관관계를 화면으로 보기 위해서 위와 같이 [분석] ➔ [기초통계] ➔ [상관관계분석(C)] ➔ [화면표시]를 클릭합니다. 그러면 아래와 같은 창이 생성됩니다.

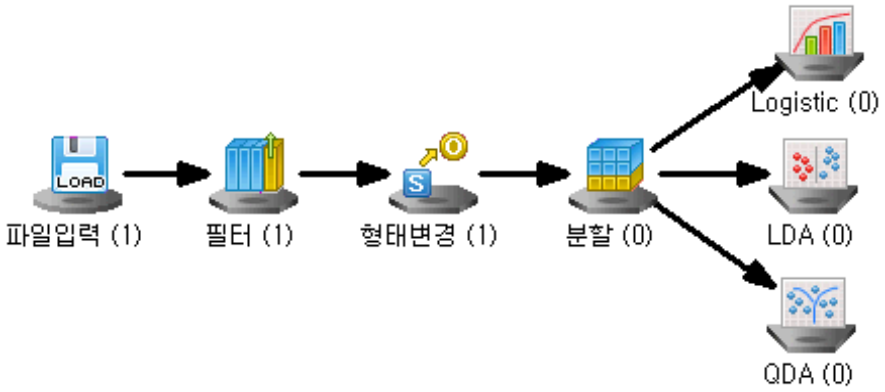


변수들간의 상관관계와 p-value 값이 나와 있습니다. 여기서도 'A8'변수가 'A5'변수 그리고 'A7'변수와 상관계수가 0.880, 0.728 으로 높게 나타나는 것을 볼 수 있습니다.

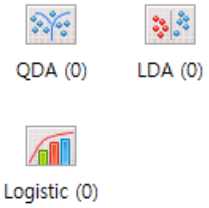
'A8'과 'A5'이 만나는 셀을 클릭하게 되면 두 변수들간의 기본차트가 나타나게 되어 두 그룹간의 상관 관계를 눈으로 확인 가능하게 됩니다.



변수들간의 상관관계가 높은 것은 다중 공선성이 생기게 됩니다. 이런 경우, 높은 상관관계를 갖는 변수들 중 일부 변수만을 사용하여 모형을 세워도 예측력에서는 별 차이가 없습니다. 그러므로 여기에서 'A8'을 제거합니다. 필터에서 'A8'을 제거하고 '형태변경'노드에서 target 변수가 될 '이탈여부' 변수를 종속변수로 지정합니다. 그리고 '이탈여부' 변수를 이산형 변수로 지정하여 줍니다. 모형의 예측력을 평가하기 위하여 '분할'노드를 이용하여 모형을 만드는 'Training Set'과 생성된 모형을 테스트 할 수 있는 'Test Set'을 7:3 의 비율로 나누어 줍니다.



각각의 방법(Logistic, LDA, QDA)을 통하여 3 가지의 예측 모형을 만듭니다. 각각의 예측 모형이 왼쪽편의 '생성된 모델' 창에 나타나는 것을 볼 수 있습니다.



예측모형의 자세한 정보를 보기 위하여 각 생성된 모델을 프로젝트 창으로 가져옵니다. 프로젝트 창에서 자세한 정보를 볼 수 있습니다.

새로 만들어진 모델들을 비교하기 위하여 각각의 '평가용 데이터'의 오분류 확률을 사용합니다.

Logistic

● 모델링용 데이터 (Training Set)

	[예측] 0	[예측] 1
0	3037 (78.97 %)	809 (21.03 %)
1	324 (6.92 %)	4360 (93.08 %)

오분류 수: 1133
오분류율: 13.28%

LDA

● 모델링용 데이터 (Training Set)

	[예측] 0	[예측] 1
0	2981 (77.51 %)	865 (22.49 %)
1	312 (6.66 %)	4372 (93.34 %)

오분류 수: 1177
오분류율: 13.80%

QDA

● 모델링용 데이터 (Training Set)

	[예측] 0	[예측] 1
0	2214 (57.57 %)	1632 (42.43 %)
1	129 (2.75 %)	4555 (97.25 %)

오분류 수: 1761
오분류율: 20.64%

세 모형 중 Logistic 모형을 이용한 모델이 가장 예측력이 높은 것을 알 수 있습니다.

생성된 모델을 이용하여 예측을 하기 위하여 먼저 예측할 데이터를 불러옵니다. 생성된 모델 중 가장 예측력이 좋은 Logistic 모형을 이용하여 새로 들어오는 고객의 데이터를 분석합니다. 그리고 그 결과로 고객의 이탈 여부를 예측합니다. 예측된 데이터를 이탈여부가 높은 순서대로 정렬합니다.

	4	5	6	7	8	9	10	11
	A4	A5	A6	A7	A9	이탈여부	LRN3_YHAT	LRN3_POS
	B	181,39500	42,00000	213	A	0	0	0,97883
	A	218,38000	191,40000	554	A	1	1	0,85442
	G	166,95200	128,40000	215	A	0	0	0,69361
	H	137,70700	102,00000	421	A	1	1	0,97468
	D	131,15100	36,50000	515	A	1	1	0,98173
	A	177,08500	0,00000	144	A	1	1	0,92423
	B	80,23190	81,60000	161	A	1	0	0,96647
	H	199,45000	78,00000	478	A	1	1	0,94378
	H	205,54300	22,80000	406	A	1	1	0,96295
	G	182,79500	24,00000	485	A	1	0	0,56525
	A	156,26800	28,80000	202	A	1	1	0,94536
	K	100,79200	77,50000	527	A	1	1	0,90048
	B	99,35700	1,80000	531	A	0	0	0,94255
	F	96,22250	0,00000	679	A	1	1	0,98742
	B	92,01710	3,50000	361	A	0	0	0,96268
	D	127,64100	94,00000	544	A	1	1	0,98989
	F	111,21000	2,00000	729	A	1	1	0,98703
	K	79,26930	91,80000	287	A	1	1	0,90285

'LRN3_YHAT' ==> 고객의 이탈 여부 추정 값

'LRN3_POS' ==> 이탈 여부에 대한 사후확률

고객 이탈 예측 값과 이탈 확률 값을 가지고 여러 가지 마케팅이나 이벤트를 할 수 있습니다.

고객 이탈분석은 데이터 마이닝 기법 중 Classification 방법을 사용할 수 있습니다. Classification 방법은 교사 학습(supervised learning)으로 종속변수(y 값)와 독립변수(x 값)가 존재 하여야 됩니다. 이러한 예측 분석은 이전의 데이터를 이용하여 향후에 들어올 고객들의 성향을 예측하는 방법이며 마케팅 전략 수립에 중요한 자료가 될 수 있습니다.

4.3 상품 연관성 분석

분석 목적

'어떤 제품들이 동시에 구매되는가?' 를 알고자 할 경우, 구매한 제품 사이의 연관성을 분석함으로써 동시에 구매되는 관계(즉, 관련성이 있는 상품)를 알아내어 이를 묶음이나 교차판매(Cross Selling)를 위한 정보로 활용하기 위한 방법입니다. 주로 '장바구니 분석(Market Basket Analysis)'이라고도 불립니다.

분석 방법

모델링 노드에 있는 '연관성 분석' 사용합니다.

데이터

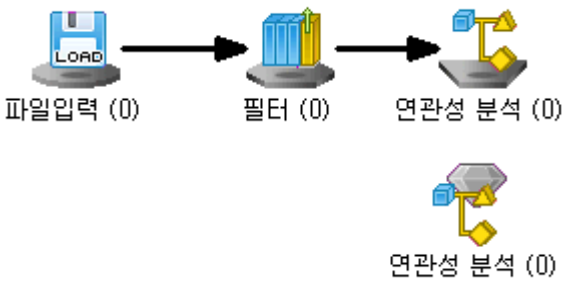
상품 연관성 분석	
예제 프로젝트 파일	'상품연관성.ecm'
데이터 파일	'상품연관성.csv'

데이터의 설명

각 고객별로 사과, 배, 감, 귤, 대추, 바나나, 파인애플의 제품 구매를 구매 시 '1', 비구매 시 '0'으로 표시, 연관성 분석에 쓰이는 알고리즘인 'APRIORI' 방법을 적용하기 위해서는 구매, 비구매 형식으로만 데이터가 구성되어야 합니다.

필요 NODE 설명 및 처리 방법

- 연관성 분석 시 전체 필요 노드



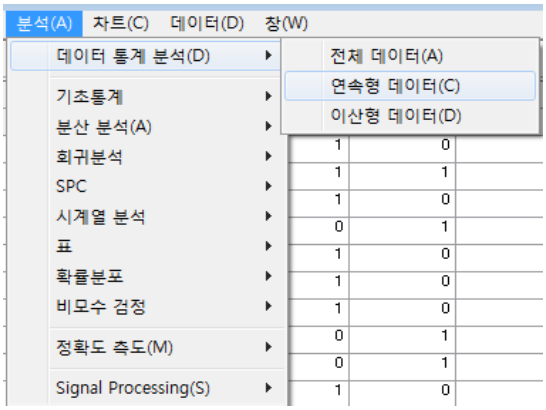
상품 연관성 관련 데이터로 엑셀 파일과 csv 파일이 있습니다.

엑셀 파일을 이용할 경우 데이터의 형태를 알아보기 위하여 '엑셀 데이터'노드를 더블 클릭합니다. csv 파일을 이용할 경우는 '파일 입력'을 더블 클릭합니다.

	1	2	3	4	5	6	7	8
	고객	사과	배	감	귤	대추	바나나	파인애플
1	1	1	1	1	0	0	0	0
2	2	1	1	1	1	1	1	1
3	3	0	1	1	0	0	0	0
4	4	0	0	0	1	1	1	0
5	5	0	0	1	0	0	1	1
6	6	0	1	1	0	1	0	1
7	7	1	0	1	0	1	0	1
8	8	0	1	0	1	0	1	0
9	9	1	0	0	1	0	0	0
10	10	0	0	1	0	0	0	0
11	11	1	1	1	0	1	0	1
12	12	1	0	1	0	1	0	0
13	13	1	1	1	1	1	0	0
14	14	1	1	1	0	0	1	0
15	15	0	0	1	1	1	1	1
16	16	0	1	0	1	1	1	0
17	17	0	1	1	1	0	0	0

'연관성 분석' 방법은 구매, 비구매 형식으로 0, 1 로만 인식을 합니다.

더 자세한 파일 형식을 알아보기 위하여 '파일입력'창의 '분석(A)' ➔ '데이터 통계분석(D)' ➔ '연속형 데이터(C)'를 클릭합니다.



변수명	변수형	데이터수	결측치수	총합	평균	박스플롯	최소값	최대값	범
고객	연속형	65,535	0	2,147,450,880	32,768,00000		1	65,535	
사과	연속형	65,535	0	19,664	0,30005		0	1	
배	연속형	65,535	0	37,136	0,56666		0	1	
감	연속형	65,535	0	45,876	0,70002		0	1	
귤	연속형	65,535	0	41,502	0,63328		0	1	
대추	연속형	65,535	0	30,584	0,46668		0	1	
바나나	연속형	65,535	0	26,214	0,40000		0	1	
파인애플	연속형	65,535	0	24,030	0,36667		0	1	

고객 번호를 제외한 다른 변수들이 0, 1 로 구성되어 있는 것을 볼 수 있습니다. 연관성 분석에서 '고객번호'는 필요가 없으므로 '필터' 노드를 통하여 변수를 걸러냅니다.

연관성 분석을 시행 후 '모델보고서'에는 변수들의 일반적인 정보와 모델 정보가 나타납니다. '모델정보'에는 'Set 정보'와 '규칙정보'가 나타납니다.

순번	연관규칙	아이템 수	신뢰도(%)	향상도	지지율(%)
1	[사과] [배] --->[감]	3	100,00	1,43	20,00
2	[사과] [배] --->[대추]	3	50,00	1,07	10,00
3	[사과] [감] --->[배]	3	75,00	1,32	20,00
4	[사과] [감] --->[대추]	3	62,50	1,34	16,67
5	[사과] [귤] --->[배]	3	66,67	1,18	6,67
6	[사과] [귤] --->[감]	3	66,67	0,95	6,67
7	[사과] [귤] --->[대추]	3	66,67	1,43	6,67
8	[사과] [대추] --->[배]	3	60,00	1,06	10,00
9	[사과] [대추] --->[감]	3	100,00	1,43	16,67
10	[사과] [대추] --->[파인애플]	3	60,00	1,64	10,00
11	[사과] [바나나] --->[배]	3	100,00	1,76	6,67

위에서 보는 것은 규칙정보입니다.

지지도 = (품목 A 와 B 를 동시에 포함하는 거래 수) / (전체 거래 수)

- 두 품목의 동시 구매가 얼마나 자주 일어나는가를 측정

신뢰도 = (품목 A 와 B 를 동시에 포함하는 거래 수) / (품목 A 를 포함하는 거래 수)

- 품목 A 를 구매하였을 경우 품목 B 를 구매하는 가능성은 얼마인가를 측정

향상도 =(품목 A 와 B 를 동시에 포함하는 거래 수) / (품목 A 를 포함하는 거래 수 * 품목 B 를 포함하는 거래 수)

- 향상도의 값이 1 에 가까울 경우: 두 품목은 독립에 가까운 사건
- 향상도의 값이 1 보다 클 경우: 두 품목이 양의 연관관계
- 향상도의 값이 1 보다 작을 경우: 두 품목이 음의 연관관계
- 의미 있는 연관성 규칙이 되기 위해선 향상도 값이 '1'이상이어야 합니다.

예를 들어 "[대추, 바나나] ==> [귤]" 규칙을 보면 신뢰도 100%, 향상도 1.579 로서, 대추와 바나나를 구매한 고객이 귤을 함께 구입할 비율이 전체 고객의 귤 구매 선택비율에 비해 1.579 배 높음을 나타냅니다.

4.4 조업 편차 분석

분석 목적

생산공정에서 설비간의 조업 편차 유무와 그 요인을 찾아내어 품질 향상 및 일정한 품질 유지를 위해 적용하는 분석 방법입니다.

분석 방법

전처리 과정을 통하여 데이터의 특성을 파악 후 PCA 방법을 사용하여 조업 현황 및 편차를 알아봅니다.

데이터

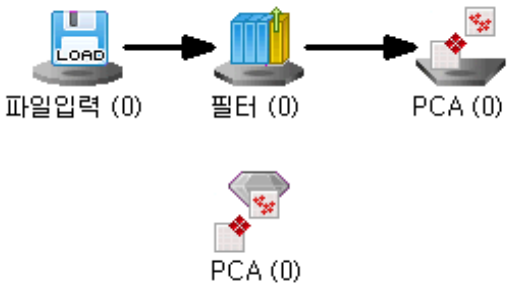
조업 편차 분석	
예제 프로젝트 파일	'조업편차분석.ecm'
데이터 파일	'조업편차분석.txt'

데이터의 설명

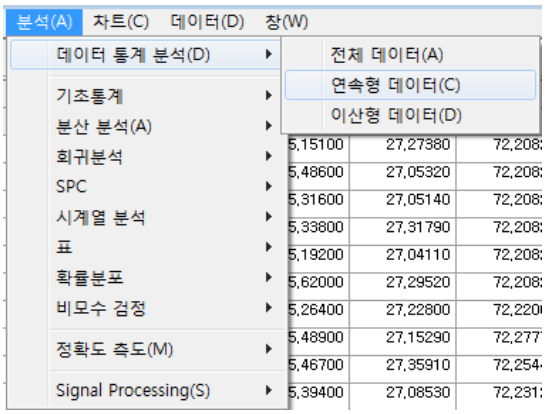
데이터 파일은 어느 공장의 54 개의 조업 공정의 데이터(모두 연속형 변수)입니다. 54 개의 독립변수만으로 구성되어 있고 7596 개의 데이터가 존재합니다.

필요 NODE 설명 및 처리 방법

▪ 조업 편차 분석 시 전체 필요 노드



▪ 데이터의 형태를 알아보기 위하여 파일입력을 더블 클릭합니다.



연속형 변수만 존재하는 것을 볼 수 있으며 데이터의 기초적인 통계량들을 살펴보기 위하여 '파일입력'창의 '분석(A)' ➔ '데이터 통계분석(D)' ➔ '연속형 데이터(C)'를 클릭합니다.

변수명	변수형	데이터수	결측치수	총합	평균	박스플롯	최소값	최대
No.	연속형	7,596	0	28,853,406	3,798,50000		1	
A1	연속형	7,596	0	830,653,10200	109,35402		104,43800	113,
A2	연속형	7,596	0	15,193,25834	2,00017		1,63106	2,
A3	연속형	7,596	0	878,024,32300	115,59035		107,88600	123,
A4	연속형	7,596	0	206,676,14450	27,20855		26,65000	27,
A5	연속형	7,596	0	548,372,69060	72,19230		72,08070	72,
A6	연속형	7,596	0	830,593,30600	109,34614		104,54300	112,
A7	연속형	7,596	0	4,085,24201	0,53781		0,35339	0,
A8	연속형	7,596	0	493,161,85820	64,92389		61,72880	67,
A9	연속형	7,596	0	467,091,77470	61,49181		49,37320	75,
A10	연속형	7,596	0	298,315,71080	39,27274		31,69960	45,
A11	연속형	7,596	0	186,229,44780	24,51678		17,61710	30,
A12	연속형	7,596	0	215,410,48240	28,35841		26,62750	30,
A13	연속형	7,596	0	117,475,04870	15,46538		15,20050	15,

각 공정 변수의 기초통계 값들을 살펴 볼 수 있습니다.

PCA(Principle Component Analysis) 분석에서는 'No.'는 필요가 없으므로 '필터' 노드에서 제거합니다. 5 개의 주성분으로 데이터를 축소시켜 살펴 보도록 합니다. 주성분을 지정하기 위해서는 선택사항에서 주성분의 수를 5 개로 설정합니다. 주성분의 수는 Eigenvalue 를 확인하고 정하게 되는데, EigenValue(고유치)가 급강하를 보이고 난 후의 EigenValue 에 해당하는 인자 수를 선택합니다. (사용자의 판단에 따라서 달라짐)

▪ PCA 분석을 시행 후 생성된 모델의 정보를 살펴봅니다.

생성된 모델 노드를 더블 클릭해 봅니다. 더블클릭을 하면 PCA 모형에 대한 정보가 나타납니다. 여기서는 변수들의 일반적인 정보가 나타나는 'General Info'탭과 주성분에 대한 정보와 각 변수의 설명력을 나타내는 수치를 포함한 'Model Info'탭, 추출된 주성분들의 관계를 도표로 볼 수 있는 'Loading Plot'탭, EigenValue 를 도표로 나타낸 'Scree Plot'탭이 있습니다.

'Model Info' 탭

● 로드

변수	주성분1	주성분2	주성분3	주성분4	주성분5
A1	-0,24360	0,09108	-0,15877	0,12632	-0,02986
A2	0,00132	0,00041	-0,02010	-0,01280	-0,05525
A3	-0,24335	0,05163	-0,05123	0,17902	0,07103
A4	0,02542	0,22263	0,05792	0,07172	0,16654
A5	-0,04617	0,09046	0,17638	0,21294	0,11003
A6	-0,24455	0,09214	-0,16095	0,12395	-0,03353
A7	-0,23241	-0,02547	-0,14758	0,14832	-0,05053
A8	0,11767	0,19428	-0,17512	0,02119	0,20550
A9	-0,01627	0,10795	-0,30890	-0,12819	-0,01167
A10	0,00410	-0,17923	0,04477	-0,03155	-0,34649
A11	-0,07708	0,15456	-0,07202	-0,11157	-0,05891

로드란 각 변수를 이용해 주성분을 생성하는 선형결합을 표시하고 있는 값으로 선형결합 시 계수값을 의미합니다. 즉, 주성분 1 은 다음과 같이 생성됩니다.

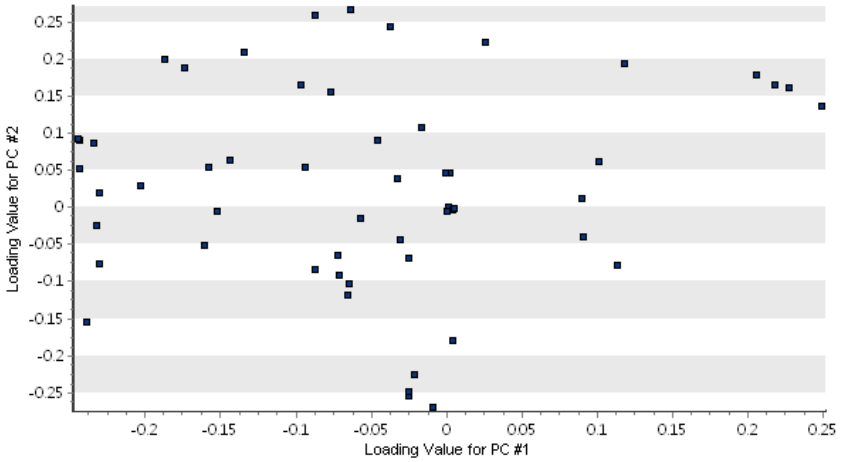
Ex) 주성분 1 = -0.2436A1 + 0.001321A2 - 0.2434A3 + ……

● 분산 설명력 (관심 주성분)

	Eigen Value of Cov(X)	X기여율(%)	X누적기여율(%)
주성분1	10,87868	20,14571	20,14571
주성분2	7,49760	13,88444	34,03015
주성분3	5,21985	9,66640	43,69655
주성분4	4,42837	8,20068	51,89723
주성분5	3,20759	5,93997	57,83720

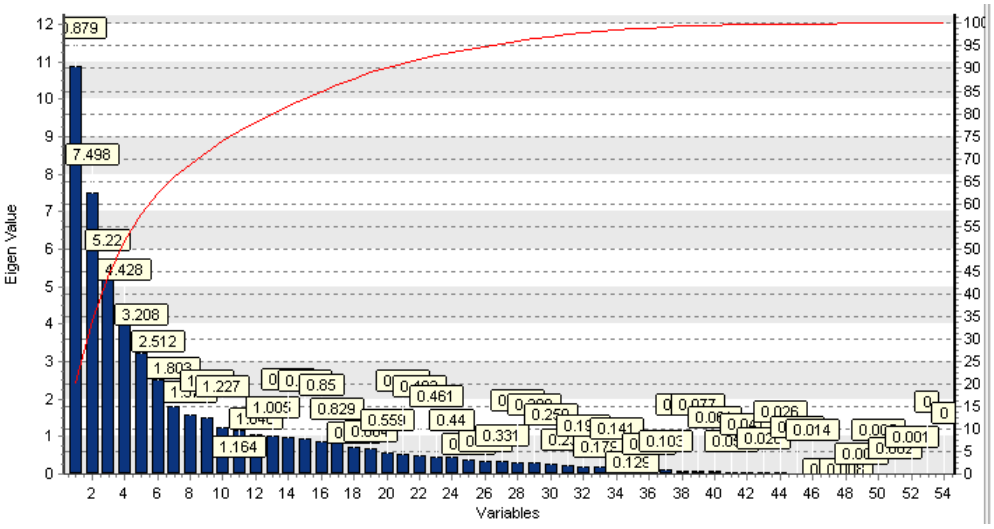
분산설명력 표에서는 각 주성분의 Eigen Value 와 모델에 영향을 미치는 기여율이 나타납니다. 기여율은 각 주성분이 전체데이터에 대한 정보를 어느 정도 가지고 있는지를 표시하고 있습니다. 즉, 각 주성분이 데이터 변동성을 설명하는데 얼마만큼의 기여를 했는지를 보여줍니다.

'Loding Plot' 탭



두 개의 주성분 간의 관계를 살펴보기 위한 plot 입니다. 탭 상단에 있는 콤보박스로 설정을 바꾸면서 주성분간의 관계를 볼 수 있습니다.

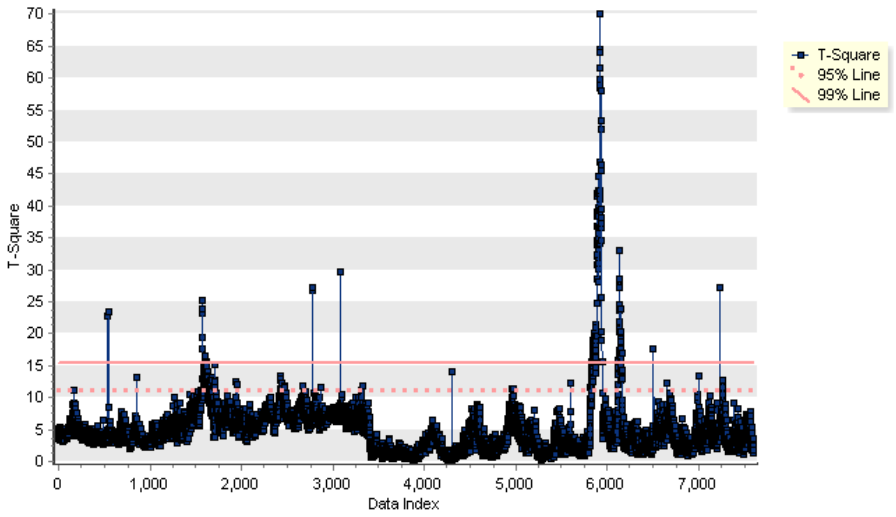
'Scree Plot' 탭



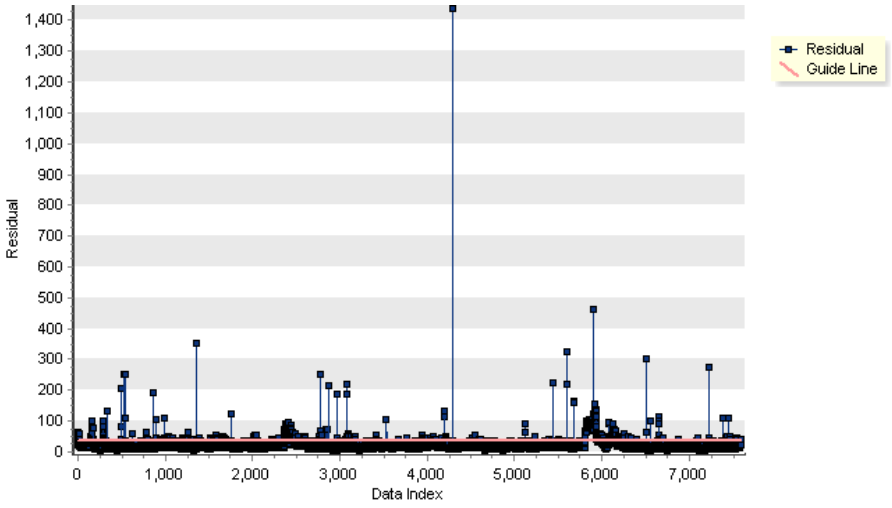
Eigen value 를 큰 순서대로 나타낸 막대그래프로 Eigen Value(고유치)가 급강하를 보이고 난 후의 Eigen Value 에 해당하는 인자 수를 최적의 주성분 수로 선택합니다. 또한 꺾은선 그래프는 주성분 개수에 따른 분산 설명력을 나타내는 그래프로, 우측에 위치한 축이 기준이 됩니다. 보통 주성분 개수를 결정할 때, 분산설명력이 80% 이상을 만족하는 최소의 주성분 수를 이용합니다.

'PCA'모델링노드에서 노드 속성창의 결과보기를 누르면 공정 데이터의 분포인 control chart 를 살펴 볼 수 있습니다.

- T-square plot

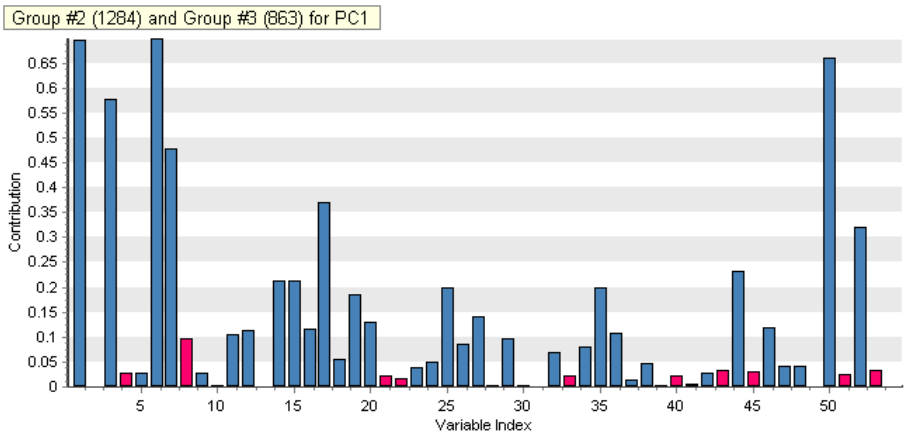


- Residual plot



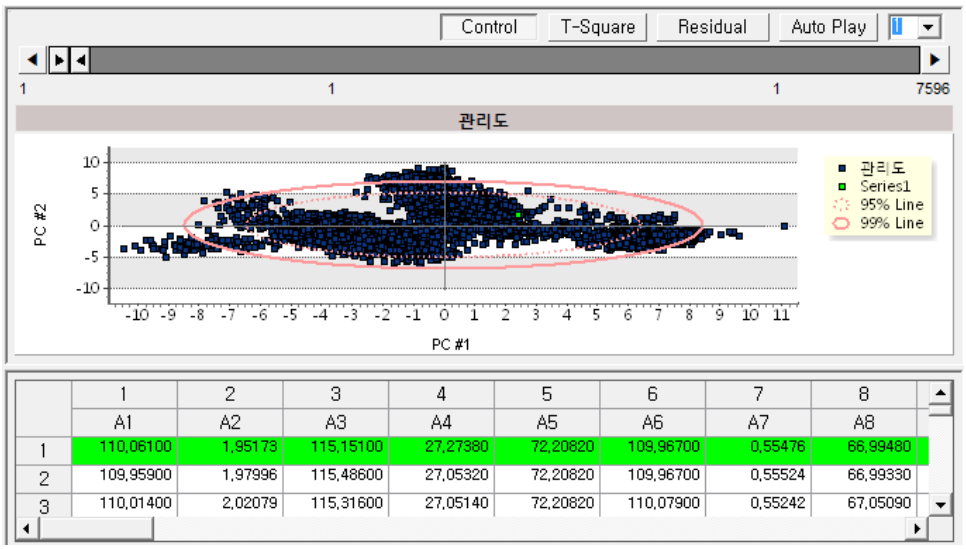
T-square plot 과 Residual plot 을 이용하여 공정상에서 이상상태를 판단할 수 있습니다. .

- Contribution



Control chart 에서 추출된 데이터이 주성분에 미치는 공헌도를 살펴 볼 수 있습니다. 이를 통해 그룹을 구분 짓는 변수를 추출해 낼 수 있습니다.

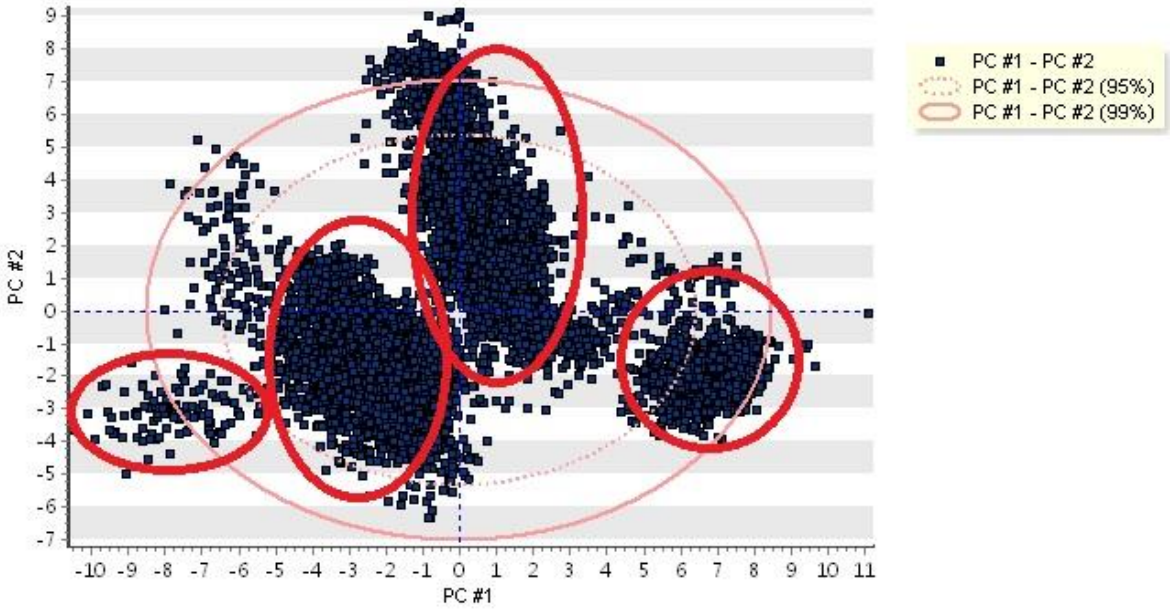
- Data chart



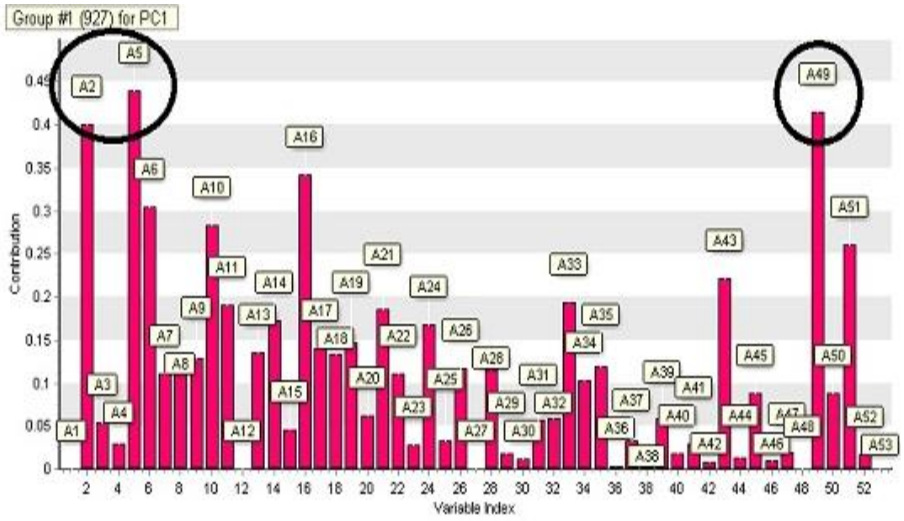
	1	2	3	4	5	6	7	8
	A1	A2	A3	A4	A5	A6	A7	A8
1	110,06100	1,95173	115,15100	27,27380	72,20820	109,96700	0,55476	66,99480
2	109,95900	1,97996	115,48600	27,05320	72,20820	109,96700	0,55524	66,99330
3	110,01400	2,02079	115,31600	27,05140	72,20820	110,07900	0,55242	67,05090

Data chart 에서 control chart, T-square plot, Residual plot 을 한눈에 볼 수 있으며, 차트와 데이터를 매칭시킬 수 있으며, 이를 이용하여 조업 진행 방향을 시각적으로 판단할 수 있습니다.

- 조업편차 영향 인자 분석



위 그림에서 4 개의 그룹으로 나누어져 있음을 볼 수 있습니다. 마우스 드래그로 관심 있는 데이터를 한 묶음으로 하고 데이터를 추출합니다. 비교하고 싶은 그룹을 직접 지정하여 공정의 차이를 볼 수 있습니다.



첫번째로 지정한 그룹 중 제 1 주성분에서 위에서 집단을 묶어서 만든 그룹을 나타낸 것입니다. A2, A5, A49 번 공정의 영향이 큰 것을 볼 수 있습니다..

조업 편차 분석은 생산공정이나 업무 프로세스간의 차이를 알아보고자 하는 분석입니다. 많은 변수들을 축소시켜 몇 개의 변수 만으로 쉽게 차이를 살펴 볼 수 있습니다. PCA 방법과

Visual 한 데이터 분석을 통하여 조업편차와 업무 프로세스의 차이를 분석할 수 있으며, 이를 통해 품질 개선을 이뤄낼 수 있습니다.

제 5 장 데이터탐색기

데이터 탐색기

데이터 탐색기는 입·출력 노드에서의 원본 데이터 탐색과 기초분석을 위한 기능입니다. 노드 종류에 따라 데이터 탐색, 입력, 수정, 전처리, 기초분석 등의 기능을 수행할 수 있습니다.

데이터 탐색기를 사용하는 노드

데이터 탐색기는 다음의 입·출력 노드에서 사용할 수 있습니다.

- **파일 입력 노드**
- **파일입력 2 노드**
- **엑셀 데이터 노드**
- **액세스 데이터 노드**
- **ODBC 입력 노드**
- **OLEDB 입력 노드**
- **화면표시 노드**
- **오라클 입력 노드**

시작 하기

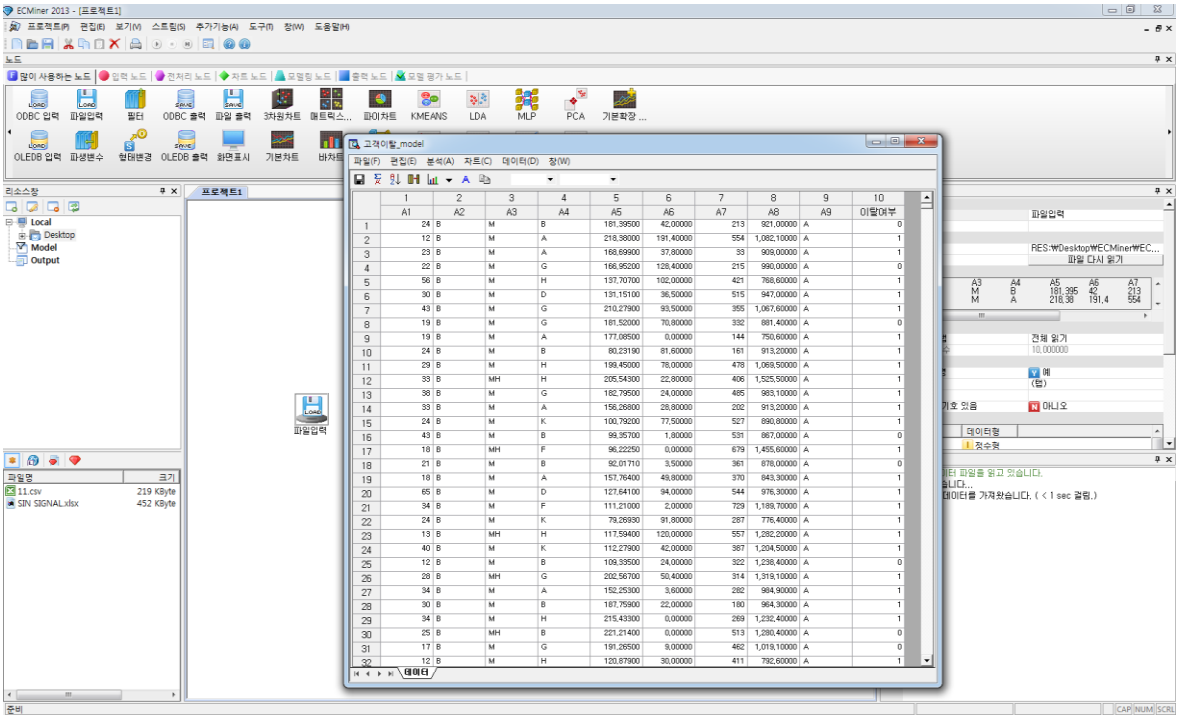
위의 노드들을 프로젝트에 추가한 후 각각의 노드의 사용법에 따라 데이터를 읽어옵니다. 데이터를 읽은 노드는 아래의 3 가지 방법을 이용하여 데이터 탐색기를 시작할 수 있습니다.

- 노드를 더블 클릭하는 방법
- 노드를 선택한 후, 마우스 오른쪽 메뉴의 **데이터 탐색기**를 선택하는 방법
- 노드를 선택한 후, 툴바의 **[보기] – [데이터 탐색기]**를 클릭하는 방법

NOTE *화면표시 노드의 경우 프로젝트 수행 시에 자동으로 실행됩니다. 사용자가 직접 실행하기 위해서는 프로젝트 수행 후에, 리소스창의 output 윈도우에서 화면표시 노드를 더블 클릭하면 됩니다.*

화면 구성

데이터 탐색기를 실행하면 다음과 같은 화면을 보실 수 있습니다.

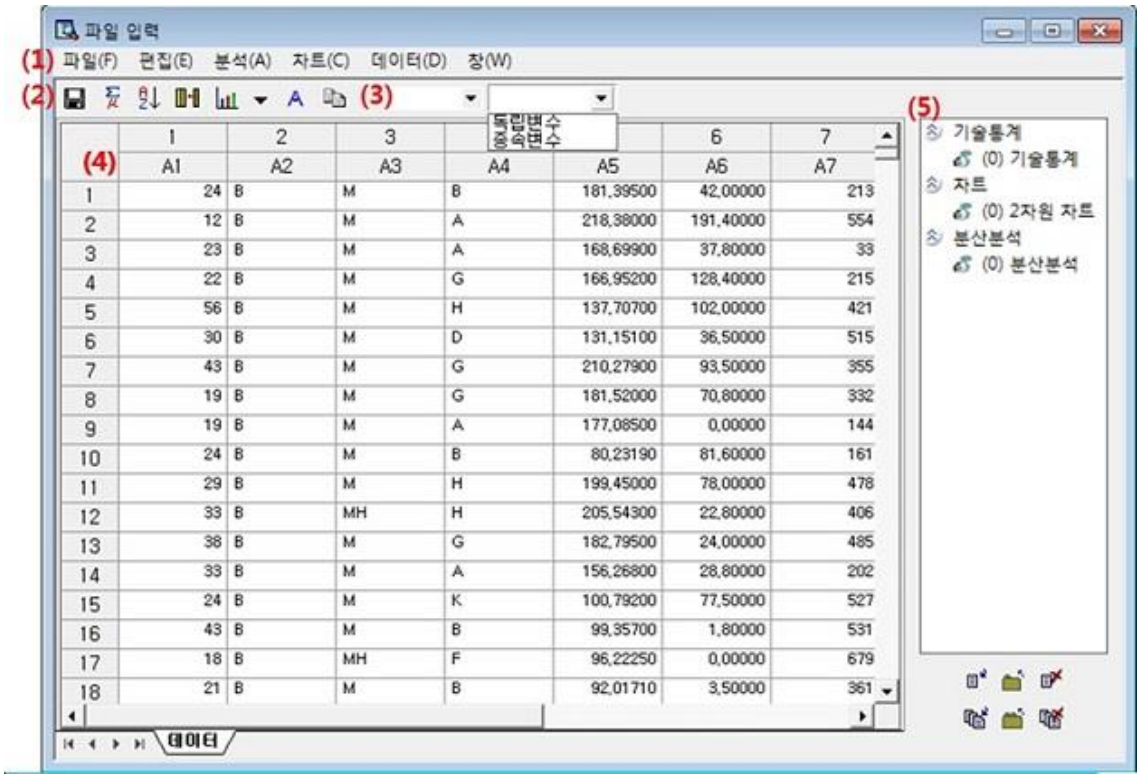


5.1 UI

5.1.1 화면구성과 윈도우 기능

데이터 탐색기 화면 구성

데이터 탐색기는 아래 그림과 같은 화면으로 구성됩니다.



윈도우 설명

윈도우 번호	윈도우명	기능	비고
1	메뉴	데이터 탐색기에서 사용할 수 있는 메뉴들입니다.	메뉴 상세
2	툴바	데이터 탐색기에서 사용할 수 있는 툴바 메뉴들입니다.	툴바 상세
3	데이터형 지정 및 변경 영역	데이터형(이산형,연속형)과 역할(독립변수,종속변수)을 지정 및 변경하는 영역입니다.	

4	데이터 영역	데이터가 실제로 보이는 영역입니다. **원본 데이터**와 분석 수행 결과인 **데이터 분포, 상관 관계** 등이 이 영역에서 표시되고, 데이터의 탐색을 할 수 있는 부분입니다.	데이터 영역 상세
5	출력 관리창	**차트, 통계 결과, 분산 분석** 내용 등의 출력 결과들을 관리하는 창입니다. 관리를 위해서는 아래 **출력 관리 툴바**를 이용하면 됩니다. 출력 관리 툴바는 (전체) 결과 보기, 숨기기, 닫기 등의 기능을 제공합니다.	출력 관리 상세

5.1.2 주메뉴

메뉴

데이터 탐색기에서는 아래와 같은 메뉴를 제공합니다. 기본적으로 아래의 메뉴를 모두 제공하지만, 노드에 따라서 일부 메뉴가 제공되지 않을 수 있습니다.

메뉴 그룹	메뉴명		설명
파일	다시읽기		데이터 탐색기를 통해 실행된 데이터 변경을 모두 무시하고, 처음의 데이터 상태로 데이터를 다시 읽어옵니다.
	저장		데이터 탐색기의 데이터를 파일로 저장합니다.
편집	복사		행, 열, 데이터 영역의 선택된 부분을 복사합니다.
	찾기		전체 데이터 영역에서 데이터를 찾습니다.
	찾기(선택필드)		선택된 필드의 데이터 영역에서 데이터를 찾습니다.
	모두선택		데이터 영역 전체를 선택합니다.
분석	데이터 통계 분석	전체 데이터	데이터 영역내의 모든 데이터에 대한 기초통계분석 결과를 한번에 보여줍니다.
		연속형	데이터 영역내의 모든 연속형 데이터에 대한

		데이터		기초통계분석 결과를 한번에 보여줍니다.
		이산형 데이터		데이터 영역내의 모든 이산형 데이터에 대한 기초통계분석 결과를 한번에 보여줍니다.
	기초통계	기술통계		선택된 필드에 대한 기술 통계량을 보여줍니다.
		상관관계분석	화면표시	모든 데이터 필드(범주형 데이터 필드 제외)에 대해 필드간의 상관 관계 값을 화면으로 보여줍니다.
			파일로 저장	모든 데이터 필드(범주형 데이터 필드 제외)에 대해 필드간의 상관 관계 값을 파일로 저장 합니다.
			네트워크	모든 데이터 필드(범주형 데이터 필드 제외)에 대해 필드간의 상관 관계를 시각적인 네트워크로 보여줍니다.
		평균비교분석	일표본	주어진 평균치와 선택된 필드 변수들의 평균값을 비교합니다.
			독립표본	상호 독립적인 각각의 필드에 대해 독립표본 T-test 를 실시합니다.
			대응표본	상호 의존적인 각각의 필드에 대해 대응표본 T-test 를 실시합니다.
			유의차 검정	요인필드에 의해 그룹간 차이가 있는지 검정합니다.
		비율검정	단일 비율 검정	선택된 필드의 데이터 값들이 이항 자료인 경우 그 비율에 대한 신뢰구간 계산 및 가설 검정합니다.
			두 비율 검정	이항 자료인 두 변수에 대하여 이벤트 비율간 차이에 대한 신뢰구간 계산 및 가설 검정합니다
			1-표본 포아송검정	선택된 필드의 데이터 값들이 포아송 자료인 경우 그 비율에 대한 신뢰구간 계산 및 가설 검정합니다
			2-표본 포아송검정	포아송 자료인 두 변수에 대하여 이벤트 비율간 차이에 대한 신뢰구간 계산 및 가설

				검정합니다.
		분산검정	단일 표본 분산 검정	연속형 변수의 분산에 대한 신뢰구간 계산 및 가설 검정합니다.
			두 표본 분산 검정	두 연속형 변수의 분산 비에 대한 신뢰구간 계산 및 가설 검정합니다
		정규성검정		선택된 필드의 데이터 값들이 정규 분포를 따르는지 따르지 않는지를 통계적으로 검정합니다.
		포아송검정		수집된 자료가 포아송 분포를 따르는지 검정합니다.
	분산분석	일원 배치법 Stack		Stack 데이터를 이용하여 일원 분산분석을 수행 합니다.
		이원 배치법 Stack		Stack 데이터를 이용하여 이원 분산분석을 수행 합니다.
		GLM(일반 선형 모델)		데이터의 가장 일반적인 형태의 분산분석을 수행합니다.
	회귀분석	직교 회귀분석		실제 관측치와 회귀 직선간 최간거리의 제곱합을 최소화하는 회귀분석을 수행합니다.
		Nonlinear Regression		비선형 데이터를 이용하여 회귀분석을 수행합니다.
	SPC	공정능력분석		연속형 데이터 필드에 대해 명세서 상의 데이터 분포와 실제 데이터 분포를 비교하여, 공정 능력을 분석합니다.
		공정능력 요약		데이터가 원하는 공정 영역 내에서 이루어지는지 여부를 요약 리포트 형태로 보여줍니다.
		합격표본추출	계수형 합격 샘플링	전체 로트(lot) 또는 배치(batch)를 대상으로 품질 기준에 의해 로트의 합격, 불합격 판정을 위한 계수형 샘플링 계획을 생성합니다.
			계량형 합격 샘플링	전체 로트(lot) 또는 배치(batch)를 대상으로 품질 기준에 의해 로트의 합격, 불합격 판정을 위한 계량형 샘플링 계획을 생성합니다.

		공차구간		최소 모집단의 비율과 신뢰 수준을 설정하며, 샘플로부터 얻어진 통계량을 기초로 지정된 모집단에서의 최소비율과 신뢰 수준을 만족하는 구간을 제시합니다.
	시계열 분석	시계열 모델	시계열 분해	데이터를 성분별로 나누어 보여줍니다.
			이동평균	현재 시점과 과거, 미래의 몇 데이터를 평균한 값을 보여줍니다.
			지수분할	최근의 자료에 더 많은 가중값을 부여하여 예측하는 방법입니다.
			추세분석	계수를 추정하고, 추정한 계수를 통해 만든 식의 적합성을 Test 하고 마지막으로 Forecasting 까지 합니다.
			ARIMA	시계열 데이터의 기본적인 추세를 얻기 위해서 수행하는 방법
			GARCH	변동성을 측정하고 예측하여 줍니다.
			VAR	특정 변수의 집합들이 단순히 개별적으로 움직이는 것이 아니라 서로 영향을 받으며 움직이고 있다면 VAR 모델을 상정할 수 있습니다.
			ARMAX	시계열 데이터와 적합치, 그리고 적합치의 상한, 하한을 보여줍니다.
		시계열 검정	단위근 검정	하나의 시계열이 단위근(Unit root)을 가지고 있는지를 검정합니다.
			그레인저 인과관계	하나의 시계열이 다른 시계열을 예측하는 데 유용한지 그렇지 않은지를 검정합니다.
			공적분검정	두 시계열이 공적분 관계(Cointegrated)를 갖는지를 검정합니다.
			ARCH Test	하나의 시계열이 시간에 따라 변하는 변동성을 예측할 수 있는지 없는지를 검정합니다.
		시계열 상관성	교차상관	두 시계열간의 유사성을 측정하는 지표입니다.
			자기상관	시계열 데이터의 값이 과거의 데이터와

			편자기상관	어떠한 상관관계를 갖는지를 알려줍니다.
	표	빈도표		어떤 필드에 대해 특정 값의 빈도를 보여줍니다.
		교차표		각각의 필드에 대해, 두 변수의 값이 공유하고 있는 빈도수가 몇 개인지를 보여줍니다.
		일변량 카이제곱 검정		수집된 자료가 특정 비율을 갖는 다항분포를 따르는지 검정합니다
		독립성 검정		각각의 필드에 대해 요인간의 관계유무를 검정합니다.
	확률분포	모수 추정		데이터가 주어져 있을 때 그 데이터에 가장 잘 맞는 모수를 찾아줍니다. 또, 추정된 모수의 신뢰 구간을 구해 주어 추정된 모수를 어느 정도 신뢰할 수 있는지를 제시해줍니다.
		개별분포 식별		데이터의 분포를 확인하기 위하여 여러 분포에 적용하여 얻어진 Anderson-Darling 통계량과 그의 P-value 값을 토대로 데이터의 가장 유사한 분포를 확인할 수 있습니다.
	비모수 검정		일표본	주어진 중앙값과 선택된 필드의 중앙값을 비교합니다.
			독립표본	한 변수 내의 두 그룹간의 평균 또는 중앙값을 비교합니다.
			대응표본	상호 의존적인 각각의 필드에 대해 중앙값을 비교합니다.
			분산분석 - 일원배치법	3 개 이상의 중앙값 차이 검정을 진행합니다.
	정확도 측도		분류	독립변수를 통해 각 클래스의 label 을 예측합니다.

		예측	종속변수와 예측변수를 사용하여 R-square, MAPE, MAD, MSD 를 제공합니다.	
	Gage R&R	Gage 런차트	실험을 통해서 얻은 전체 관측치를 보여줍니다.	
		Gage 선형성 및 편향 연구	부품의 규격에 따라서 측정 값과 기준 값의 차이라고 할 수 있는 편향이 선형적인 관계를 가지고 있는지를 판단하기 위해 사용합니다	
		Gage R&R(교차분석)	측정시스템이 부품간을 얼마나 잘 구별하는지, 개별 부품이 개별 작업에 의해 여러 번 측정될 때 측정시스템에서 야기하는 변동을 파악할 때 사용합니다.	
		Gage R&R(내포분석)	파괴 검사와 같이 각 부품을 하나의 측정 시스템에서만 측정할 경우 사용합니다.	
차트	2 차원 차트		2 차원 차트를 보여줍니다.	
	3 차원 차트		3 차원 차트를 보여줍니다.	
	바 차트		바 차트(Bar Chart)를 보여줍니다.	
	박스 차트		박스 차트(Box Chart)를 보여줍니다.	
	매트릭스 차트		매트릭스 차트(Matrix Chart)를 보여줍니다.	
	파레토 차트		파레토 차트(Pareto Chart)를 보여줍니다.	
	파이 차트		파이 차트(Pie Chart)를 보여줍니다.	
	멀티 차트		여러 변수의 특징을 한 눈에 보여줍니다.	
데이터	정렬		선택된 필드들을 정렬합니다.	
	파생 변수		지정된 조건에 따라 새로운 파생 필드를 생성 합니다.	

	적용	데이터 탐색기에서 조작된 전처리 과정을 프로젝트의 노드를 이용하여 스트림으로 구성 합니다.	
	필터	지정된 조건에 따라 해당되는 변수를 걸러 냅니다.	
	관심변수 고정	고정할 열을 선택하여 관심변수를 지정합니다.	
	Box-Cox 변환	정규분포를 따르지 않는 관측치들을 정규분포를 따르도록 변환 수식을 제공합니다	
	Johnson 변환	정규분포를 따르지 않는 관측치들을 정규분포를 따르도록 변환 수식을 제공합니다	
창	결과 관리	결과로 출력된 차트 및 통계량 등의 윈도우를 관리합니다.	

5.1.3 부메뉴(마우스, 키보드 이용)

데이터 탐색기에서 제공하는 기타 메뉴로, 마우스 오른쪽 클릭 시에 나타나는 메뉴입니다. 기본적으로 아래의 메뉴를 모두 제공하지만, 노드의 종류에 따라 일부 메뉴가 제공되지 않을 수 있습니다.

열 삭제

- 실행 방법: 데이터 영역 열 머리글의 열 번호 또는 필드명에서 마우스 오른쪽 버튼을 누르면 메뉴가 나타납니다.
- 실행 결과: 선택된 열이 데이터 탐색기상에서 삭제됩니다. 열 삭제 후 적용이 실행되면, 스트림에 **필터노드**가 추가되어, 선택된 열이 제거되게 됩니다.

행 삭제

- 실행 방법: 데이터 영역의 행 머리글의 행 번호에서 마우스 오른쪽 버튼을 누르면 메뉴가 나타납니다.
- 실행 결과: 선택된 행이 데이터 탐색기상에서 삭제됩니다. 행 삭제 후 적용이 실행되면, 스트림에 **선택 노드**가 추가되어, 선택된 행이 제거되게 됩니다.

열 숨기기

- 실행 방법: 데이터 영역 열 머리글의 열 번호 또는 필드명에서 마우스 오른쪽 버튼을 누르면 메뉴가 나타납니다.
- 실행 결과: 선택된 열이 데이터 탐색기상에서 숨겨지게 됩니다. 열 삭제와는 달리, 단지 숨겨지기만 했기 때문에 데이터 스트림에는 영향을 주지 않습니다.

열 숨기기 취소

- 실행 방법: 하나 이상의 열이 숨겨졌을 경우, 데이터 영역 열 머리글의 열 번호 또는 필드명에서 마우스 오른쪽 버튼을 누르면 숨겨진 열 번호와 필드명이 나타납니다.
- 실행 결과: 선택된 열이 다시 나타나게 됩니다.

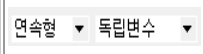
기타

- 복사/컬럼 이름과 함께 복사/전체선택/Excel 로 데이터 내보내기 기능을 추가로 제공합니다.
- 열삭제, 행삭제 등을 통해 수정된 데이터는 '적용'을 통해 즉시 데이터 스트림에 적용할 수 있습니다.

5.1.4 툴바

데이터 탐색기의 툴바로 제공되는 메뉴입니다

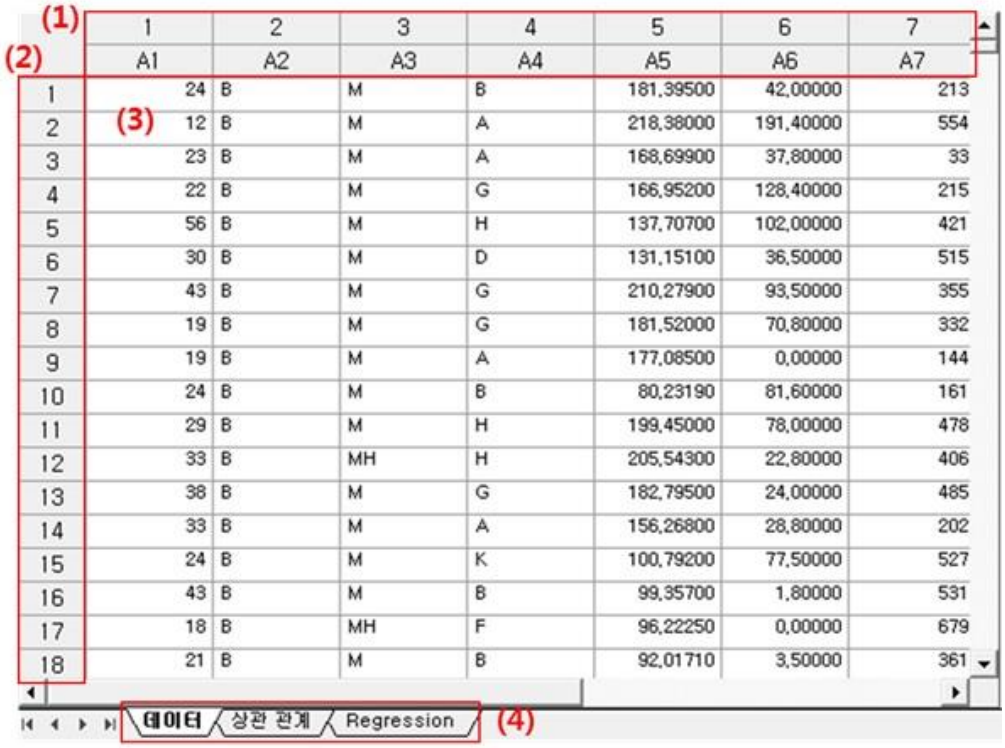
데이터 탐색기 툴바 메뉴

메뉴	아이콘	설명	기타	비고
저장		데이터를 파일로 저장합니다.		
기술통계		기술 통계량을 봅니다.		
정렬		다중 정렬을 수행합니다.		
파생 변수		파생 변수를 생성합니다.		
차트		차트를 선택합니다.		
적용		전처리 결과를 프로젝트에 적용합니다.		
결과 관리		결과 창을 관리합니다.		
형태 변경		데이터의 형태를 알 수 있으며, 변경할 수 있습니다.		

5.1.5 데이터 영역(Grid)

데이터 및 분석 결과가 보여지는 영역으로, 원 데이터 시트에서는 데이터 탐색 및 행, 열 편집이 가능합니다.

윈도우 배치



윈도우 설명

윈도우 번호	윈도우명	기능	비고
1	열 머리글	데이터 영역의 위쪽 부분으로, **열 번호, 필드명**이 보여집니다. 마우스 오른쪽 버튼을 클릭하여 **열삭제, 열숨기기, 복사, 컬럼이름과 함께 복사, 전체선택, Excel 로 데이터 보내기**가 가능 합니다. 열의 경계 부분을 더블 클릭하면 열 간격이 자동으로 맞춰집니다.	
2	행 머리글	데이터 영역의 왼쪽 부분으로, **행 번호**가 보여집니다. 마우스 오른쪽 버튼을 클릭하여 **행 삭제, 복사, 컬럼이름과 함께 복사, 전체선택, Excel 로 데이터 보내기**가 가능합니다.	
3	데이터 영역	실제로 데이터가 표시되는 부분입니다. 처음에는 **데이터시트**만 존재하나, 데이터 분석에 따라, **데이터 분포, 이산형 분포`, 상관 관계** 등의 분석결과 시트가 추가 됩니다. 마우스 오른쪽 버튼을 클릭하여 **복사, 붙여넣기,**	

		삭제, 전체선택 등이 가능합니다.(데이터 영역의 종류에 따라 메뉴는 조금씩 달라집니다.)	
4	시트 탭	데이터 영역을 관리하는 탭입니다. 선택을 통해 원하는 데이터 창을 볼 수 있을 뿐만 아니라, 마우스 드래그를 사용하여 탭의 위치를 바꿀 수 있습니다.	

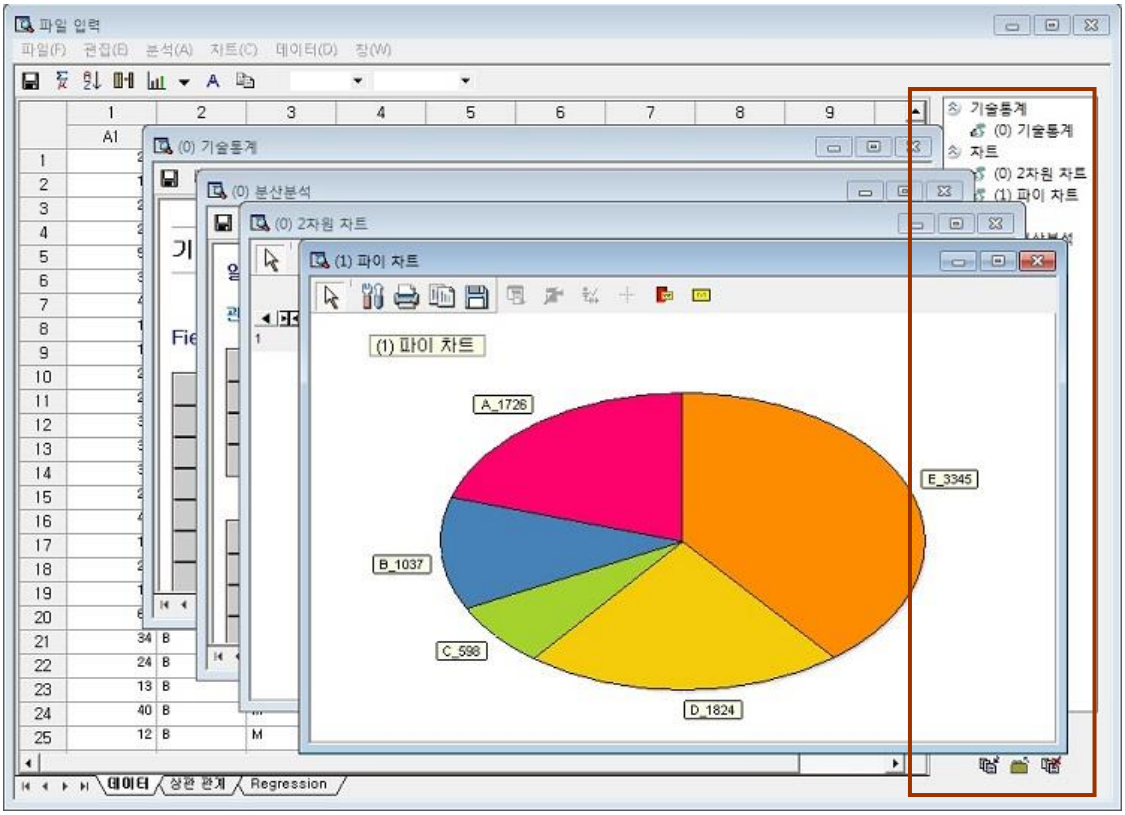
5.1.6 결과 관리창

결과 관리창은 데이터 탐색기에서는 수행된 여러 결과(차트, 통계량)들을 관리하기 위한 기능입니다. 새로운 결과가 수행되어 새 창이 생기면, 이 결과에 해당되는 항목이 자동으로 **결과 관리창**에 추가되게 됩니다. 추가된 결과 항목은 결과창 아래의 툴바를 이용하여 관리할 수 있습니다. 또한 결과는 그 종류에 따라 "차트", "기술통계", "분산 분석"등의 그룹으로 나뉘어서 관리됩니다.

실행 방법

툴바의 **[창]-[결과관리]** 아이콘을 클릭하시면 **결과 관리창**이 나타납니다.

실행 화면



데이터 탐색기 결과 관리 윈도우 메뉴

메뉴	아이콘	설명	기타	비고
보이기		선택된 결과창만을 보여줍니다.		
모두 보이기		모든 결과창을 보여줍니다.		
숨기기		선택된 결과창을 숨깁니다.		
모두 숨기기		모든 결과창을 숨깁니다.		
닫기		선택된 결과창을 닫습니다.		
모두 닫기		모든 결과창을 닫습니다.		

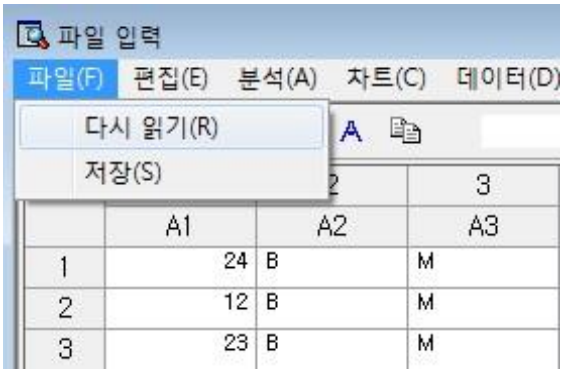
5.2 파일

5.2.1 다시읽기

데이터 탐색기에서 실행한 여러 조작을 무시하고, 파일 또는 DB 에서 데이터를 다시 읽어 오는 기능입니다.

실행 방법

[파일] 메뉴의 **[다시 읽기]** 메뉴를 선택하여 실행합니다.



NOTE *필드 및 데이터 삭제, 파생 필드 추가, 정렬 등의 작업이 모두 무시되고 초기상태로 되돌아 갑니다.*

5.2.2 저장하기

데이터 탐색기에서는 **ECMiner™**의 기본확장자인 ecl 로 저장하는 기능을 제공하고 있습니다. 그 밖에 txt, csv 확장자도 지원합니다.

실행 방법

[파일]의 **[저장]** 메뉴 또는 툴바의 **저장 아이콘**을 선택하여 저장할 수 있습니다.

또는 

저장 형식

텍스트 파일 형태로는 확장자 "txt", "csv"가 지원되고 **ECMiner™** 데이터 파일인 ecl 파일로 저장이 가능합니다.

5.3 분석

5.3.1 데이터 통계분석

(1) 전체 데이터

데이터 탐색기에서는 **데이터 통계 분석 – 전체 데이터** 기능을 사용하여, 데이터 영역내의 모든 데이터에 대한 통계량과 분포를 볼 수 있습니다.

실행 방법

[분석] – [데이터 통계 분석] – [전체 데이터]를 선택하면, 데이터 분포 시트가 추가됩니다.

예제 데이터

	1	2
	FIELD1	FIELD2
1	A	19,400
2	A	32,600
3	A	27,000
4	A	32,100
5	A	33,000
6	B	17,700
7	B	24,800
8	B	27,900
9	B	25,200
10	B	24,300
11	C	17,000
12	C	19,400

전체 데이터 통계분석 실행 결과

변수명				범주형				
	변수형	데이터수	결측치수	변수값	빈도	누적 빈도	백분율(%)	유효백분율(%)
FIELD1	범주형	30	0	A	5	5	16,666667	16,666667
				F	5	10	16,666667	16,666667
				E	5	15	16,666667	16,666667
				D	5	20	16,666667	16,666667
				C	5	25	16,666667	16,666667
				B	5	30	16,666667	16,666667
FIELD2	연속형	30	0					

NOTE *연속형 데이터 통계 분석 결과해석과 이산형 데이터 통계 분석 결과해석을 참조하십시오.*

(2) 데이터 분포(연속형)

데이터 탐색기에서는 **데이터 통계 분석 – 연속형 데이터** 기능을 사용하여, 데이터 영역내의 모든 연속형 데이터에 대한 통계량을 한번에 볼 수 있습니다.

실행 방법

[분석] – [데이터 통계 분석] – [연속형 데이터]를 선택하면, 연속형 분포 시트가 추가 됩니다.

연속형 데이터 통계분석 실행 결과

변수명	변수형	데이터수	결측치수	총합	평균	박스플롯	최소값	최대값	범위	
A1	연속형	8,530	0	259,500	30.42204		12	82	70	
A5	연속형	8,530	0	516,894.76180	177.83057		59.94000	619.32700	559.38700	6
A6	연속형	8,530	0	426,118.57368	49.95528		0.00000	195.50000	195.50000	1
A7	연속형	8,530	0	2,871.005	336.57737		0	1,977	1,977	64
A8	연속형	8,530	0	537,734.78555	1,000.90677		0.00000	3,423.30000	3,423.30000	308
이탈여부	연속형	8,530	0	4,684	0.54912		0	1	1	

지원되는 통계량

총합, 평균, 박스 플롯, 결측치 수, 최소값, 최대값, 범위, 분산, 표준 편차, 첨도, 왜도, 중간값, 사분위값이 지원됩니다.

NOTE *박스 플롯의 양쪽 끝의 숫자는 outlier 의 수를 나타냅니다. outlier 의 수가 100 개를 넘으면 '*'로 나타납니다.*

(3) 데이터 분포(이산형)

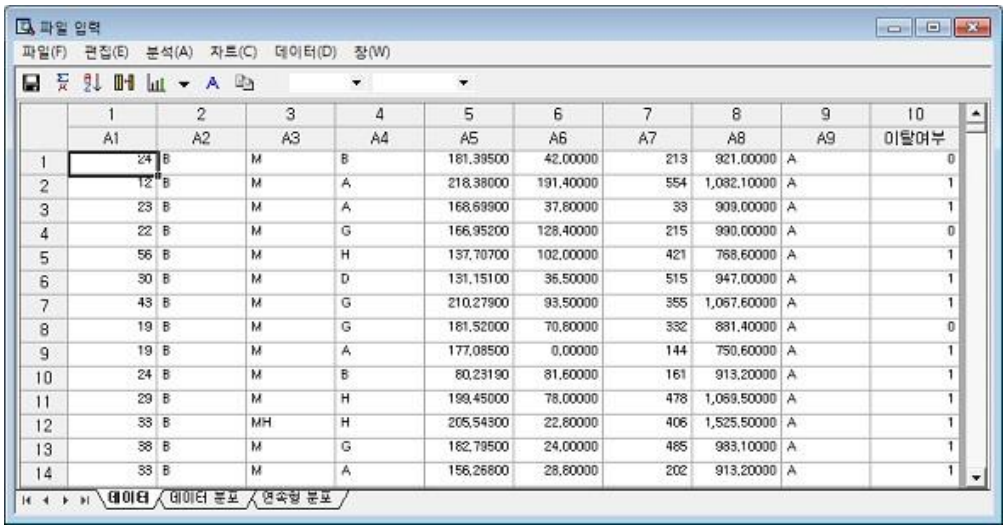
데이터 탐색기에서는 **데이터 통계 분석 – 이산형 데이터** 기능을 사용하여, 데이터 영역내의 모든 이산형 데이터에 대해 그 분포를 한번에 볼 수 있습니다.

실행 방법

[분석] – [데이터 통계 분석] – [이산형 데이터]를 선택하면, 데이터 분포 시트가 추가 됩니다.

예제 데이터

NOTE *A2~A4 부분은 이산형 데이터 필드입니다.*



	1 A1	2 A2	3 A3	4 A4	5 A5	6 A6	7 A7	8 A8	9 A9	10 이탈여부
1	24	B	M	B	161.39500	42.00000	213	921.00000	A	0
2	12	B	M	A	218.38000	191.40000	554	1,082.10000	A	1
3	23	B	M	A	168.69900	37.80000	33	909.00000	A	1
4	22	B	M	G	166.95200	128.40000	215	990.00000	A	0
5	56	B	M	H	137.70700	102.00000	421	768.60000	A	1
6	30	B	M	D	131.15100	36.50000	515	947.00000	A	1
7	43	B	M	G	210.27900	93.50000	355	1,067.60000	A	1
8	19	B	M	G	161.52000	70.80000	332	881.40000	A	0
9	19	B	M	A	177.08500	0.00000	144	750.60000	A	1
10	24	B	M	B	80.23190	81.60000	161	913.20000	A	1
11	29	B	M	H	199.45000	78.00000	478	1,069.50000	A	1
12	33	B	MH	H	205.54300	22.80000	406	1,525.50000	A	1
13	38	B	M	G	162.79500	24.00000	485	983.10000	A	1
14	33	B	M	A	156.26800	28.80000	202	913.20000	A	1

실행 결과

변수명	변수형	데이터수	결측치수	변수값	빈도	누적 빈도	백분율(%)	유효백분율(%	누적백분율(%
A2	범주형	8,530	0	E	3345	3,345	39,21454	39,21454	39,21454
				D	1824	5,169	21,38335	21,38335	60,59789
				A	1726	6,895	20,23447	20,23447	80,83236
				B	1037	7,932	12,15709	12,15709	92,98945
				C	598	8,530	7,01055	7,01055	100
A3	범주형	8,530	0	M	4583	4,583	53,72802	53,72802	53,72802
				MH	1878	6,461	22,01641	22,01641	75,74443
				ML	1260	7,721	14,77140	14,77140	90,51583
				H	542	8,263	6,35404	6,35404	96,86987
				L	263	8,526	3,08324	3,08324	99,95311
				N	4	8,530	0,04689	0,04689	100

결과 설명

- **변수 개수와 빈도**: 해당 변수값의 출현 빈도수를 나타냅니다.
- **누적 빈도**: 가장 많이 출현한 변수부터의 누적 빈도수입니다.
- **백분율**: 결측값을 포함한 백분율입니다.
- **유효 백분율**: 결측값을 제외한 백분율입니다.
- **누적 백분율**: 유효 백분율을 누적시킨 값입니다.

5.3.2 기초통계

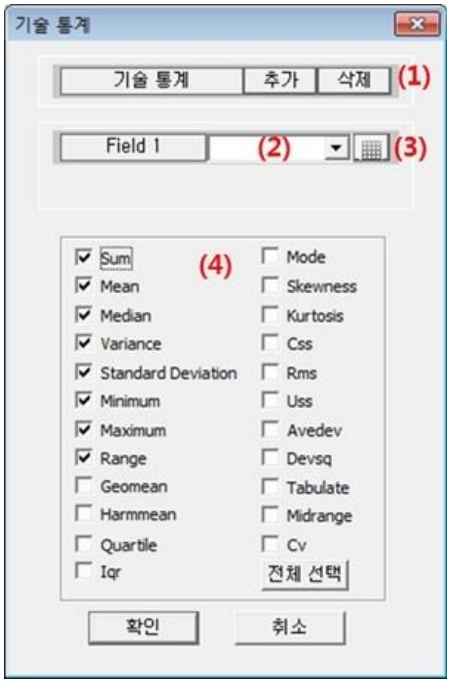
- 5.3.2.1 기술통계

데이터 탐색기에서는 **기술 통계**기능을 통해, 선택된 필드에 대한 기술 통계량을 볼 수 있습니다.

실행 방법

[분석] – [기술통계]를 선택하거나, 또는 툴바의 **기술통계** 아이콘을 클릭하시면 기술통계 다이얼로그가 나타납니다.

기술통계 다이얼로그



1 : 여러 필드의 기술 통계량을 볼 때 사용합니다. 추가 버튼을 누르면 필드를 추가할 수 있는 리스트가 추가되고, 삭제 버튼을 누르면 마지막 필드 리스트가 제거됩니다. 만약 필드 리스트 원소가 1 개일 때는 삭제 버튼을 눌러도 더 이상 리스트 원소가 줄어들지 않습니다.

2 : 기술 통계량을 볼 필드를 선택할 수 있는 콤보 박스입니다.

3 : 필드 선택 버튼. 필드를 선택하는 다른 방법으로, 버튼을 누른 후 데이터 영역의 열 머리글을 클릭하면 필드가 선택됩니다.

4 : 전체 선택 버튼. 통계량 선택 시, 체크박스로 일일이 선택하지 않고, 모든 통계량을 한번에 선택 또는 해제 할 수 있는 버튼 입니다.

NOTE *수학적인 통계량을 얻을 수 없는 범주형 데이터필드는 선택할 수 없습니다.(필드 선택 콤보 박스에 범주형 필드는 나타나지 않습니다)*

실행 결과

Sum	1516894,7618
Mean	177,83057
Min	59,94
Max	619,327
Range	559,387
Variance	6866,26352
Standard Deviation	82,86292
Kurtosis	1,61591
Skewness	1,09217
Median	165,1835
Quartile Q1	113,913
Quartile Q2	165,1835
Quartile Q3	222,191

결과 설명

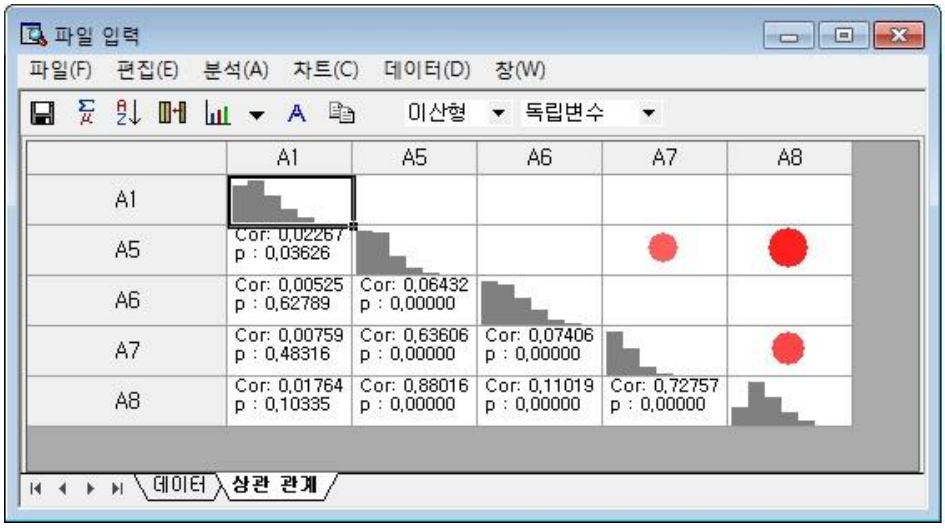
선택된 필드들에 대해 통계량을 보여줍니다. 지원되는 통계량으로는 총합, 평균, 최소값, 최대값, 범위, 분산, 표준편차, 왜도, 첨도, 중앙값, 4 분위수등이 있습니다.

- **5.3.2.2 상관 관계 분석**

(1) 화면표시

화면 표시 기능은 상관관계 테이블을 화면에 바로 보여주는 기능입니다. 실행 되면 상관 관계를 표시하는 시트가 데이터 탐색기에 추가됩니다.

실행 결과



결과 설명

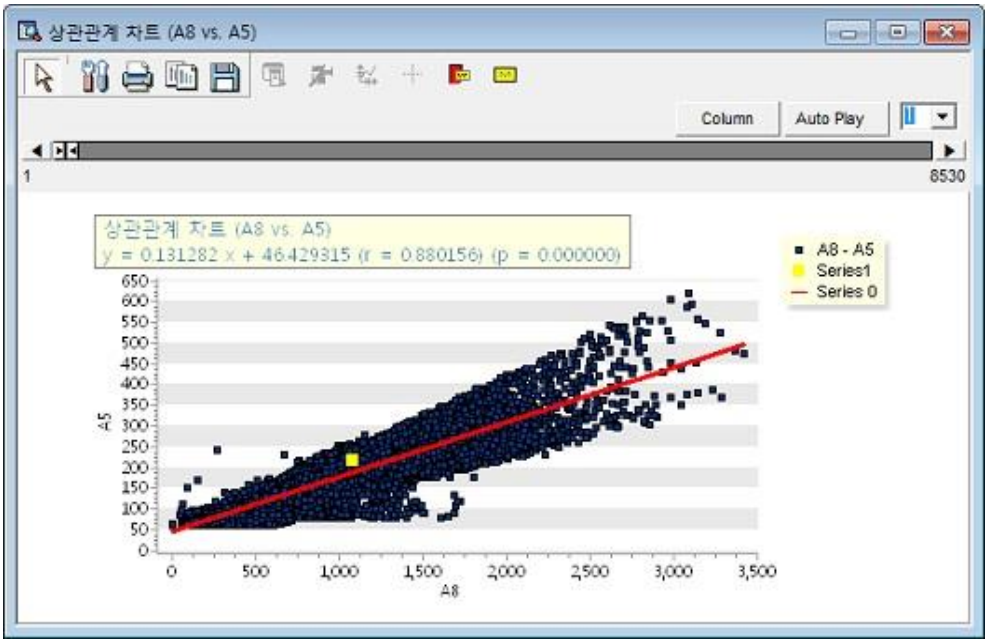
Cor : 상관관계(Correlation)값입니다.

p : 상관관계 값에 해당되는 p-Value 입니다.

상관 관계 차트

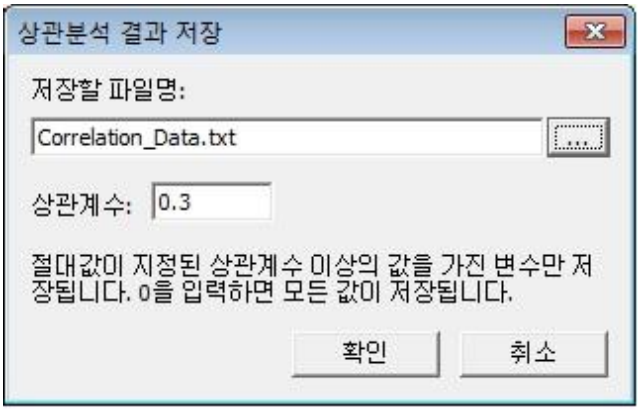
상관 관계 시트에서는 마우스 클릭을 통해, 선택된 필드들의 상관 관계 차트를 볼 수 있습니다.

두 필드가 만나는 셀을 더블 클릭하시면, 두 필드에 대한 분포도 차트와 추세선을 볼 수 있습니다.



(2) 파일로 저장

파일로 저장 기능은 상관관계 테이블을 파일로 저장하는 기능입니다. 실행되면 아래와 같은 다이얼로그가 새로 보여지면서 **파일명** 지정 및 **상관 계수**를 지정할 수 있습니다. 결과는 적어도 하나의 상관 관계에서 지정된 **상관 계수**보다 큰 값을 가지는 변수들만 저장됩니다.

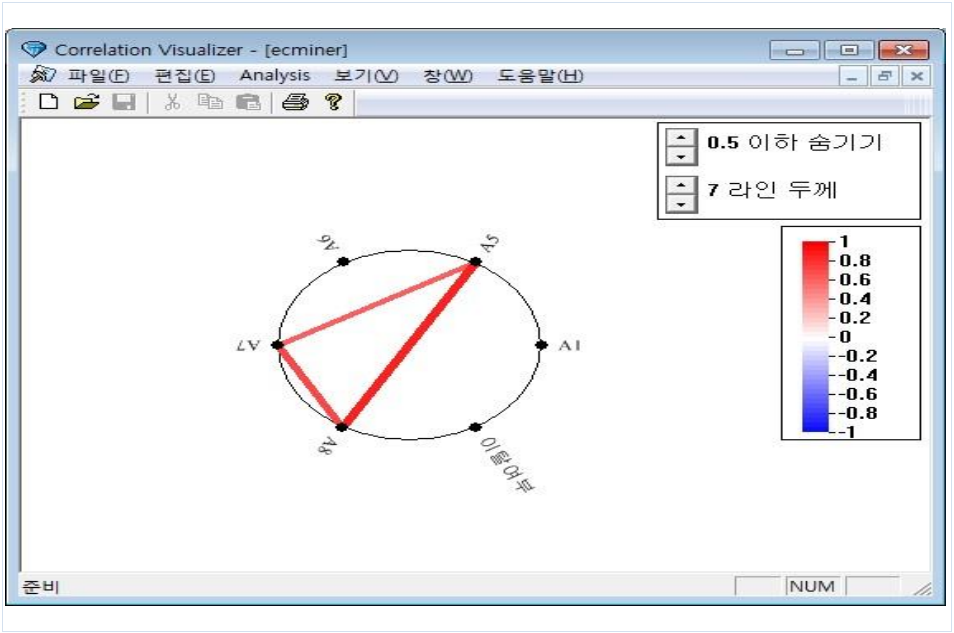


실행 결과

(3) 네트워크

네트워크 기능은 위의 두 가지 경우와 달리 상관관계를 테이블로 표시하지 않고, 선의 색깔과 굵기를 이용하여 변수간의 상관관계를 보여주는 기능입니다. 빨간색은 양의 상관관계, 파란색은 음의 상관관계를 의미하며, 굵기가 굵을수록 상관계수의 절대값이 큼을 의미합니다. 또한 상관 계수 조절을 통해 지정된 상관 계수값 이상을 가지는 관계만을 보여주는 기능이 있어, 의미 있는 상관계수를 보이는 변수들만의 상관 관계를 선택적으로 볼 수도 있습니다.

실행 결과

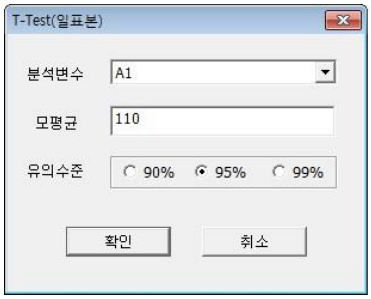


- 5.3.2.3 평균 비교 분석

(1) 일표본

데이터 탐색기에서는 각각의 필드에 대해 **평균 비교 분석 - 일표본** 기능을 사용하여, 주어진 평균치와 변수들의 평균값을 비교할 수 있습니다.

실행 방법



[분석] – [평균 비교 분석] – [일표본]을 선택하면, **일표본** 윈도우가 나타납니다. 분석할 필드를 선택하고 모평균을 입력합니다. 유의수준을 정합니다.

결과

일표본 통계량과 검정결과를 보여줍니다.

t-Test(일표본)

일표본 통계량

변수명	데이터개수	평균	표준편차	평균의 표준오차
A5	8530	177,83057	82,86292	0,89719

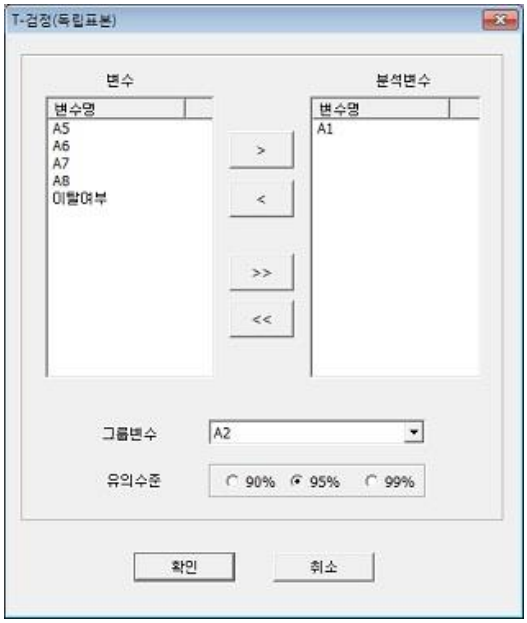
일표본 검정

	t	자유도	유의확률 (양쪽)	평균차	차이의 95% 신뢰구간
A5	198,20778	8529	0,00000	177,83057	(176,071, 179,590)

(2) 독립표본

데이터 탐색기에서는 상호 독립적인 각각의 필드에 대해 **평균 비교 분석 - 독립표본** 기능을 사용하여, 독립표본 T-test 를 실시합니다.

실행 방법



[분석] – [평균 비교 분석] – [독립표본]을 선택하면, **독립표본** 윈도우가 나타납니다. 분석변수, 독립변수를 선택하고 유의수준을 정합니다.

결과

집단 통계량과 독립표본검정 결과를 보여줍니다.

t-Test(독립표본)

집단통계량

	이탈여부	데이터 개수	평균	표준편차	평균의 표준 오차
A5	0	3846	174.845	89.3079	1.4401
	1	4684	180.282	77.0929	1.1264

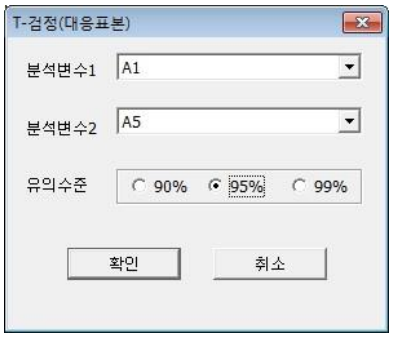
독립표본검정

필드명	가정	F	유의확률 (Levene)	t	자유도	유의확률(양쪽)	평균차	차이의 표준오차	차이의 95% 신뢰구간
A5	등분산 가정	104.2609	< 0.0001	-3.0168	8528	0.0026	-5.437	1.8023	(-8.971, -1.903)
	등분산 가정되지 않음			-2.9738	7640.9236	0.0030	-5.437	1.8283	(-9.022, -1.852)

(3) 대응표본

데이터 탐색기에서는 상호 의존적인 각각의 필드에 대해 **평균 비교 분석 - 대응표본** 기능을 사용하여, 대응표본 T-test 를 실시합니다.

실행 방법



[분석] – [평균 비교 분석] – [대응표본]을 선택하면, **대응표본** 윈도우가 나타납니다. 분석변수 1, 분석변수 2 를 선택하고 유의수준을 정합니다.

결과

대응표본 통계량과 대응표본 검정 결과를 보여줍니다.

t-Test(대응표본)

대응표본 통계량

	평균	데이터 개수	표준편차	평균의 표준 오차
A5	177,8306	8530	82,8629	0,8972
A6	49,9553	8530	36,3919	0,394

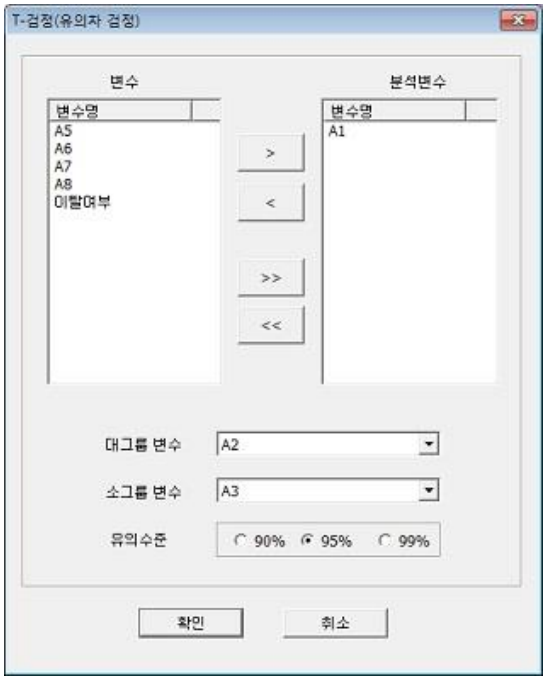
대응표본 검정

	평균	표준편차	평균의 표준 오차)	차이의 95% 신뢰구간	t	자유도	유의확률(양쪽)
A5 - A6	127,8753	88,3331	0,9564	(126,000, 129,751)	133,7019	8529	0

(4) 유의차 검정

데이터 탐색기에서는 두 개의 이산형 필드와 하나의 연속형 필드에 대해 **평균 비교 분석 – 유의차 검정** 기능을 사용하여, 유의차 검정을 실시합니다.

실행 방법



[분석] – [평균 비교 분석] – [유의차 검정]을 선택하면, **유의차 검정** 윈도우가 나타납니다. 대그룹 변수, 소그룹 변수를 선택하고 유의수준을 정합니다.

결과

등분산이 가정되었을 때와 가정되지 않았을 때의 유의차 검정 결과를 보여줍니다.

● 등분산 가정되었을 경우

대그룹 (이탈여부)	소그룹 (A3)	필드명	t	자유도	유의확률(양쪽)	판정	평균차	차이의 표준오차	차이의 95% 신뢰구간
0	H <--> M	A5	3740,1369	366	0	차이있음	227,7635	0,0609	(227,644, 227,883)
	H <--> MH	A5	813,9634	79	0	차이있음	129,8986	0,1596	(129,581, 130,216)
	H <--> ML	A5	229,7451	14	0		265,9839	1,1577	(263,500, 268,468)
	M <--> MH	A5	-4985,001	417	0	차이있음	-97,8649	0,0196	(-97,903, -97,826)
	M <--> ML	A5	514,439	352	0		38,2204	0,0743	(38,074, 38,367)
	MH <--> ML	A5	411,61	65	0		136,0852	0,3306	(135,425, 136,746)
1	H <--> L	A5	229,295	21	0		214,4217	0,9351	(212,476, 216,367)
	H <--> M	A5	5769,8447	576	0	차이있음	227,616	0,0394	(227,538, 227,693)
	H <--> MH	A5	1071,5493	109	0	차이있음	120,2721	0,1122	(120,050, 120,495)
	H <--> ML	A5	606,637	27	0		249,8528	0,4119	(249,007, 250,698)
	L <--> M	A5	262,0469	555	0		13,1943	0,0504	(13,095, 13,293)
	L <--> MH	A5	-340,4011	88	0		-94,1495	0,2766	(-94,699, -93,600)
	L <--> ML	A5	31,2327	6	0,		35,4311	1,1344	(32,656, 38,207)
	M <--> MH	A5	-8397,0023	643	0	차이있음	-107,3438	0,0128	(-107,369, -107,319)
	M <--> ML	A5	484,7528	561	0		22,2369	0,0459	(22,147, 22,327)
	MH <--> ML	A5	873,8624	94	0		129,5807	0,1483	(129,286, 129,875)

● 등분산 가정되지 않았을 경우

대그룹 (이탈여부)	소그룹 (A3)	필드명	t	자유도	유의확률(양쪽)	판정	평균차	차이의 표준오차	차이의 95% 신뢰구간
0	H <--> M	A5	50,1303	14,0144	0	차이있음	227,7635	4,5434	(218,015, 237,512)
	H <--> MH	A5	28,0969	15,0212	0,	차이있음	129,8986	4,6232	(120,041, 139,756)
	H <--> ML	A5	58,5576	0	-2		265,9839	4,5423	(270,526, 261,442)
	M <--> MH	A5	-112,7916	66,8734	0	차이있음	-97,8649	0,8677	(-99,598, -96,132)
	M <--> ML	A5	370,6347	0	-2		38,2204	0,1031	(38,323, 38,117)
	MH <--> ML	A5	157,9611	0	-2		136,0852	0,8615	(136,947, 135,224)
1	H <--> L	A5	57,2556	0	-2		214,4217	3,745	(218,167, 210,677)
	H <--> M	A5	60,7711	21,0106	0	차이있음	227,616	3,7455	(219,824, 235,408)
	H <--> MH	A5	31,5601	22,5109	0	차이있음	120,2721	3,8109	(112,366, 128,178)
	H <--> ML	A5	55,1213	25,6556	0		249,8528	4,5328	(240,514, 259,191)
	L <--> M	A5	221,5121	0	-2		13,1943	0,0596	(13,254, 13,135)
	L <--> MH	A5	-133,4223	0	-2		-94,1495	0,7057	(-93,444, -94,855)
	L <--> ML	A5	13,8747	0	-2		35,4311	2,5536	(37,985, 32,877)
	M <--> MH	A5	-151,5813	89,2578	0	차이있음	-107,3438	0,7082	(-108,751, -105,936)
	M <--> ML	A5	8,7055	6,0065	0,0001		22,2369	2,5543	(15,987, 28,486)
	MH <--> ML	A5	48,9104	6,9485	0,		129,5807	2,6493	(123,099, 136,063)

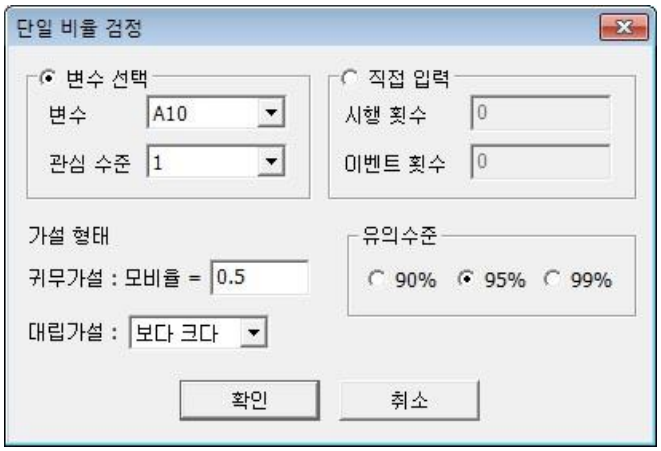
- 5.3.2.4 비율 검정

(1) 단일 비율 검정

데이터 탐색기에서는 선택된 필드의 데이터 값들이 이항 자료인 경우 그 비율에 대한 신뢰구간 계산 및 가설 검정을 하는 **단일 비율 검정** 기능을 제공합니다. ECMiner™ 에서는 정규분포에 근사 된 통계량 및 신뢰구간을 제공합니다.

실행 방법

[분석] - [기초통계] – [비율 검정] – [단일 비율 검정]을 선택하면 다음과 같은 단일 비율 검정 다이얼로그가 나타납니다.



변수 선택 방식과 직접 입력 방식 중 원하는 진행 방식을 선택 후 다음 절차로 수행합니다.

- 변수 선택 방법은 데이터 탐색창의 이항 변수를 사용하는 방법으로, 해당 변수(이항 변수여야만 함)를 선택하고 두 가지의 수준 중에 어느 수준을 이벤트로 할 것인지 결정합니다.
- 직접 입력 방법에서는 베르누이 시행의 횟수와 그 중 이벤트 시행의 횟수를 직접 입력하게 됩니다.

검증하고자 하는 모집단 비율값을 귀무가설로 설정하고 대립가설의 형태, 유의수준을 선택합니다.

실행 결과

선택한 방법에 대한 Test 결과를 다음과 같이 표로 나타내 줍니다. P-value 가 작으면 작을수록 귀무가설을 기각할 가능성이 커지게 되어 모집단의 비율을 대립가설의 형태와 같이 주장할 수 있게 됩니다.

단일 비율 검정

가설 검정 및 신뢰 구간

귀무가설 : 모비율 = 0.500000, 대립가설 : 보다 크다

변수	관측치수	이벤트수	표본비율	95% C.I	z값	p값
A10	8530	4684	0.54912	(0.540255 ,)	9.07339	0

(2) 두 비율 검정

데이터 탐색기에서는 데이터 값들이 이항 자료인 두 변수에 대하여 이벤트 비율간 차이에 대한 신뢰구간 계산 및 가설 검정을 하는 **두 비율 검정** 기능을 제공합니다. ECMiner™ 에서는 정규분포에 근사 된 통계량 및 신뢰구간을 제공합니다.

실행 방법

[분석] - [기초통계] – [비율 검정] – [두 비율 검정]을 선택하면 다음과 같은 두 비율 검정 다이얼로그가 나타납니다.

변수 선택 방식과 직접 입력 방식 중 원하는 진행 방식을 선택 후 다음 절차로 수행합니다.

- 변수 선택 방법은 데이터 탐색창의 이항 변수를 사용하는 방법으로, 해당 변수(이항 변수여야만 함)를 선택하고 두 가지의 수준 중에 어느 수준을 이벤트로 할 것인지 결정합니다.
- 직접 입력 방법에서는 베르누이 시행의 횟수와 그 중 이벤트 시행의 횟수를 직접 입력하게 됩니다.

검증하고자 하는 모집단 비율값을 귀무가설로 설정하고 대립가설의 형태, 유의수준을 선택합니다.

실행 결과

선택한 방법에 대한 Test 결과를 다음과 같이 표로 나타내 줍니다. P value 가 작으면 작을수록 귀무가설을 기각할 가능성이 커지게 되어 두 모집단의 비율 차를 대립가설의 형태와 같이 주장할 수 있게 됩니다.

두 비율 검정

가설 검정 및 신뢰 구간

귀무가설 : 모비율 차 = 0.000000, 대립가설 : 같지 않다

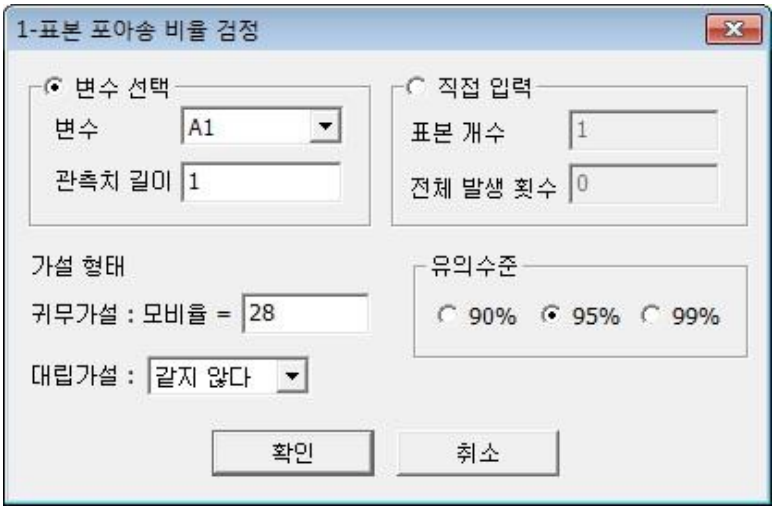
변수	관측치수	이벤트수	표본비율	95% C.I	z값	p값
A10	8530	4684	0.54912	(-0.141464 , -0.112463)	-17.16683	0.00000
A11	8530	5767	0.67608	(-0.141464 , -0.112463)	-17.16683	0.00000

(3) 1-표본 포아송 비율 검정

데이터 탐색기에서는 선택된 필드의 데이터 값들이 포아송 자료인 경우 그 비율에 대한 신뢰구간 계산 및 가설 검정을 하는 **1-표본 포아송 비율 검정** 기능을 제공합니다. ECMiner™ 에서는 정규분포에 근사 된 통계량 및 신뢰구간을 제공합니다.

실행 방법

[분석] - [기초통계] – [비율 검정] – [1-표본 포아송 검정]을 선택하면 다음과 같은 1-표본 포아송 비율 검정 다이얼로그가 나타납니다.



변수 선택 방식과 직접 입력 방식 중 원하는 진행 방식을 선택 후 다음 절차로 수행합니다.

- 변수 선택 방법은 데이터 탐색창의 정수형 변수를 사용하는 방법으로, 해당 변수를 선택하고 관측치 길이를 결정합니다.
- 직접 입력 방법에서는 시행의 횟수와 그에 대한 이벤트 발생 횟수를 직접 입력하게 됩니다.

검증하고자 하는 모집단 비율값을 귀무가설로 설정하고 대립가설의 형태, 유의수준을 선택합니다.

실행 결과

선택한 방법에 대한 Test 결과를 다음과 같이 표로 나타내 줍니다. P-value 가 작으면 작을수록 귀무가설을 기각할 가능성이 커지게 되어 모집단의 비율을 대립가설의 형태와 같이 주장할 수 있게 됩니다.

1-표본 포아송 비율 검정

가설 검정 및 신뢰 구간

귀무가설 : 모비율 = 28.000000, 대립가설 : 같지 않다

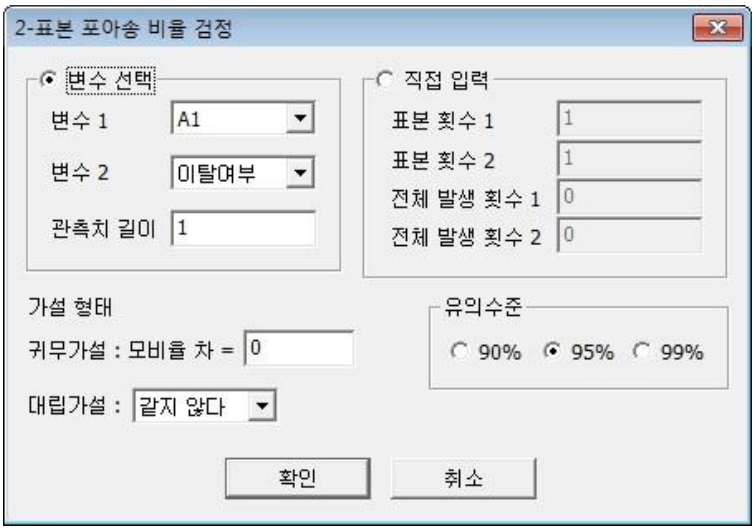
변수	표본개수	발생횟수	관측치 길이	발생율	평균발생횟수	95% C.I	z값	p값
A1	100	2854	1	28.54000	28.54000	(27.492595 , 29.587405)	-1.02050	0.30749

(4) 2-표본 포아송 비율 검정

데이터 탐색기에서는 데이터 값들이 포아송 자료인 두 변수에 대하여 이벤트 비율간 차이에 대한 신뢰구간 계산 및 가설 검정을 하는 2-표본 포아송 비율 검정 기능을 제공합니다. ECMiner™ 에서는 정규분포에 근사 된 통계량 및 신뢰구간을 제공합니다.

실행 방법

[분석] - [기초통계] – [비율 검정] – [2-표본 포아송 검정]을 선택하면 다음과 같은 2-표본 포아송 비율 검정 다이얼로그가 나타납니다.



변수 선택 방식과 직접 입력 방식 중 원하는 진행 방식을 선택 후 다음 절차로 수행합니다.

- 변수 선택 방법은 데이터 탐색창의 정수형 변수를 사용하는 방법으로, 두 개의 해당 변수를 선택하고 관측치 길이를 결정합니다.

- 직접 입력 방법에서는 비교 대상인 두 시행의 횟수와 그 각각에 대한 이벤트 발생 횟수를 직접 입력하게 됩니다.

검증하고자 하는 모집단 비율값을 귀무가설로 설정하고 대립가설의 형태, 유의수준을 선택합니다.

실행 결과

선택한 방법에 대한 Test 결과를 다음과 같이 표로 나타내 줍니다. P value 가 작으면 작을수록 귀무가설을 기각할 가능성이 커지게 되어 두 포아송 자료에 대한 모집단 비율 차를 대립가설의 형태와 같이 주장할 수 있게 됩니다.

2-표본 포아송 비율 검정

가설 검정 및 신뢰 구간

귀무가설 : 모비율 차 = 0.000000, 대립가설 : 같지 않다

변수	표본개수	발생횟수	발생율	평균발생횟수	비율차	95% C.I	z값	p값
A1	100	2854	28.54000	28.54000	-1	(-2.494174 , 0.494174)	-1.31216	0.18947
A10	100	2954	29.54000	29.54000	-1	(-2.494174 , 0.494174)	-1.31216	0.18947

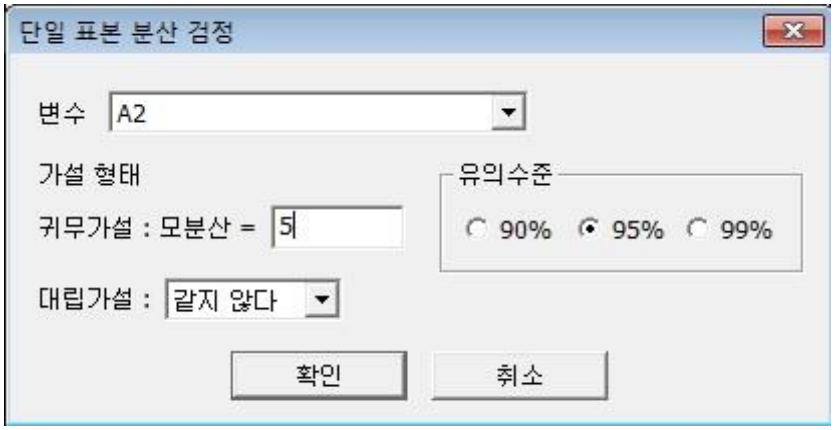
- 5.3.2.5 분산 검정

(1) 단일 표본 분산 검정

데이터 탐색기에서는 연속형 변수의 분산에 대한 신뢰구간 계산 및 가설 검정을 하는 **단일 표본 분산 검정** 기능을 제공합니다.

실행 방법

[분석] - [기초통계] – [분산 검정] – [단일 표본 분산 검정]을 선택하면 다음과 같은 단일 표본 분산 검정 다이얼로그가 나타납니다.



단일 표본 분산 검정을 하기 위해 대상 변수를 선택합니다.
그리고 검증하고자 하는 모집단 분산값을 귀무가설로 설정하고 대립가설의 형태, 유의수준을 선택합니다.

실행 결과

선택한 방법에 대한 Test 결과를 다음과 같이 표로 나타내 줍니다. P value 가 작으면 작을수록 귀무가설을 기각할 가능성이 커지게 되어 모집단의 분산을 대립가설의 형태와 같이 주장할 수 있게 됩니다.

단일 표본 분산 검정

가설 검정 및 신뢰 구간

귀무가설 : 모분산 = 5.000000, 대립가설 : 같지 않다

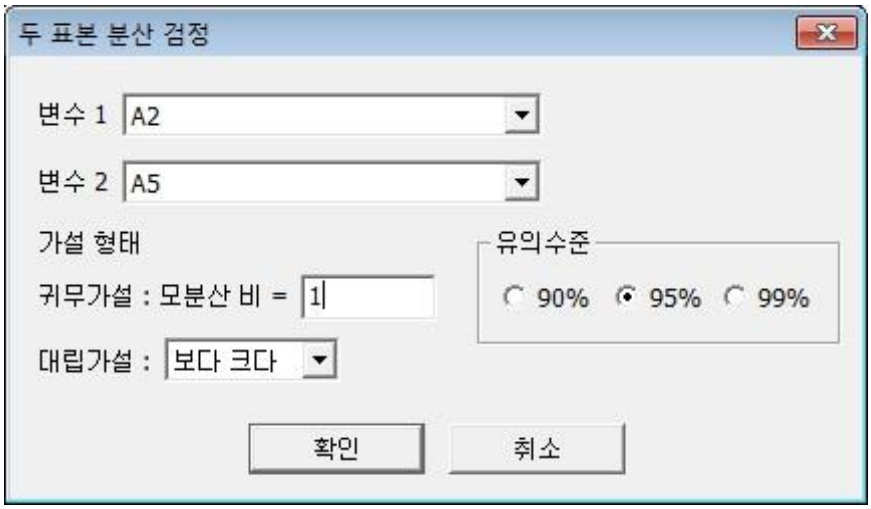
변수	관측치수	표준편차	분산	95% C.I	자유도	x2값	p값
A2	4462	2.39155	5.71951	(5.489377 , 5.964490)	4461	5102.94429	0.00000

(2) 두 표본 분산 검정

데이터 탐색기에서는 두 연속형 변수의 분산 비에 대한 신뢰구간 계산 및 가설 검정을 하는 **두 표본 분산 검정** 기능을 제공합니다.

실행 방법

[분석] - [기초통계] – [분산 검정] – [두 표본 분산 검정]을 선택하면 다음과 같은 단일 표본 분산 검정 다이얼로그가 나타납니다.



두 표본 분산 검정을 하기 위해 대상이 되는 두 변수를 선택합니다.
그리고 검증하고자 하는 모집단 분산 비율을 귀무가설로 설정하고 대립가설의 형태, 유의수준을 선택합니다.

실행 결과

선택한 방법에 대한 Test 결과를 다음과 같이 표로 나타내 줍니다. P value 가 작으면 작을수록 귀무가설을 기각할 가능성이 커지게 되어 두 모집단의 분산 비를 대립가설의 형태와 같이 주장할 수 있게 됩니다.

두 표본 분산 검정

가설 검정 및 신뢰 구간

귀무가설 : 모분산 비 = 1.000000, 대립가설 : 보다 크다

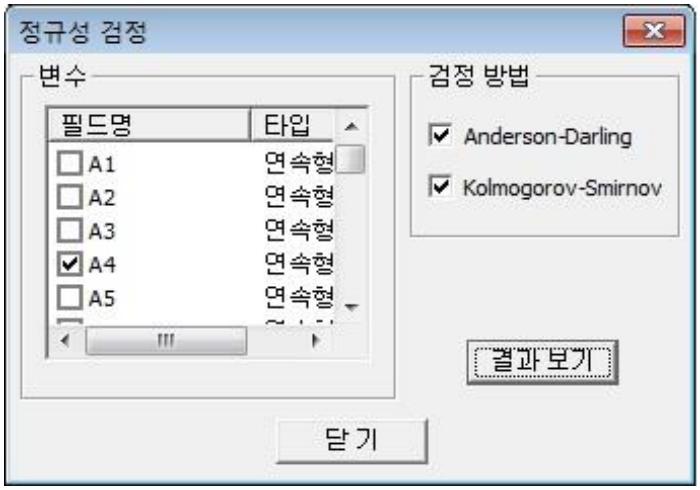
변수	관측치수	표준편차	분산	분산비율	95% C.I	자유도1	자유도2	f값	p값
A2	4462	2.39155	5.71951	1.29266	(1.230523 ,)	4461	4461	1.29266	0
A5	4462	2.10348	4.42462	1.29266	(1.230523 ,)	4461	4461	1.29266	0

- **5.3.2.6 정규성 검정**

정규성 검정 기능은 선택된 필드의 데이터 값들이 정규 분포를 따르는지 여부를 통계적으로 검정해 주는 기능을 제공합니다. ECMiner™ 에서는 Anderson-Darling, Kolmogorov-Smirnov Test 를 제공합니다.

실행 방법

[분석] – [기초통계] - [정규성 검정]을 선택하면 다음과 같은 정규성 검정 다이얼로그가 나타납니다.

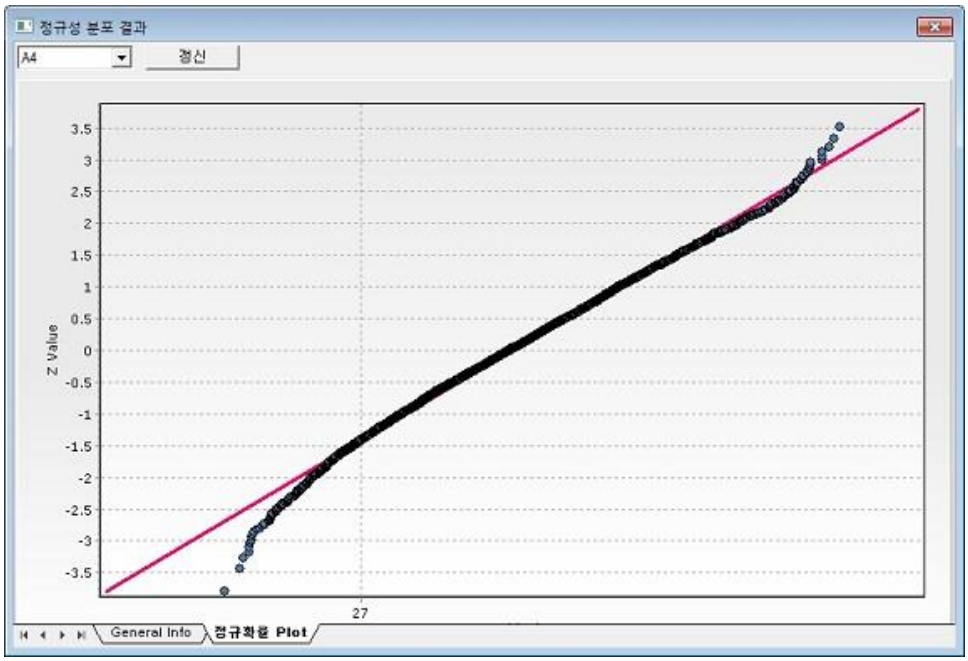


변수 목록에서 정규성을 테스트하고자 하는 변수를 선택하고(복수 선택 가능), 검정 방법에서 어떠한 Test 를 사용할지를 선택합니다.(복수 선택 가능)

실행 결과

General Information: 선택한 방법에 대한 Test 결과를 다음과 같이 표로 나타내 줍니다. P value 가 작으면 작을수록 정규 분포가 아님을 나타낸다고 할 수 있습니다.

정규 확률 Plot: 빨간 선을 기준으로 빨간 선 근방에 데이터가 많이 분포하면 데이터가 정규성을 따른다고 할 수 있고, 빨간 선에 많이 벗어날수록 정규성 가정에 위배됩니다.



•

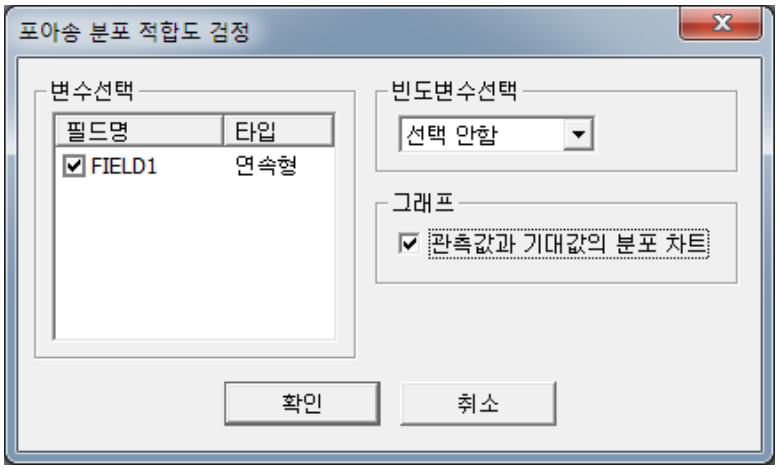
- 5.3.2.7 포아송 검정

포아송 적합도 검정은 수집된 자료가 포아송 분포를 따르는지 검정하는 기법입니다.

실행 방법

1. **[분석] – [기초통계] – [포아송 검정]**을 선택하면, 포아송 분포 적합도 검정 윈도우가 나타납니다.

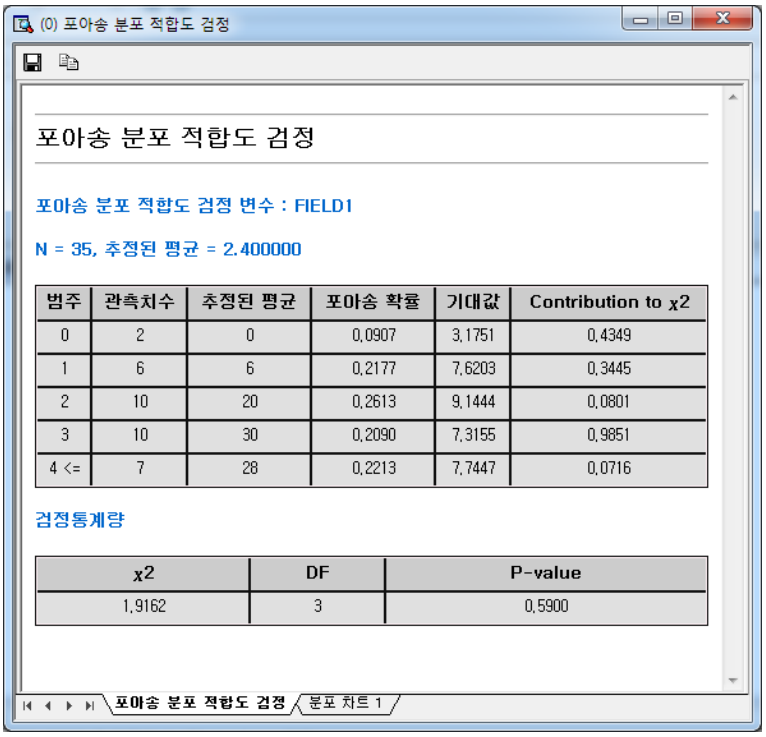
2. 검정할 변수를 선택합니다. 추가적으로 관측값과 기대값의 분포 차트를 선택할 수 있습니다.



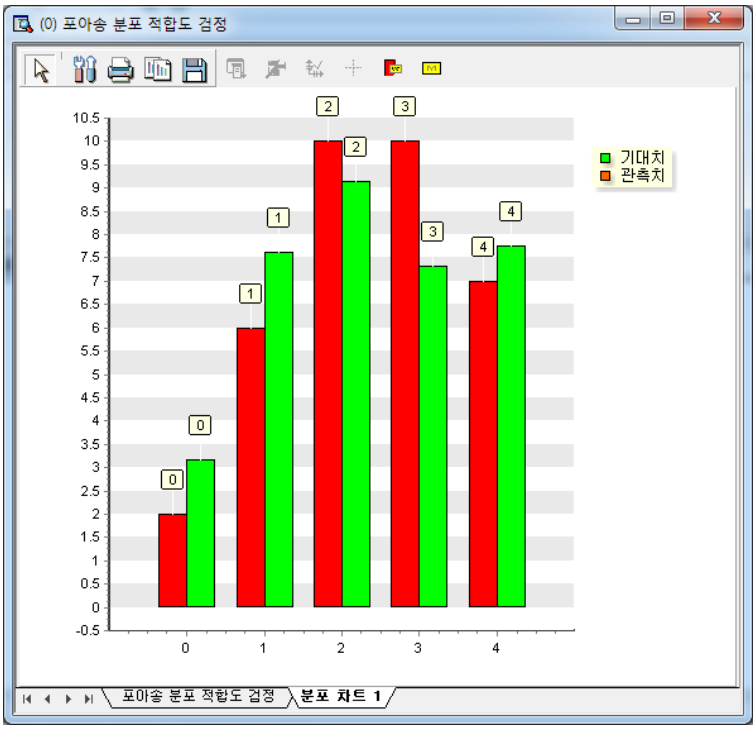
결과

포아송 분포 적합도 검정 결과가 다음과 같이 나타납니다.

각 범주별 포아송 확률, 기대값, 그리고 카이제곱 검정 통계량에 대한 contribution 이 출력되며, p-value 를 통해 기대빈도에 대한 자료의 적합도를 검정할 수 있습니다.



결과창의 포아송분포 탭을 선택하면 다음과 같이 관측값과 기대값의 분포차트를 확인할 수 있습니다..



5.3.3 분산분석

- **5.3.3.1 일원배치**

일원배치 분산 분석은 하나의 인자에 대한 분산 분석 방법입니다. 인자와 관측치를 선택하면 분산분석을 하실 수 있습니다.

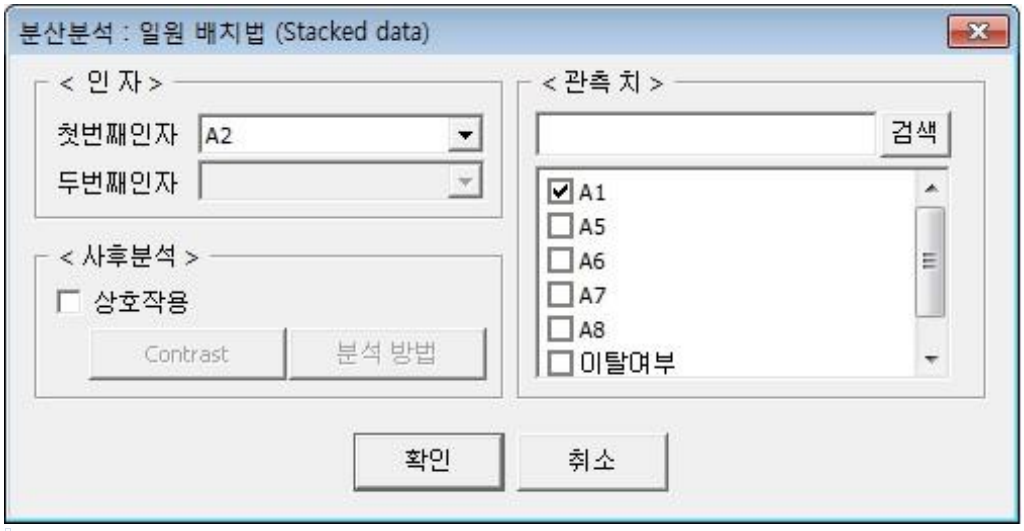
분산 분석 다이얼로그 열기

메뉴의 **[분석]** - **[분산분석]**을 다시 선택하면 3 가지 분산 분석 종류가 나열됩니다. 그 중에서 **[일원배치법]**을 선택합니다.

일원 배치: Stack 데이터

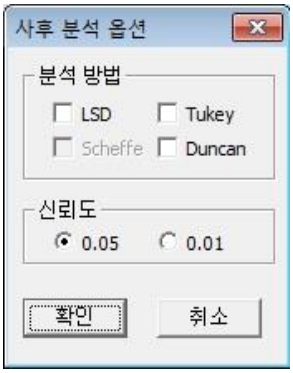
인자: 그룹을 구분 짓는 이산형인 변수를 선택합니다.

관측치: 그룹별로 일원배치 분산분석을 실시할 변수를 선택합니다.



사후 분석 및 신뢰도

일원배치의 사후 분석 방법으로는 3 가지 사후 분석 방법이 제공되고 있습니다. 필요에 따라 원하는 사후 분석방법을 체크 하시면 됩니다. 신뢰도는 0.05 와 0.01 두 가지가 있습니다. 필요에 따라 원하는 신뢰 구간을 선택 하시면 됩니다.



LSD : The Least Significant Difference Method

Tukey : Tukey's Test (by Tukey(1953))

Duncan : Duncan's Multiple Range Test

분석 결과

일원 배치 분산 분석을 실행 하면 아래와 같은 결과를 볼 수 있습니다. 보여진 결과는 결과 윈도우의 왼쪽 위에 있는 저장 아이콘을 이용하여 HTML 파일로 저장을 할 수 있습니다.

(1) 분산분석

일원배치 분산분석표

관측 변수명: A1

출처	자유도	변동	평균변동	F Value	Pr > F
A2 (Factor)	4	1736.82797	434.20699	2.64882	0.03158
Error	8525	1397455.82854	163.92444		
Total	8529	1399192.65651			

LSD method

Alpha : 0.0500
Least Significant Difference : 1.0121

*같은 그룹에 속해 있는 수준의 평균은 유효하게 다르지 않습니다.

LSD Grouping	LSD Grouping	Mean	N	Variable
Group1		31.06360	1824	D
Group1	Group2	30.58356	3345	E
	Group2	30.05014	1037	B
	Group2	29.86622	598	C
	Group2	29.84705	1726	A

분산 분석 : 일원 배치

- 5.3.3.2 이원배치 : stack

이원 배치 분산분석은 두 가지 인자에 대한 분산 분석입니다. 인자 1 과 인자 2, 관측치를 선택하면 분석을 하실 수 있습니다.

분산 분석 다이얼로그 열기

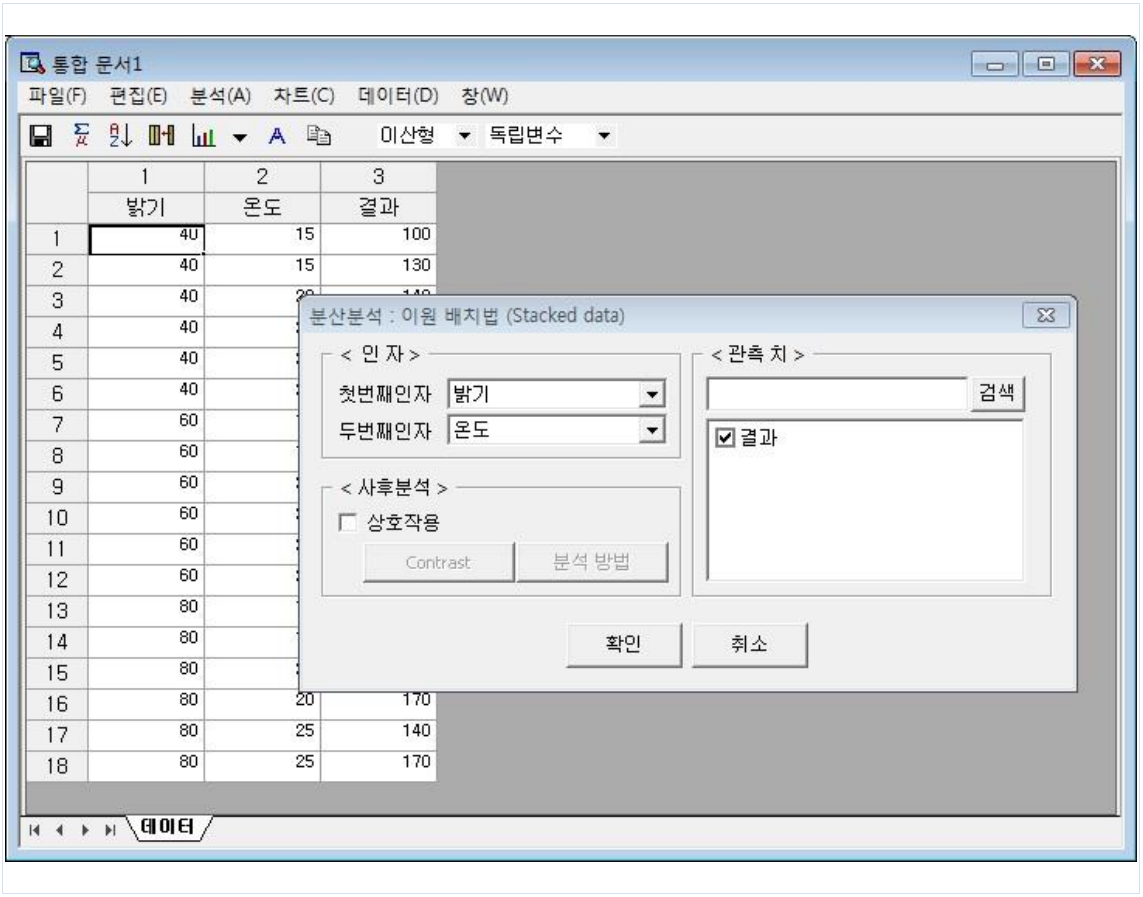
메뉴의 **[분석]** - **[분산분석]**을 다시 선택하면 3 가지 분산 분석 종류가 나열됩니다. 그 중에서 **[이원 배치법]**을 선택하시면 됩니다.

이원 배치: Stack 데이터

인자 1: "밝기" 필드 --> 40, 60, 80 lux 의 밝기 변화를 주었다.

인자 2: "온도" 필드 --> 15, 20, 25 의 온도 변화를 주었다

관측치: "결과" 필드 --> 세번째 결과 필드가 관측치이다.



분석 결과

이원 배치 분산 분석을 실행 하면 아래와 같은 결과를 볼 수 있습니다. 보여진 결과는 결과 윈도우의 왼쪽 위에 있는 저장 아이콘을 이용하여 HTML 파일로 저장을 할 수 있습니다.

이원배치 분산 분석표

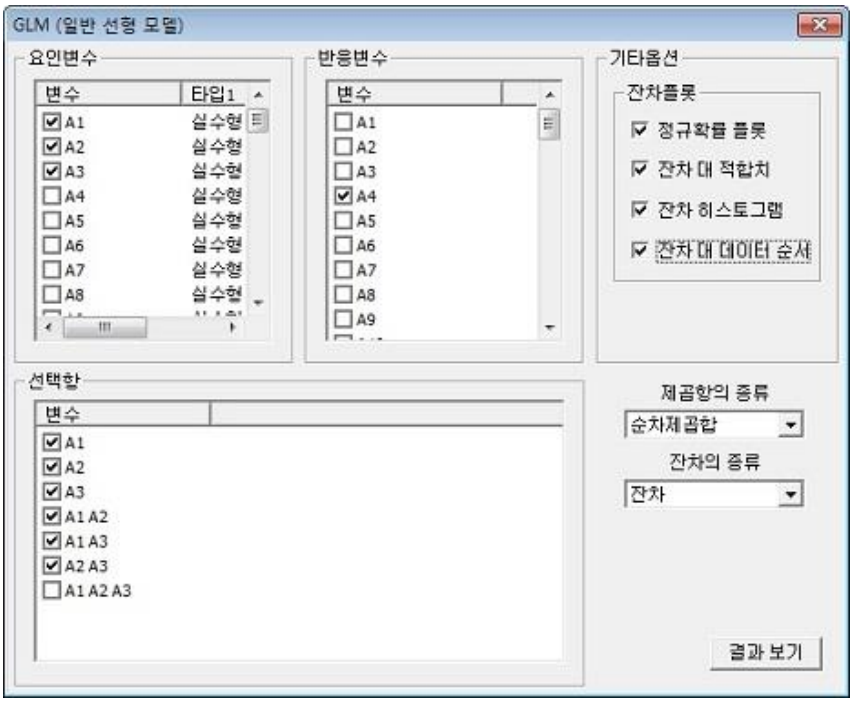
관측 변수명: 결과

출처	자유도	변동	평균변동	F	P value
밝기 (Factor A)	2	5233,33333	2616,66667	2,50532	0,13647
온도 (Factor B)	2	9033,33333	4516,66667	4,32447	0,04829
상호작용	4	8133,33333	2033,33333	1,94681	0,18676
잔차오차	9	9400,00000	1044,44444		
전체	17	31800,00000			

- 5.3.3.3 GLM(일반 선형 모형)

일반 선형 모형은 가장 일반적인 형태의 분산 분석 절차라고 할 수 있습니다. ECMiner™에서 제공하는 분산 분석은 불균형 데이터를 처리할 수 없고 공변량을 처리할 수 없습니다. 하지만 일반 선형 모형에서는 실제 상황에서 많이 나타나는 불균형 데이터를 처리할 수 있고 공변량 또한 처리할 수 있습니다.

실행 방법

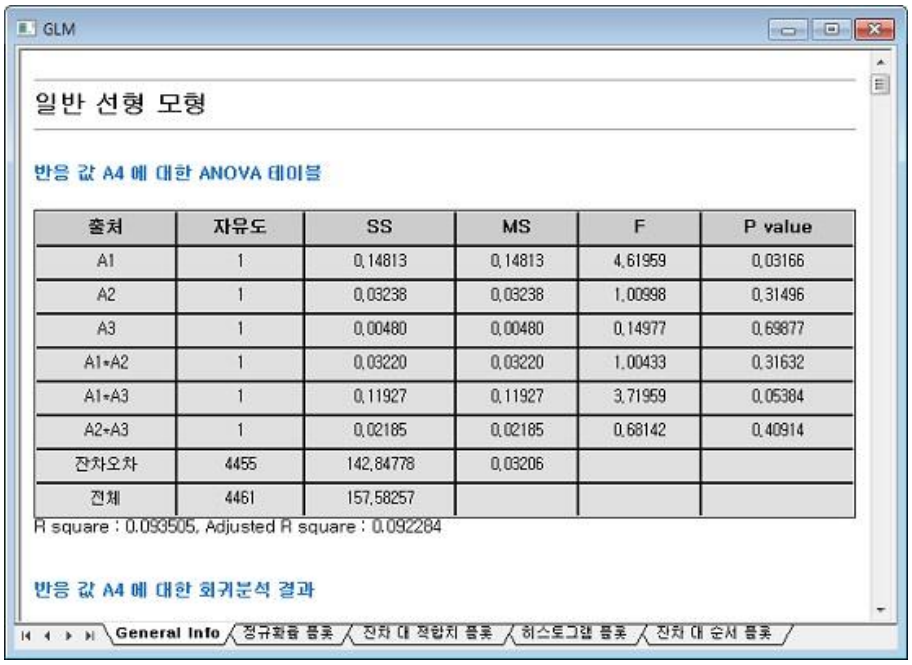


[분석] – [분산 분석] – [일반 선형 모형]을 선택하면, **일반 선형 모형** 윈도우가 나타납니다.

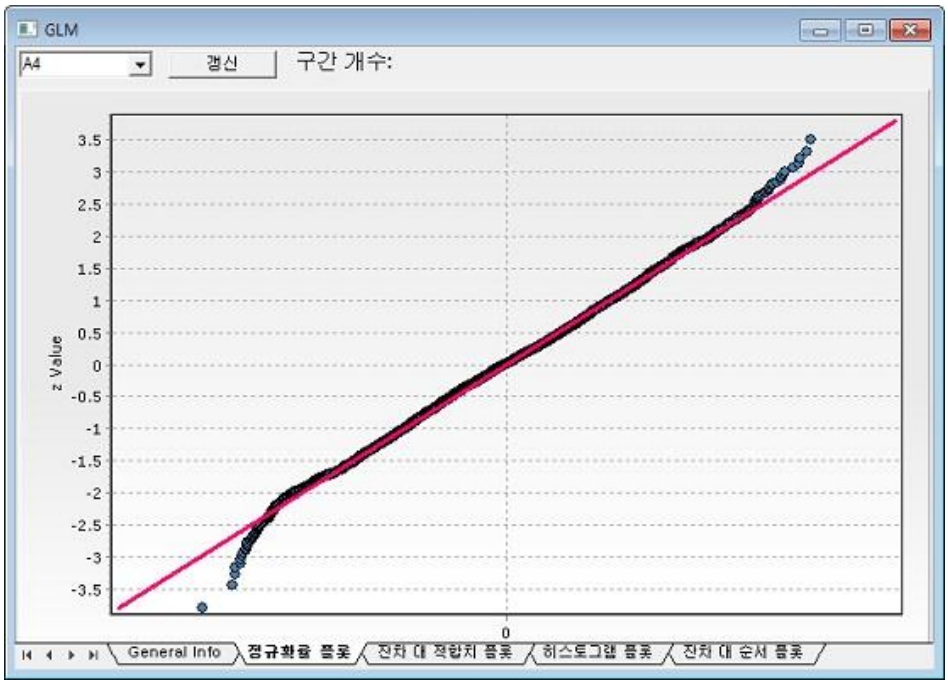
- 요인 변수: 실험에서 사용된 요인에 해당하는 변수를 입력합니다. 이 때 선택된 연속형 데이터는 자동적으로 공변량으로 처리되게 됩니다.
- 반응 변수: 반응 변수는 실험에서 반응 값에 해당하는 변수를 말합니다. 복수 선택이 가능하며 복수로 선택하였을 경우 여러 반응 변수에 대한 분산 분석 및 선형 회귀 모형에 대한 결과를 얻을 수 있습니다.
- 선택항: 선택한 요인에 대해 어떤 선택항을 선택할지를 결정합니다. 주요인, 2 요인 상호작용, 3 요인 상호작용을 선택할 수 있습니다.
- 잔차 플롯: 어떠한 형태의 잔차 플롯을 선택할지를 선택합니다. ECMiner™ 에서는 정규 확률 플롯, 잔차 대 적합치, 잔차 히스토그램, 잔차 대 순서 플롯을 제공합니다.
- 제곱항의 종류: 수정제곱항을 사용할 것인지 순차 제곱항을 사용할 것인지를 선택합니다.
- 잔차의 종류: 잔차, 표준화 잔차, 외표준화 잔차 중 어떠한 형태의 잔차를 사용할 것인지를 결정합니다.

결과

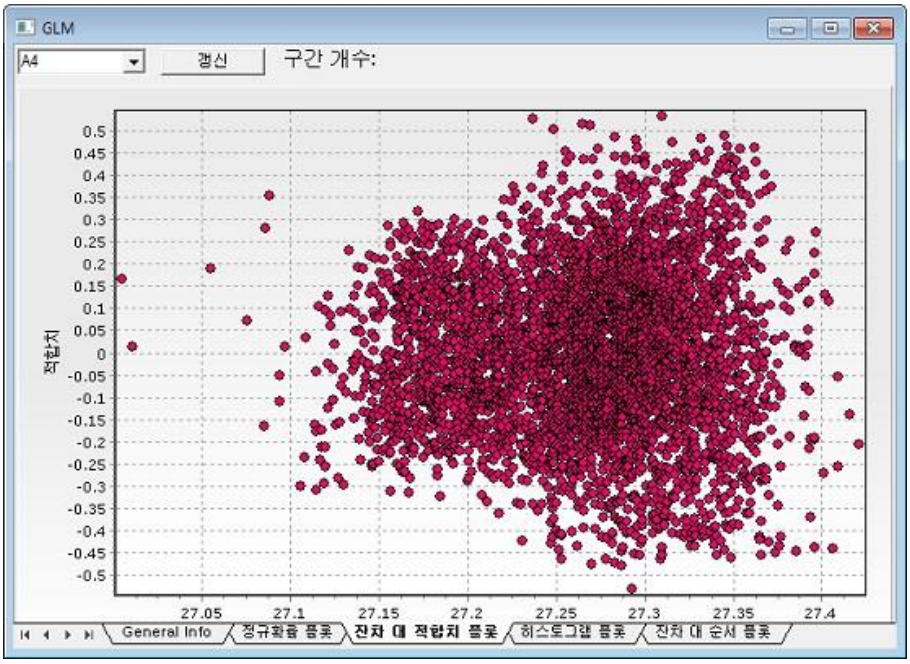
- General Information: General Information 을 통해서 설계에 대한 ANOVA Table, 회귀분석 결과, 비정상적 관측치 결과를 보여줍니다.



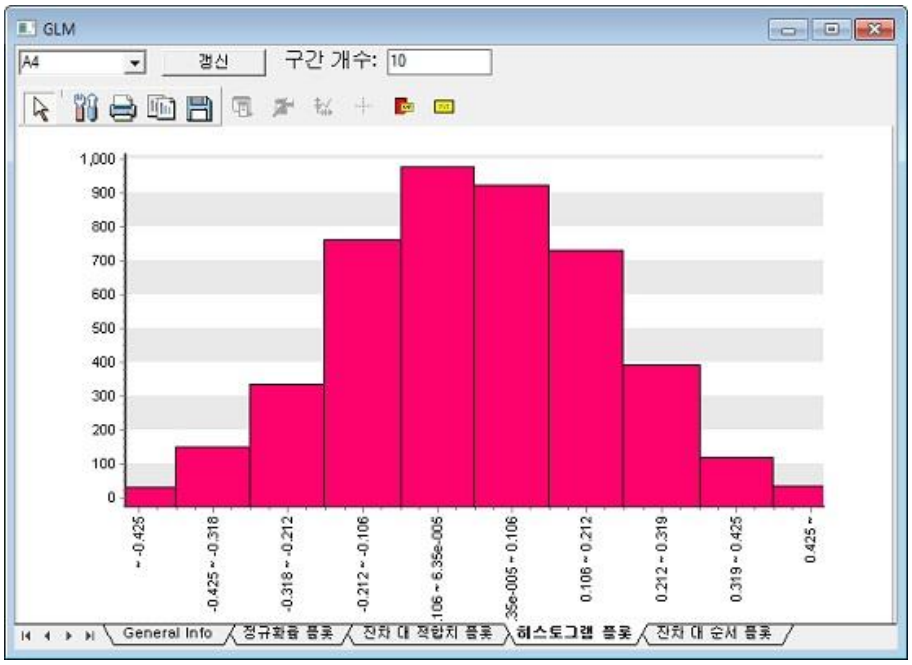
- 정규 확률 플롯: 잔차에 대한 정규 확률 플롯을 보여줍니다. 붉은 선 근처에 데이터 점이 많을수록 잔차가 정규 분포에 가까움을 나타냅니다.



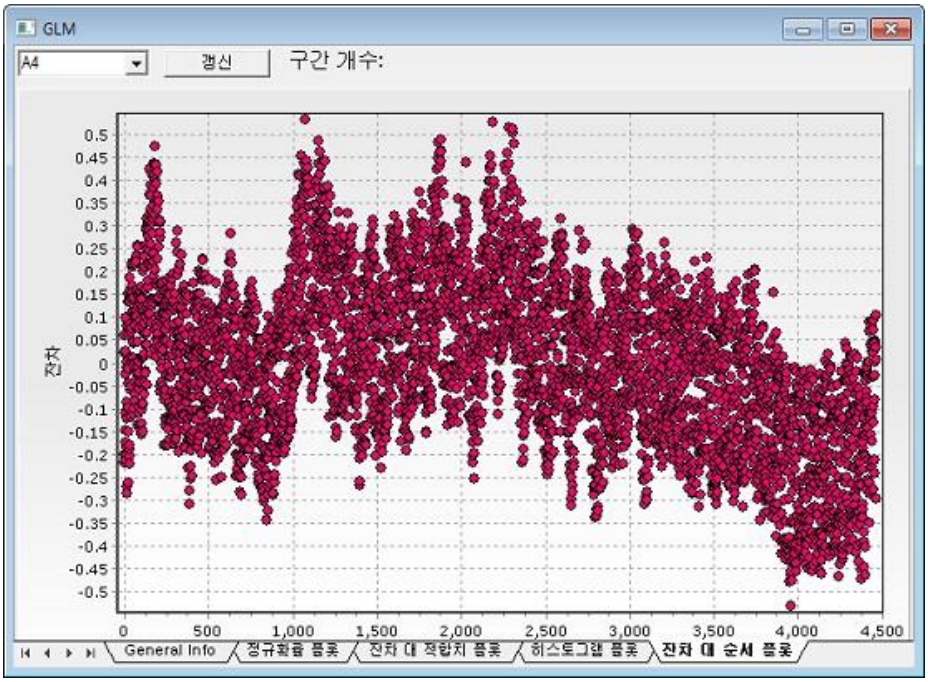
▪ 잔차 대 적합치 플롯: 가로 축을 잔차, 세로 축을 적합치로 한 플롯을 표시해 줍니다.



▪ 히스토그램 플롯: 잔차를 이용하여 히스토그램을 표시해줍니다.



- 잔차 대 순서 플롯: 순서 별로 잔차를 플롯해 줍니다.



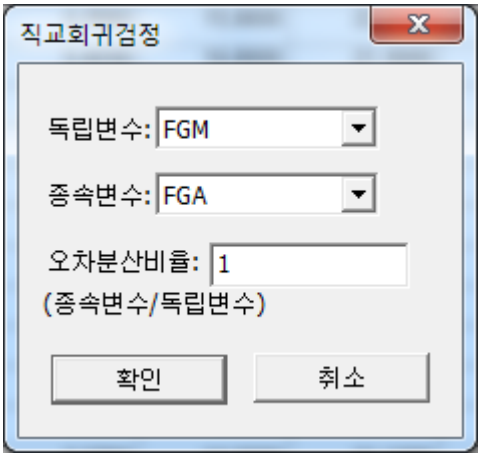
5.3.4 회귀분석

- **5.3.4.1 직교 회귀분석**

  일반적 회귀분석과 달리 직교회귀분석은 반응변수와 예측변수 모두에 대한 측정오차를 포함하기 때문에 오차를 포함하는 예측변수에 대해 오차 분산 비를 지정하여 직교 회귀분석을 수행합니다.

실행 방법

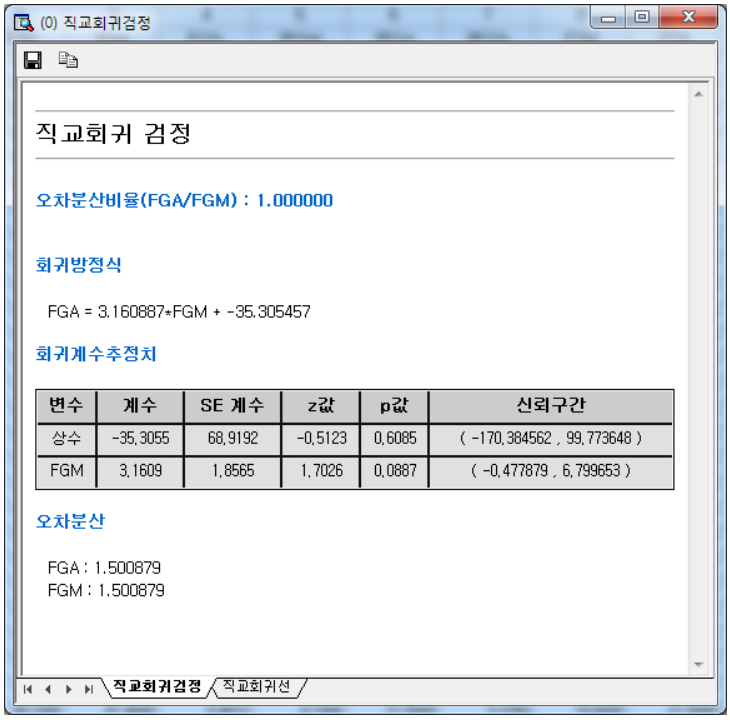
1. **[분석] – [회귀분석] – [직교회귀분석]**을 선택하면, 직교회귀검정 윈도우가 나타납니다.
2. 독립변수와 종속변수를 선택하고 오차분산비율(종속변수/독립변수)을 선택합니다.



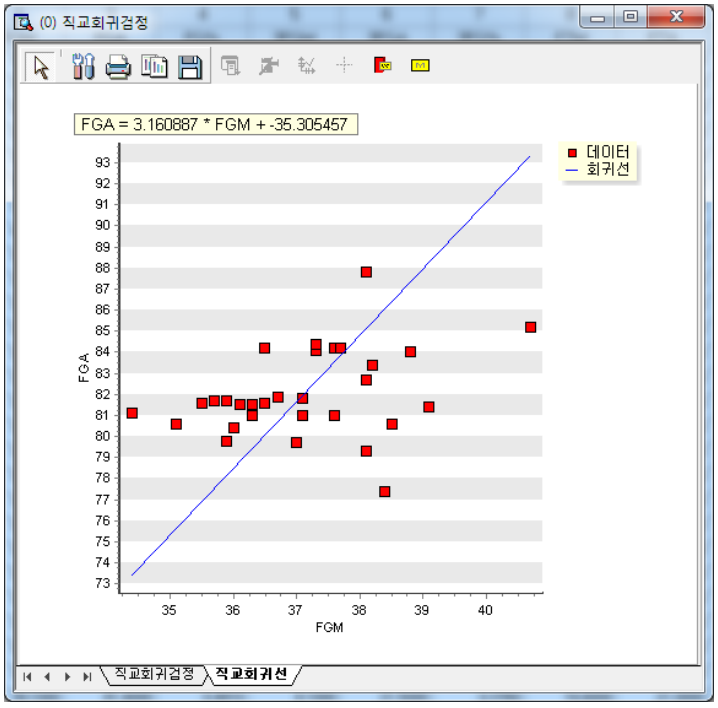
결과

직교회귀분석결과가 다음과 같이 나타납니다.

직교회귀식과 추정치 테이블, 각각의 오차분산이 출력됩니다.



결과창의 직교회귀선 탭을 선택하면 다음과 같이 관측값과 직교회귀선 그래프를 확인할 수 있습니다..



- 5.3.4.2 Nonlinear Regression

개요

▪ **비선형 최적화 알고리즘(Nonlinear Optimization Algorithm)**

일반적으로 함수 f 가 다음과 같은 성질을 가질 때 우리는 그 함수를 선형함수(Linear Function 혹은 Linear Map)이라고 합니다.

F(x+y)=F(x)+F(y)… (a)

F(ax)=aF(x) … (b)

정의상은 그렇지만 우리는 통상적으로 함수가 입력 변수에 대한 1 차식으로 이루어져 있으면 선형함수라고 합니다. 비선형 함수는 선형 함수가 아닌 모든 함수를 가리킵니다.(더욱 일반적) 이러한 비선형 함수를 최적화 하는 것이 이번 Nonlinear Regression 의 목적이라고 할 수 있습니다.

▪ **비선형 최소 제곱 회귀분석(Nonlinear Least Square Regression)**

다음과 같은 함수가 있다고 합시다.

$$F(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - M(\mathbf{x}|t_i))^2 = \frac{1}{2}\sum_{i=1}^{n}(f_i(\mathbf{x}))^2 = \frac{1}{2}\| \mathbf{f}(\mathbf{x}) \|^2$$

x 와 진한 글씨로 나타낸 것은 그것이 vector 임을 나타냅니다. 현재 (t_i, y_i)는 데이터로 주어진 상태에서 위의 F 함수를 최소화하는 것을 비선형 최소 제곱 회귀분석(Nonlinear Least Square Regression)이라고 합니다.

위의 f 가 비선형 함수일 때 문제는 어려워집니다. 함수 f 가 선형일 경우 F 를 최소화하는 x 벡터는 Closed Form 으로 구해집니다. 하지만 현재 선형 및 모든 함수를 포괄하는 방법을 찾아야 하는데 우리는 이를 위해 일반적인 비선형 최적화 알고리즘에 의존합니다. 이 중에서도 가장 널리 쓰이고 안정적이며 우수한 성능을 나타내는 것이 바로 Levenberg Marquardt 알고리즘입니다. ECMiner™ Nonlinear Regression 에서는 이 알고리즘을 이용하여 위의 문제를 해결합니다.

정형 수식과 사용자 정의 수식

ECMiner™ 에서 사용할 Curve Fitting 의 구성을 간단히 정리하면 다음과 같습니다.

▪ **정형 수식 회귀분석**

지수함수

다항함수

로그함수

거듭제곱함수 $y = ax^b$

선형함수

위의 기본적인 함수라는 것에는 공통점이 있습니다. 모든 함수를 소위 선형화시킬 수 있다는 것입니다. 다항함수, 로그함수, 선형함수는 모두 선형화되어 있습니다. Parameter 가 서로 일차식의 관계를 갖기 때문입니다. 하지만 지수함수와 거듭제곱함수는 그렇지 않은데 이는 다음과 같은 방법으로 쉽게 변형함으로써 해결할 수 있습니다.

$$y = ae^{bx} ?? \ln y = \ln a + l = bx$$

$$y = ax^b \rightarrow \ln y = \ln a + b\ln x$$

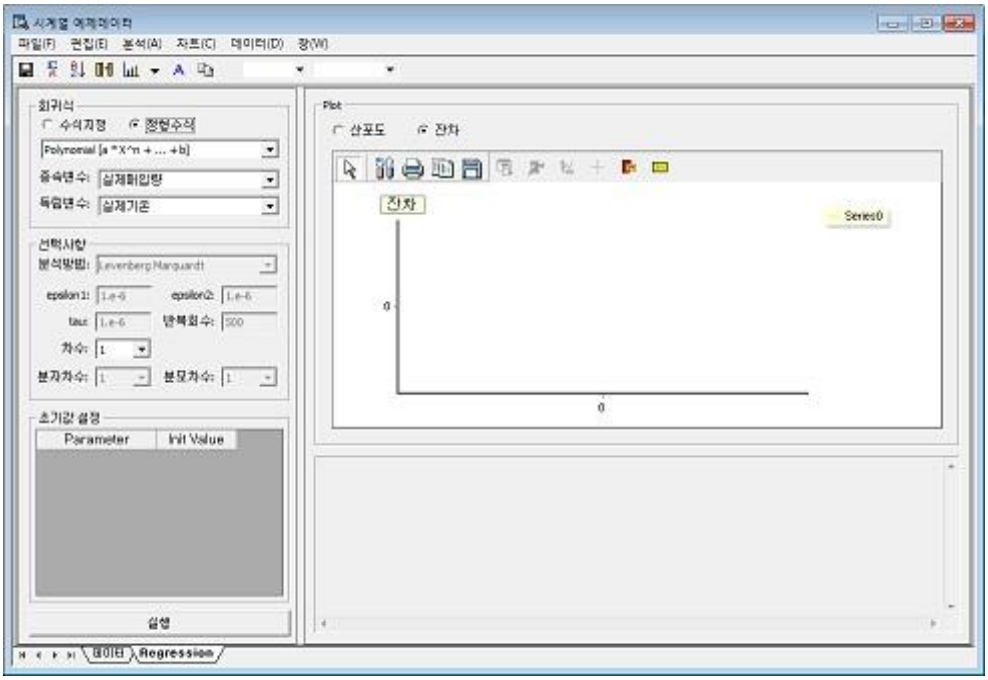
이런식으로 하면 쉽게 Parameter 를 추정할 수 있습니다. 하지만 이는 정확한 Least Square 가 아니라 정확한 결과와는 차이가 있을 수 있습니다. 하지만 꽤나 잘 맞는 예측이라고 받아들일 수는 있겠습니다.

▪ 사용자 정의 회귀분석

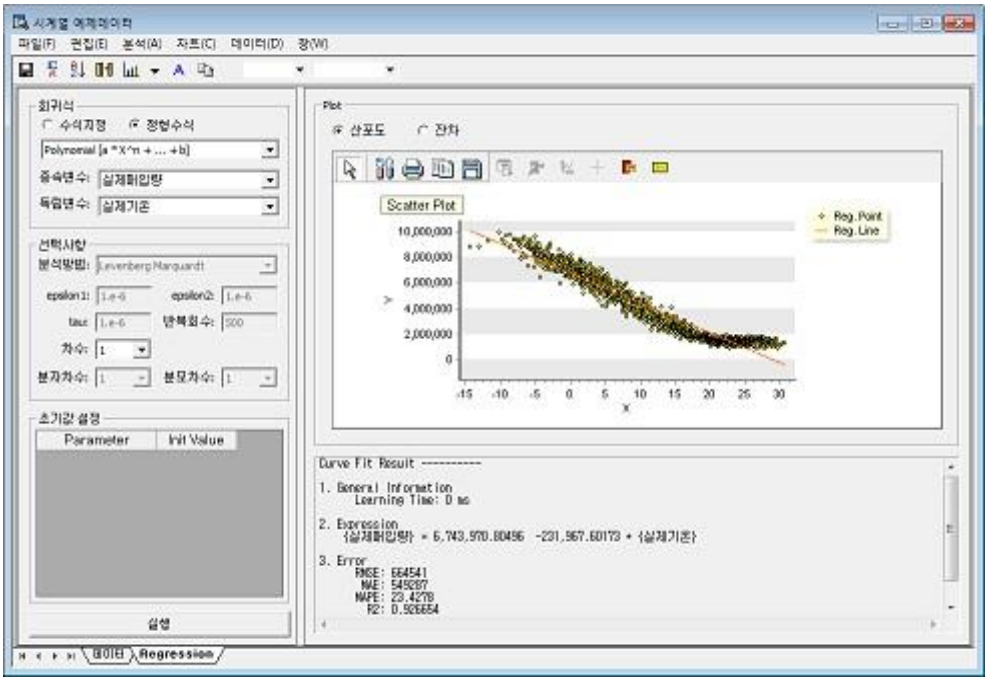
이는 사용자가 형태를 지정한 복잡한 모양에 대한 회귀분석을 하기 위한 것입니다. 사용자는 어떤 형태로든 원하는 함수의 모양을 입력할 수 있고 알고리즘은 이 함수에 가장 잘 맞는 Parameter 를 찾아줍니다.

실행방법

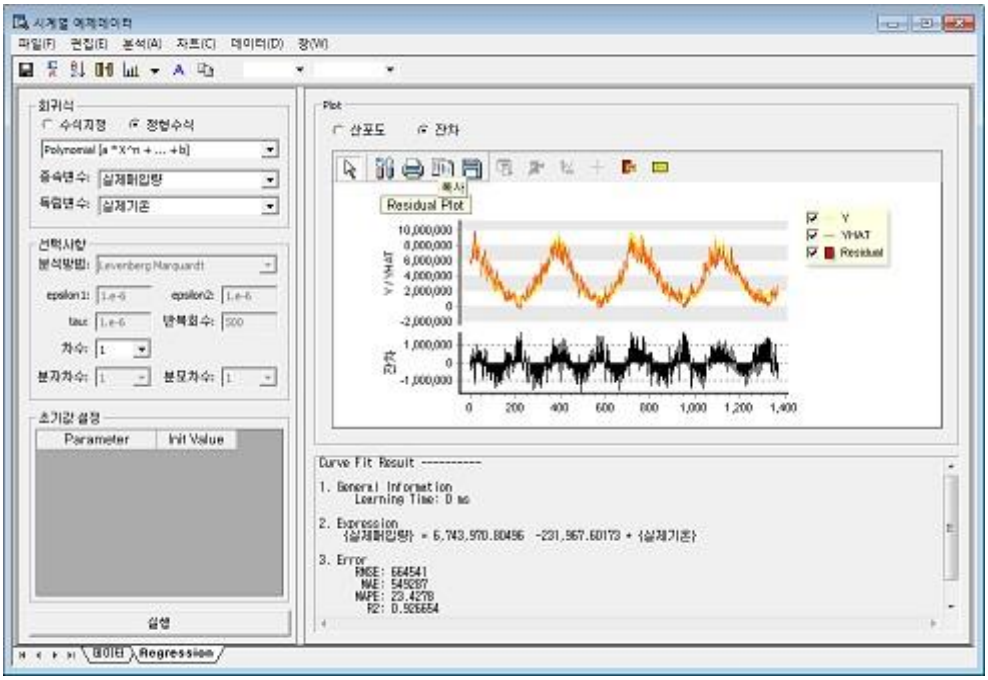
[분석] – [Nonlinear Regression] 를 선택하면 **[Nonlinear Regression]** 윈도우가 나타납니다. 다음과 같은 메인 화면에서 회귀식을 찾기 위해서 사용자 정의 수식 지정을 할 수 있고, 정형 수식을 선택할 수 있습니다. 정형 수식을 선택하면 다음과 같이 독립변수와 종속변수를 지정할 수 있습니다.



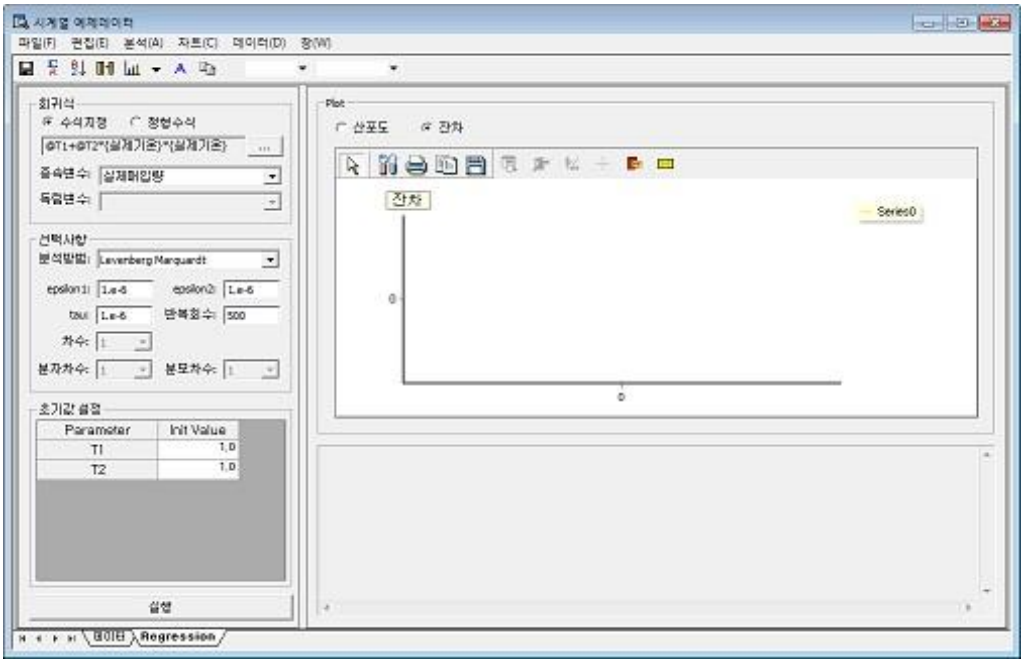
변수를 지정하고 함수의 형태를 선택하여 실행 단추를 누르면 다음과 같이 Estimation 된 결과(플롯 및 회귀 분석 결과)를 볼 수 있습니다.



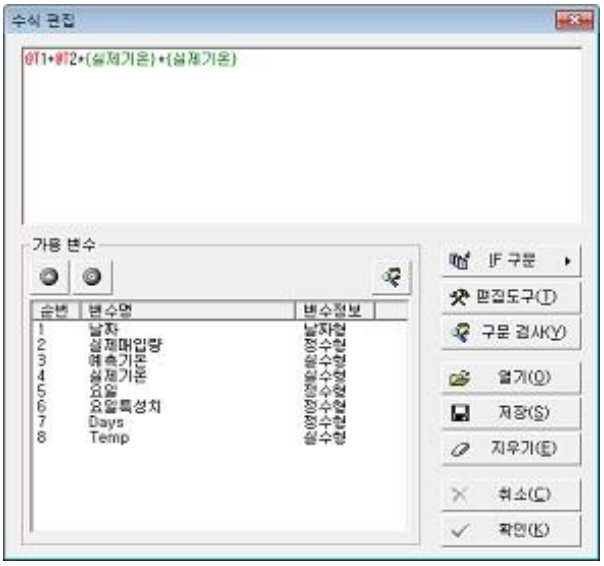
그리고 잔차를 선택하면 실제 시계열 데이터와 적합치를 한번에 볼 수 있고 그 아래의 잔차도를 통해 잔차는 어떠한 형태를 띠는지 알 수 있습니다.



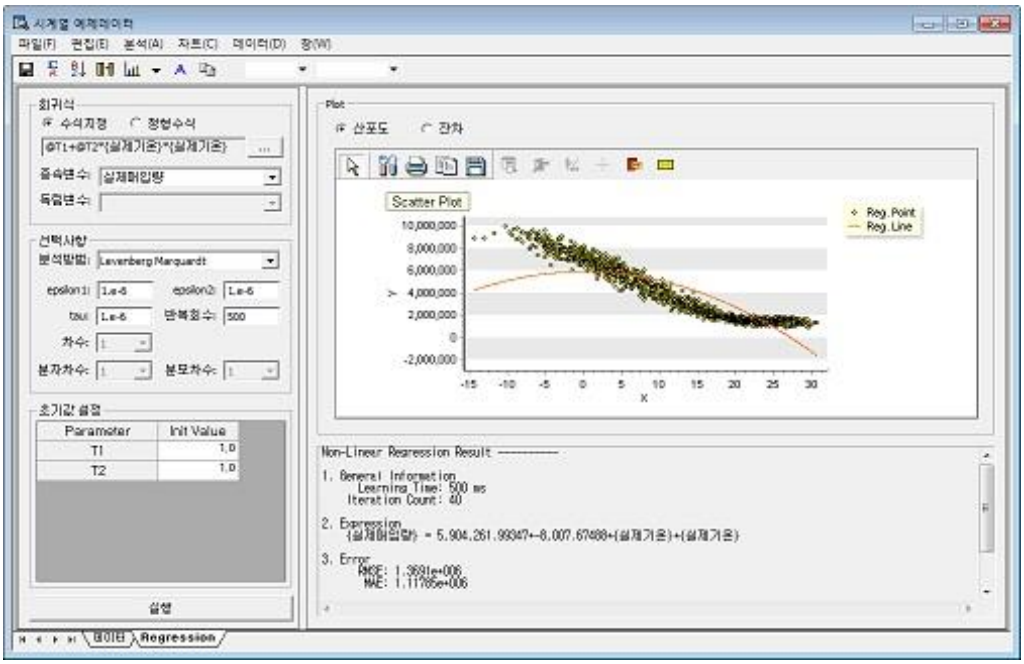
사용자 정의 수식을 위해서 수식 지정을 선택합니다.



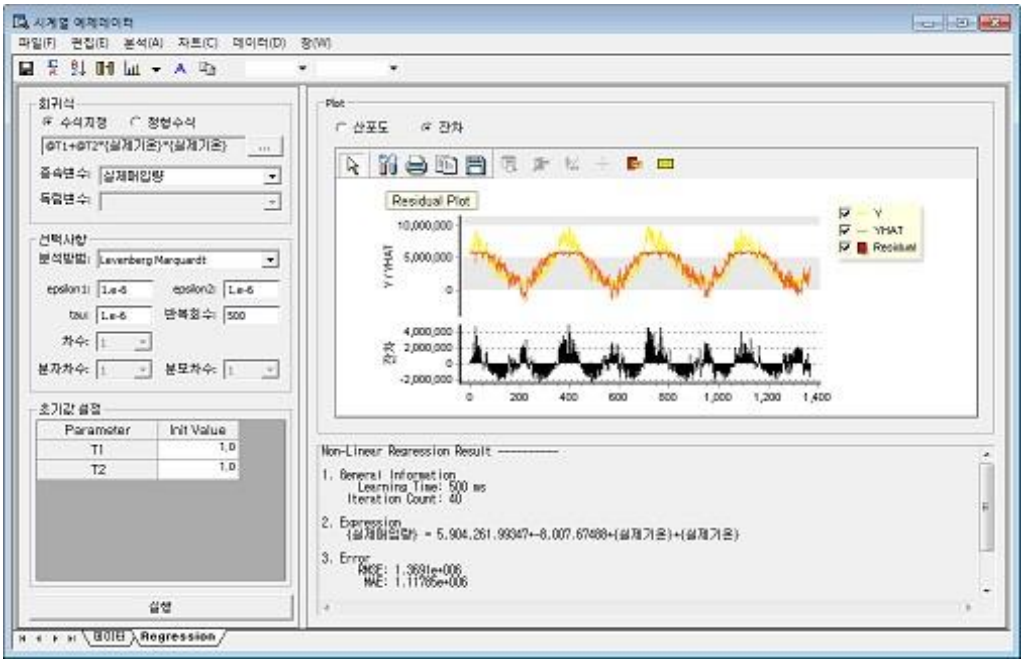
버튼을 클릭하면 다음과 같은 수식 편집기 창이 나타나게 되는데 여기서 사용자가 원하는 형태의 수식을 지정할 수 있습니다. (이 때 추정하고자 하는 Parameter 의 이름은 @T1, @T2,와 같은 형태로 써 주어야 함에 주의합니다.)



Main 화면에서 초기 값을 설정하고 실행 버튼을 클릭합니다. 만약 위에서 독립변수로 사용하는 변수를 두 개로 하면 다음과 같은 3 차원 플롯을 결과로 얻을 수 있습니다.



잔차 라디오 버튼을 선택하면 다음과 같이 실제 시계열 데이터와 적합치, 그리고 잔차의 추이를 볼 수 있습니다.



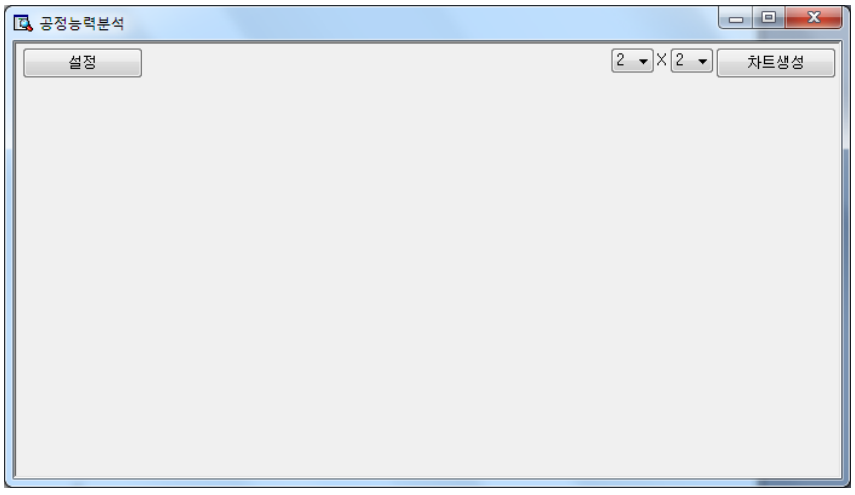
5.3.5 SPC(Statistical Process Control)

- **5.3.5.1 공정능력분석**

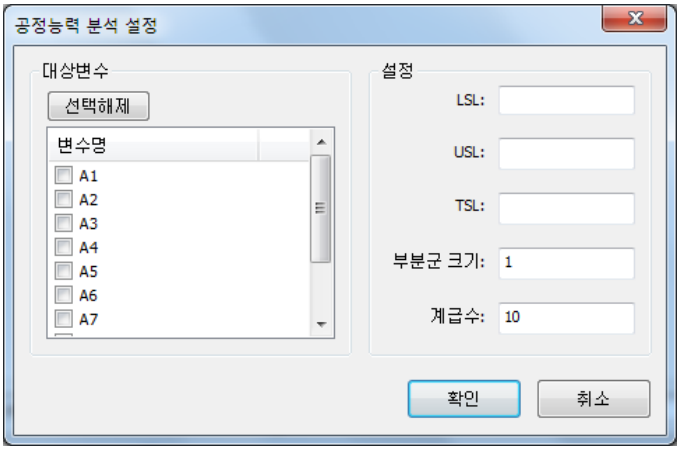
  데이터 탐색기에서는 각각의 필드에 대해 **공정 능력 관리** 기능을 사용하여, 데이터가 원하는 공정 영역 내에서 분포하는지를 알 수 있습니다.

실행 방법

1. **[분석] - [SPC] – [공정능력분석]**을 선택하면 다음과 같은 공정능력분석 다이얼로그가 나타납니다.

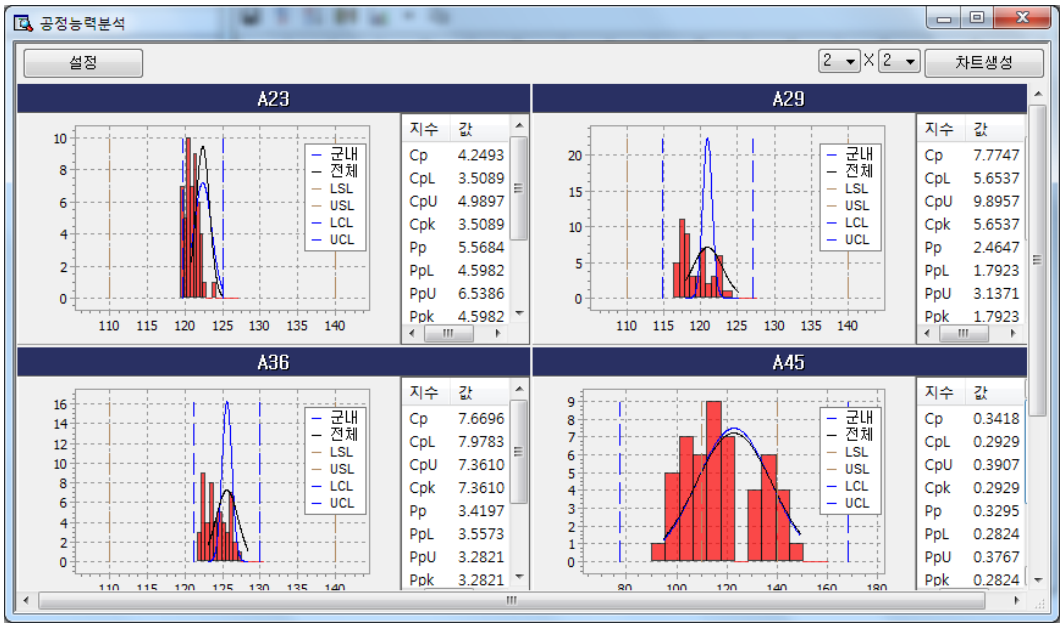


2. 대상 변수를 선택하기 위하여 [설정]버튼을 클릭합니다.



3. 분석 대상 변수를 선택하고(다중 선택 가능) 이에 대한 LSL, USL, TSL, 부분군 크기, 히스토그램 작성을 위한 계급수를 설정합니다.

실행 결과 및 용어



명칭	설명	기타

히스토그램	선택된 필드의 히스토그램 차트입니다.	속성창의 계급수에 따라 차트가 달라집니다.
군내	선택된 필드의 평균과 부분군 내의 표준편차를 이용한 정규 분포 곡선입니다.	부분군 크기에 따라 달라집니다.
전체	선택된 필드의 평균과 표준편차를 이용한 정규 분포 곡선입니다.	
LSL	Lower specification limit. 공정 명세서 상의 하한값으로 사용자의 입력이 필요합니다.	속성창의 LSL 값에 따라 선이 이동합니다.
USL	Upper specification limit. 공정 명세서 상의 상한값으로 사용자의 입력이 필요합니다.	속성창의 USL 값에 따라 선이 이동합니다.
LCL	Lower control limit. 결과 데이터 상의 하한값. LCL = 평균 - 3 * 표준편차	
UCL	Upper control limit. 결과 데이터 상의 상한값. UCL = 평균 + 3 * 표준편차	
Cp	PCI(Process capability index : 공정능력지수)를 가리키는 값. Cp = (USL - LSL)/(6 * 군내 표준편차)	
CpL	부분군내 평균이 규격 하한에 얼마나 근접해 있는지를 나타내는 값. CpL = (평균 – LSL)/(3 * 군내 표준편차)	
CpU	부분군내 평균이 규격 상한에 얼마나 근접해 있는지를 나타내는 값. CpU = (USL – 평균)/(3 * 군내 표준편차)	
Cpk	부분군내 평균과 규격 한계와의 차이이며 CpL 과 CpU 중에 작은 값을 나타냄	
Pp	부분군을 고려하지 않은 전체 공정에 대한 공정능력지수. Pp = (USL – LSL)/(6 * 표준편차)	
PpL	공정 전체 평균이 규격 하한에 얼마나 근접해 있는지를 나타내는 값. PpL = (평균 – LSL)/(3 * 표준편차)	
PpU	공정 전체 평균이 규격 상한에 얼마나 근접해 있는지를 나타내는 값. PpU = (USL –	

	평균)/(3 * 표준편차)	
Ppk	공정 전체 평균과 공정 한계와의 차이이며 PpL 과 PpU 중에 작은 값을 나타냅니다.	
Cpm	공정이 목표값에 얼마나 근접하였는지를 나타내는 공정능력지수. 공정 평균과 목표값과의 편차로 계산됨	

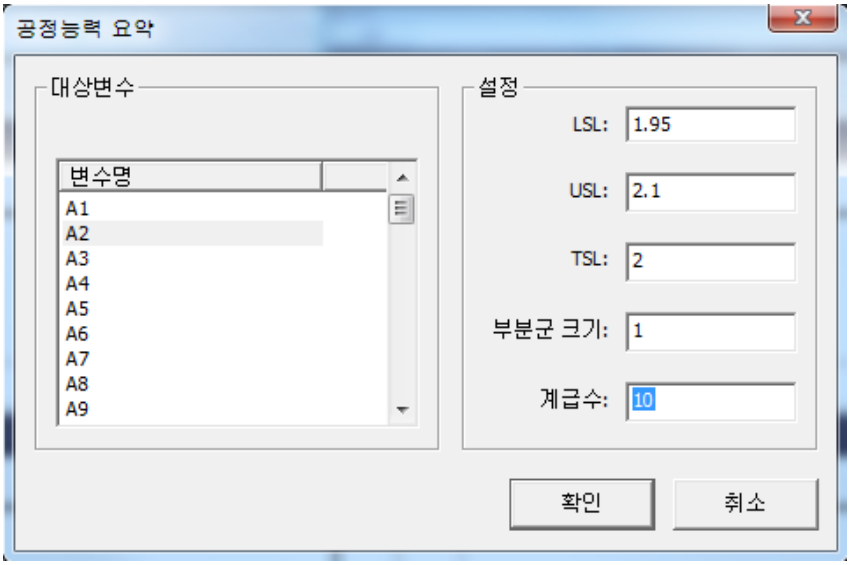
- 5.3.5.2 공정능력요약

  데이터 탐색기에서는 각각의 필드에 대해 공정능력요약 기능을 사용하여, 데이터가 원하는 공정 영역 내에서 이루어지는지 여부를 요약 리포트 형태로 제공됩니다.

실행 방법

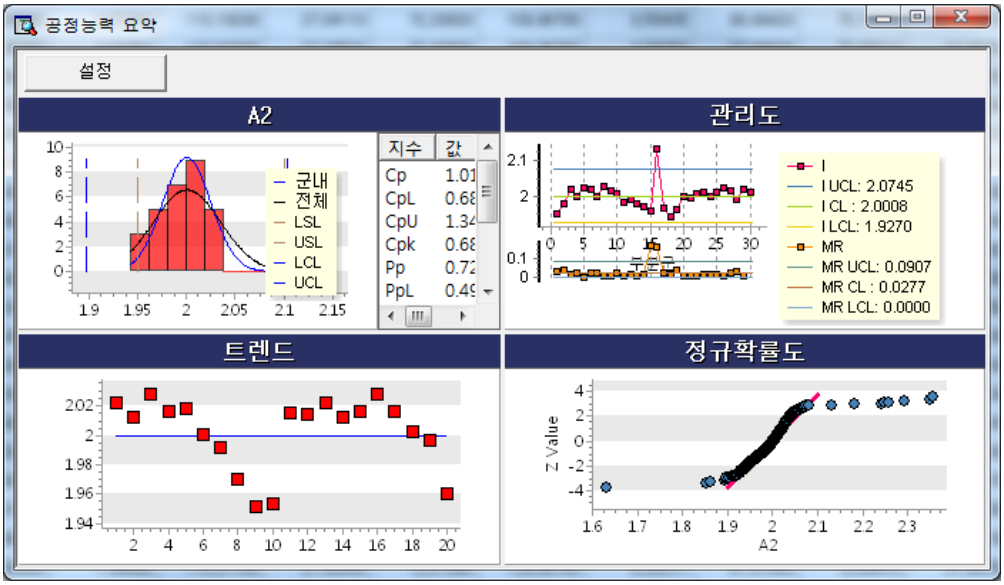
1. **[분석] - [SPC] – [공정능력요약]**을 선택하면 다음과 같은 공정능력요약 다이얼로그가 나타납니다.

2. 대상 변수를 선택하기 위하여 [설정]버튼을 클릭합니다.



3. 분석 대상 변수를 선택하고 이에 대한 LSL, USL, TSL, 부분군 크기, 히스토그램 작성을 위한 계급수를 설정합니다.

실행 결과



- **5.3.5.3 합격표본추출**

합격 표본추출은 전체 로트(lot) 또는 배치(batch)를 대상으로 품질 기준에 의해 로트의 합격, 불합격을 판정하는 기법입니다.

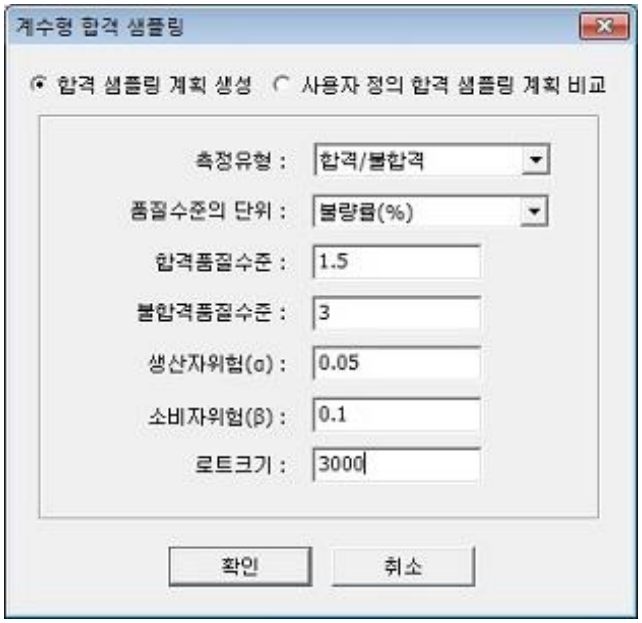
ECMiner 에서는 주어진 품질 기준을 기반으로 최적의 표본추출 검사 계획을 생성하고, 서로 다른 검사 계획을 비교할 수 있습니다.

실행 방법 - 계수형

1. **[분석] – [SPC] – [합격표본추출] – [계수형 합격 샘플링]**을 선택하면, 계수형 합격 샘플링 윈도우가 나타납니다.

2. 계수형 합격 샘플링을 이용하여 불합격 판정 기준이 불량개수인 경우 표본추출 검사 계획을 생성할 수 있습니다.

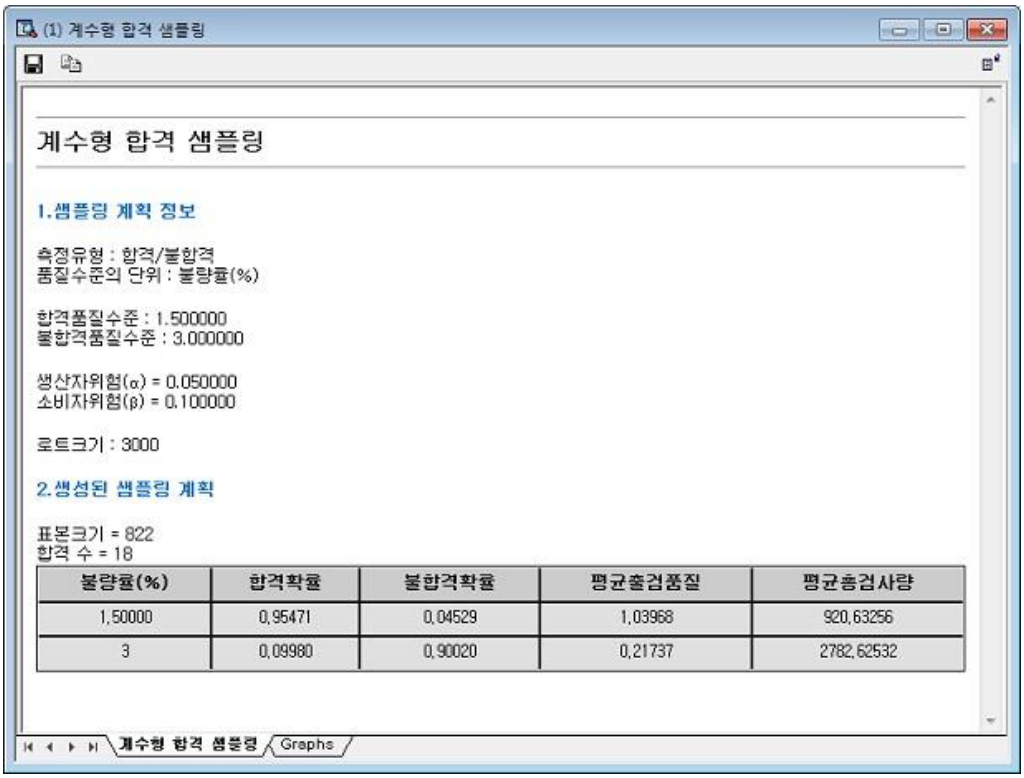
3. 측정유형 및 품질수준의 단위를 선택하고, 각각의 품질 기준을 입력합니다.



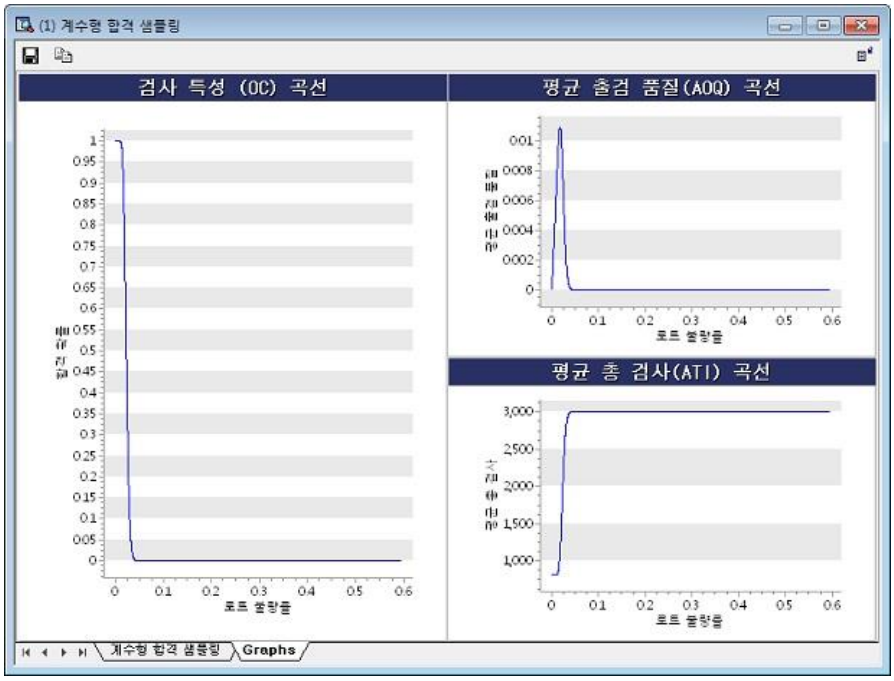
결과 – 계수형

계수형 합격 샘플링 결과가 다음과 같이 나타납니다.

주어진 품질 기준에 따라 최적의 표본추출 검사계획이 생성됩니다.



결과창의 Graphs 탭을 선택하면 다음과 같이 OC 곡선, AOQ 및 ATI 곡선 차트를 확인할 수 있습니다..

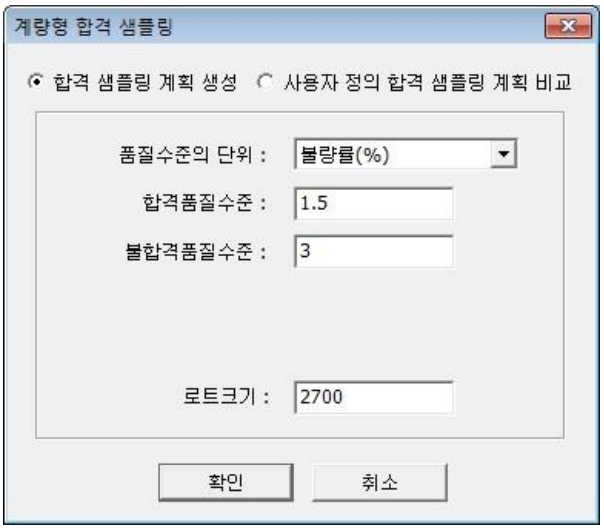


실행 방법 – 계량형

1. **[분석] – [SPC] – [합격표본추출] – [계량형 합격 샘플링]**을 선택하면, 계량형 합격 샘플링 윈도우가 나타납니다.

2. 계량형 합격 샘플링을 이용하여 불합격 판정 기준이 계량치(특성치)인 경우 표본추출 검사 계획을 생성할 수 있습니다.

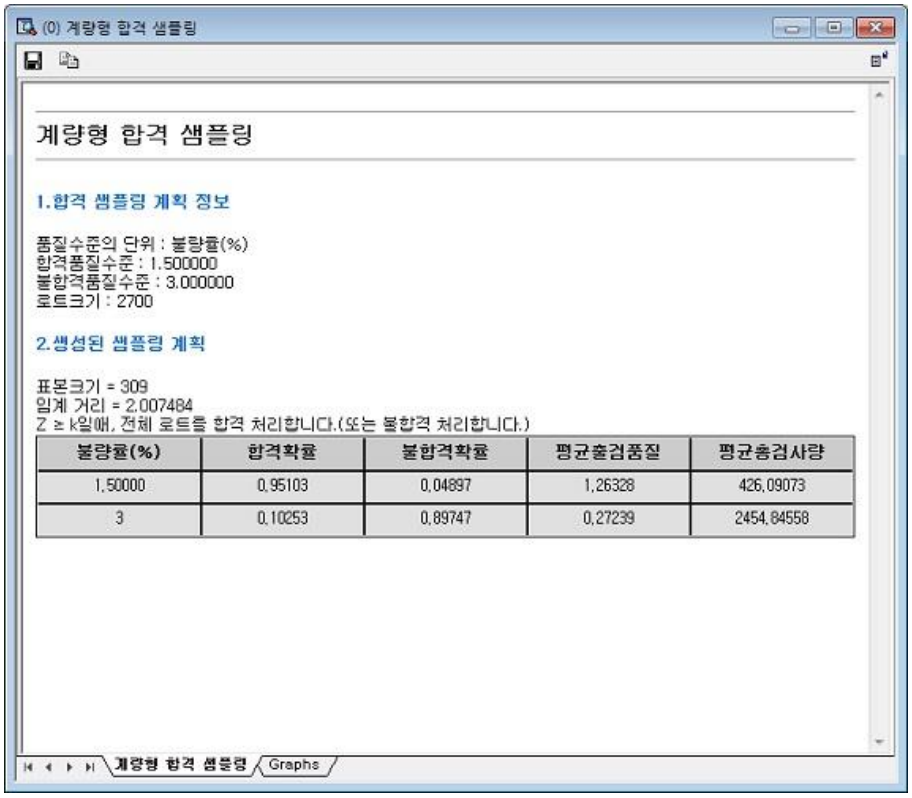
3. 품질수준의 단위를 선택하고, 각각의 품질 기준을 입력합니다.



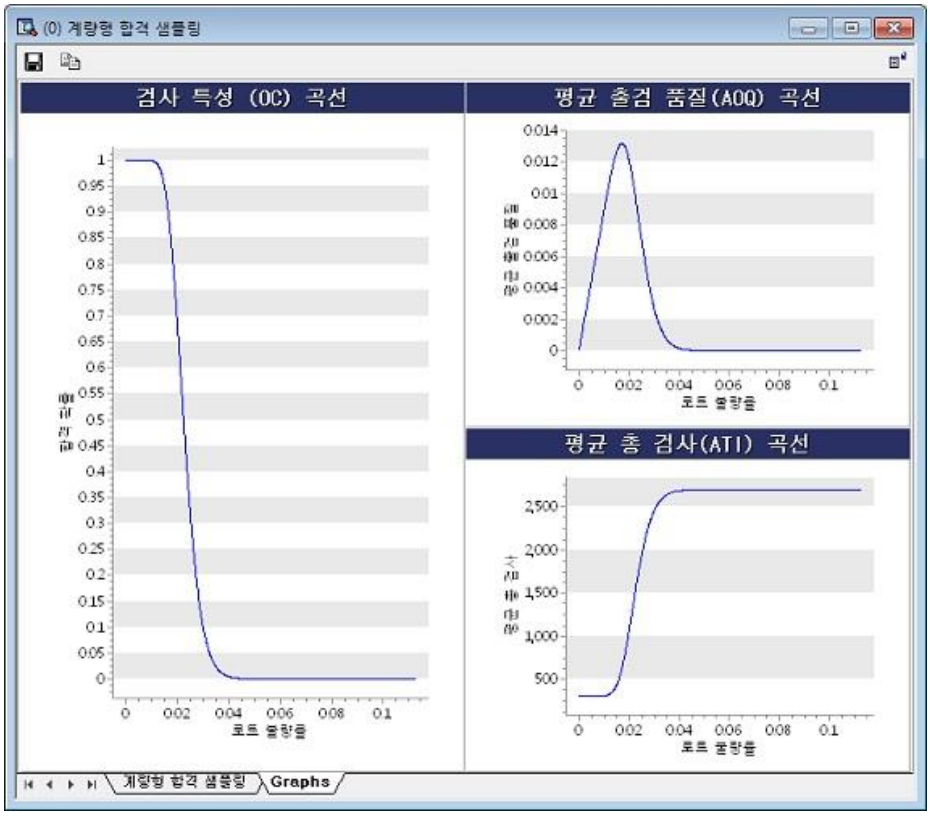
결과 – 계량형

계량형 합격 샘플링 결과가 다음과 같이 나타납니다.

주어진 품질 기준에 따라 최적의 표본추출 검사계획이 생성됩니다.



결과창의 Graphs 탭을 선택하면 다음과 같이 OC 곡선, AOQ 및 ATI 곡선 차트를 확인할 수 있습니다..

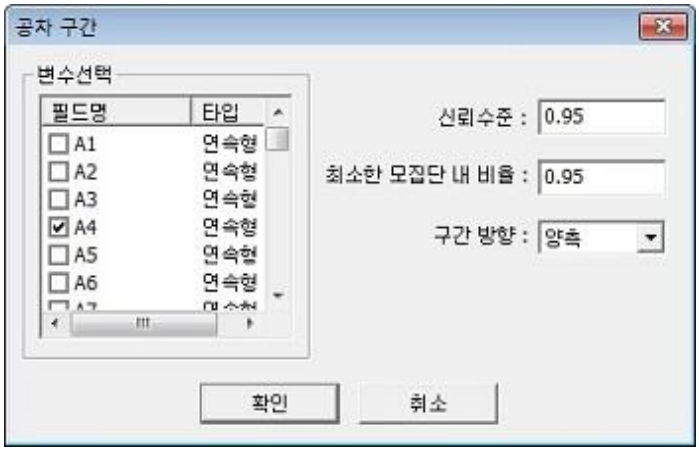


- 5.3.5.4 공차구간

  공차구간은 최소 모집단의 비율과 신뢰 수준을 설정하며, 샘플로부터 얻어진 통계량을 기초로 지정된 모집단에서의 최소비율과 신뢰 수준을 만족하는 구간을 제시합니다. 얻어진 공차구간과 고객의 요구사항을 비교하여 현재 공정상태를 파악할 수 있습니다.

실행 방법

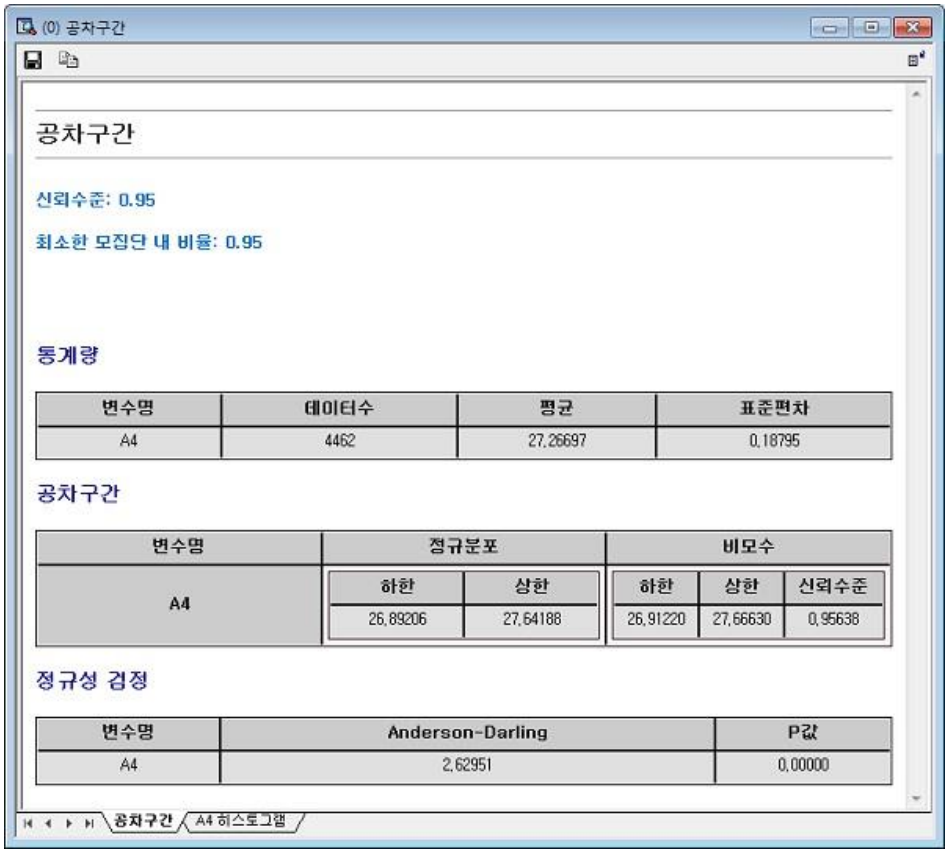
[분석] – [SPC] – [공차구간]을 선택하면 다음과 같은 공차구간 다이얼로그가 나타납니다.



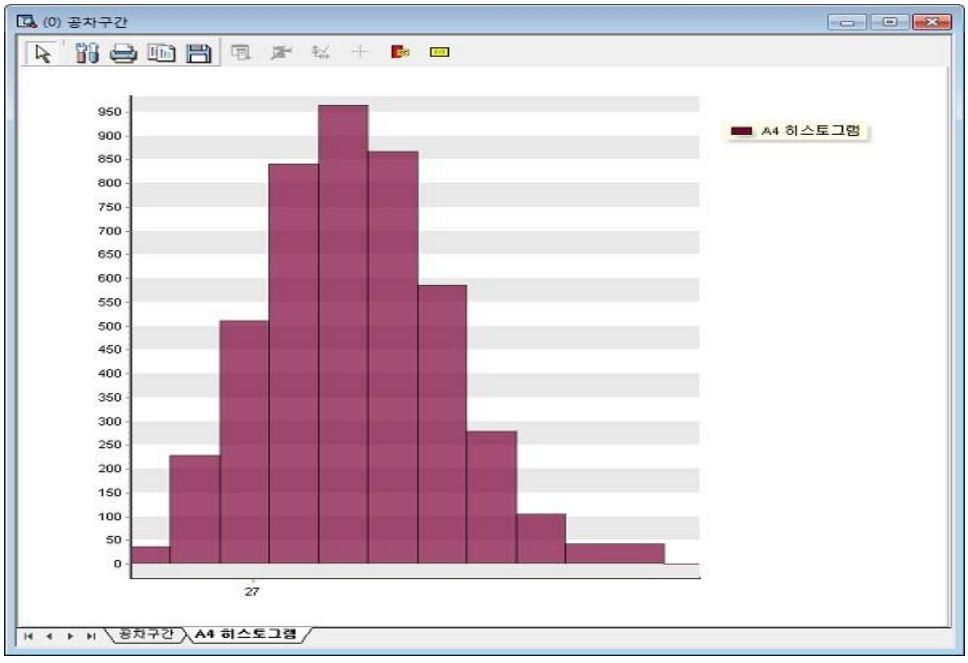
변수 목록에서 공차구간을 구하고자 하는 변수를 선택하고(복수 선택 가능), 신뢰수준과 최소한 모집단 내 비율을 설정합니다(0~1 값). 마지막으로 구간 방향을 선택합니다(양측/상한/하한).

결과

- 통계량: 선택한 변수의 기초 통계량들을 제공합니다.
- 공차구간: 선택한 변수로부터 지정된 신뢰수준과 최소합 모집단 비율을 만족하는 공차구간을 정규분포 또는 비모수 방법으로 계산하여 보여줍니다.
- 정규성 검정: 선택한 변수가 정규분포에 얼마나 따르는지 보여주는 Anderson-Darling 통계량과 P-value 를 제공합니다. P-value 가 0.05 보다 크면, 변수는 정규분포를 따른다고 할 수 있습니다.



▪ 히스토그램: 선택한 변수의 분포를 알 수 있는 히스토그램을 그려줍니다.



5.3.6 시계열분석

- 5.3.6.1 시계열 모델

(1) 시계열 분해

개요

시계열 데이터를 분해(Decomposition)한다는 것은 데이터를 성분별로 나눈다는 것입니다. 구체적으로 말해서 Time Series data 는 추세, 계절성 등의 성분을 가질 수 있는데 이러한 성분을 뽑아내어 분리하는 것 이라고 할 수 있습니다. Classical Decomposition 에는 Multiplicative Decomposition(승법분해)와 Additive Decomposition(가법분해)가 있습니다.

Classical Decomposition 의 구성은 다음과 같습니다.

Multiplicative Decomposition(추세 포함)
Multiplicative Decomposition(추세 제거)
Additive Decomposition(추세 포함)
Additive Decomposition(추세 제거)

그리고 Decomposition 을 통해서 얻게 되는 통계량을 정리하면 다음과 같습니다.

Trend Data($\mathrm{tr_t}$)
Detrended Data
Seasonal Index Data($\mathrm{S_t}$)
Deseasonal Index Data($\mathrm{y_t - S_t}$)
Fitted Value($\hat{y}_t$)
Residual

승법 분해(추세 포함)

승법 분해 모델(Multiplicative Decomposition Model)은 다음과 같은 형태입니다.

$$y_t = TR_t \times SN_t \times CL_t \times IR_t$$

yt: t 시점에서의 관측 값.

TRt: t 시점에서의 trend component.

SNt: t 시점에서의 seasonal component.

CLt: t 시점에서의 cyclical component.

IRt: t 시점에서의 irregular component.

가법 분해(추세 포함)

가법 분해 모델(Additive Decomposition Model)은 다음과 같은 형태입니다.

$$y_t = TR_t + SN_t + CL_t + IR_t$$

yt: t 시점에서의 관측 값.

TRt: t 시점에서의 trend component.

SNt: t 시점에서의 seasonal component.

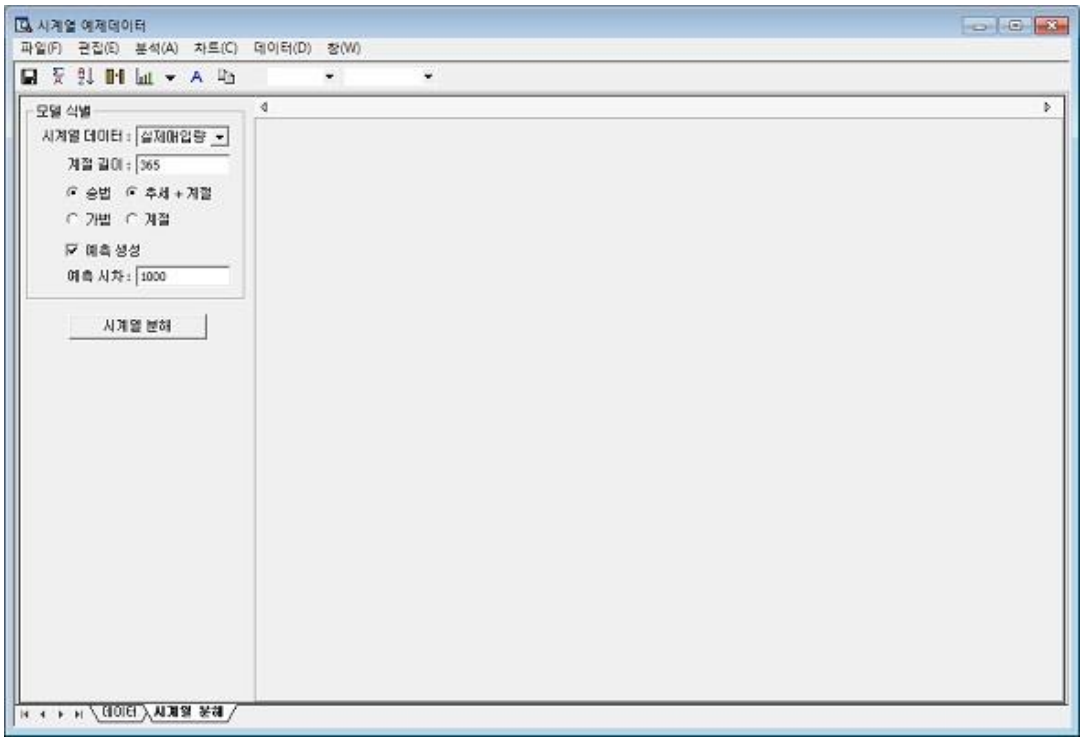
CLt: t 시점에서의 cyclical component.

IRt: t 시점에서의 irregular component

이 때 추세는 분석에서 제외할 수 있습니다.

실행방법

[분석] – [시계열 분석] – [시계열 모델] – [시계열 분해]를 선택하면 **[시계열 분해]** 윈도우가 나타납니다.



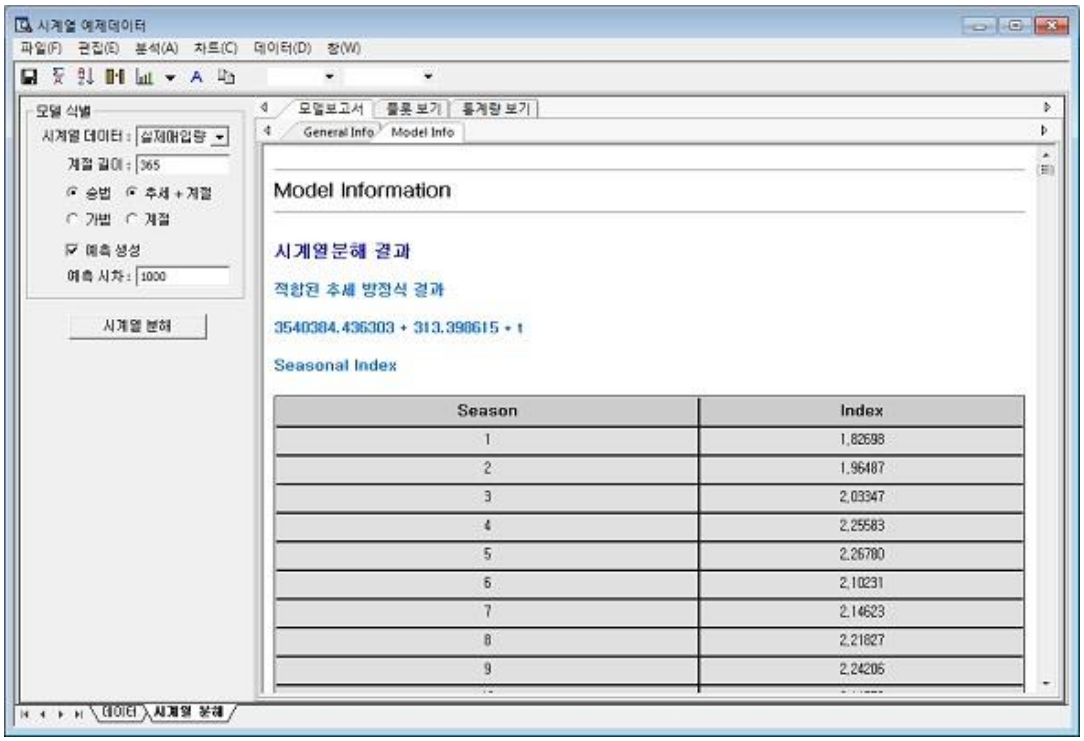
Main 화면에서 시계열 데이터를 선택하고 계절의 길이, 그리고 승법 분해를 사용할 것인지 가법분해를 사용할 것인지 선택한 후 추세+계절, 추세 중 하나를 선택합니다. 예측을 원할 경우 예측 생성을 누르고 예측 시차를 입력합니다.

결과

- **모델 보고서**

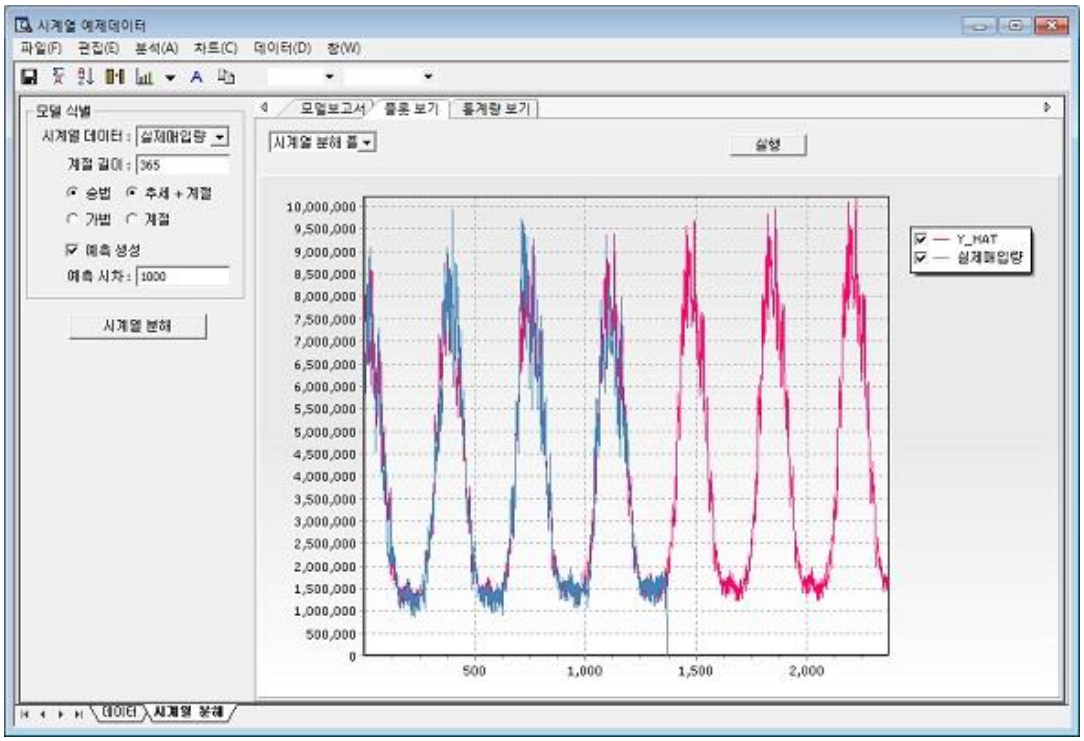
  General Info: 시계열 데이터의 기본적인 정보를 보여줍니다.

  Model Info: 적합 된 추세 방정식, Seasonal Index, 그리고 예측 생성을 선택하였을 시 예측 결과를 보여줍니다.

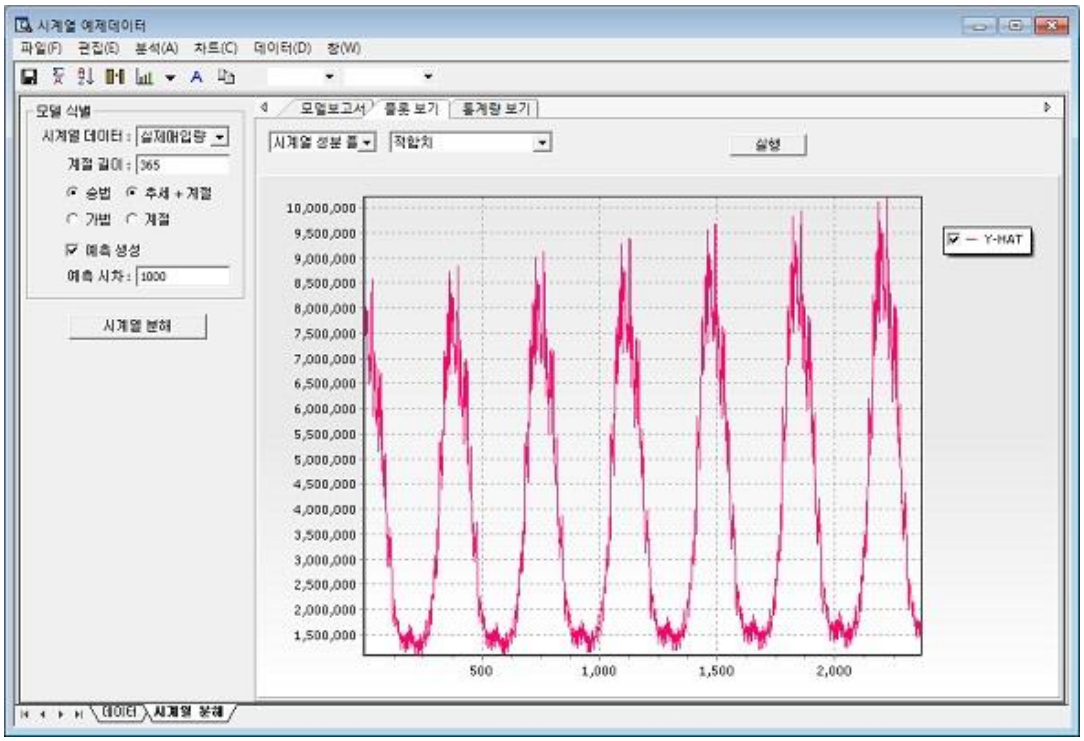


- **플롯 보기**

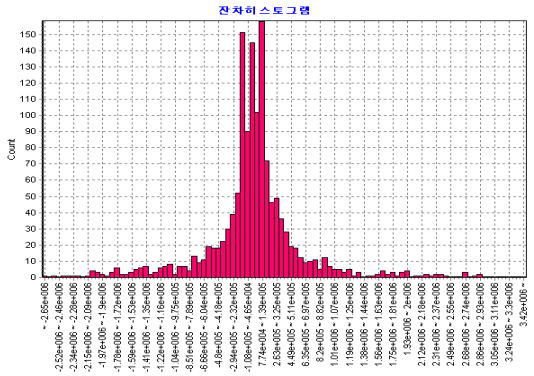
  시계열 분해를 통해서 얻은 데이터의 시각적인 해석을 위해 제공되는 기능입니다. 시계열 분해에서는 시계열 분해 플롯, 시계열 성분 플롯, 잔차 관련 플롯을 제공합니다.

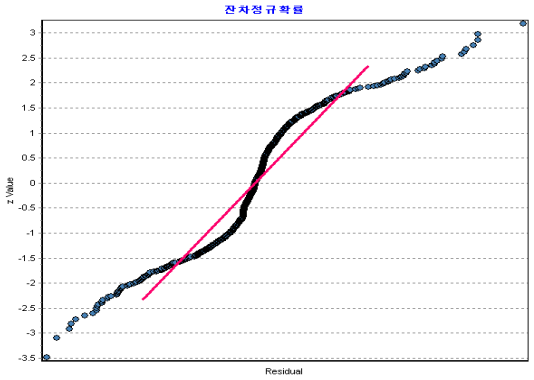


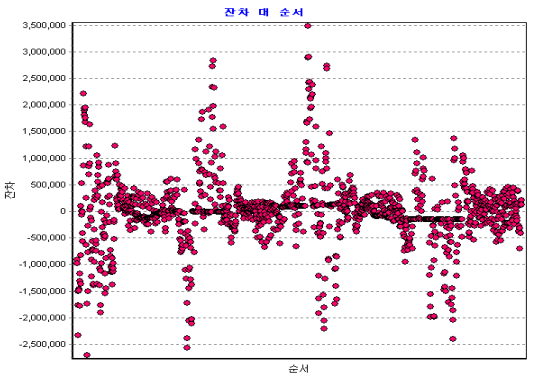
시계열 분해 플롯을 통해서는 실제 시계열 데이터와 적합치, 그리고 예측값을 보여줍니다.

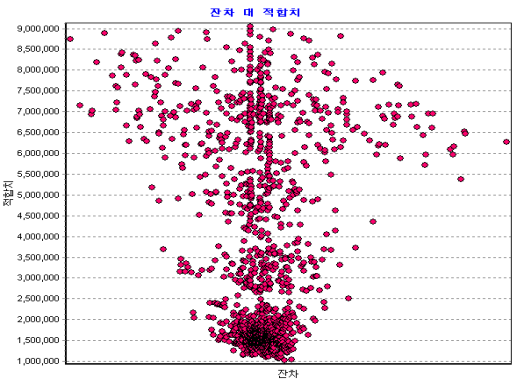


시계열 성분 플롯을 통해서는 분해된 성분 각각을 플롯으로 보여줍니다. 모든 데이터를 한번에 볼 수도 있고 각 데이터 하나 하나에 따라 따로 그래프를 볼 수도 있습니다.









잔차 관련 플롯(잔차 히스토그램, 잔차 정규 확률 플롯, 잔차 대 순서, 잔차 대 적합치)을 통해서 분해 결과 얻어진 잔차를 분석 할 수 있습니다.

- **통계량 보기**

  시계열 분해 분석을 통해서 얻어진 데이터를 Table 형태로 보여줍니다. 이를 저장할 수 있는 기능도 제공합니다.

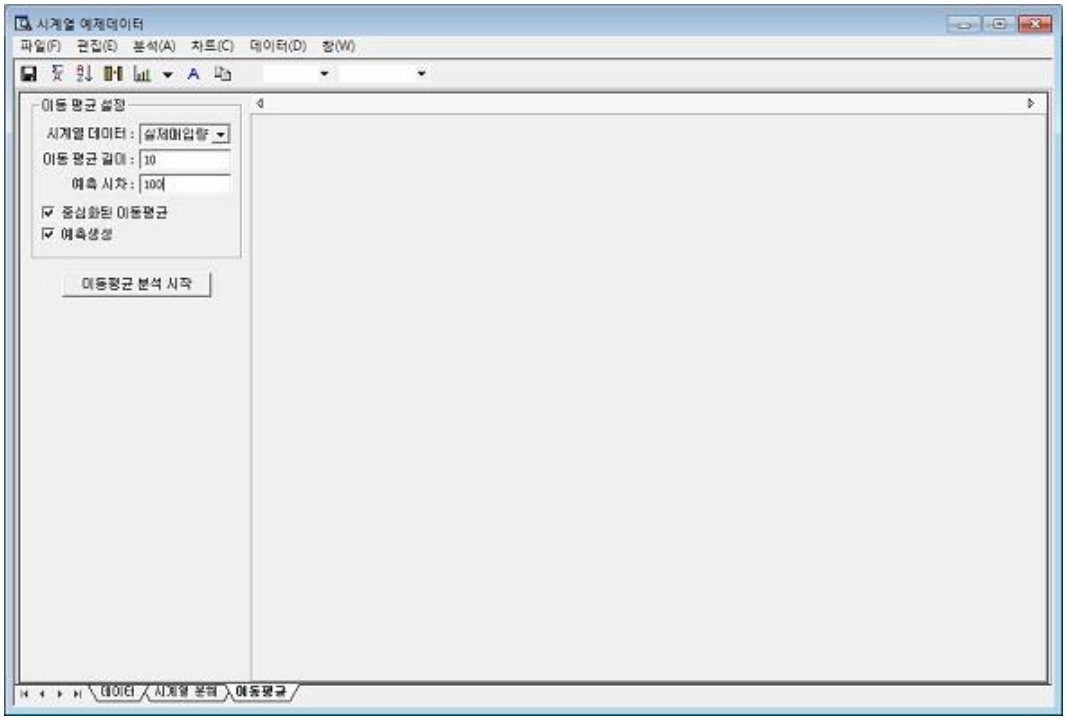
(2) 이동 평균

개요

이동 평균은 현재 시점과 과거, 미래의 몇 데이터를 평균한 값을 말합니다. 이동 평균은 이를 확장한 중심화된 이동 평균과 함께 가장 단순한 Form 으로 많이 사용되는 방법입니다.

실행방법

[분석] – [시계열 분석] – [시계열 모델] – [이동 평균] 를 선택하면 **[이동 평균]** 윈도우가 나타납니다.



Main 화면에서 시계열 데이터를 선택하고 이동평균의 길이, 중심화된 이동평균을 사용할지의 여부, 예측을 원할 경우 예측시차를 입력합니다. 입력을 마치고 [이동평균 분석 시작] 버튼을 클릭합니다.

결과

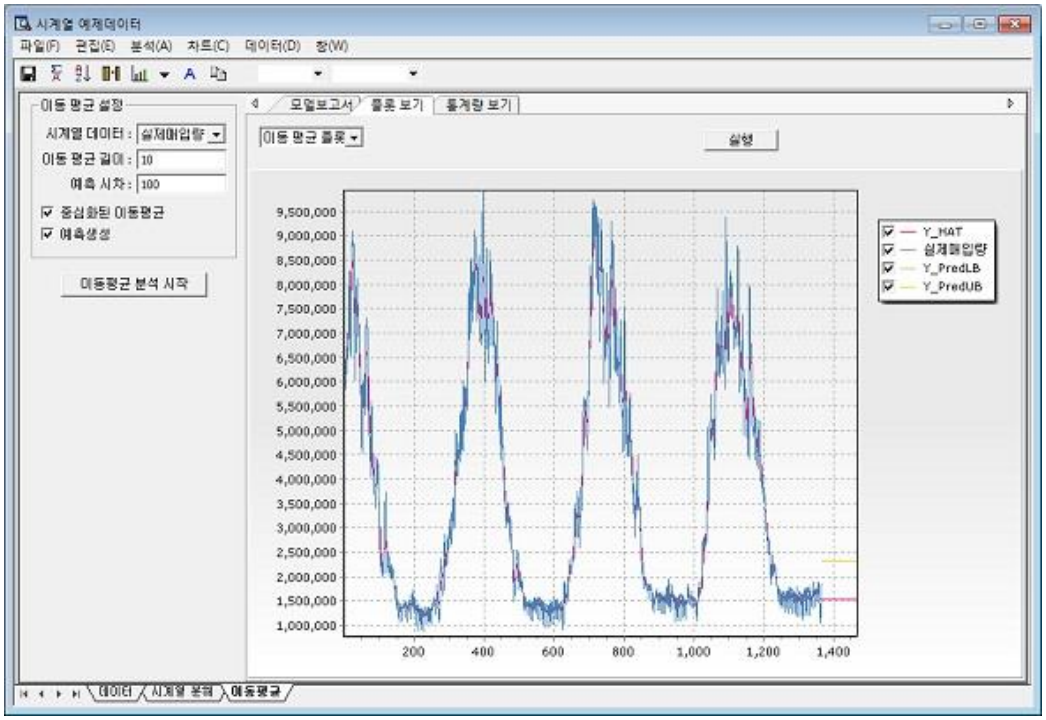
- **모델 보고서**

General Info: 시계열 데이터에 대한 기본적인 정보를 제공합니다.

Model Info: 이동 평균 분석 결과에 대한 정보를 제공합니다. 정확도 측도, 예측 결과를 볼 수 있습니다.

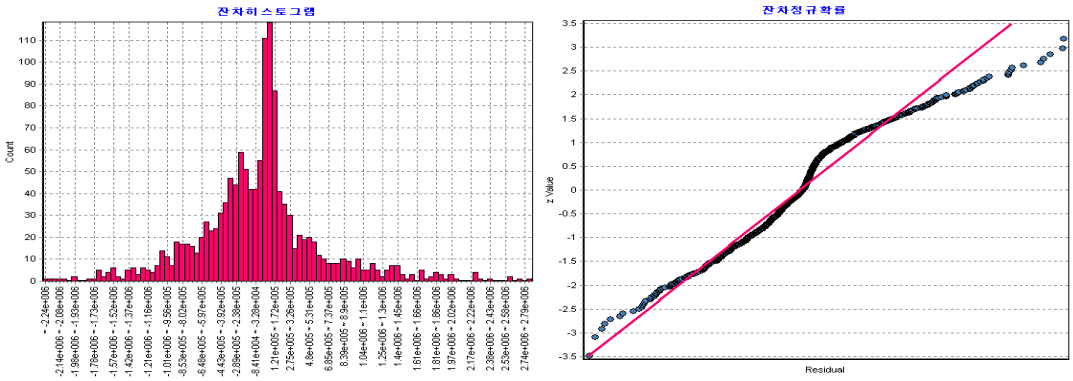
- **플롯 보기**

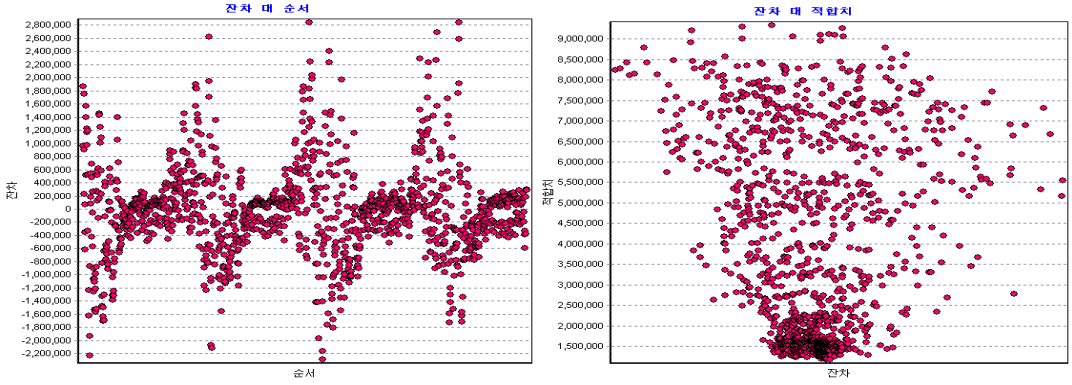
  이동평균 분석 결과 얻어진 데이터를 시각적으로 볼 수 있습니다.



이동 평균 플롯을 통해서는 시계열 데이터와 적합치 그리고 예측 관련 통계량을 보여줍니다.

잔차 관련 플롯(잔차 히스토그램, 잔차 정규확률 플롯, 잔차 대 순서, 잔차 대 적합치)을 통해서는 얻어진 잔차를 분석할 수 있습니다.





- **통계량 보기**

  이동 평균 분석을 통해서 얻어진 데이터를 Table 형태로 보여줍니다. 이를 저장할 수 있는 기능도 제공합니다.

(3) 지수평활

개요

- **단순 지수 평활**

  다음과 같은 trend 가 없는 model 이 있을 때

$$y = \beta_0 + \epsilon_t$$

  과 같은 model 을 상정할 수 있습니다. Least squares 방법을 이용하면

$$\hat{\beta}_0 = \bar{y} = \sum_{t=1}^{n} \frac{y_t}{n}$$

  과 같은 방법으로 estimation 이 이루어집니다. 위의 식을 보면 β_0의 estimate 을 구하는 과정에서 모든 관측치에 같은 크기의 weight(1/n)이 부여된 것을 볼 수 있는데 이는 합리적이지 않은 것 같습니다. 그래서 최근 데이터에 더 많은 weight 을 주고, 오래된 데이터일수록 작은 weight 를 주려고 하는 것이 지수 평활의 idea 입니다. 이러한 지수

평활 방법론 중 추세도 없고, 계절성도 없는 데이터를 처리하는 방법론이 바로 단순 지수 평활입니다.

▪ **이중 지수 평활**

다음과 같은 time series model 을 생각해보도록 합시다.

$$y_t = \beta_0 + \beta_1 t + \epsilon_t$$

이 때 β_0, β_1이 t 에 상관없이 상수라면 위의 model 은 linear regression 을 통해서 계수 추정을 할 수 있습니다. 반면에 β_0, β_1이 시간에 따라 변할 수 있다면 이러한 model 을 설명하기 위해서 제시된 것이 이중 지수 평활입니다.

▪ **Winters' Method** : Multiplicative Winters' Method

Winters' method 는 수준, 추세뿐 아니라 계절성을 고려한 time series model 입니다. Multiplicative Winters' Method 는 다음과 같은 equation 으로 표현되는 time series 의 forecasting 에 적당하다고 알려져 있습니다.

$$y_t = (\beta_0 + \beta_1 t) * SN_t + \epsilon_t$$

▪ **Winters' Method:** Additive Winters' Method

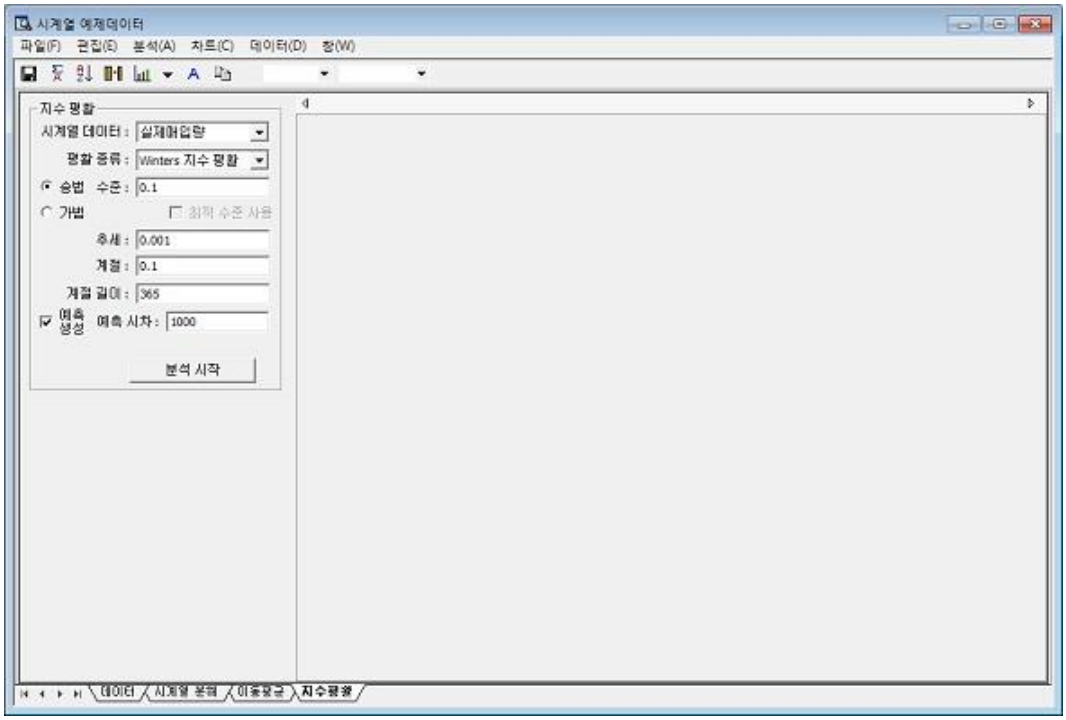
Additive Winters' method 의 경우 다음과 같은 equation 을 만족하는 time series data 의 forecasting 에 가장 적합하다고 알려져 있습니다.

$$y_t = (\beta_0 + \beta_1 t) + SN_t + \epsilon_t$$

Additive Winters's Method 에 대한 공식은 Multiplicative 공식을 약간 변형함으로써 얻을 수 있습니다.

실행방법

[분석] – [시계열 분석] – [시계열 모델] – [지수 평활]를 선택하면 **[지수 평활]** 윈도우가 나타납니다.



Main 화면에서는 시계열 데이터를 선택하고 어떠한 평활 방법을 사용할 것인지를 선택합니다. 이 때 단순 지수 평활을 사용할 경우, 수준 값을 사용자가 직접 입력하거나, '최적 수준 사용'을 체크하여 자동으로 0~1 사이의 최적 수준 값을 입력할 수 있습니다. 이중 지수 평활을 선택할 경우에는 수준과 추세 평활화 상수를 입력하고, winters 의 방법을 사용할 경우에는 승법과 가법을 선택하고 수준, 추세, 계절 평활화 상수 및 계절 길이를 입력해 줍니다. 예측을 하고자 할 경우 예측 시차를 입력해 주도록 합니다.

결과

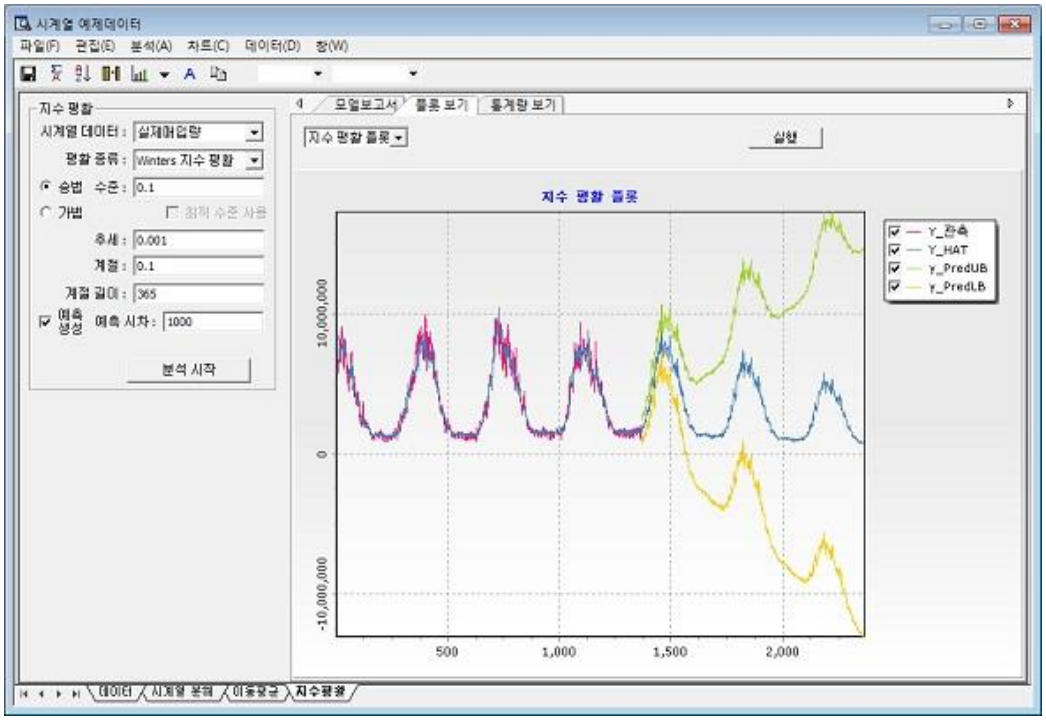
- **모델 보고서**

  General Info: 시계열 데이터에 대한 기본적인 정보를 보여줍니다.

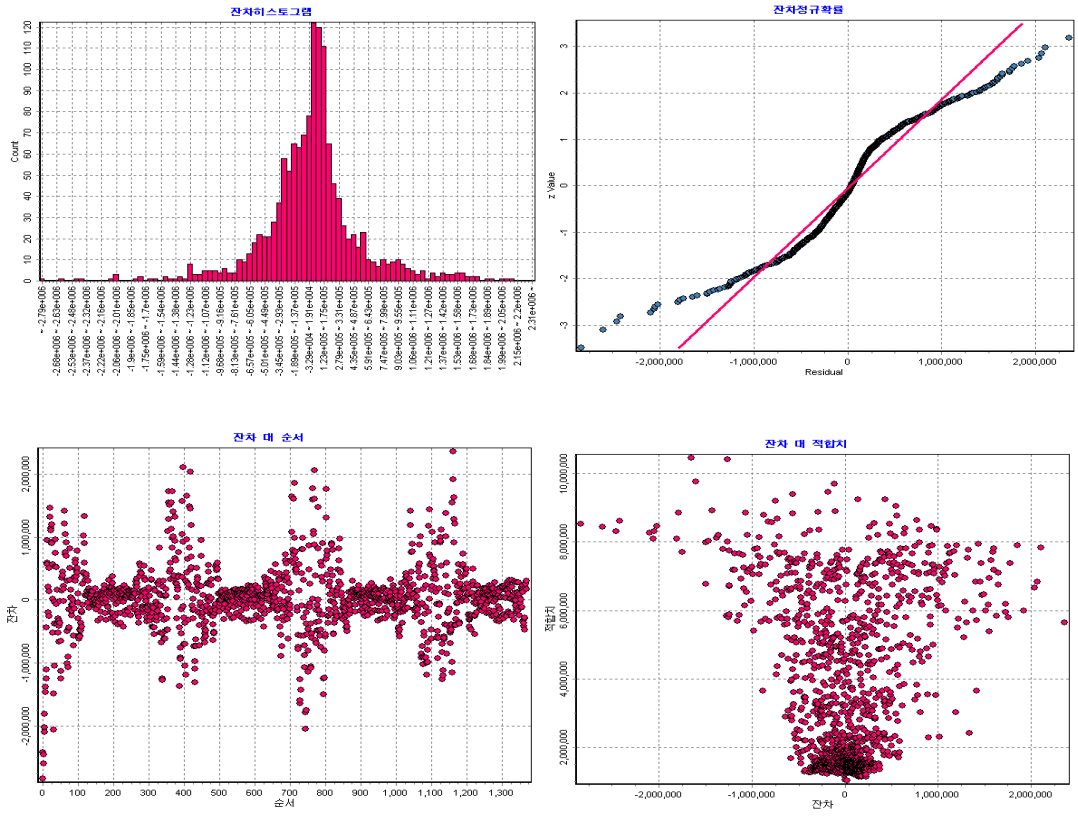
  Model Info: 지수 평활 분석을 통해서 얻어진 정보를 보여줍니다. 정확도 측도, 예측 결과를 볼 수 있습니다.

- **플롯 보기**

  지수 평활 분석 결과 얻어진 데이터를 시각적으로 볼 수 있습니다.



지수 평활 플롯을 통해서는 시계열 데이터와 적합치, 예측 관련 값을 시각적으로 볼 수 있습니다.



잔차 관련 플롯(잔차 히스토 그램, 잔차 정규 확률 플롯, 잔차 대 순서, 잔차 대 적합치)를 통해서는 지수 평활 분석 후에 얻어진 잔차를 분석할 수 있습니다.

통계량 보기

지수 평활 분석 결과 얻어진 통계량을 Table 에서 볼 수 있습니다. 이를 저장하는 기능 또한 제공합니다.

(4) ARIMA

개요

- **ARIMA 모델**

ARIMA 모델은 Univariate Time Series Analysis 에서 가장 많이 쓰이는 모델로 다음과 같은 equation 을 만족시킵니다. ARIMA(p, d, q)의 경우

$$\phi(B)(1-B)^d z_t = \theta(B)a_t \quad \text{where } a_t \text{ is i.i.d normal white noise}$$

$$\left(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p\right)(1-B)^d z_t = \left(1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q\right)a_t$$

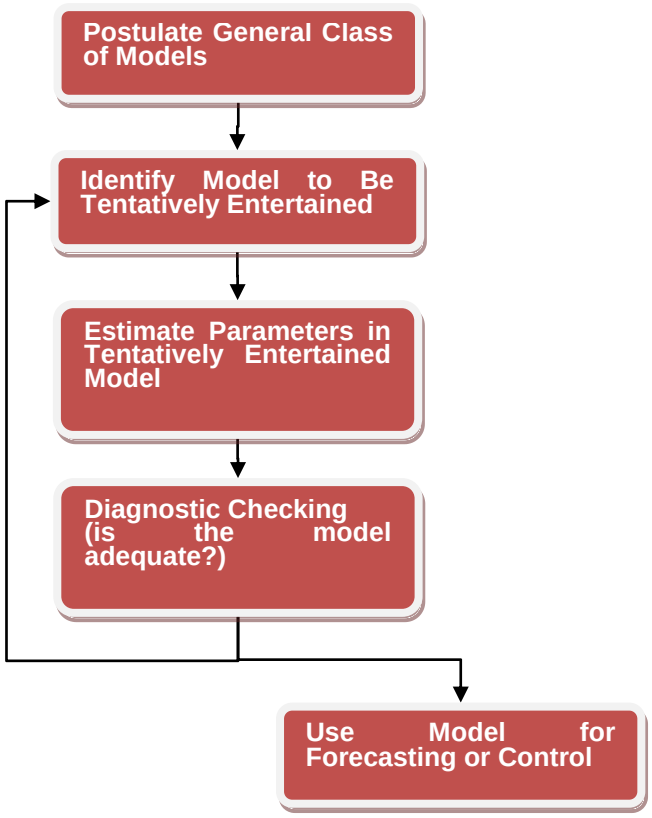
$$\left(1-\varphi_1 B-\varphi_2 B^2-\cdots-\varphi_{p+d} B^{p+d}\right)z_t = \left(1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q\right)a_t$$

위와 같은 Equation 에서 계수를 추정하고, 추정한 계수를 통해 만든 식의 적합성을 Test 하고 마지막으로 Forecasting 까지 하는 것이 목적입니다.

▪ **Box-Jenkins 의 방법론**

Box-Jenkins 는 Time Series Analysis 의 일련의 과정을 다음의 그림으로 설명했습니다.

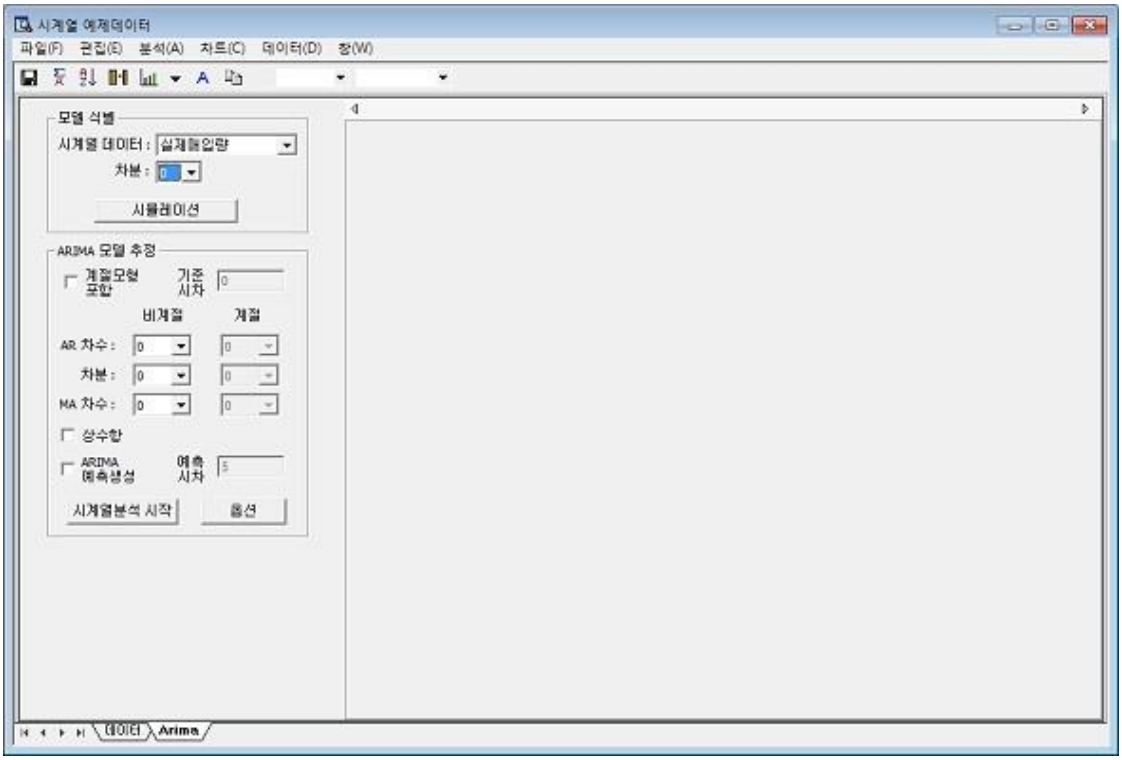
1. From the interaction of theory and practice, a useful class of models for the purposes at hand is considered.
2. Because this class is too expensive to be conveniently fitted directly to data, rough methods for identifying subclass of these models are developed. Such methods of model identification employ data and knowledge of the system to suggest an appropriate parsimonious subclass of models which may tentatively entertained. In addition, the identification process can be used to yield rough preliminary estimates of the parameters in the model.
3. The tentatively entertained model is fitted to data and its parameters estimated. The rough estimates obtained during the identification stage can now be used as starting values in more refined iterative methods for estimating the parameters.
4. Diagnostic checks are applied with the object of uncovering possible lack of fit and diagnosing the cause. If no lack of fit is indicated, the model is ready to use. If any inadequacy is found, the iterative cycle of identification, estimation, and diagnostic checking is repeated until a suitable representation is found.



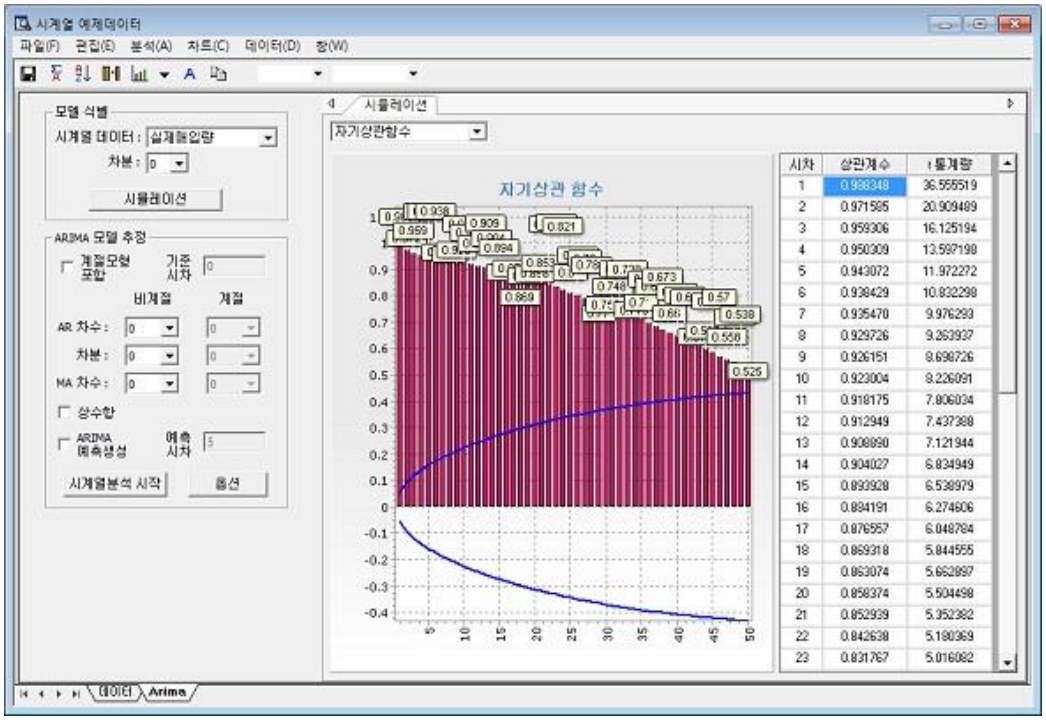
간단히 말하면 1. 일반적인 모델(ARIMA, ARCH 등)을 상정하고, 2. 그 일반적인 모델 중에서 부분 모델(ARIMA(1,1,1)과 같은 것)을 정하고, 3. 이렇게 정한 모델의 Parameter 를 Estimation 하고 4. 적합성 검증을 한 후 Forecasting 과 같은 이후의 작업을 진행하는 것이라고 할 수 있습니다.

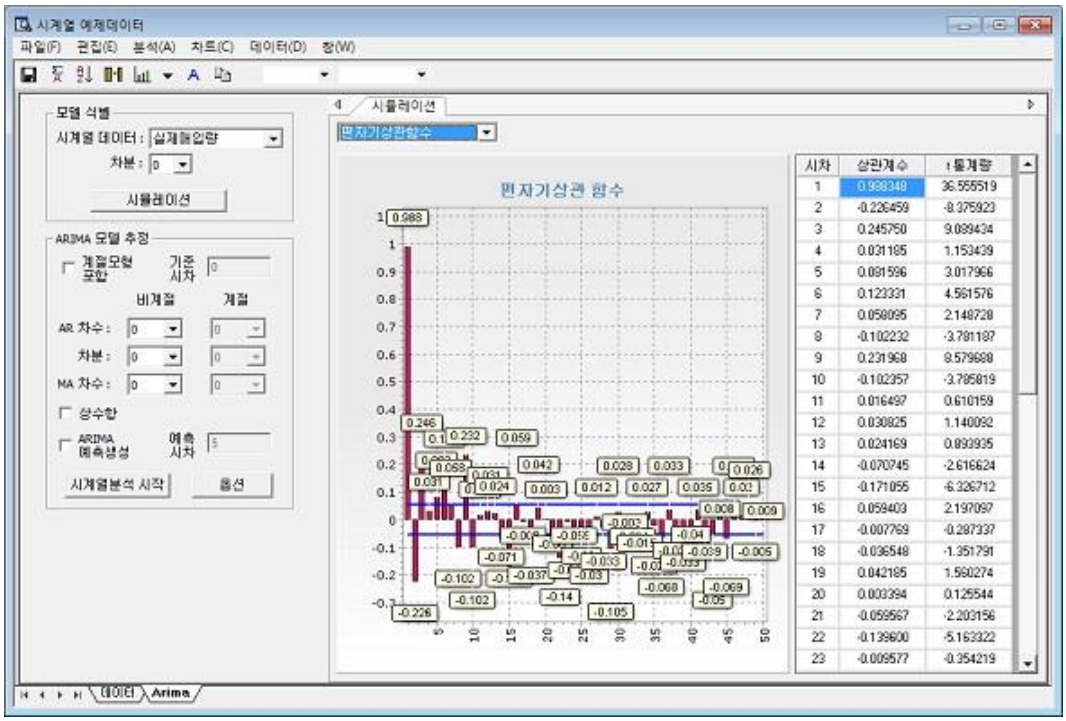
실행방법

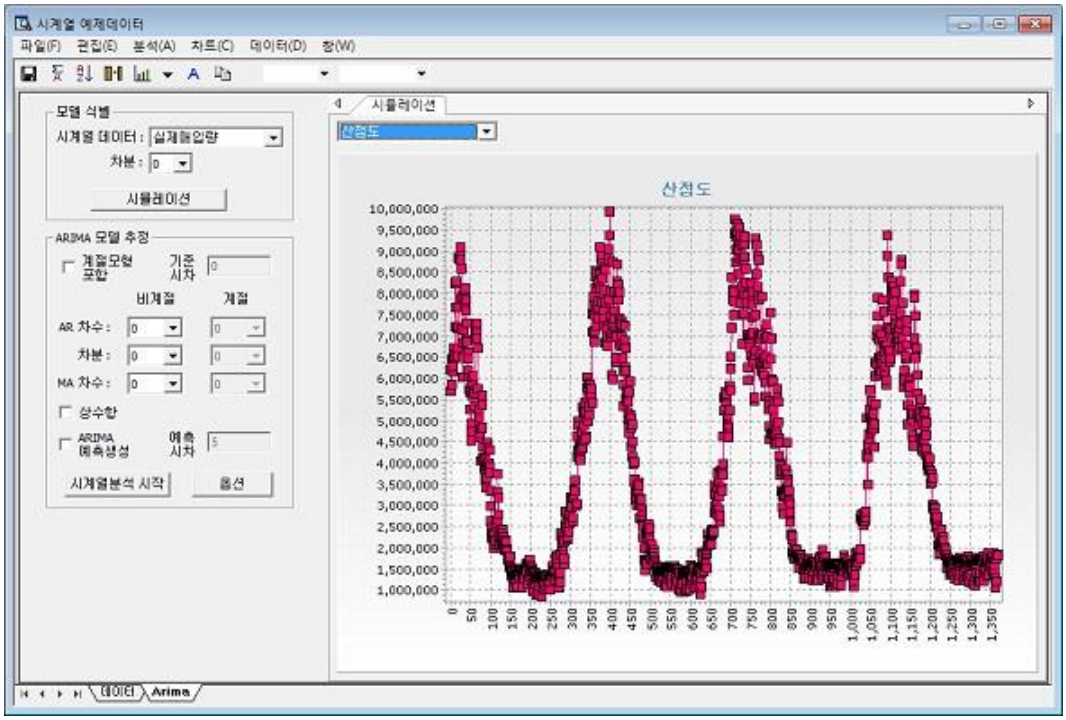
[분석] – [시계열 분석] – [시계열 모델] – [ARIMA] 를 선택하면 **[ARIMA]** 윈도우가 나타납니다.



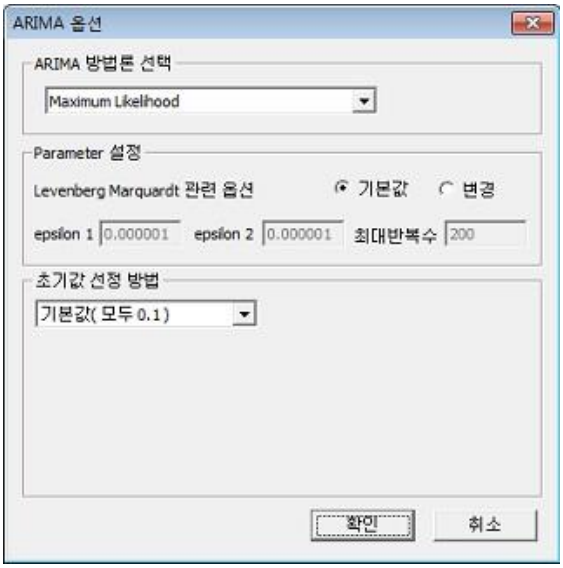
ARIMA 메인화면에서는 모델 식별 과정과 ARIMA 모델 추정 과정을 실행할 수 있습니다. 모델 식별을 통해서는 차분된 시계열 데이터의 자기 상관함수와 편자기 상관함수 그리고 산점도를 볼 수 있는데, 이를 통해서 사용자는 ARIMA 모델을 추정할 때 필요한 ARIMA 모델의 차수를 결정하는데 도움을 받을 수 있습니다. 아래는 자기 상관함수, 편자기 상관함수, 산점도의 예시입니다.


시계열 예제데이터
파일(F) 편집(E) 분석(A) 차트(C) 데이터(D) 창(W)
모델 식별
시계열 데이터 : 실제매입량
차분 : 0
시뮬레이션
ARIMA 모델 추정
계절모형 포함
기준 시차
비계절
계절
AR 차수 :
차분 :
MA 차수 :
상수항
ARIMA 예측생성
예측 시차
시계열분석 시작
옵션
시뮬레이션
자기상관함수
자기상관 함수
시차
상관계수
통계량
데이터
Arima


시계열 예제데이터
파일(F) 편집(E) 분석(A) 차트(C) 데이터(D) 창(W)
모델 식별
시계열 데이터 : 실제매입량
차분 : 0
시뮬레이션
ARIMA 모델 추정
계절모형 포함
기준 시차
비계절
계절
AR 차수 :
차분 :
MA 차수 :
상수항
ARIMA 예측생성
예측 시차
시계열분석 시작
옵션
시뮬레이션
편자기상관함수
편자기상관 함수
시차
상관계수
통계량
데이터
Arima



이렇게 자기 상관함수, 편자기 상관함수, 산점도를 통해 모델을 식별하고 차수를 결정합니다. 옵션 버튼을 클릭하면 다음과 같이 ARIMA 방법론(Conditional Least Square, Maximum Likelihood)를 선택할 수 있고 Parameter Optimization 을 위해서 사용하는 Levenberg Marquardt 에 관련된 상수 설정, 초기값 선정 방법 등을 입력 하도록 합니다.



그리고 시계열 분석을 시작하면 다음과 같은 결과들을 얻을 수 있습니다.

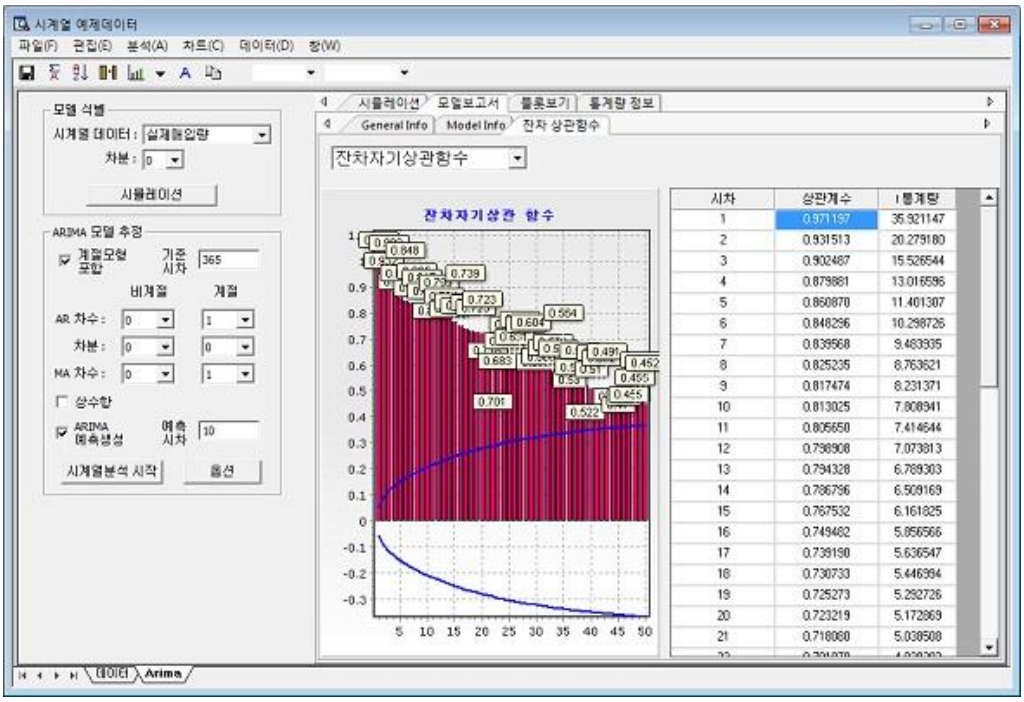
결과

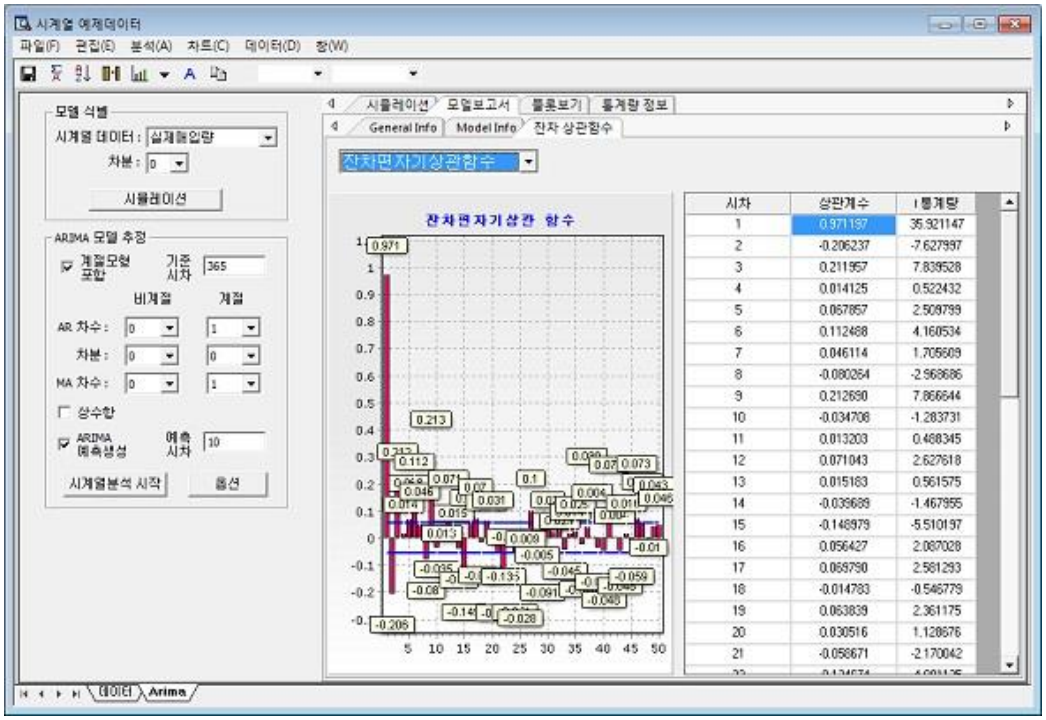
- **모델 보고서**

General Info: 시계열 데이터에 대한 기본적인 정보를 보여줍니다.

Model Info: 시계열 분석 결과를 보여 줍니다. 모델 형태 정보, 모수 정보, Parameter Optimization 결과, 예측 결과를 얻을 수 있습니다.

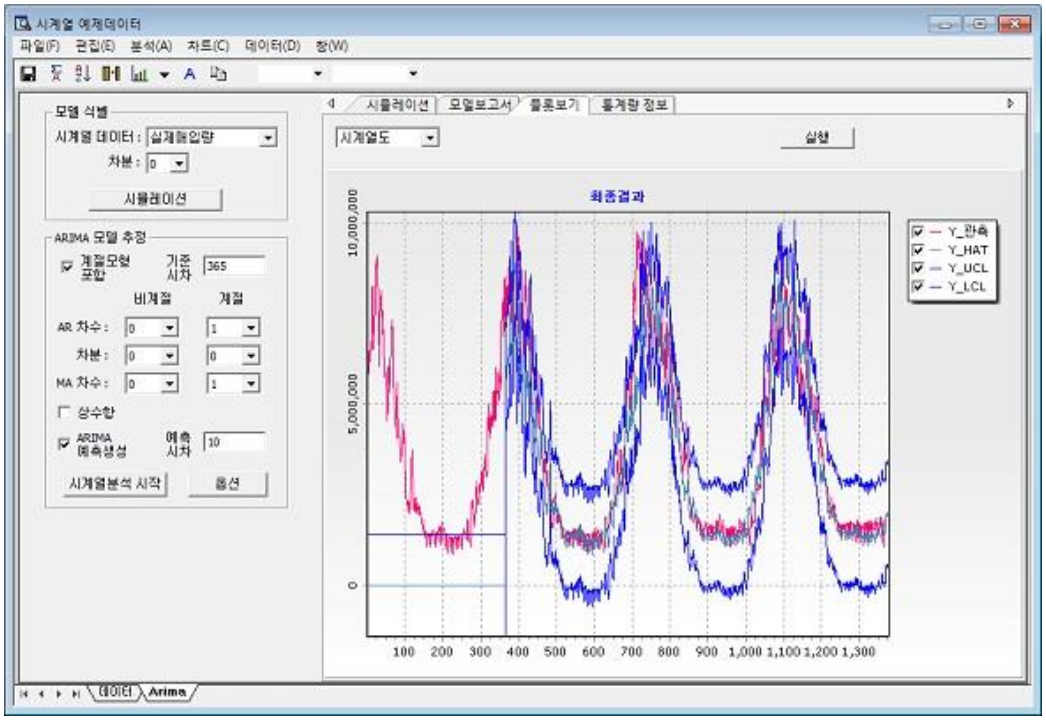
잔차 자기 상관 함수 및 잔차 편자기 상관함수: 잔차 자기 상관 함수 및 편자기 상관함수를 통해서 Diagnostic Check 를 해 볼 수 있습니다.



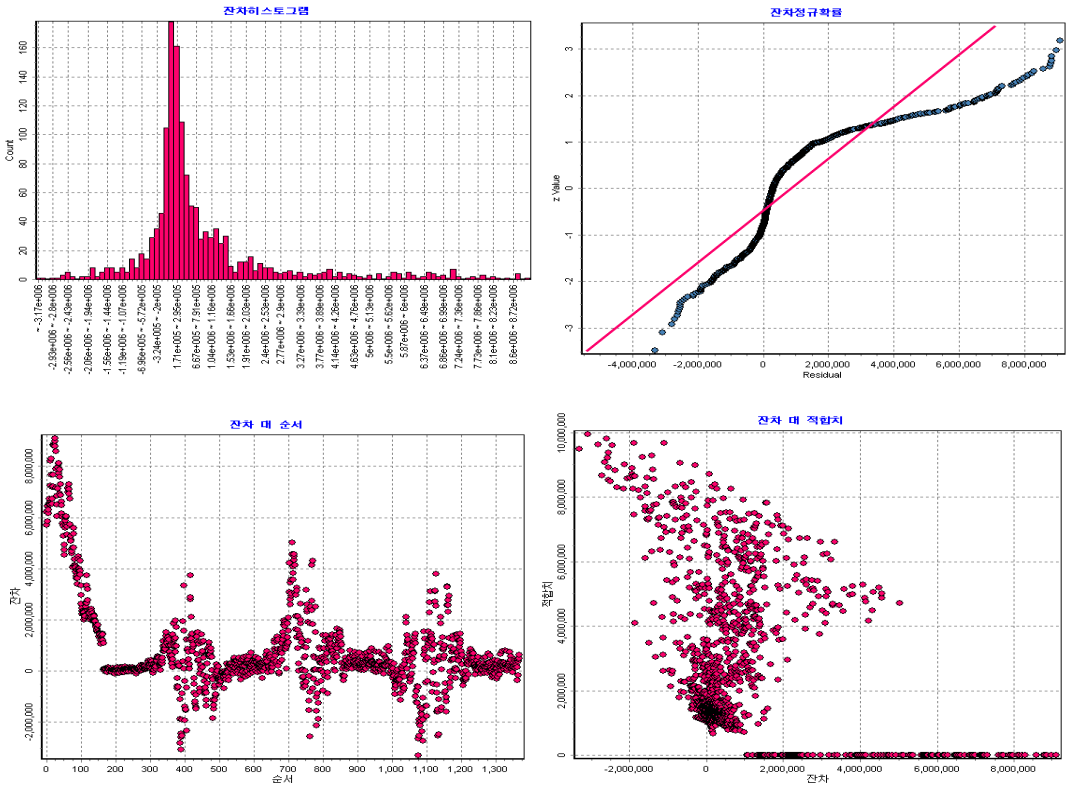


- **플롯 보기**

  ARIMA 모델을 통해서 얻어진 데이터를 시각적으로 분석할 수 있습니다. 시계열도를 통해서는 관측값, 적합치, 예측값, 예측 상한, 하한을 볼 수 있습니다.



다음의 잔차 관련 플롯들을 통해서는 Diagnostic Check 를 위한 잔차 분석을 할 수 있습니다.



- **통계량 정보**

  통계량 정보를 통해서는 모델을 생성한 후에 얻어지는 데이터를 Table 에서 볼 수 있습니다. 이와 같이 얻어진 정보를 저장할 수 있는 기능 또한 제공합니다.

(5) 추세분석

개요

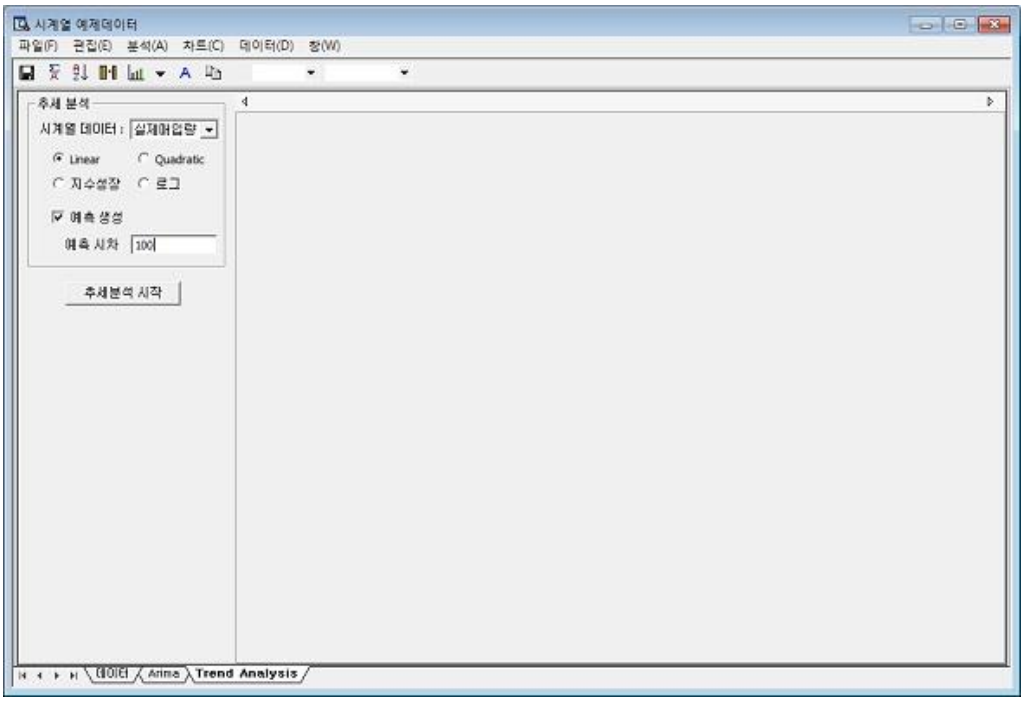
추세 분석은 시계열 데이터의 기본적인 추세를 얻기 위해서 수행하는 방법론입니다. 시점별로 관측되는 시계열 데이터는 오직 시간을 수평축으로 하고, 관측 값을 수직축으로 하는 함수 관계에 의해서 설명된다는 논리 하에서 만들어진 것입니다. 이러한 함수 관계는 다음과 같은 것으로 한정하도록 합니다.

선형
2 차 곡선
Exponential Growth
로그

각 함수에서 추정되는 Parameter 는 모두 Least Square 방법을 통해서 Closed Form 으로 구해지게 됩니다.

실행방법

[분석] – [시계열 분석] – [시계열 모델] – [추세분석]을 선택하면 **[추세분석]** 윈도우가 나타납니다. 창에서 어떠한 형태의 함수에 적합을 할 것인지 선택하고 예측을 하고자 할 경우 예측 시차를 입력하고 추세 분석 시작 버튼을 클릭합니다.



결과

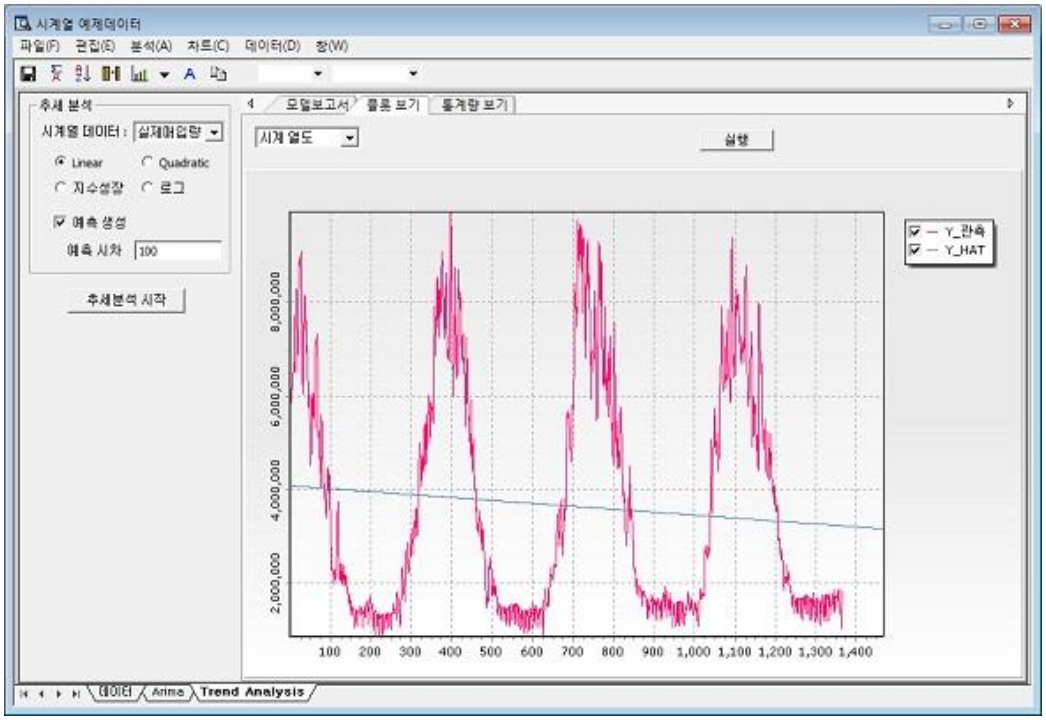
- **모델 보고서**

  General Info: 시계열 데이터의 기본적인 정보를 보여줍니다.

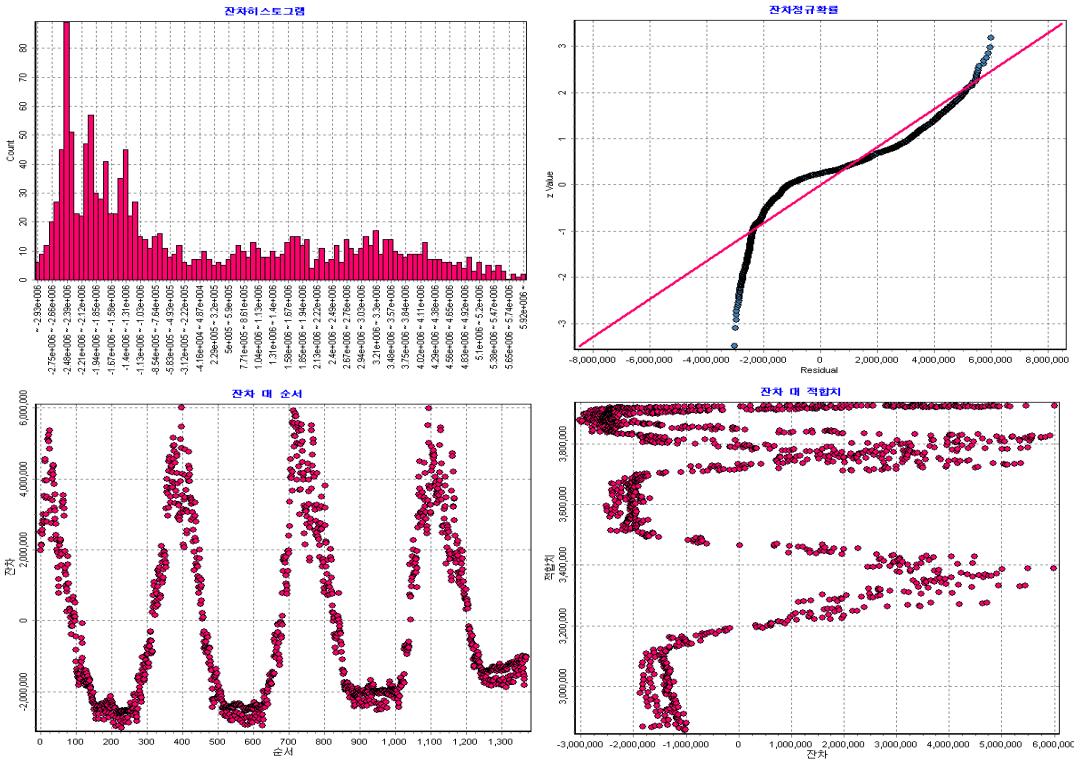
  Model Info: 추세 분석을 통해서 얻어진 회귀식과 정확도 측도, 예측 결과를 보여줍니다.

- **플롯 보기**

  추세분석을 통해서 얻어진 데이터를 시각적으로 표시해 줍니다.



시계열도를 통해 시계열 데이터와 그것을 추정하는 회귀식을 한 눈에 볼 수 있습니다.



잔차 관련 플롯(잔차 히스토그램, 잔차 정규확률 플롯, 잔차 대 순서, 잔차 대 적합치)를 통해서는 추세 분석 후에 얻어진 잔차를 분석할 수 있습니다.

- **통계량 보기**

  추세 분석을 통해서 얻어진 통계량을 Table 에서 볼 수 있습니다. 이와 함께 Table 을 저장하는 기능도 제공합니다.

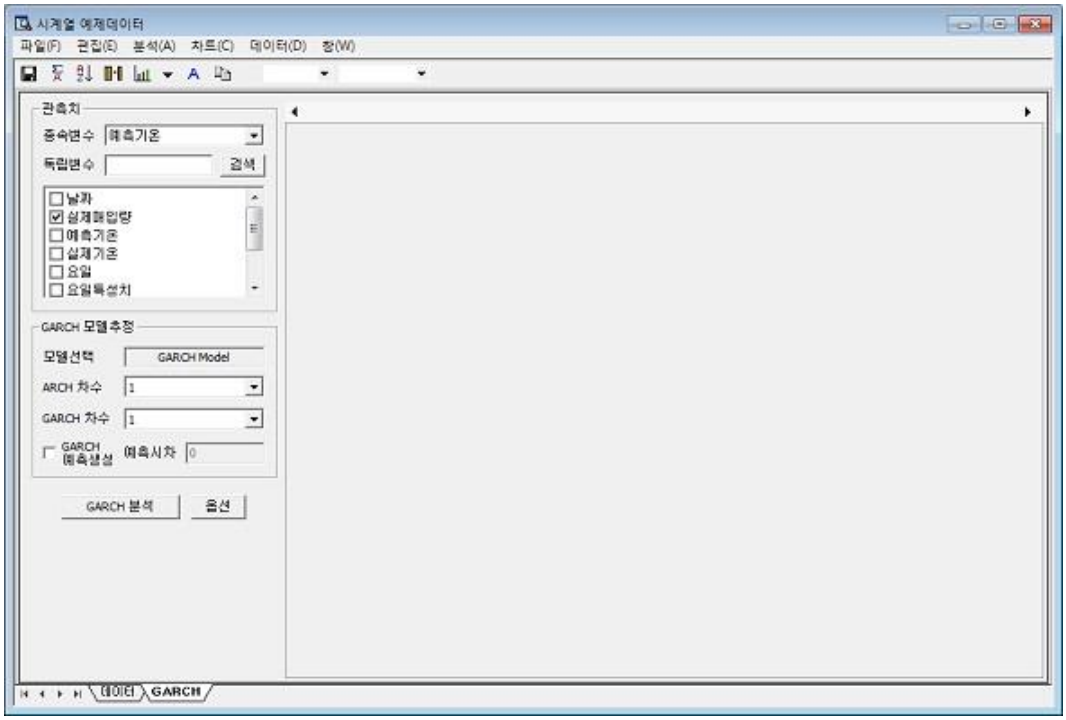
(6) GARCH

개요

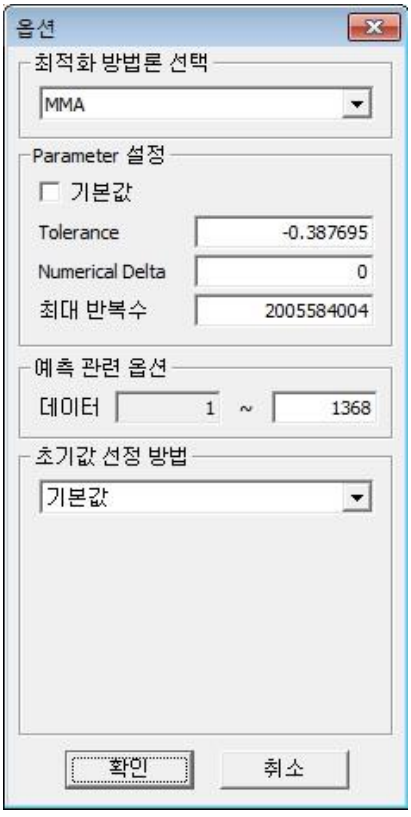
ARCH (AutoRegressive Conditional Hetroskedasticity) 방법론은 1980 년대에 제시된 방법론으로 이 방법을 제시한 Robert Engle 은 2003 년도 노벨 경제학상을 수상하였습니다. 본 방법론의 혁신적인 면은 그 동안의 Time Series 분석에서 Random Shock 이 일정한 분산을 갖는다는 가정을 수정하여 Random Shock 은 비조건부적으로는 일정한 분산을 갖지만 조건부적으로는 변화하는 분산을 갖는다는 가정을 도입했다는 것입니다. Robert Engle 이 1984 년에 ARCH Model 을 제시한 이후로 일반화된 ARCH 모델 즉 GARCH 모델이 제시되었고, 이후에 EGARCH, GARCH-M, GJR 등의 여러 변형된 모델이 제시되어 왔습니다. 이렇듯 ARCH 의 변형된 모델이 많이 제시되고 또 많은 각광을 받은 이유는 금융 경제학의 발전과 더불어 단순히 Time Series 의 예측뿐 아니라 변동성(분산)의 측정과 예측이 매우 중요해졌기 때문입니다. 많은 자본 자산 가격 결정 모형에서 기초 자산의 변동성은 자산의 가격에 영향을 미치는 중요한 요인으로 여겨져 왔는데 이러한 변동성을 측정하고 예측하는 방법론이 ARCH 이전에는 발전되어 있지 않았습니다. 1980 년대뿐 아니라 그 이후, 그리고 최근에도 이와 관련된 연구는 꾸준히 지속되고 있으며 이제는 금융 경제학 이외에도 여러 공학 분야 특히 Network Traffic 분석에 GARCH 모형이 활용되고 있습니다.

실행방법

[분석] – [시계열 분석] – [시계열 모델] – [GARCH]을 선택하면 **[GARCH]** 윈도우가 나타납니다. 메인 화면에서는 GARCH 분석에 필요한 기본적인 것들을 입력하도록 합니다. 종속변수(반응변수)와 독립변수(설명변수)를 선택하고 ARCH 차수와 GARCH 차수를 선택해줍니다. 마지막으로 예측 생성 여부를 선택하면 모든 설정이 끝납니다.



더욱 구체적인 설정은 아래의 옵션창에서 설정할 수 있습니다.



GARCH 의 Parameter Estimation 은 매우 까다롭습니다. 경우에 따라서는 Estimation 이 국부 최적해(Local Optimum)에서 종료하는 경우가 빈번하게 발생합니다. 따라서 ECMiner™에서는 여러 최적화 방법론을 제시하여 사용자가 선택할 수 있도록 하였습니다. 사용자는 여러 방법론으로 최적화를 수행해 보고 가장 좋은 해를 취할 수 있습니다. Parameter 설정은 최적화 알고리즘을 수행하는데 있어서 설정할 옵션들입니다. 예측 관련 옵션을 통해서는 어떠한 데이터를 모델링에 사용할지를 정합니다. 초기값의 선정방법을 통해서는 어떠한 방법으로 초기값을 선정할 것인지를 결정합니다.

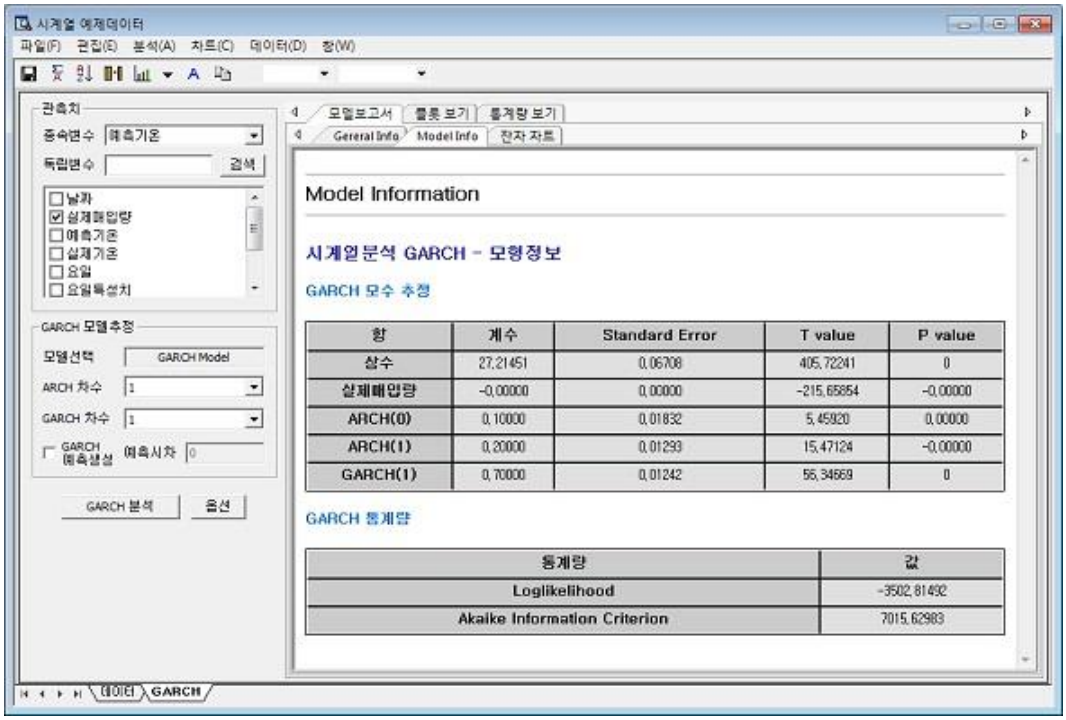
결과보기

- **모델 보고서**

  General Info: 시계열 데이터의 기본적인 정보를 보여줍니다.

  Model Info: GARCH 분석을 통해서 얻어진 모수와 통계량, 그리고 예측 결과를 보여줍니다.

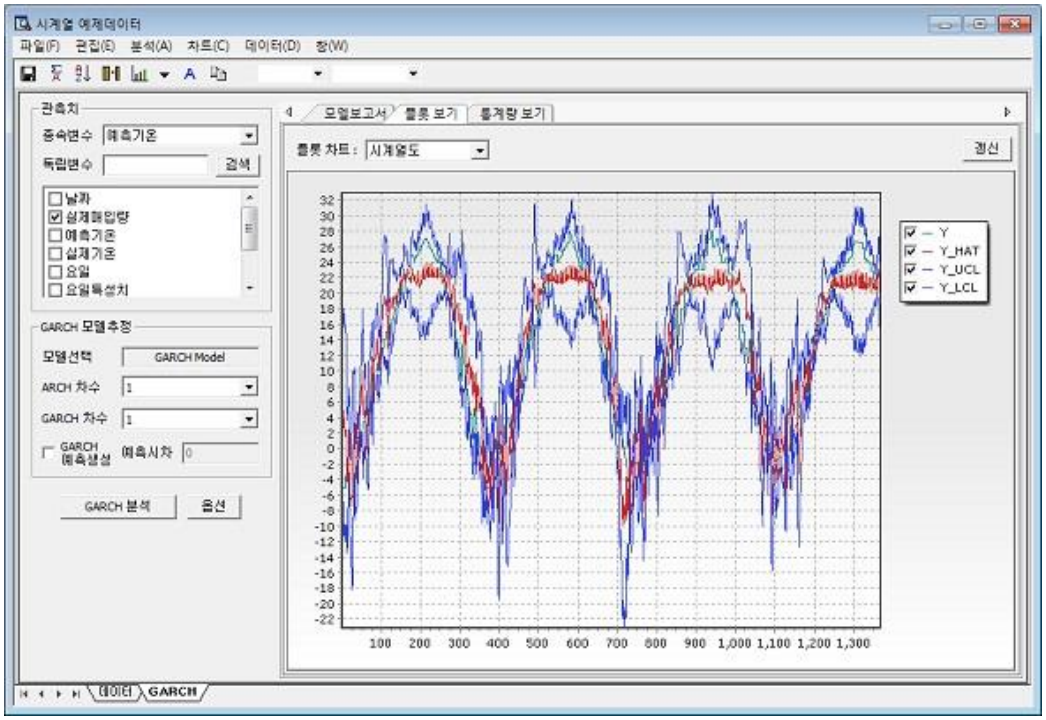
  잔차 차트: 잔차 자기 상관함수, 잔차 자기 상관 함수를 보여줍니다.



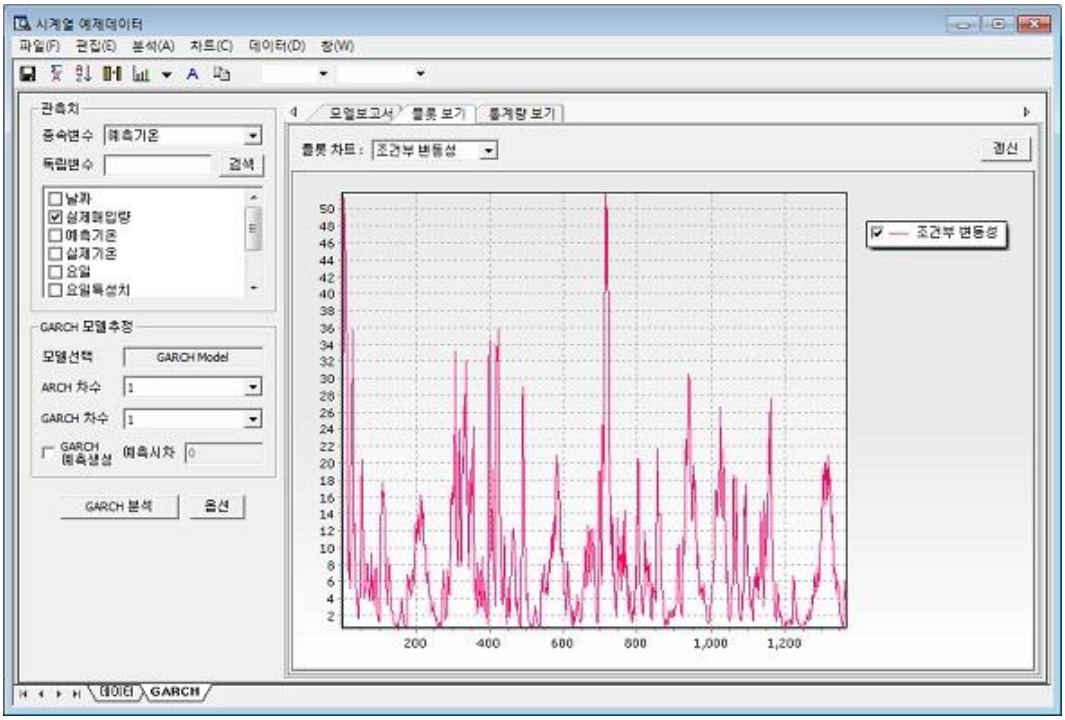
- **플롯 보기**

GARCH 분석을 통해서 얻어진 데이터를 시각적으로 보여줍니다. 시계열도, 잔차제곱, 조건부 변동성, 그리고 잔차에 대한 플롯(잔차 히스토그램, 잔차 정규확률 플롯, 잔차 대 순서, 잔차 대 적합치)을 보여줍니다.

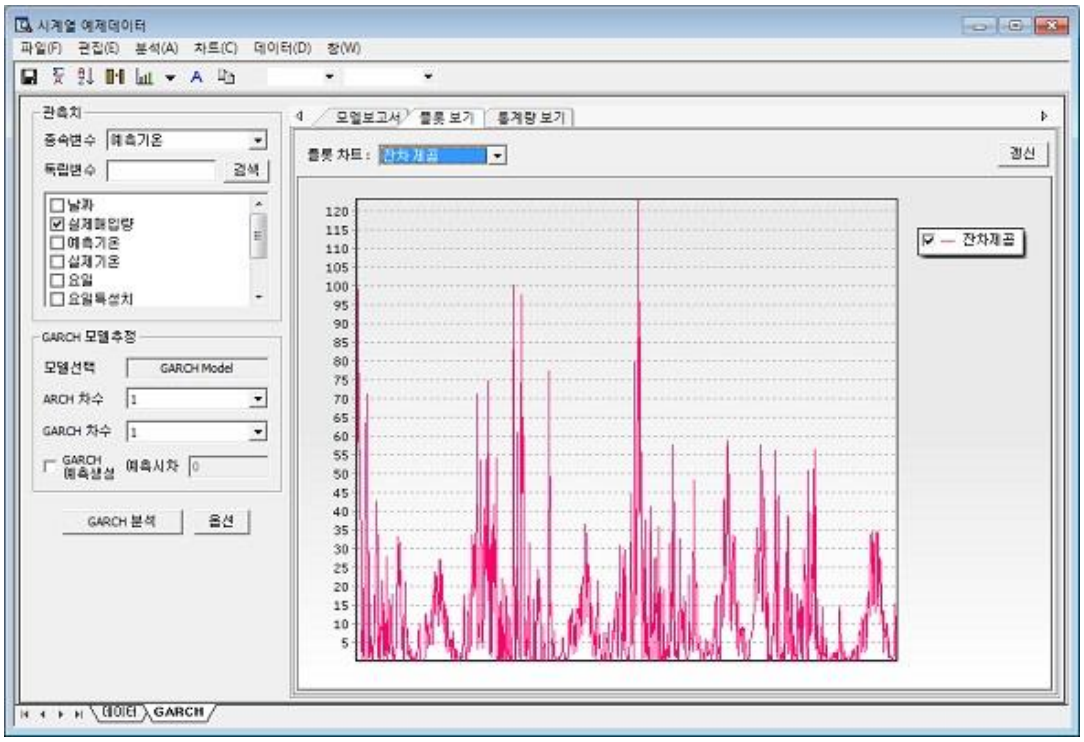
아래는 시계열도의 예시입니다. 시계열 데이터와 적합치, 그리고 적합치의 상한, 하한을 한 눈에 볼 수 있습니다.



아래 플롯은 조건부 변동성 플롯입니다. 모델링의 결과 나타난 조건부 변동성의 값을 볼 수 있습니다. (만약 예측을 할 경우 조건부 변동성의 예측 값도 나타나게 됩니다.)



아래 플롯은 잔차 제곱 플롯입니다. 모델링의 결과 나타난 잔차 제곱의 값을 볼 수 있습니다.



이와 함께 네가지의 잔차 관련 플롯을 통해서 정규성을 확인해 볼 수 있습니다.

- **통계량 보기**

  GARCH 분석을 통해서 얻어진 통계량을 Table 에서 볼 수 있습니다. 이와 함께 Table 을 저장하는 기능도 제공합니다.

(7) VAR

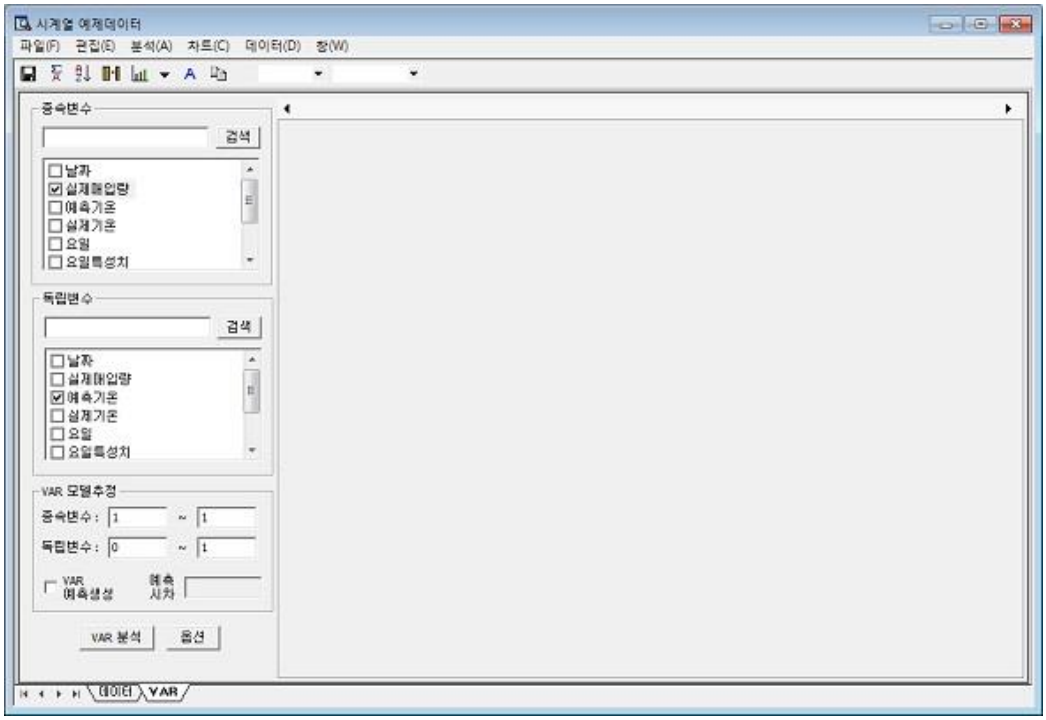
개요

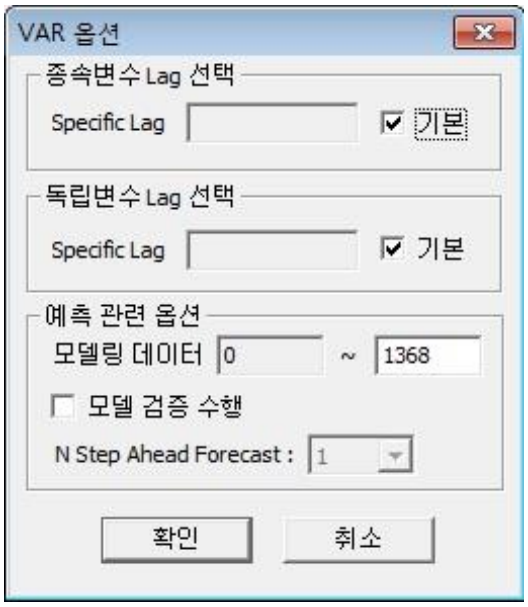
실증 분석에서는 종종 두 개 이상의 시계열을 동시에 모형화하는 것이 유리한 경우가 많습니다. 특정 변수의 집합들이 단순히 개별적으로 움직이는 것이 아니라 서로 영향을 받으며 움직이고 있다면 다음과 같이 모델을 상정할 수 있습니다. 이를 VAR 모델이라고 합니다.

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p x_{t-p} + B_0 x_t + B_1 x_{t-1} + B_q x_{t-q} + \epsilon_t$$

실행방법

[분석] – [시계열 분석] – [시계열 모델] – [VAR]을 선택하면 **[VAR]** 윈도우가 나타납니다. 메인 화면에서는 VAR 분석에 필요한 기본적인 것들을 입력하도록 합니다. 종속변수(반응변수)와 독립변수(설명변수)를 선택할 시에 종속 변수로 여러 개의 변수를 선택할 수 있고, 독립변수로는 선택을 할 수도 있고 안 할 수 도 있습니다. 종속 변수의 차수와 독립변수의 차수를 입력할 때는 시작 차수와 끝차수를 입력합니다.





종속변수 Lag 선택 시 기본을 선택하면 Main 화면에서 설정한 차수가 설정됩니다. 이 때 기본을 선택하지 않으면 Specific Lag 를 사용자가 입력할 수 있습니다. (이 때 차수의 구분은 space 로 합니다.)

독립변수 Lag 선택 시 기본을 선택하면 마찬가지로 Main 화면에서 설정한 차수가 설정됩니다. 이 때 기본을 선택하지 않으면 Specific Lag 를 사용자가 입력할 수 있습니다. (이 때 차수의 구분은 space 로 합니다.)

예측 관련 옵션에서는 모델링에 사용할 데이터를 정하고 모델 검증 수행을 선택할 경우 특정 앞 step 를 선택함으로써 N step ahead forecast 를 할 수 있습니다. ECMiner™ 에서는 모델링 데이터의 바로 다음 시점부터 N step ahead forecast 를 수행합니다. 이를 통해 분석자는 모델링을 통해서 만든 모델이 이후의 예측에 유용한지를 가늠해 볼 수 있습니다.

결과보기

- **모델 보고서**

  General Info: 시계열 데이터의 기본적인 정보를 보여줍니다.

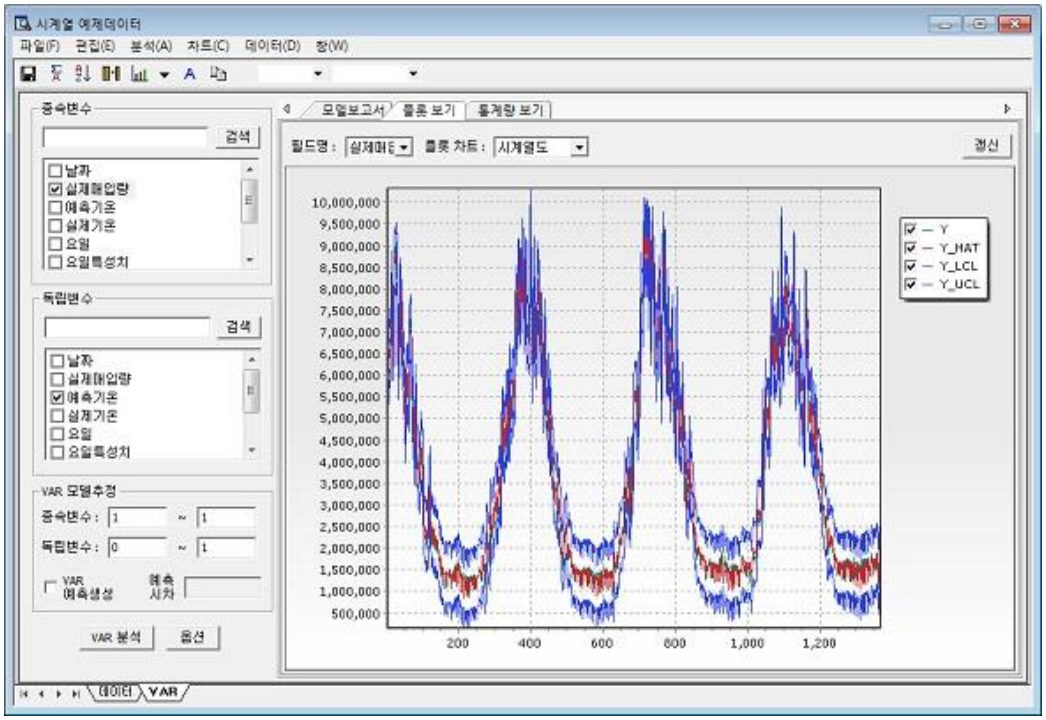
  Model Info: VAR 분석을 통해서 얻어진 모수와 통계량, 그리고 예측 결과를 보여줍니다.

  잔차 차트: 잔차 자기 상관함수, 잔차 자기 상관 함수를 보여줍니다.

- **플롯 보기**

  VAR 분석을 통해서 얻어진 데이터를 시각적으로 보여줍니다. 시계열도 그리고 잔차에 대한 플롯(잔차 히스토그램, 잔차 정규확률 플롯, 잔차 대 순서, 잔차 대 적합치)을 보여줍니다.

아래는 시계열도의 예시입니다. 시계열 데이터와 적합치, 그리고 적합치의 상한, 하한을 한 눈에 볼 수 있습니다.



- **통계량 보기**

  VAR 분석을 통해서 얻어진 통계량을 Table 에서 볼 수 있습니다. 이와 함께 Table 을 저장하는 기능도 제공합니다.

(8) ARMAX

개요

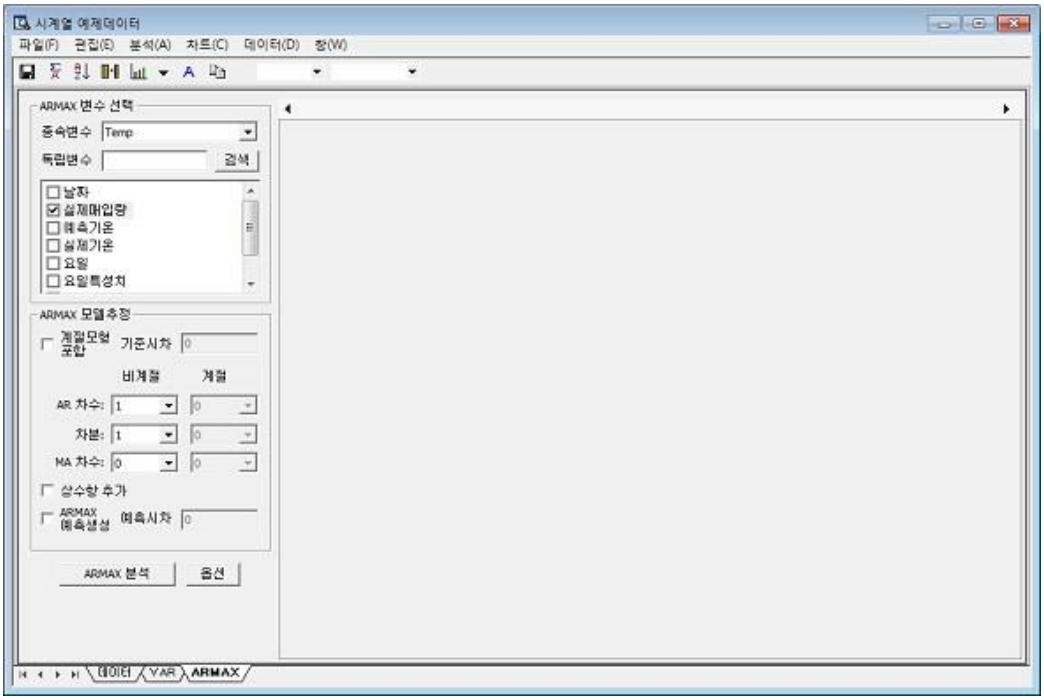
ARMAX 는 ARMA with Exogenous Variable 의 약자로 이미 언급한 ARMA 모형에 Exogenous Variable 를 추가한 모델 형태입니다. 이 모델을 다음과 같이 기술할 수 있습니다.

$$\phi(B)\left(y_t - \mu - x_{t1}\beta_1 - \cdots - x_{tm}\beta_m\right) = \theta(B)a_t$$

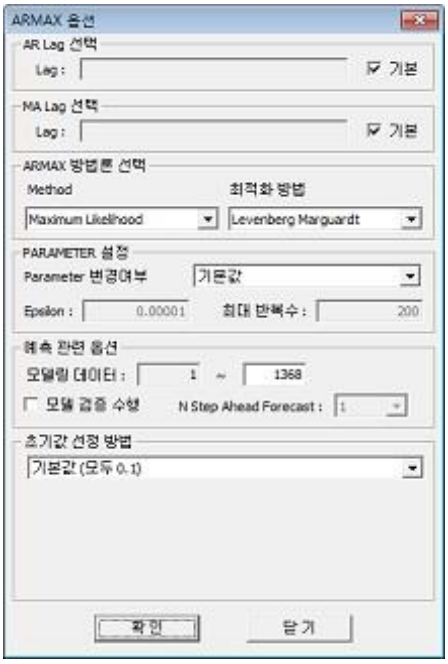
Parameter Estimation 과 Forecasting 모두 기존 ARMA 의 방법과 동일합니다. 단지 기존의 y_t 혹은 $y_t - \mu$를 $y_t - x_{t1}\beta_1 - \cdots - x_{tm}\beta_m$ 혹은 $y_t - \mu - x_{t1}\beta_1 - \cdots - x_{tm}\beta_m$ 로 대체하면 모든 과정은 동일하다고 할 수 있습니다. 이는 단순히 ARMA 만으로 설명할 수 없는 시계열을 Exogenous Variable 을 사용하여 더 잘 설명할 수 있도록 도와주는 모델이라고 할 수 있습니다.

실행 방법

[분석] – [시계열 분석] – [시계열 모델] – [ARMAX]을 선택하면 **[ARMAX]** 윈도우가 나타납니다.



메인 화면에서는 ARMAX 분석에 필요한 기본적인 것들을 입력하도록 합니다. 종속변수(반응변수)와 독립변수(설명변수)를 선택할 시에 독립 변수는 선택하지 않을 수 있습니다. 이 경우 기존의 ARIMA 와 동일한 결과를 얻게 됩니다. 분석자는 기존과 ARIMA 와 같이 계절성을 추가할 수도 있고 차분을 넣을 수도 있습니다. Main 화면에서 나타난 여러 차수 설정을 통해서 ARMAX 분석을 할 수 있습니다.



이 뿐 아니라 옵션 창을 통해서 보다 심화된 설정을 할 수 있습니다. AR 차수를 분석자가 지정하는 차수로 입력하고 싶을 경우 체크 박스에서 기본에 해당하는 체크를 제거합니다. 그러면 space 단위로 하여 AR 차수를 지정할 수 있습니다. MA 차수를 분석자가 지정하는 차수로 입력하고 싶을 경우 체크 박스에서 기본에 해당하는 체크를 제거합니다. 그러면 space 단위로 하여 MA 차수를 지정할 수 있습니다.

ARMAX 의 모수를 추정하기 위해서 어떠한 방법을 사용할지에 대해서는 Maximum Likelihood 와 Conditional Least Square 중에서 하나를 선택할 수 있습니다. 최적화 방법으로는 Levenberg Marquardt 와 Quasi Newton 을 제공하는데 사용자는 이 두 가지 방법을 사용해 봄으로써 둘 중에서 더 좋은 Parameter 를 사용할 수 있습니다.

최적화에 사용되는 Parameter 를 변경하고 싶을 경우 '변경'을 선택해서 하면 됩니다. 이를 통해서 좀 더 최적화를 자세히 할지 그렇지 않을지를 결정할 수 있습니다.

예측 관련 옵션을 통해서는 모델링에 사용되는 데이터를 선택하고 모델 검증을 수행할지를 결정합니다. 모델 검증을 선택할 시에 모델링 데이터 이후부터 예측값을 계산합니다. 이를 통해서 모델링을 통해 얻은 모델이 얼마나 정확한지를 가늠할 수 있습니다.

초기값 선택은 Nonlinear Optimization 이 갖는 한계를 보완하기 위한 것입니다. 다양한 초기값 설정을 통해 최적의 해를 구하는데 도움이 됩니다.

결과 보기

- **모델 보고서**

  General Info : 시계열 데이터의 기본적인 정보를 보여줍니다.

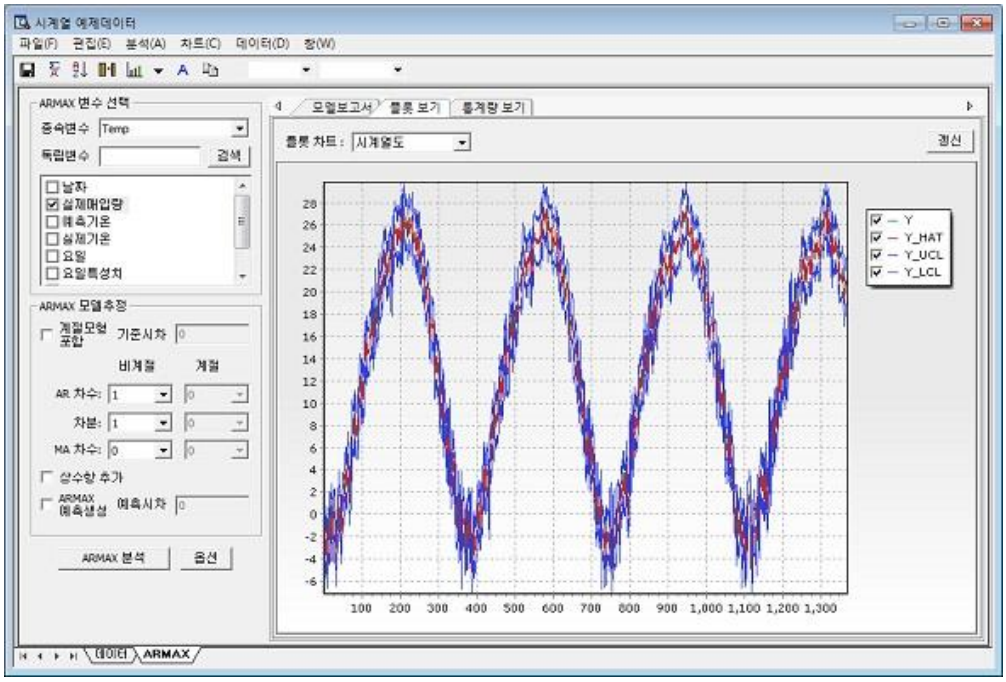
  Model Info: ARMAX 분석을 통해서 얻어진 모수와 통계량, 그리고 예측 결과를 보여줍니다.

  잔차 차트: 잔차 자기 상관함수, 잔차 자기 상관 함수를 보여줍니다

- **플롯 보기**

  ARMAX 분석을 통해서 얻어진 데이터를 시각적으로 보여줍니다. 시계열도 그리고 잔차에 대한 플롯(잔차 히스토그램, 잔차 정규확률 플롯, 잔차 대 순서, 잔차 대 적합치)을 보여줍니다.

  아래는 시계열도의 예시입니다. 시계열 데이터와 적합치, 그리고 적합치의 상한, 하한을 한 눈에 볼 수 있습니다.



- **통계량 보기**

  VAR 분석을 통해서 얻어진 통계량을 Table 에서 볼 수 있습니다. 이와 함께 Table 을 저장하는 기능도 제공합니다.

- 5.3.6.2 시계열 검정

(1) 단위근 검정

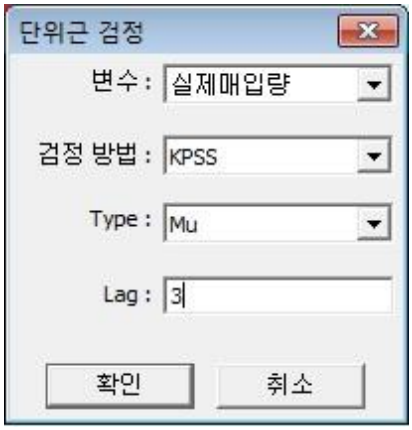
개요

불안정한 시계열(Non-stationary time series)은 자기회귀모형(autoregressive model)으로 나타낼 경우 특성근이 1 입니다. 즉 단위근(unit root)을 갖습니다. 시계열 검정 메뉴에서는 KPSS(Kwiatkowski & Phillips & Schmidt & Shin) 검정, ADF (Augmented Dickey-Fuller) 검정을 지원하는데, ADF 은 변수에 단위근이 존재 한다는 귀무가설 이며, KPSS 검정은 변수에 단위근이 존재하지 않는다는 것이 귀무가설입니다. 검정통계량 (TAU)과 유의확률(p-value)등을 통해 귀무가설을 채택할 것인지 기각할 것인지를 결정합니다.

실행 방법

[분석] – [시계열 분석] – [시계열 검정] – [단위근 검정]을 선택하면 **[단위근 검정]** 윈도우가 나타납니다. 창에서 어떠한 검정 방법을 사용할 것인지 선택하고, 각 방법에 필요한

parameter 를 선택/입력한 후 검정할 변수를 선택합니다. 확인 버튼을 클릭하면 시계열 검정이 수행됩니다.



▪ **KPSS(Kwiatkowski & Phillips & Schmidt & Shin) Test**

KPSS 의 test statistics value 를 구하는 식은 아래와 같습니다.

$$KPSS = N^{-2}\sum_{t=1}^{N} S_t^2 \big/ \hat{\sigma}^2(p)$$

시계열을 $x_t = r_t + \beta t + \varepsilon_t$ 와 같이 random work + deterministic trend + stationary error 의 회귀식으로 나타낼 때, $S_t = \sum_{j=1}^{t} e_j$ 입니다. (e_t (t = 1, 2, … , N)는 residuals)

Type 을 mu 로 입력한 경우 test statistics 계산 시의 residual = y – mean(y) 이고, tau 로 입력한 경우 y 를 종속 변수로 하고 time trend 를 독립 변수로 linear regression 한 결과의 residuals 를 사용합니다.

SelectLags 를 nil 로 입력한 경우,

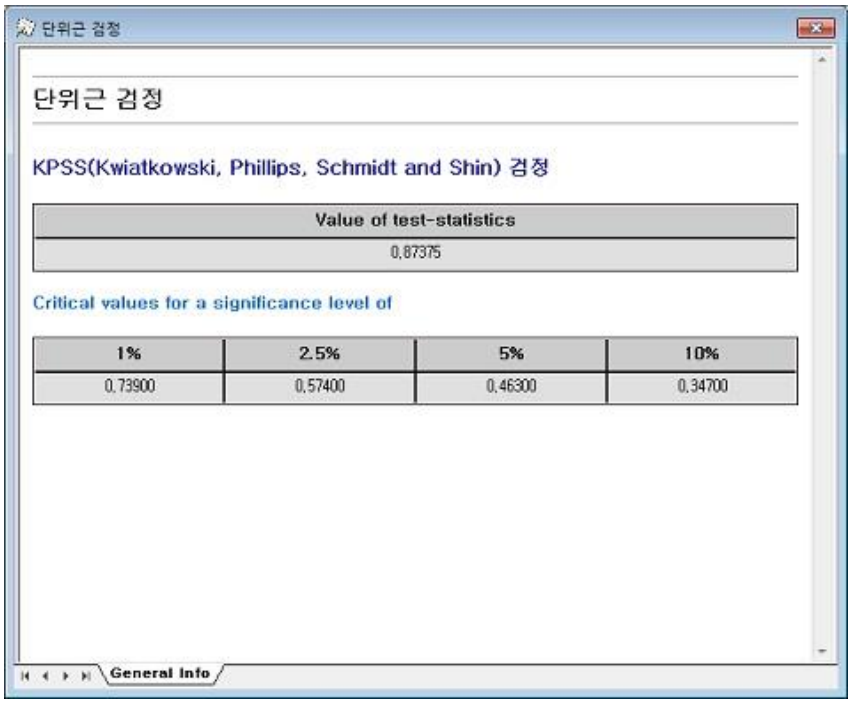
$$\hat{\sigma}^2(p) = \frac{1}{N}\sum_{t=1}^{N} e_t^2$$

로 계산하고, short, long 일 때는

$$\hat{\sigma}^2(p) = \frac{1}{N}\sum_{t=1}^{N} e_t^2 + \frac{2}{N}\sum_{j=1}^{p} w_j(p) \sum_{t=j+1}^{N} e_t e_{t-j}$$

로 계산합니다. Short 일 경우 lag 값이 4 * (0.01 * N)$^{1/4}$ 의 정수 부분을 취하고, long 일 경우 12 * (0.01 * N)$^{1/4}$ 의 정수 부분을 취합니다.

결과



이 검정 방법 따른 검정 통계량과 각 significance level(1%, 2.5%, 5%, 10%)에 대한 critical value 값을 제공합니다.

▪ **ADF(Augmented Dichey-Fuller) Test**

시계열 데이터는 상수와 확정적 추세(trend)의 포함 여부에 따라 다른 형태로 나타낼 수 있습니다. ADF Test 의 선택항목 Type 은 검정할 시계열 데이터의 형태에 맞게 입력해야 합니다.

상수와 확정적 추세를 포함하지 않는 경우(Type = none)의 검정모형은

$$y_t = \gamma y_{t-1} + \sum_{i=1}^{p} \nabla y_{t-i} + e_t$$

입니다. 상수만을 포함하는 경우(Type = Drift)의 검정 모형은 아래와 같이 나타낼 수 있습니다.

$$y_t = \alpha + \gamma y_{t-1} + \sum_{i=1}^{p} \nabla y_{t-i} + e_t$$

상수와 확정적 추세 모두를 포함하는 경우(Type = Trend)의 검정모형은

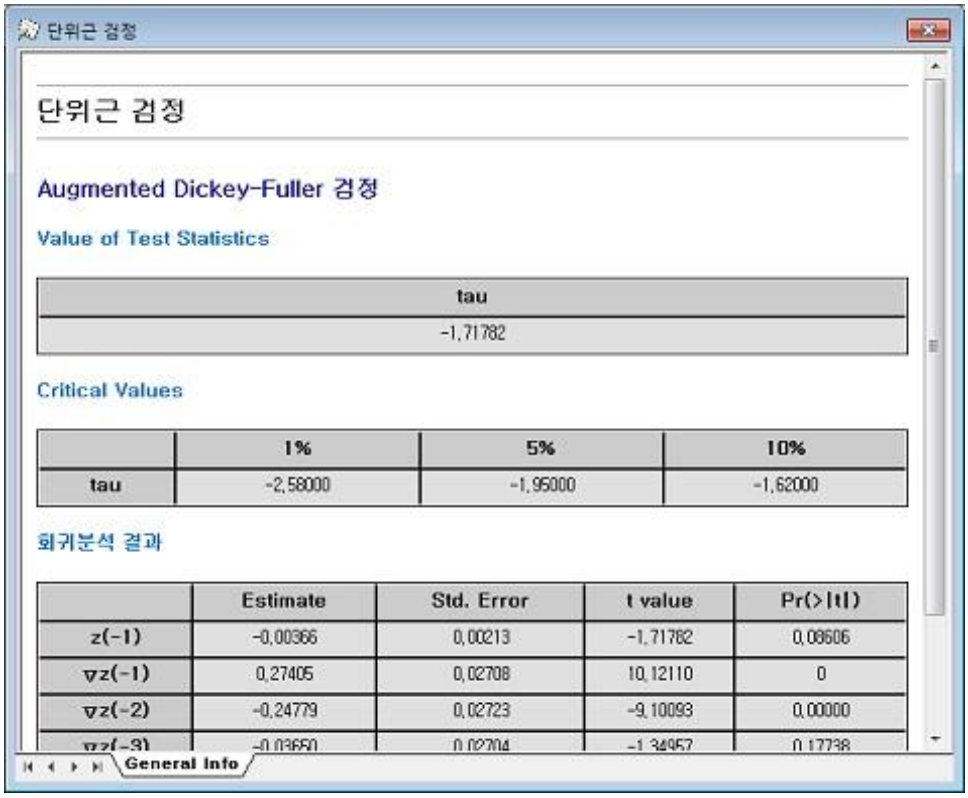
$$y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \nabla y_{t-i} + e_t$$

입니다.

Lag 은 검정모형의 회귀식에 포함되는 1 차 자기회귀항의 차수로, 위 회귀식의 p 값입니다. 따라서 Lag 값이 1 이상일 때 올바른 결과를 얻을 수 있습니다.

SelectLags 는 likelihood 값에 따라 lag order 자동 설정 여부에 관련된 입력입니다. SelectLags 를 Fixed 로 입력하면 직접 입력한 Lag 값을 사용하고, AIC, BIC 를 선택하면 Akaike Information Criterion 인 *-2 * log(likelihood) + k * npar* 에 의해 lag 값을 구합니다. k=log(n) (n 은 observation 수) 일 때 BIC(Bayesian Information Criterion)라고 합니다. (npar = number of parameter).

결과



각 검정 방법에 따른 Test-statistics 의 값, Critical value, Coefficient 등과 Multiple R-square, Adjusted R-square, F-statistics, p-value 등의 통계량을 제공합니다. Coefficient 에서 Intercept 는 상수항, tt 는 확정적 추세, z.lag.1 은 y, z.diff.lag1 은 y 의 first difference 에 대한 값입니다.

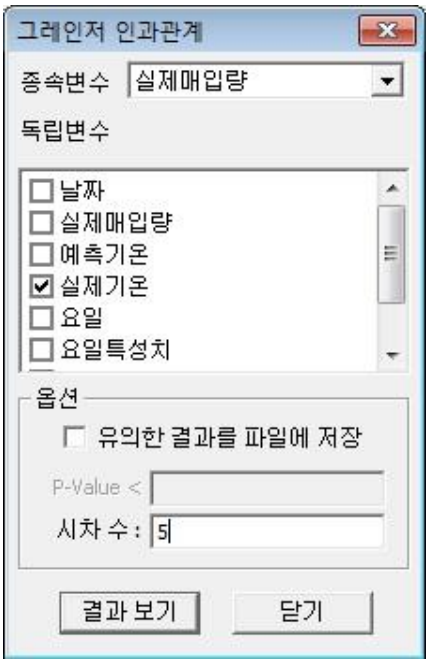
(2) 그레인저 인과 관계

개요

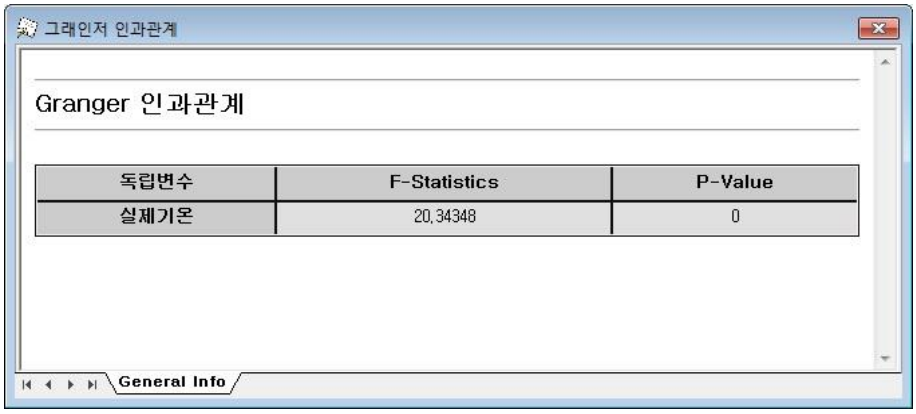
Granger causality test 는 statistical hypothesis test 로, 하나의 시계열이 다른 시계열을 예측하는 데 유용한지 그렇지 않은지를 검정하는 이론입니다. t-test 와 F-test 를 통해서 독립변수가 종속변수의 미래에 대해 통계적으로 중요한 정보를 제공하는지를 검정합니다.

실행 방법

[분석] – [시계열 분석] – [시계열 검정] – [그레인저 인과관계]를 선택하면 **[그레인저 인과관계]** 윈도우가 나타납니다. 창에서 종속 변수와 독립 변수를 선택합니다. 옵션 부분의 유의한 결과를 파일에 저장을 체크하면, 검정 결과인 p-value 가 사용자가 입력한 값 이하인 독립 변수의 리스트와 p-value 값을 텍스트 파일로 저장할 수 있습니다. 확인 버튼을 클릭하면 그레인저 인과관계가 수행됩니다.



결과



인과관계 검정 결과인 F-Statistics value 와 p-Value 를 제공합니다.

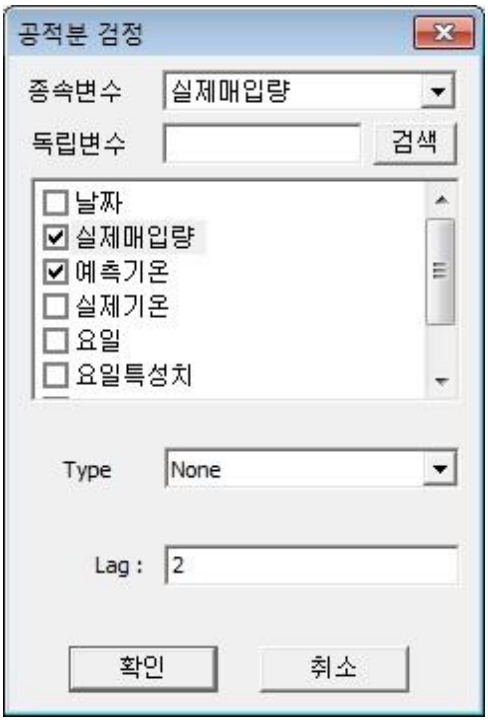
(3) 공적분 검정

개요

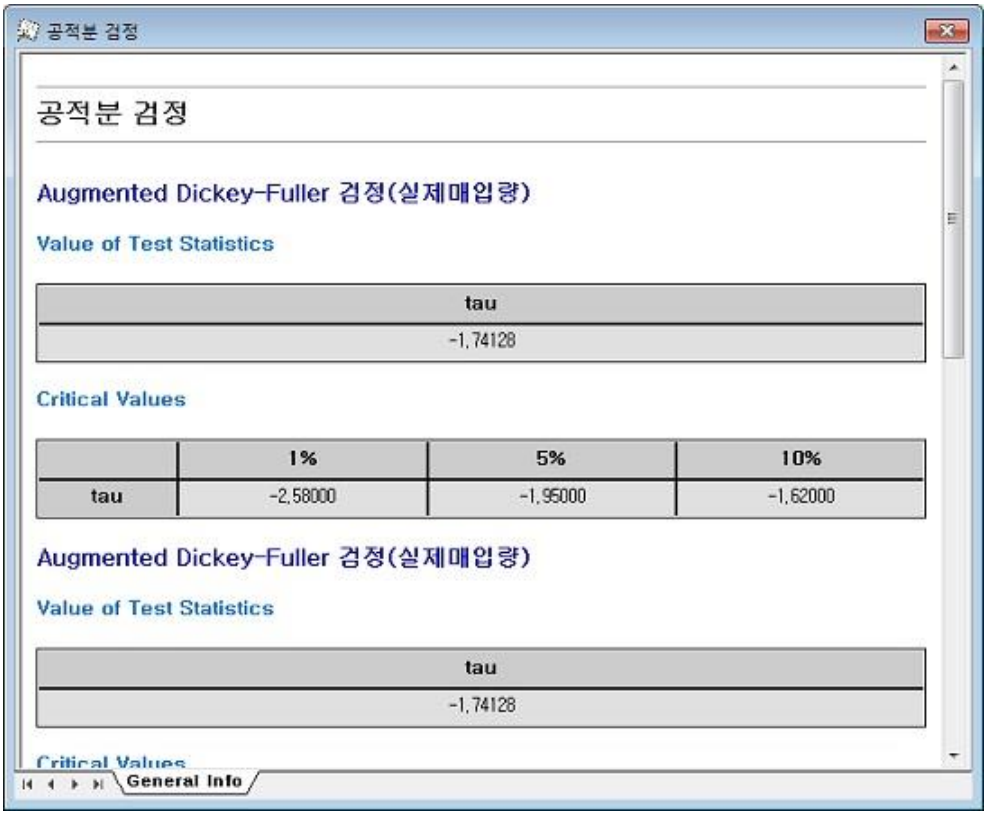
두 개 이상의 시계열이 individually integrated 하지만 그들의 linear combination 중 일부가 낮은 차수일 때, 이 시계열들이 cointegrated 되어있다고 표현합니다. 공적분 검정은 두 개 이상의 시계열이 얼마나 cointegrated 되어있는지를 검정합니다. The Engle-Granger two-step method, The johansen's procedure, Phillips-Ouliaris Cointegration Test 등의 방법이 있는데, 공적분 검정 메뉴에서는 The johansen's procedure 를 사용합니다.

실행 방법

[분석] – [시계열 분석] – [시계열 검정] – [공적분 검정]을 선택하면 **[공적분 검정]** 윈도우가 나타납니다. 창에서 종속 변수와 독립 변수를 선택합니다. Type 메뉴에서는 None, Constant, Trend 를 선택할 수 있습니다..



결과



Cointegration test 의 결과값인 test statistics 의 값을 제공합니다.

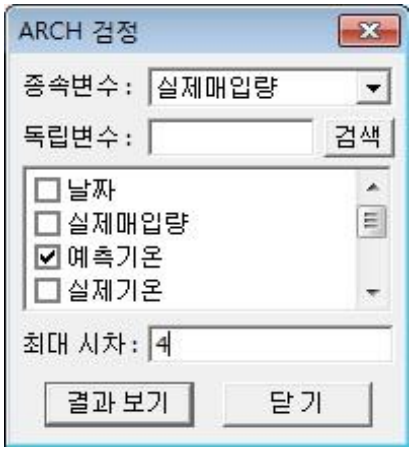
(4) Arch Test

개요

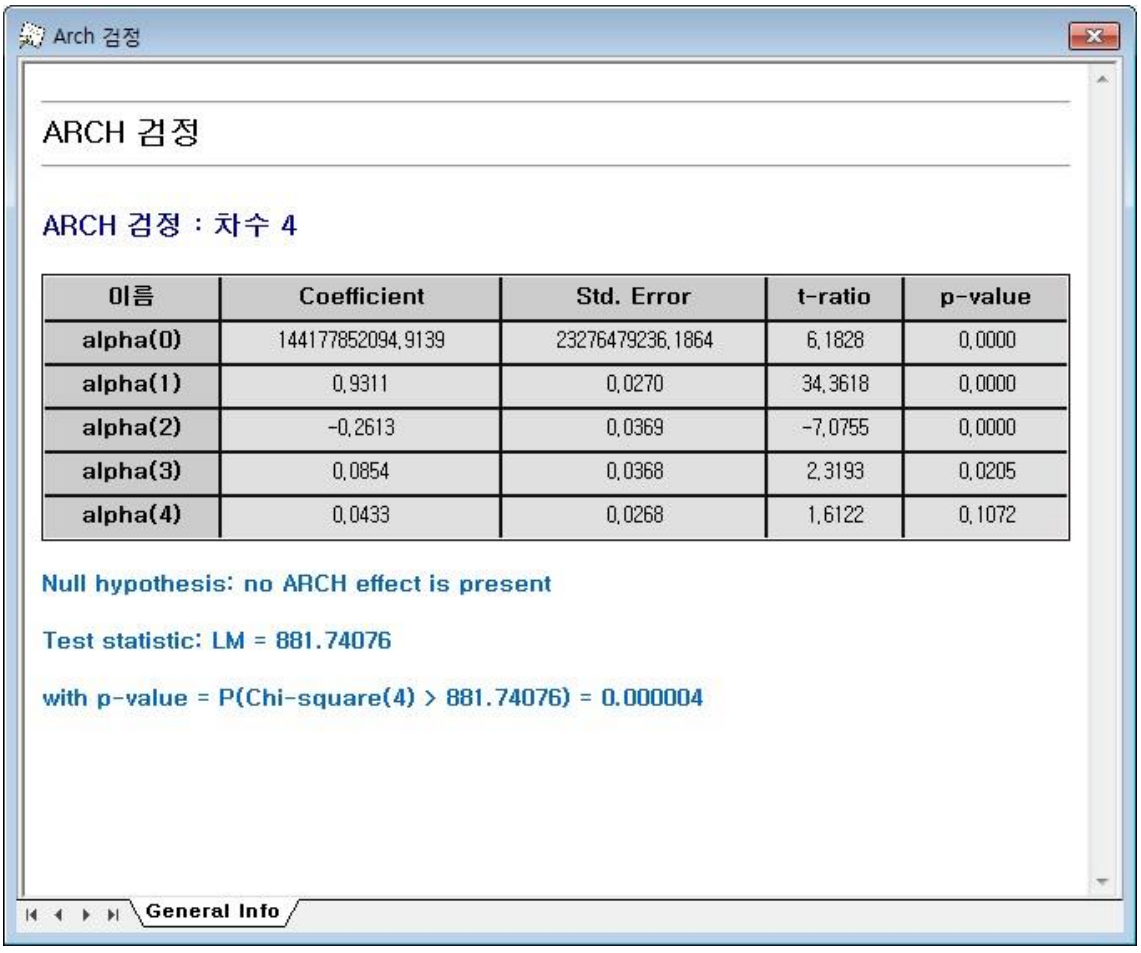
Arch 모델이란, 하나의 시계열이 시간에 따라 변하는 변동성을 예측하기 위한 이분산 조건부 자기 회귀 모형입니다. Arch 검정은 현재의 분산이 과거의 분산으로 예측이 가능한지 안 한지 판단하기 위해 사용합니다.

실행 방법

[분석] – [시계열 분석] – [시계열 검정] – [Arch Test]를 선택하면 **[Arch Test]** 윈도우가 나타납니다. 창에서 종속 변수와 독립 변수를 선택합니다. 최대 시차 값을 입력합니다. 결과 보기 버튼을 클릭하면 Arch Test 가 수행됩니다.



결과



Arch 검정의 Test statistic 이 제공됩니다.

- 5.3.6.3 시계열 상관성

(1) 교차상관

개요

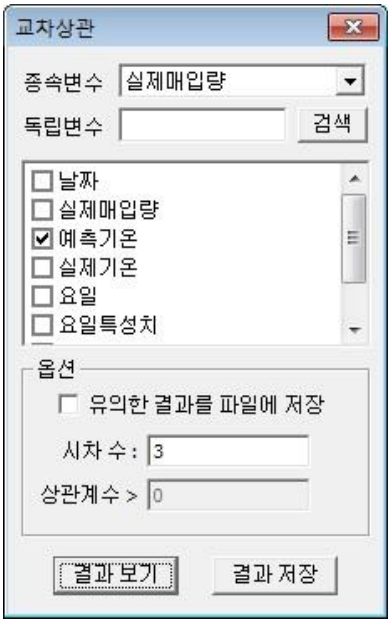
교차상관은 두 시계열간의 유사성을 측정하는 지표입니다. 교차상관은 주로 long-duration signal 에서 짧고 알려진 특징을 발견하기 위해 사용됩니다. 두 시계열 X, Y 가 존재할 때,

두 시계열 간의 k 시차 교차상관계수 r_k의 계산법은 다음과 같습니다.

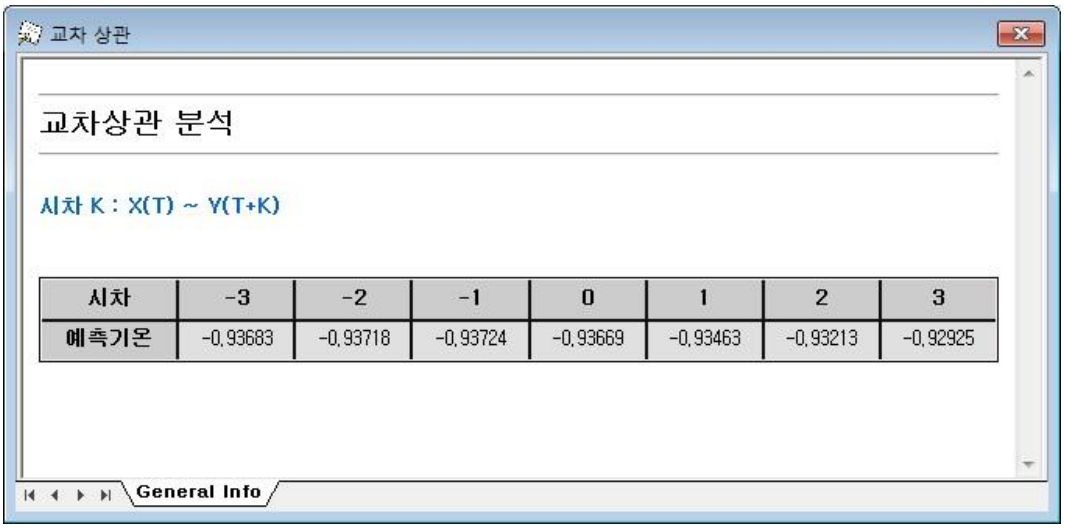
$$r_k = \frac{\sum(X_t - \bar{X})(Y_{t+k} - \bar{Y})}{\sqrt{\sum(X_t - \bar{X})^2 \sum(Y_t - \bar{Y})^2}}$$

실행 방법

[분석] – [시계열 분석] – [시계열 상관성] – [교차상관]을 선택하면 **[교차상관]** 윈도우가 나타납니다. 창에서 종속 변수와 독립 변수를 선택합니다. 옵션 부분의 유의한 결과를 파일에 저장을 체크하면, 검정 결과인 상관계수가 사용자가 입력한 값 이상인 독립 변수의 이름과 해당하는 차수, 상관계수 값을 텍스트 파일로 저장할 수 있습니다. 확인 버튼을 클릭하면 교차상관이 수행됩니다.



결과



해당 시차에서 독립변수와 종속변수 사이의 교차상관 계수를 제공합니다.

(2) 자기상관, 편자기상관

개요

자기 상관성은 현재의 시계열 데이터의 값이 과거의 데이터와 어떠한 상관관계를 갖는지를 나타내주는 척도입니다. 자기 상관성이 높으면 높을수록 자신의 과거 데이터를 통해서 현재의 자신의 값을 예측할 수 있는 가능성이 높아집니다.

편자기 상관성을 설명하기 위해서 다음과 같은 모델을 예로 들겠습니다.

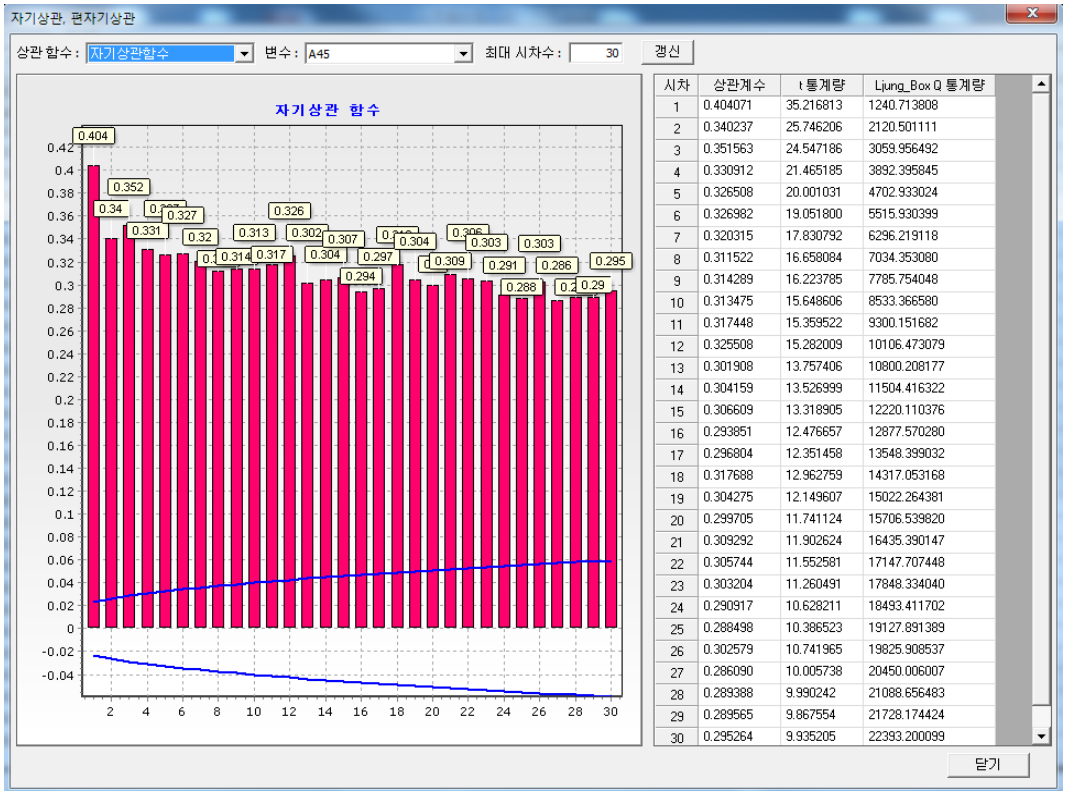
$$z_t = \phi_1 z_{t-1} + a_t \quad a_t \sim \text{iid } N(0, \sigma^2)$$

위의 시계열에서 2 차 자기 상관계수를 구해보면 유의한 값이 나오게 됩니다. 만약 그렇다면 분석자는 2 차에서 자기 상관계수가 유의한 것이 위의 시계열이 AR(2)에서 만들어졌기 때문, 혹은 AR(3)에서 만들어졌기 때문이라고 생각할 수 있습니다. 따라서 자기 상관성 외에 편자기 상관이라는 개념이 필요한 것입니다. 예를 들어 제 2 차 편자기 상관계수를 구한다고 하면 1 차까지의 영향을 제외한 후의 상관계수를 구합니다. 이렇게 할 경우 2 차 자기 상관계수의 값은 높게 나올 수 있지만 2 차 편자기 상관함수의 값은 작게 나오게 됩니다.

실행 방법

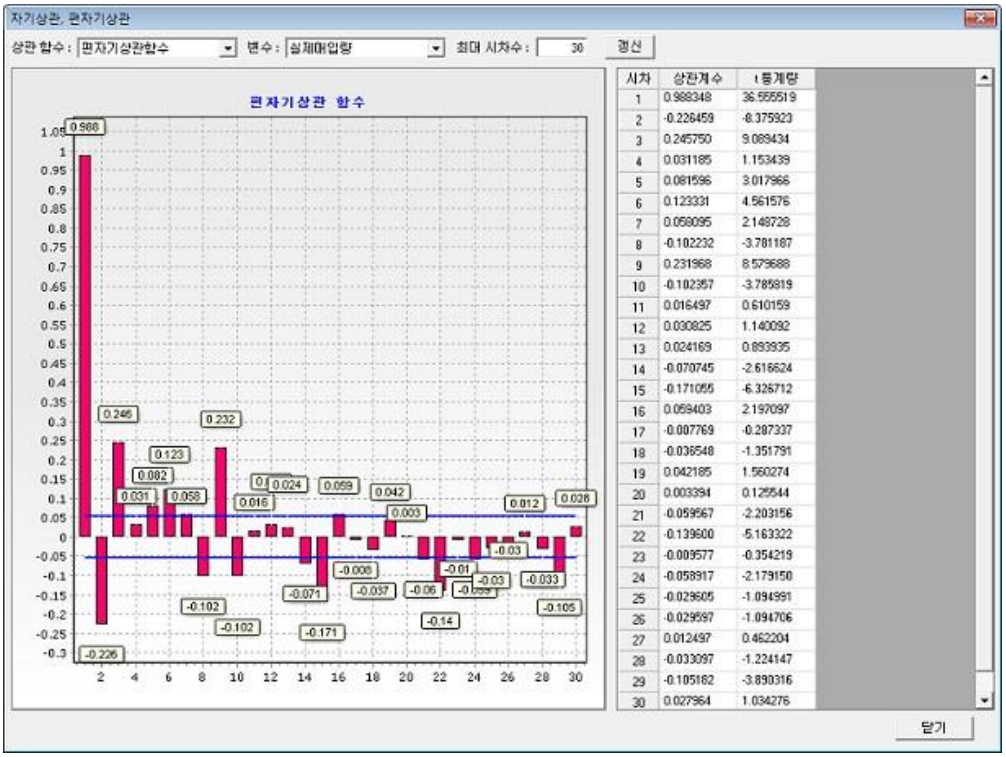
[분석] – [시계열 분석] – [시계열 상관성] – [자기상관, 편자기상관]을 선택하면 **[자기상관, 편자기상관]** 윈도우가 나타납니다. 나타나는 윈도우에서 상관함수의 종류를 선택합니다. 분석자는 자기 상관함수, 혹은 편자기 상관함수를 선택할 수 있고 이와 함께 분석하고자 하는 변수를 선택합니다. 분석자는 최대 시차수를 정해 줌으로써 어느 정도의 시차까지 상관함수를 구할지를 정할 수 있습니다.

결과



시차	상관계수	t 통계량	Ljung_Box Q 통계량
1	0.404071	35.216813	1240.713808
2	0.340237	25.746206	2120.501111
3	0.351563	24.547186	3059.956492
4	0.330912	21.465185	3892.395845
5	0.326508	20.001031	4702.933024
6	0.326982	19.051800	5515.930399
7	0.320315	17.830792	6296.219118
8	0.311522	16.658084	7034.353080
9	0.314289	16.223785	7785.754048
10	0.313475	15.648606	8533.366580
11	0.317448	15.359522	9300.151682
12	0.325508	15.282009	10106.473079
13	0.301908	13.757406	10800.208177
14	0.304159	13.526999	11504.416322
15	0.306609	13.318905	12220.110376
16	0.293851	12.476657	12877.570280
17	0.296804	12.351458	13548.399032
18	0.317688	12.962759	14317.053168
19	0.304275	12.149607	15022.264381
20	0.299705	11.741124	15706.539820
21	0.309292	11.902624	16435.390147
22	0.305744	11.552581	17147.707448
23	0.303204	11.260491	17848.334040
24	0.290917	10.628211	18493.411702
25	0.288498	10.386523	19127.891389
26	0.302579	10.741965	19825.908537
27	0.286090	10.005738	20450.006007
28	0.289388	9.990242	21088.656483
29	0.289565	9.867554	21728.174424
30	0.295264	9.935205	22393.200099

위는 자기 상관함수 예시입니다. 자기 상관함수에서는 상관계수, t 통계량 뿐 아니라 Ljung-Box Q 검정 통계량 값을 보여줍니다. Q 검정 통계량 값은 데이터가 독립적으로 분포되었는지 여부에 대한 가설 검정에 사용됩니다. 데이터가 독립적으로 분포한다는 의미는 데이터에서 발견되는 상관관계는 랜덤 샘플링에 의해 생겼음을 의미합니다.



위는 편자기 상관함수의 예시입니다.

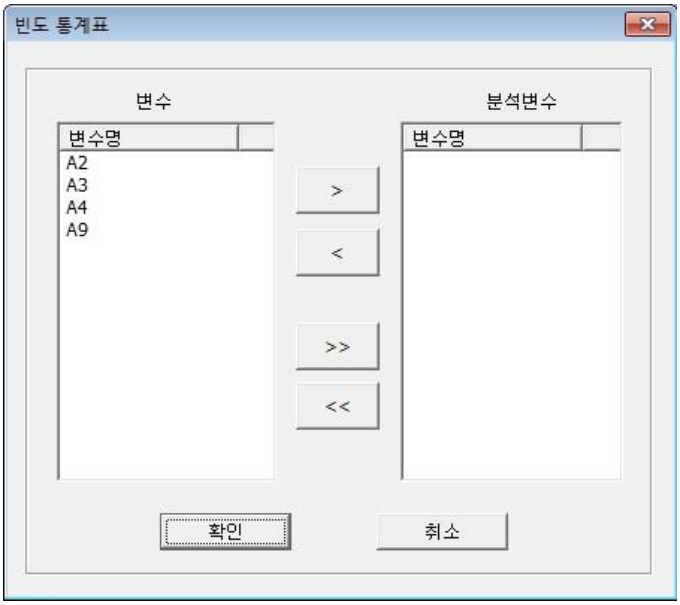
5.3.7 표

- **5.3.7.1 빈도표**

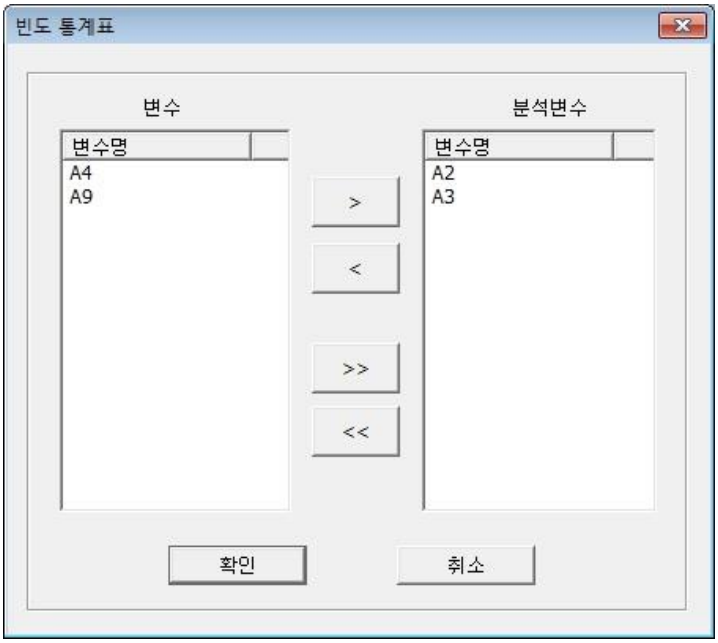
데이터 탐색기에서는 각각의 필드에 대해 **빈도 통계표** 기능을 사용하여, 한 변수의 값에 빈도가 몇 개 인지를 알 수 있습니다.

실행 방법

1. **[분석] –[표] – [빈도표]**를 선택하면, **빈도 통계표** 윈도우가 나타납니다.



2. '변수' 영역에서 분석하고자 하는 변수를 선택합니다.



결과

빈도 통계표

※ 빈도 통계표 (Class)

Class (연속형)	빈도	백분율	누적빈도	누적백분율
0	130	66.67%	130	66.67%
1	65	33.33%	195	100.00%
계	195	100%	195	100%

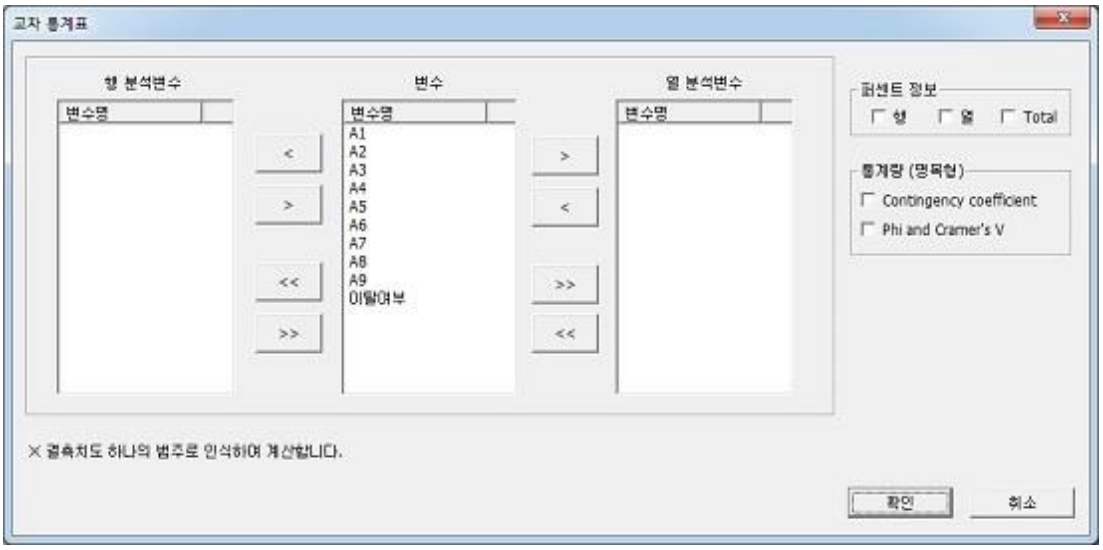
선택된 변수의 빈도 통계 정보가 나타납니다.

- 5.3.7.2 교차표

데이터 탐색기에서는 각각의 필드에 대해 **교차표** 기능을 사용하여, 두 변수의 값이 공유하고 있는 빈도수가 몇 개인지 구할 수 있습니다.

실행 방법

1. **[분석]- [표]- [교차표]**를 선택하면, 교차 통계표 윈도우가 나타납니다.

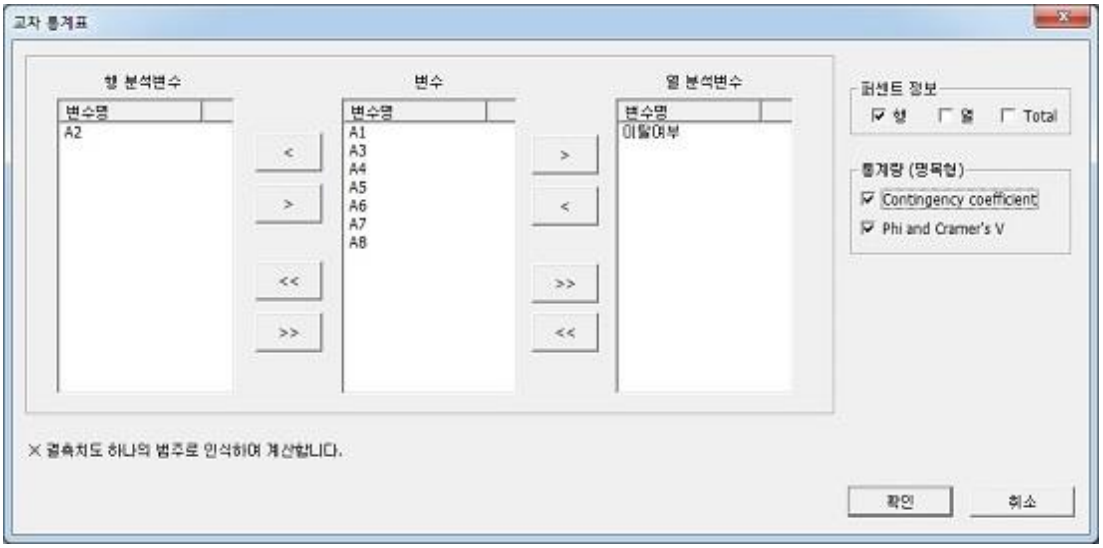


2. '변수' 영역에서 분석하고자 하는 변수를 행 단위, 열 단위로 선택합니다.

3. 오른쪽의 속성에서 '퍼센트 정보' 와 '통계량' 옵션을 선택할 수 있습니다.

'퍼센트 정보'는 해당 변수의 빈도가 행, 열로 선택된 변수 내에서 차지하는 비율을 표시하는 정보입니다. Total 일 경우 전체 데이터에 대한 빈도수를 비율로 표시합니다.

'통계량(명목형)'은 카이제곱을 기초하여 행과 열 변수 간의 연관성의 강도를 나타내는 척도로써, Contigency coefficient 와 Phi and Cramer's V 를 제공합니다. 두 통계량 모두 0~1 사이의 값으로 1 에 가까울수록 관련성이 큽니다.



결과

교차 통계 정보와 명목형 변수간의 연관성을 나타내는 통계량이 표 형태로 나타나게 됩니다.

※ 교차 통계표 (A2 vs. 이탈여부)

		0	1	계
B	개수 % in A2	113 48,50%	120 51,50%	233 100%
D	개수 % in A2	231 35,43%	421 64,57%	652 100%
E	개수 % in A2	91 40,44%	134 59,56%	225 100%
계	개수 % in A2	435 39,19%	675 60,81%	1110 100%

※ 명목형 통계량

		Value
명목형 - 명목형	Phi	0,10607
	Cramer's V	0,10607
	Contingency Coefficient	0,10548
전체 데이터 수		1110

* 범주 개수가 서로 같지 않을 때, 명목형 통계량으로 Cramers'V 통계량이 가장 적합합니다.

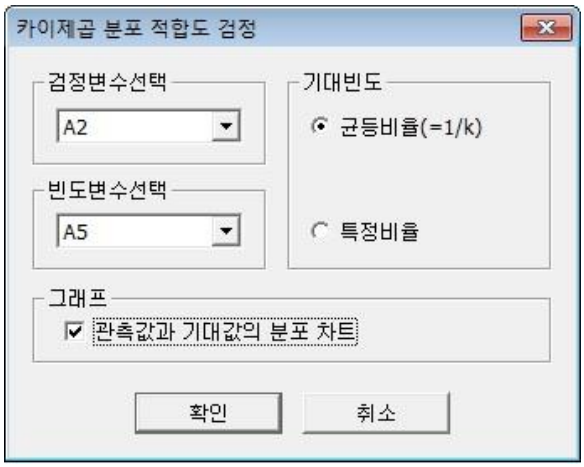
- **5.3.7.3 일변량 카이제곱 검정**

카이제곱 적합도 검정은 수집된 자료가 특정 비율을 갖는 다항분포를 따르는지 검정하는 기법입니다.

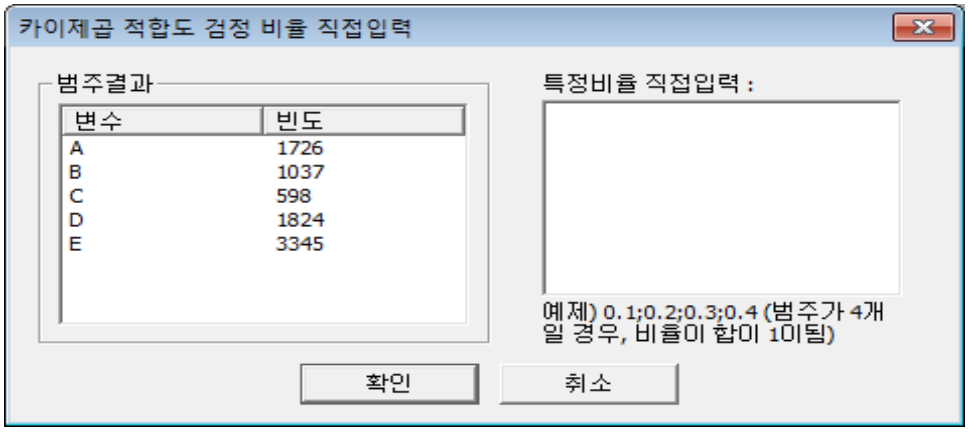
실행 방법

1. **[분석] – [표] – [일변량 카이제곱 검정]**을 선택하면, 카이제곱 적합도 검정 윈도우가 나타납니다.

2. 범주변수와 빈도변수를 선택하고 검정할 비율(균등비율 또는 특정비율)을 선택합니다. 추가적으로 관측값과 기대값의 분포 차트를 선택할 수 있습니다.



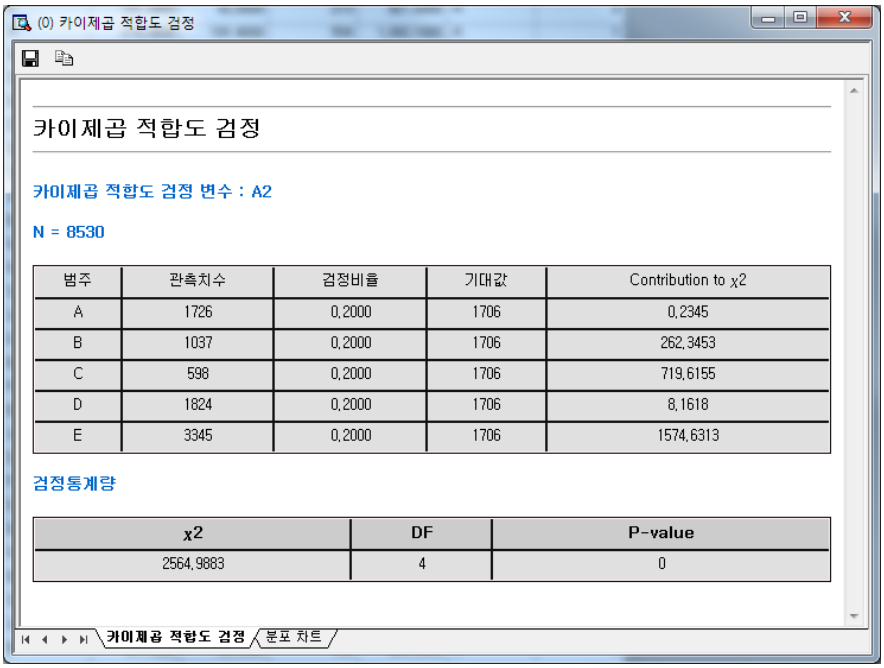
기대빈도 란에서 '특정비율' 선택 후 실행 시 범주별 빈도를 확인할 수 있으며, 특정 비율을 직접 입력할 수 있습니다. 특정 비율은 범주 수만큼 기입해야 하며, ';'을 이용하여 범주별 비율을 구분합니다.



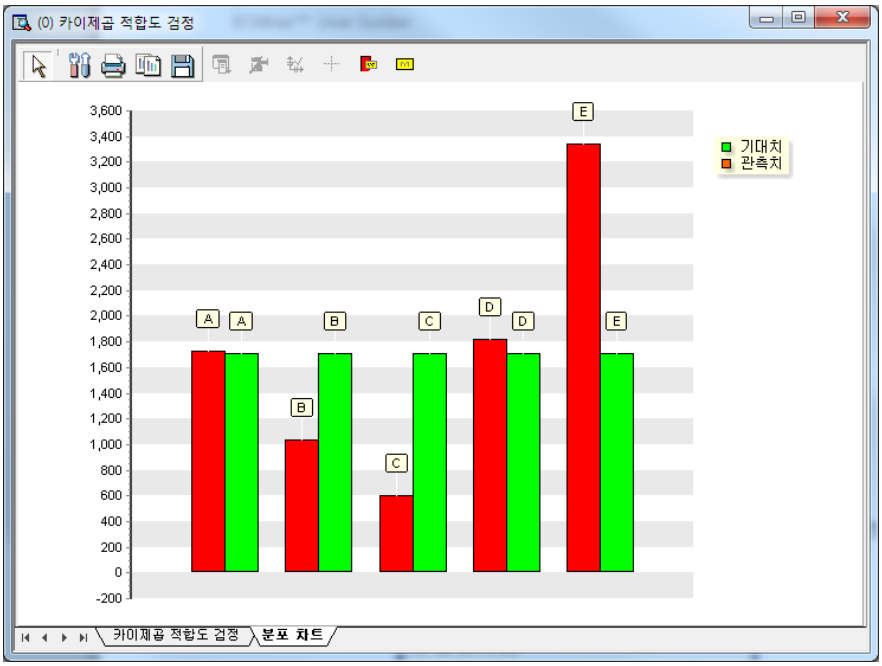
결과

카이제곱 적합도 검정 결과가 다음과 같이 나타납니다.

각 범주별 검정비율, 기대값, 그리고 카이제곱 검정 통계량에 대한 contribution 이 출력되며, p-value 를 통해 기대빈도에 대한 자료의 적합도를 검정할 수 있습니다.



결과창의 분포차트 탭을 선택하면 다음과 같이 관측값과 기대값의 분포차트를 확인할 수 있습니다..



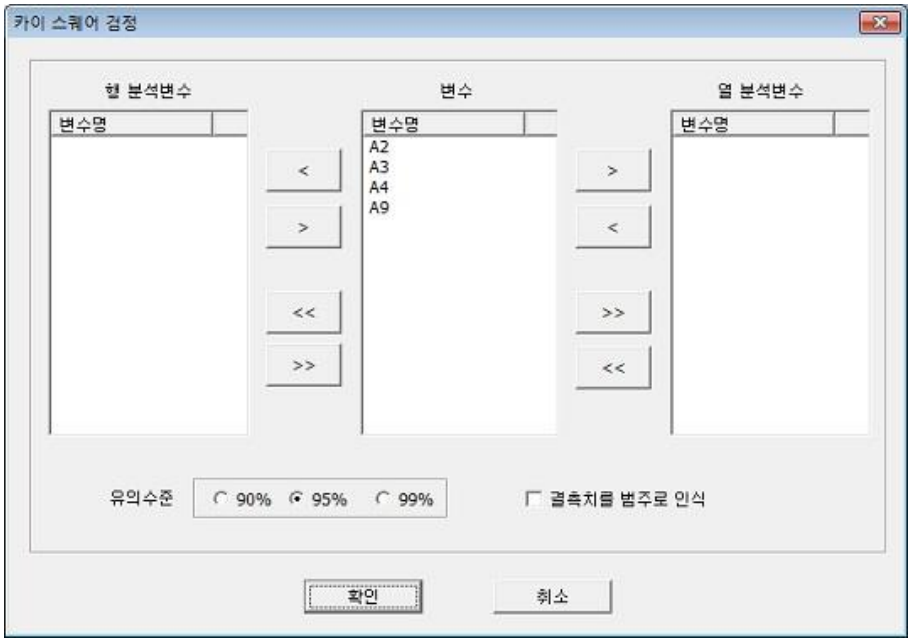
- 5.3.7.4 독립성 검정

데이터 탐색기에서는 각각의 필드에 대해 독립성 검정 - 카이 스퀘어 검정 기능을 사용하여, 두 요인간의 관계유무를 검정할 수 있습니다.

실행 방법

1. **[분석] – [표] – [독립성 검정]**를 선택하면, 카이 스퀘어 검정 윈도우가 나타납니다.

2. 필드명 콤보 박스에서 열 단위, 행 단위로 분석을 원하는 필드를 선택하고 유의 수준을 선택합니다.



결과

열, 행 단위 필드들의 교차표와 통계 정보를 보여줍니다.

카이제곱 검정

※ 교차 통계표 (이탈여부 vs. A9)

		A	B	C	D	계
0	개수 % in 이탈여부 % in A9 % in Total	3780 98,28% 44,68% 44,31%	35 0,91% 94,59% 0,41%	7 0,18% 100,00% 0,08%	24 0,62% 92,31% 0,28%	3846 100% 45,09%
1	개수 % in 이탈여부 % in A9 % in Total	4680 99,91% 55,32% 54,87%	2 0,04% 5,41% 0,02%	0 0,00% 0,00% 0,00%	2 0,04% 7,69% 0,02%	4684 100% 54,91%
계	개수 % in 이탈여부 % in A9 % in Total	8460 100% 99,18%	37 100% 0,43%	7 100% 0,08%	26 100% 0,30%	8530 100%

※ 통계량 (이탈여부 vs. A9)

통계량	값	자유도	유의확률(p-value)
Pearson Chi-Square	69,13335	3	< 0,0001
Likelihood Chi-Square	80,84457	3	< 0,0001

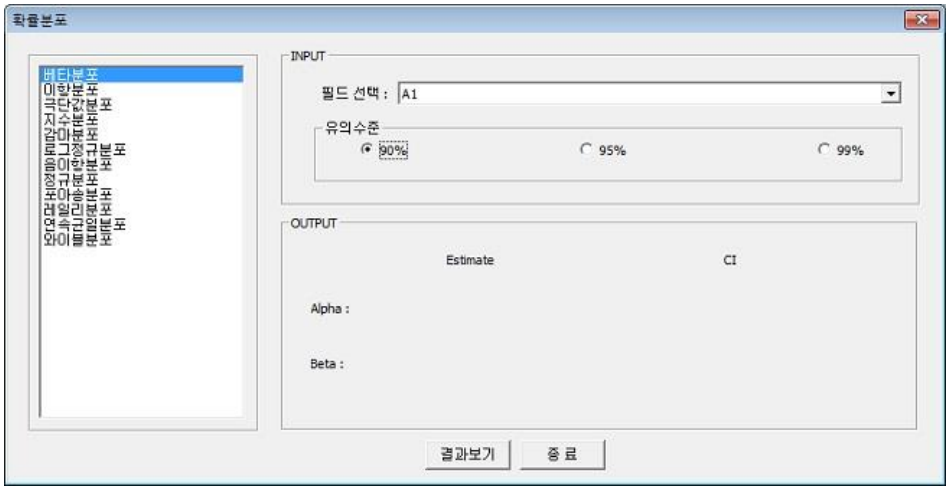
5.3.8 확률분포

- ### 5.3.8.1 모수추정

확률 분포의 모수를 추정하는 방법에는 Method of Moment, Maximum Likelihood 등 여러 가지 방법이 있습니다. 하지만 가장 많이 사용하는 방법은 Maximum Likelihood 인데 그 이유는 Maximum Likelihood Estimator 는 모수의 Asymptotic Normality 와 같은 좋은 특징을 갖는 Estimator 이기 때문입니다. ECMiner™ 의 확률 분포 모수 추정은 모두 데이터가 주어져 있을 때 그 데이터에 가장 잘 맞는 모수를 Maximum Likelihood 방법으로 찾아 줍니다. 그리고 이와 함께 추정된 모수의 신뢰 구간을 구해 주어 추정된 모수를 어느 정도 신뢰할 수 있는지를 제시해줍니다.

실행방법

[분석] – [확률분포] – [모수 추정]을 선택하면 **[확률 분포]**윈도우가 나타납니다.



위 화면에서 선택할 수 있는 분포는 12 가지 입니다. 사용자는 데이터를 적합하고자 하는 분포를 선택하고 데이터가 어떤 필드에 있는지 그리고 어떠한 유의 수준에서 추정을 하고 싶은지를 선택하고 결과 보기를 클릭하면 됩니다.

(1) 베타 분포

추정 방법

베타 분포의 pdf 는 다음과 같습니다.

$$f_X(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{(0,1)}(x)$$

이에 대한 Likelihood Function 은 다음과 같습니다.

$$L(\alpha,\beta\,|X) = \prod_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_i^{\alpha-1}(1-x_i)^{\beta-1}$$

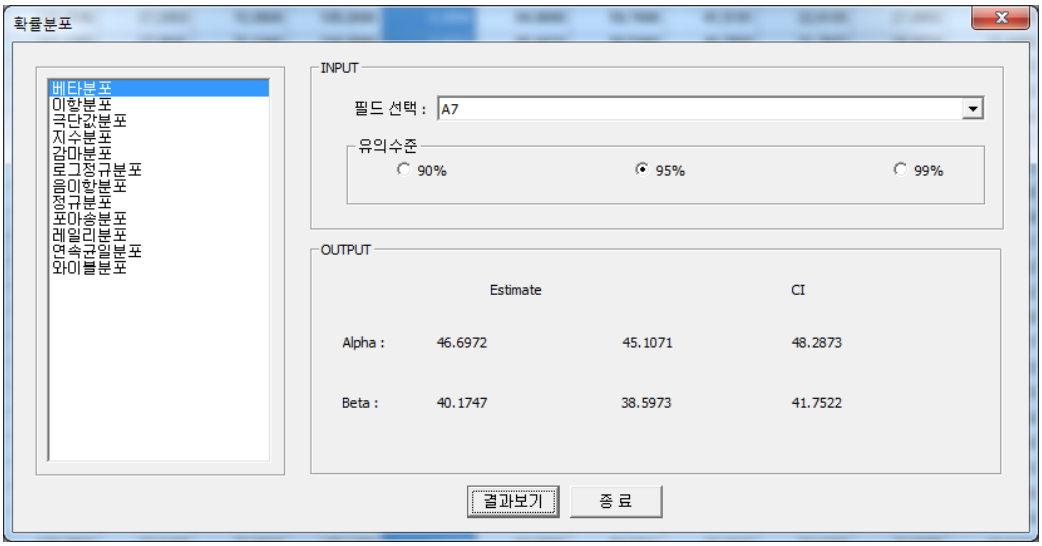
이를 최대화하는 α,β 가 Maximum Likelihood Estimator 입니다. 위의 식을 좀더 간단히 하기 위해서 로그를 취하면

$$\ln L(\alpha,\beta|X) = n\ln\Gamma(\alpha+\beta) - n\ln\Gamma(\alpha) - n\ln\Gamma(\beta) + (\alpha-1)\sum_{i=1}^{n}\ln(x_i) + (\beta-1)\sum_{i=1}^{n}\ln(1-x_i)$$

이 됩니다. 위의 식을 최대화 하는 방법은 여러 가지가 있습니다. 위의 식을 direct 하게 최대화하는 경우와 위의 식을 변형하여 푸는 것입니다. 일단 위의 식을 direct 하게 최대화하는 것은 Nelder and Mead 의 Simplex Method 를 사용합니다. 이와 함께 Newton Raphson 의 방법을 쓸 수 있습니다. 이 두 방법 중에서 ECMiner™는 Nelder and Mead 의 Simplex 방법으로 Likelihood Function 을 최대화 합니다.

주의: 베타 분포에 사용되는 모든 데이터는 0 과 1 사이의 값이어야 합니다.

예시



(2) 이항 분포

추정 방법

베르누이 분포의 함수는 정의역으로 0 혹은 1 을 갖습니다. 그것의 pdf 는 다음과 같습니다.

$$f_X(x;p) = p^x(1-p)^{1-x} \quad (x = 0,1)$$

이 때 Likelihood Function 을 잡으면 다음과 같습니다.

$$L(p) = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

그리고 위의 식에 Log 를 취하여 Log Likelihood 를 구하면 다음과 같습니다.

$$\ln L(p) = \left(\sum_{i=1}^{n} x_i\right)\ln(p) + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-p)$$

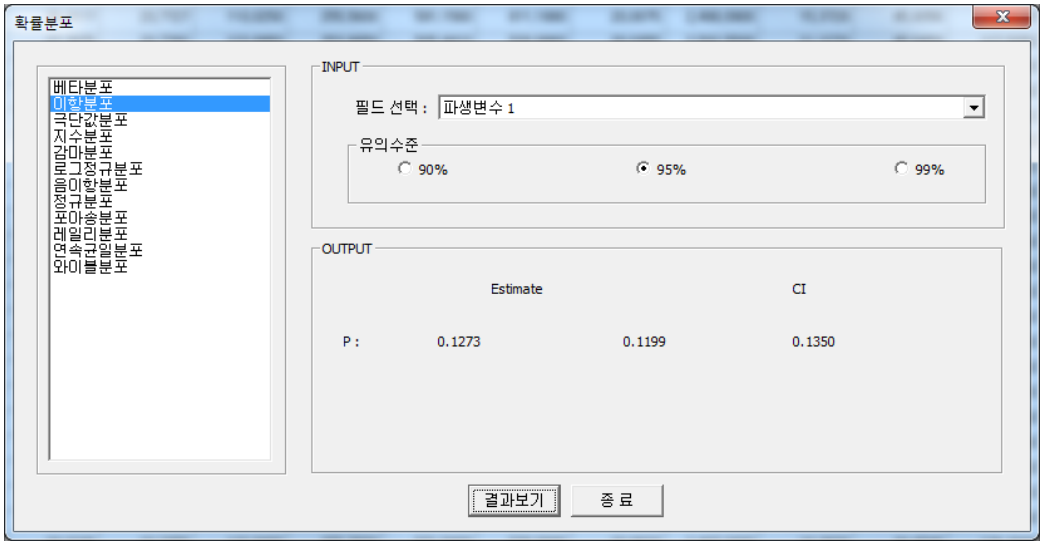
이를 p 에 대해서 미분하고 0 이 되는 것을 찾으면

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

이 됩니다. 즉 X 의 평균이 바로 Maximum Likelihood Estimator 가 되는 것입니다.

주의: 이항 분포의 Estimation 에 사용되는 데이터는 모두 0 혹은 1 의 값을 가져야 합니다.

예시



(3) 극단값 분포

추정 방법

극단 값 분포의 pdf 는 다음과 같습니다.

$$f(x|\mu,\sigma) = \frac{1}{\sigma}\exp\left(\frac{x-\mu}{\sigma}\right)\exp\left(-exp\left(\frac{x-\mu}{\sigma}\right)\right)$$

이 때 Log Likelihood Function 은 다음과 같은 Form 을 갖게 됩니다.

$$l(\mu,\sigma) = \sum_{i=1}^{n} \ln\left[\frac{1}{\sigma}\exp\left(\frac{x_i-\mu}{\sigma}\right)\exp\left(-\exp\left(\frac{x_i-\mu}{\sigma}\right)\right)\right.$$

$$= -n\ln(\sigma) + \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right) - \sum_{i=1}^{n}\exp\left(\frac{x_i-\mu}{\sigma}\right)$$

각 Parameter 에 대해서 편미분을 취하여 0 으로 두면 다음과 같이 됩니다.

$$\frac{\partial l(\mu,\sigma)}{\partial\mu} = -\frac{n}{\sigma} + \frac{1}{\sigma}\sum_{i=1}^{n}\exp\left(\frac{x_i-\mu}{\sigma}\right) = 0$$

$$\frac{\partial l(\mu,\sigma)}{\partial\sigma} = -\frac{n}{\sigma} - \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) + \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma^2}\right)\exp\left(\frac{x_i-\mu}{\sigma}\right) = 0$$

첫번째 식을 정리하면 다음과 같습니다.

$$\mu = \sigma\ln\left[\frac{1}{n}\sum_{i=1}^{n}exp\left(\frac{x_i}{\sigma}\right)\right]$$

그리고 이 결과를 두 번째 식에 대입하여 정리하면 다음과 같이 됩니다.

$$\sigma + \sum_{i=1}^{n} \frac{x_i}{n} - \frac{\sum_{i=1}^{n} x_i \exp\left(\frac{x_i}{\sigma}\right)}{\sum_{i=1}^{n} \exp\left(\frac{x_i}{\sigma}\right)} = 0$$

따라서 전략은 σ 에 대한 Equation 을 먼저 풀고 이를 μ 에 대한 식에 넣는 것입니다. σ 에 대한 Equation 은 Newton Raphson 의 방법을 사용하여 해결합니다. 이를 위해서 다음과 같은 함수를 정의합니다.

$$f(\sigma) = \sigma + \sum_{i=1}^{n} \frac{x_i}{n} - \frac{\sum_{i=1}^{n} x_i \exp\left(\frac{x_i}{\sigma}\right)}{\sum_{i=1}^{n} \exp\left(\frac{x_i}{\sigma}\right)}$$

$$f'(\sigma) = 1 + \frac{\left(\sum_{i=1}^{n} x_i^2 \exp\left(\frac{x_i}{\sigma}\right)\right)\left(\sum_{i=1}^{n} \exp\left(\frac{x_i}{\sigma}\right)\right) - \left(\sum_{i=1}^{n} x_i \exp\left(\frac{x_i}{\sigma}\right)\right)^2}{\sigma^2 \left(\sum_{i=1}^{n} \exp\left(\frac{x_i}{\sigma}\right)\right)^2}$$

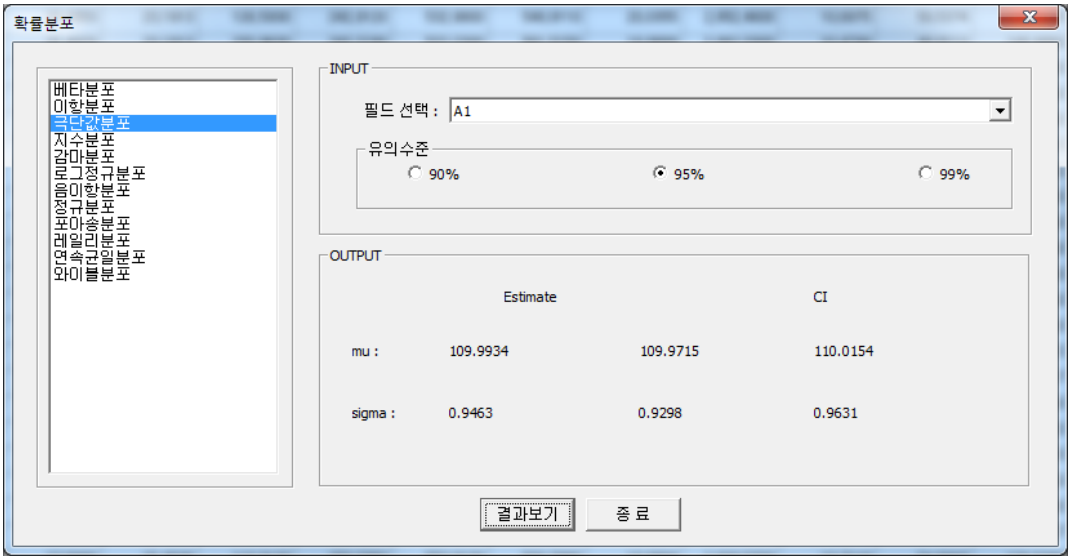
이와 같은 상황에서 Newton Raphson 의 식을 도출하면

$$\sigma_{k+1} = \sigma_k - \frac{f(\sigma_k)}{f'(\sigma_k)}$$

입니다. 이렇게 σ 에 대한 MLE 를 구하고 이를 이용하여 μ 에 대한 MLE 또한 구해 줍니다.

예시

▫



(4) 지수 분포

추정 방법

Maximum Likelihood 방법으로 Estimation 합니다. 먼저 Likelihood Function 을 잡으면

$$L(\lambda) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) = \lambda^n \exp(-\lambda n \bar{x})$$

$$\text{where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

위의 본 Likelihood function 을 최대화하는 것보다 Log Likelihood 를 최대화하는 것이 더 편한 방법입니다.

$$\frac{d}{d\lambda}\ln(L(\lambda)) = \frac{d}{d\lambda}(n\ln(\lambda) - \lambda n \bar{x}) = \frac{n}{\lambda} - n\bar{x} \begin{cases} > 0 & if\ 0 < \lambda < 1/\bar{x} \\ = 0 & if\ \lambda = 1/\bar{x} \\ < 0 & if\ \lambda > 1/\bar{x} \end{cases}$$

위를 통해서 볼 때 의 maximum likelihood estimate 은 다음과 같습니다.

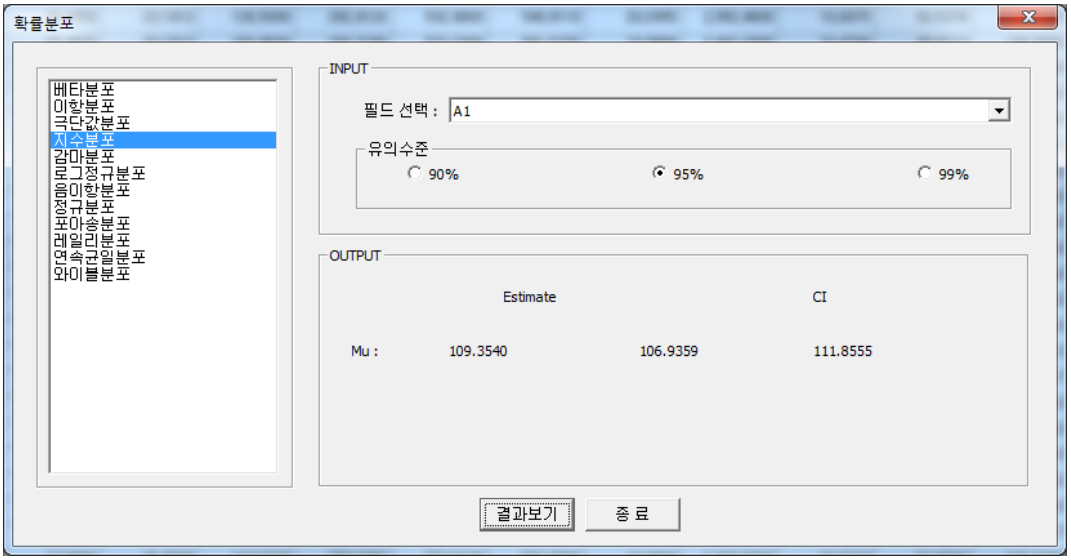
$$\hat{\lambda} = \frac{1}{\bar{x}}$$

보통 Maximum Likelihood estimation 에서는 sample 의 수가 많을 때 MLE 의 Asymptotic Property 를 이용하여 parameter 의 confidence interval 을 구하는데 특수한 경우 finite sample 에 대해서 exact confidence interval 을 구할 수 있습니다. 그 결과는 다음과 같습니다.

$$\frac{1}{\hat{\lambda}}\frac{2n}{\chi^2_{2n;\frac{\alpha}{2}}} < \frac{1}{\lambda} < \frac{1}{\hat{\lambda}}\frac{2n}{\chi^2_{2n;1-\frac{\alpha}{2}}}$$

주의: 지수 분포 모수 추정에 사용되는 모든 데이터의 값은 0 보다 커야 합니다.

예시



(5) 감마 분포

추정 방법

감마 분포의 pdf 는 다음과 같습니다.

$$f(x; k, \theta) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k \Gamma(k)} \quad \text{for } x > 0 \text{ and } k, \theta > 0.$$

데이터가 있을 때 이를 감마분포에 적합하기 위해서 MLE 구한다고 할 때 Likelihood Function 을 잡으면 다음과 같습니다.

$$L(k, \theta) = \prod_{i=1}^{n} f(x_i; k, \theta)$$

이를 단번에 최대화하는 것은 힘들기 때문에 Log 를 취하여 Log Likelihood Function 을 잡으면 다음과 같습니다.

$$l(k,\theta) = (k-1)\sum_{i=1}^{n} \ln(x_i) - \sum_{i=1}^{n} \frac{x_i}{\theta} - nk\ln(\theta) - n\ln(\Gamma(k))$$

이를 최대화 하기 위해서 먼저 θ 에 대해서 편미분하여 0 으로 두면 다음과 같은 식을 얻을 수 있습니다.

$$\hat{\theta} = \frac{1}{kn}\sum_{i=1}^{n} x_i$$

이를 위의 식에 대입하면 함수는 1 변수 함수가 됩니다.

$$l(k) = (k-1)\sum_{i=1}^{n} \ln(x_i) - nk - nk\ln\left(\frac{1}{kn}\sum_{i=1}^{n} x_i\right) - n\ln(\Gamma(k))$$

이제 이를 최대화하면 되는 것인데 이를 위해서 derivative 를 구하면 다음과 같습니다.

$$\frac{dl(k)}{dk} = \sum_{i=1}^{n} \ln(x_i) - n - n\ln\left(\frac{1}{kn}\sum_{i=1}^{n} x_i\right) - nk\,\frac{-\frac{1}{k^2 n}\sum_{i=1}^{n} x_i}{\frac{1}{kn}\sum_{i=1}^{n} x_i} - n\psi(k)$$

$$= \sum_{i=1}^{n} \ln(x_i) - n\ln\left(\frac{1}{kn}\sum_{i=1}^{n} x_i\right) - n\psi(k) = \sum_{i=1}^{n} \ln(x_i) + n\ln(k) - n\ln\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - n\psi(k)$$

Second derivative 를 구하면

$$\frac{d^2 l(k)}{dk^2} = \frac{n}{k} - n\psi'(k)$$

여기서 Newton Raphson 의 방법을 사용하면

$$k \leftarrow k + \frac{l'(k)}{l''(k)} = k + \frac{\sum_{i=1}^{n} \ln(x_i) + n\ln(k) - n\ln\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - n\psi(k)}{\frac{n}{k} - n\psi'(k)}$$

이 때 s 라는 값을 다음과 같이 정의하면

$$s = \ln\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - \frac{1}{n}\sum_{i=1}^{n} \ln(x_i)$$

Newton Raphson 의 식은 다음과 같이 다시 쓸 수 있습니다.

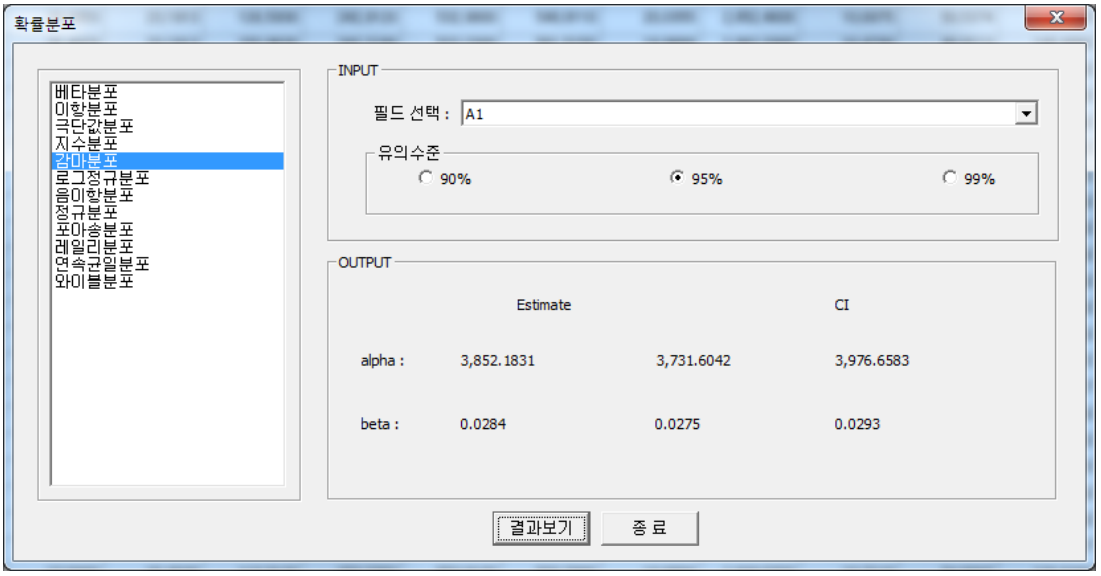
$$k \leftarrow k - \frac{\ln(k) - \psi(k) - s}{\frac{1}{k} - \psi'(k)}$$

이 때 알고리즘을 시작하기 위한 초기값은 다음과 같습니다.

$$k \approx \frac{3 - s + \sqrt{(s-3)^2 + 24s}}{12s}$$

주의: 감마 분포 모수 추정에 사용되는 모든 데이터는 0 보다 커야 합니다.

예시



(6) 로그 정규 분포

추정 방법

로그 정규 분포의 확률 변수 X 는 다음과 같습니다.

$$Y \sim \log(X) \quad Y \; is \; normal$$

즉 로그를 취하였을 때 normal random variable 을 만드는 것이 바로 로그 정규 분포의 확률 변수 입니다. pdf 는 다음과 같습니다.

$$f_X(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(ln(x)-\mu)^2}{2\sigma^2}\right), x > 0$$

이로부터 Log Likelihood Function 을 구하면

$$\ln\left(L(\mu,\sigma^2)\right) = -\sum_{i=1}^{n}\ln(x_i) - n\ln\left(\sigma\sqrt{2\pi}\right) - \sum_{i=1}^{n}\frac{(\ln(x_i)-\mu)^2}{2\sigma^2}$$

각 parameter 에 대해서 편미분을 취하여 위의 Log Likelihood Function 을 최대화하는 Parameter 조합을 구하면 다음과 같습니다.

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln(x_i)\ ,\quad \hat{\sigma}^2 = s^2 = \frac{1}{n}\sum_{i=1}^{n}(\ln(x_i)-\hat{\mu})^2$$

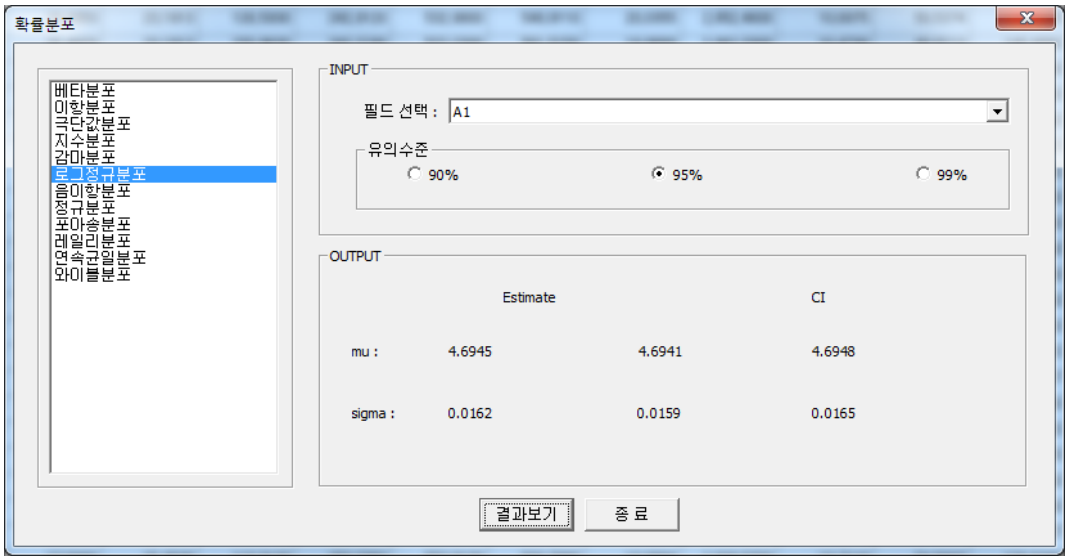
이에 대한 신뢰 구간(confidence interval)은

$$\hat{\mu} - \frac{1}{\sqrt{n}}\left|z_{\frac{\alpha}{2}}\right| s < \mu < \hat{\mu} + \frac{1}{\sqrt{n}}\left|z_{\frac{\alpha}{2}}\right| s$$

$$\hat{\sigma}\sqrt{\frac{n-1}{\chi^2_{1-\frac{\alpha}{2},n-1}}} < \sigma < \hat{\sigma}\sqrt{\frac{n-1}{\chi^2_{\frac{\alpha}{2},n-1}}}$$

입니다. 신뢰 구간에 정규 분포와 카이 제곱 분포의 quantile 이 나오는 이유는 log(x)가 정규 분포를 따르기 때문입니다.

예시



(7) 음이항 분포

추정 방법

음이항 분포의 pmf 는 다음과 같습니다.

$$f(x) = \frac{\Gamma(x+r)}{x!\,\Gamma(r)} p^r (1-p)^x \quad x = 0,1,2,\ldots$$

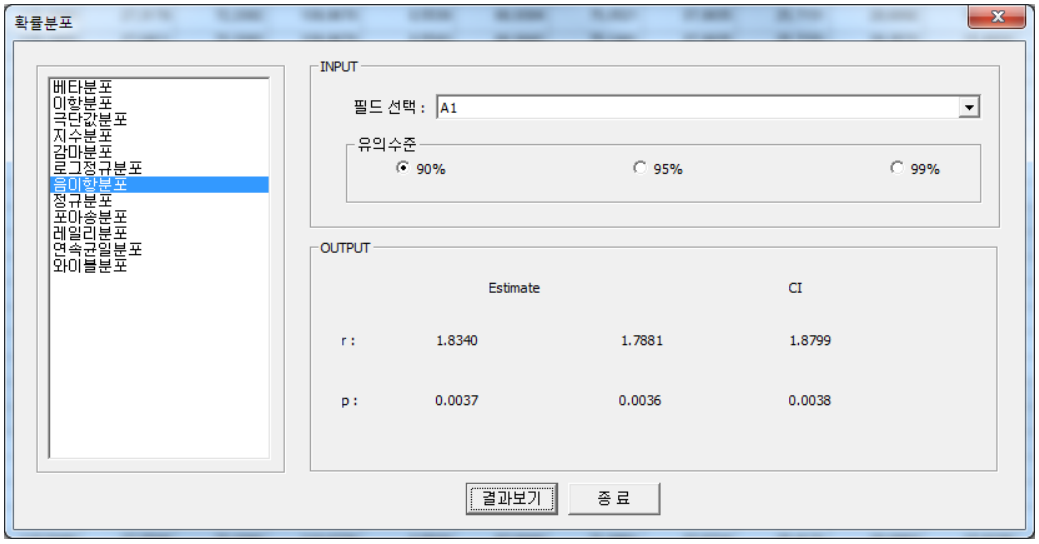
그리고 이를 통해서 만든 Log Likelihood Function 은 다음과 같습니다.

$$l(p,r) = \sum_{i=1}^{n} \ln\left(\Gamma(x_i + r)\right) - n\ln\left(\Gamma(r)\right) + \left(\sum_{i=1}^{n} x_i\right)\ln(1-p) + nr\ln(p)$$

위의 함수를 최대화 하는 Parameter 를 찾기 위해서 Nelder and Mead 의 Simplex 알고리즘을 사용합니다.

주의: 모든 데이터는 0 이상의 양의 정수이어야 합니다.

예시



(8) 정규 분포

추정 방법

정규 분포의 pdf 는 다음과 같습니다.

$$f(x;\ \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

이를 이용하여 Log Likelihood function 을 잡으면

$$\ln\left(L(\mu, \sigma^2)\right) = \sum_{i=1}^{n} \ln\left(f(x_i; \mu, \sigma^2)\right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2} ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

입니다. 위와 같은 함수에 각 Parameter 에 대해서 편미분을 취하여 MLE 를 찾으면 다음과 같습니다.

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \, , \hat{\sigma}^2 = s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

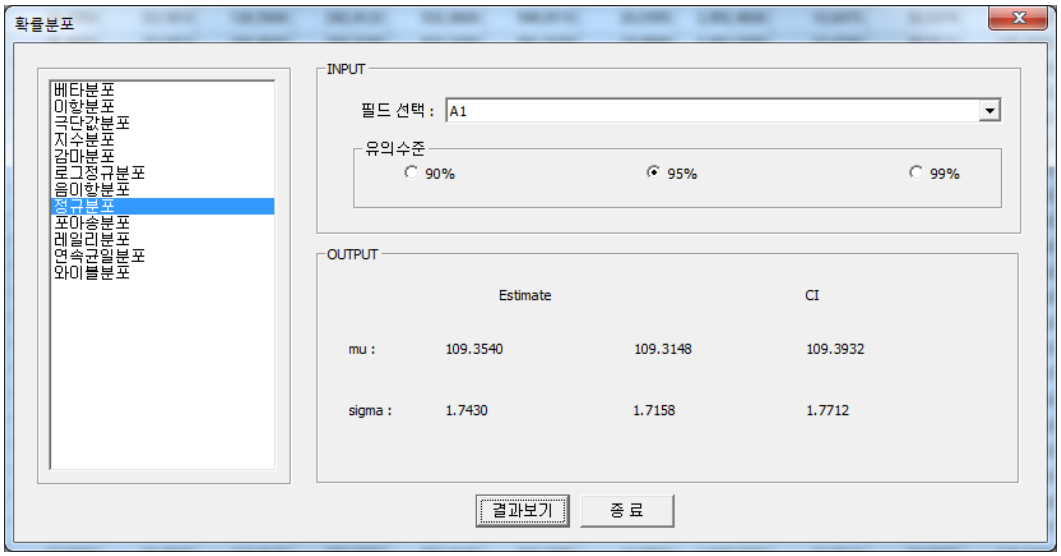
이에 대한 Confidence Interval 은 다음과 같습니다.

$$\hat{\mu} - \left|z_{\frac{\alpha}{2}}\right|\frac{1}{\sqrt{n}}s < \mu < \hat{\mu} + \left|z_{\frac{\alpha}{2}}\right|\frac{1}{\sqrt{n}}s$$

$$s^2 - \left|z_{\frac{\alpha}{2}}\right|\frac{\sqrt{2}}{\sqrt{n}}s^2 < \sigma^2 < s^2 + \left|z_{\frac{\alpha}{2}}\right|\frac{\sqrt{2}}{\sqrt{n}}s^2$$

여기서 z_α 는 정규 분포의 α quantile 입니다.

예시



(9) 포아송 분포

추정 방법

포아송 분포의 pdf 는 다음과 같습니다.

$$f(n;\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$

이 때 Maximum Likelihood function 은 다음과 같습니다.

$$L(\lambda) = \ln\left(\prod_{i=1}^{n} f(k_i|\lambda)\right) = \sum_{i=1}^{n} \ln\left(\frac{e^{-\lambda}\lambda^{k_i}}{k_i!}\right)$$

$$= -n\lambda + \left(\sum_{i=1}^{n} k_i\right)\ln(\lambda) - \sum_{i=1}^{n} \ln(k_i!)$$

$$\frac{d}{d\lambda}L(\lambda) = 0 \leftrightarrow -n + \frac{1}{\lambda}\left(\sum_{i=1}^{n} k_i\right) = 0$$

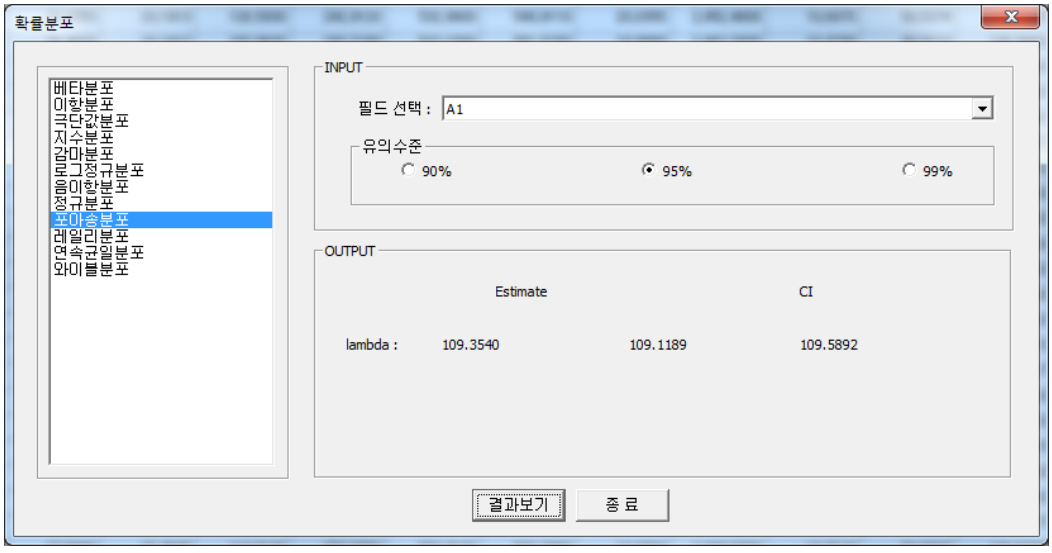
이를 풀면 다음과 같은 추정 값을 얻을 수 있습니다.

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} k_i$$

이고 이에 대한 confidence interval 은 다음과 같습니다.

$$\frac{\chi^2_{\frac{\alpha}{2},2\sum_{i=1}^{n} k_i}}{2n} < \lambda < \frac{\chi^2_{1-\frac{\alpha}{2},2(\sum_{i=1}^{n} k_i+1)}}{2n}$$

예시



(10) 레일리 분포

추정 방법

레일리 분포의 pdf 는 다음과 같습니다.

$$f(x;\sigma) = \frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right)$$

이를 가지고 Likelihood function 을 구하면

$$L(\sigma) = \frac{x_1 x_2 \cdots x_n}{\sigma^{2n}}\exp\left(-\frac{\sum_{i=1}^{n} x_i^2}{2\sigma^2}\right)$$

여기에 로그를 취해서 log likelihood function 을 구하면

$$\ln L(\sigma) = \sum_{i=1}^{n} \ln x_i - n\ln(\sigma^2) - \frac{\sum_{i=1}^{n} x_i^2}{2\sigma^2}$$

이 됩니다. 위 식을 σ 에 대해서 미분하고 미분된 함수가 0 이 되도록 σ 를 구하면 MLE 는

$$\widehat{\sigma^2} = \frac{1}{2n}\sum_{i=1}^{n} x_i^2$$

이 됩니다. Confidence Interval 을 구하기 위해서 다음과 같은 Fact 를 사용합니다.

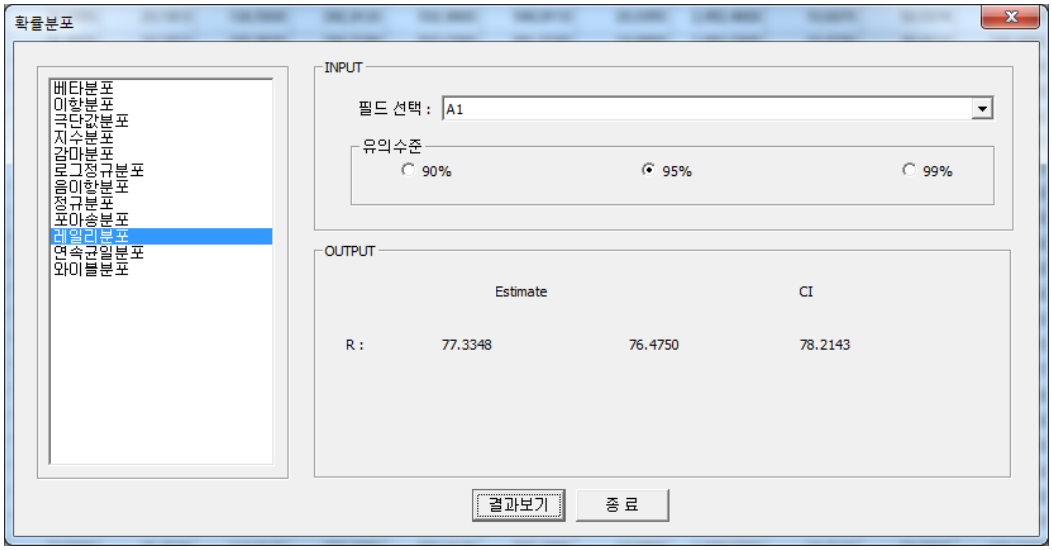
$$\text{if } X_i \sim \text{Rayleigh}(\sigma), \text{then } \sum_{i=1}^{n} X_i^2 \text{ has gamma distribution with parameter n and } 2\sigma^2$$

$$\frac{1}{2n}\Sigma_{i=1}^{n} X_i^2 = \frac{1}{2n}\text{Gamma}(n, 2\sigma^2) = \frac{\sigma^2}{2n} Gamma\,(n, 2) = \frac{\sigma^2}{2n}\chi^2(2n)$$

따라서 Confidence Interval 은 다음과 같습니다.

$$\frac{2n}{\chi^2_{1-\frac{\alpha}{2}}(2n)}\widehat{\sigma^2} < \sigma^2 < \frac{2n}{\chi^2_{\frac{\alpha}{2}}(2n)}\widehat{\sigma^2}$$

예시



(11) 연속 균일 분포

추정 방법

연속 균일 분포의 pdf 는 다음과 같습니다.

$$f_X(x; a, b) = \frac{1}{b-a} I_{(a,b)}(x)$$

그리고 a 와 b 에 대한 MLE 는 다음과 같습니다.

$$\hat{a} = \min(x_1, x_2, \ldots, x_n)$$

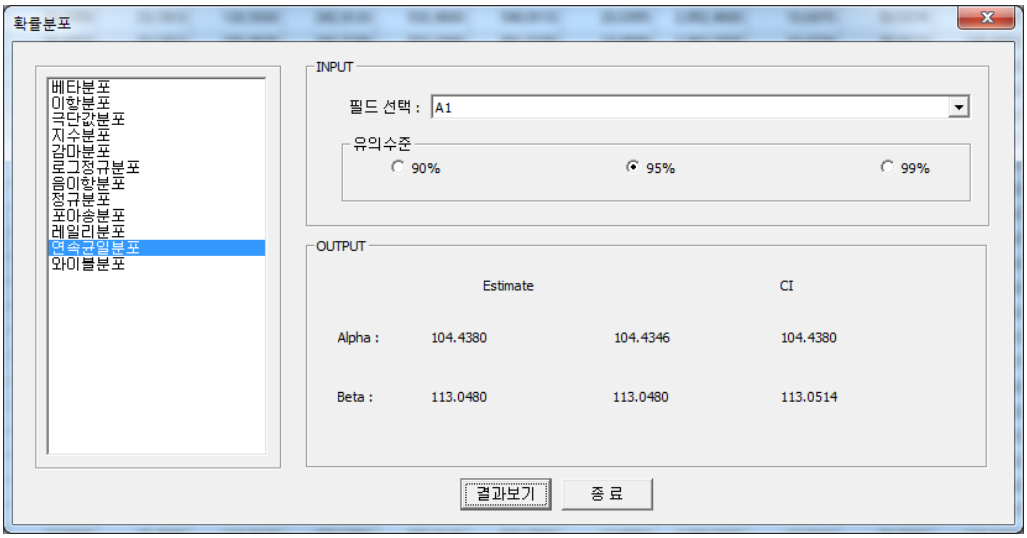
$$\hat{b} = \max(x_1, x_2, \ldots, x_n)$$

이는 직관적으로도 쉽게 이해할 수 있습니다. 얻어진 데이터에서 가장 작은 값이 a 에 가장 가까울 것이기 때문에 a 의 추정 값이 되고, 가장 큰 값이 b 에 가장 가까울 것이기 때문에 b 의 추정 값이 됩니다. 이에 대한 confidence interval 은 다음과 같습니다.

$$\hat{b} - \frac{\hat{b} - \hat{a}}{\alpha^{1/n}} < a < \hat{a}$$

$$\hat{b} < b < \hat{a} + \frac{\hat{b} - \hat{a}}{\alpha^{1/n}}$$

예시



(12) 와이블 분포

추정 방법

와이블 분포의 pdf 는 다음과 같습니다.

$$f_X(x;a,b) = \left(\frac{b}{a}\right)\left(\frac{x}{a}\right)^{b-1} \exp\left(-\left(\frac{x}{a}\right)^{b}\right) \quad where\ x \geq 0, a > 0, b > 0$$

이 때 Log Likelihood Function 을 구하면 다음과 같습니다.

$$\log L(a,b|X) = n\log(b) - nc\log(a) + (b-1)\sum_{i=1}^{n}\log(x_i) - \sum_{i=1}^{n}\left(\frac{x_i}{a}\right)^{b}$$

이를 최대화하기 위해서 a, b 에 대해서 편미분을 하고 0 으로 놓으면

$$\frac{\partial \ln L(a,b|X)}{\partial a} = -\frac{nb}{a} + \sum_{i=1}^{n} b x_i^b a^{-b-1} = 0$$

$$\frac{\partial \ln L(a,b|X)}{\partial b} = \frac{n}{b} - n log\,(a) + \sum_{i=1}^{n} \left(\frac{x_i}{a}\right)^b \log\left(\frac{x_i}{a}\right) = 0$$

이는 다음과 같습니다.

$$a = \left[\frac{\sum_{i=1}^{n} x_i^b}{n}\right]^{\frac{1}{b}}$$

$$\frac{\sum_{i=1}^{n} x_i^b \log(x_i)}{\sum_{i=1}^{n} x_i^b} - \frac{1}{b} = \frac{1}{n}\sum_{i=1}^{n} \log(x_i)$$

위의 두 equation 을 푸는 것이 목표입니다. 이를 위해서 보통 Newton – Raphson 방법을 사용합니다. 두 번째 식을 고치면 다음과 같이 됩니다.

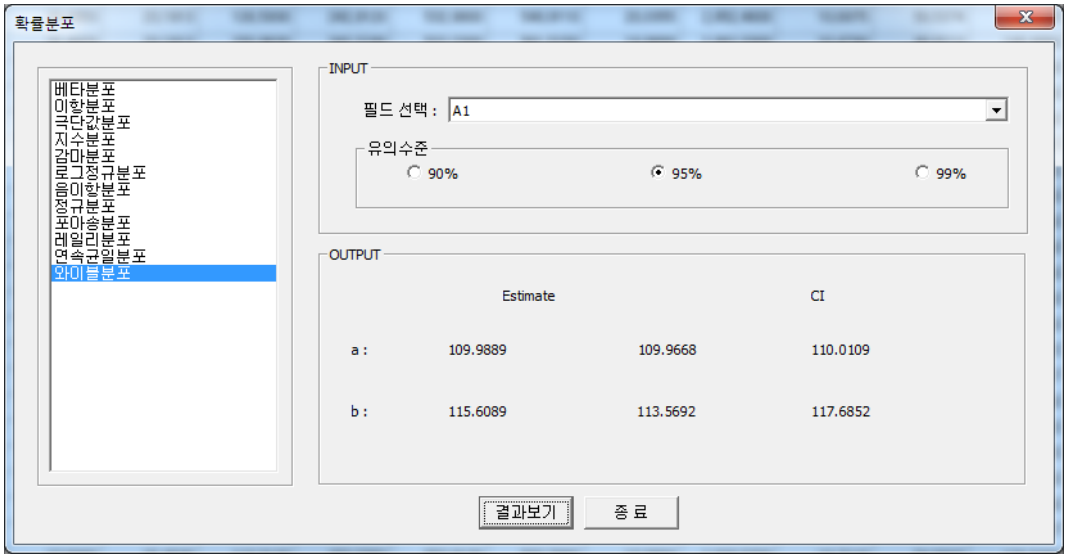
$$\frac{b}{n}\left(\sum_{i=1}^{n} \log(x_i)\right)\left(\sum_{i=1}^{n} x_i^b\right) - b\sum_{i=1}^{n} x_i^b \log(x_i) + \sum_{i=1}^{n} x_i^b = 0$$

이 식을 $g(b)$ 라고 놓으면 이를 통해서 다음과 같은 Newton Raphson 알고리즘 식을 세워 모수를 추정합니다.

$$b_{k+1} = b_k - \frac{g(b_k)}{g'(b_k)}, \qquad for\ k = 0,1,2,\ldots$$

$$\text{where } g'(b) = \frac{1}{n}\left(\sum_{i=1}^{n} \log x_i\right)\left(\sum_{i=1}^{n} x_i^b + b\sum_{i=1}^{n} x_i^b \log x_i\right) - b\sum_{i=1}^{n} x_i^b (\log x_i)^2$$

예시

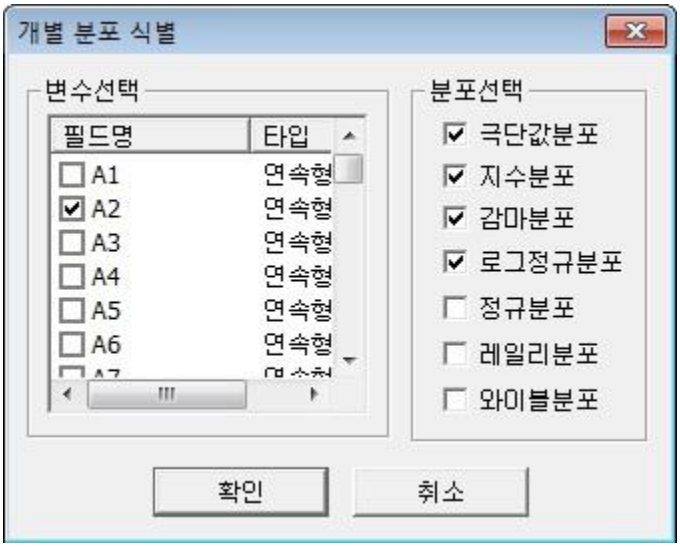


- 5.3.8.2 개별분포 식별

  데이터의 분포를 확인하기 위하여 여러 분포에 적용하여 얻어진 Anderson-Darling 통계량과 그의 P-value 값을 토대로 데이터의 가장 유사한 분포를 확인할 수 있습니다.

실행 방법

[분석] – [확률분포] – [개별분포식별]을 선택하면 다음과 같은 개별분포식별 다이얼로그가 나타납니다.



변수 목록에서 분포을 식별하고자 하는 변수를 선택하고(복수 선택 가능), 분포 선택에서 어떠한 분포에 적용할지를 선택합니다.(복수 선택 가능)

결과

- 변수 정보: 선택한 변수의 기초 통계량들을 제공합니다.
- 분포 모수추정: 해당 변수로부터 선택한 분포의 모수들을 추정하고 추정치들을 보여줍니다.
- 검정통계량: 해당 변수가 선택한 분포에 얼마나 적합한지 보여주는 Anderson-Darling 통계량과 P-value 를 제공합니다. P-value 가 0.05 보다 크면, 변수는 분포를 따른다고 할 수 있습니다.

5.3.9 비모수 검정

비모수적 방법은 모집단의 분포가 특정함수의 형태를 가정하는 모수적 방법과는 달리, 모집단의 분포함수에 대하여 특정형태를 가정하지 않는 통계적 방법을 뜻합니다. 모집단의 분포가 정규분포에 가까운 경우나, 표본이 많아서 중심극한정리(Central Limit Theorem)에 의해 정규분포에 근사 될 경우는 모수적 방법을 통해 중심위치에 대한 추정이 가능합니다. 하지만 모집단에 대한 구체적인 분포함수의 가정이 무리가 있을 경우의 모수적 방법은 오류의 가능성이 높으며, 비효율적입니다. 이때 비모수적 방법이 대안이 될 수 있습니다.
비모수적 방법에서는 주로 관측값의 부호(sign)와 순위(rank) 또는 순위에 기초한 점수(score) 를 흔히 사용합니다.

실행방법

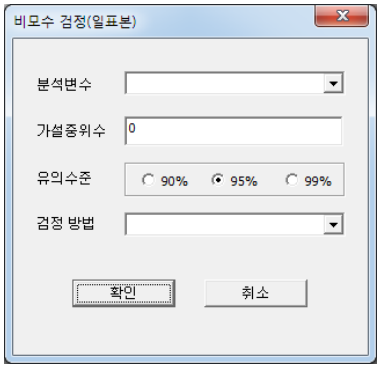
[분석] – [비모수 검정]의 하위 메뉴에서 선택이 가능합니다.

- #### 5.3.9.1 일표본

데이터 탐색기에서는 각각의 필드에 대해 **비모수 검정 - 일표본** 기능을 사용하여, 주어진 위치모수(가설중위수,Median)에 대한 검정을 진행 할 수 있습니다. 현재 비모수 일표본 검정에서는 부호 검정(Sign test)와 윌콕슨 부호순위검정(Wilcoxon signed rank test)을 지원하고 있습니다.

부호 검정(Sign test)는 귀무가설 하에서 위치모수의 값보다 큰 관측값의 개수만을 이용하여 검정하는 방법이며, 윌콕슨 부호순위검정(Wilcoxon signed rank test)는 부호 검정에 추가로 관측값의 순서를 고려하는 검정방법으로, 비모수 일표본 분제에서 가장 널리 사용되는 방법입니다.

실행 방법



[분석] – [비모수 검정] – [일표본]을 선택하면, **일표본** 윈도우가 나타납니다. 분석할 필드를 선택하고 가설중위수(Median)를 입력합니다. 유의수준을 정합니다. 마지막으로 두 가지의 검정 방법 중 하나를 택합니다.

결과

일표본 통계량과 검정결과를 보여줍니다.

일표본 비모수 검정

일표본 통계량

변수명	데이터개수	평균	표준편차	평균의 표준오차
A1	8530	30,42204	12,80825	0,13868

일표본 검정

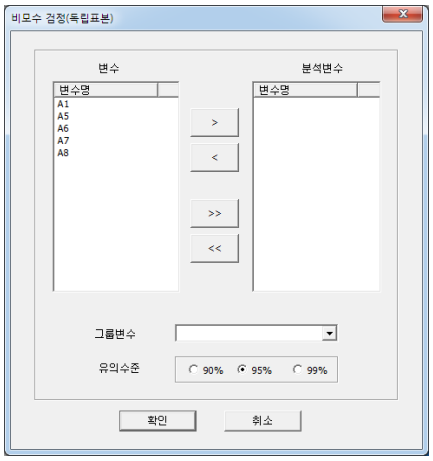
귀무가설	가설 검정 방법	유의확률(양쪽)	결론
A1의 중앙값은 28,0000000이다.	Wilcoxon Signed Rank Test	0	귀무가설을 기각합니다.

* 검정 통계량 연산 시, 대표본 근사를 사용하였습니다.

- 5.3.9.2 독립표본

데이터 탐색기에서는 각각의 필드에 대해 **비모수 검정 - 독립표본** 기능을 사용하여, 한 변수 내의 두 그룹간의 평균 또는 중앙값 비교에 사용되는 검정방법입니다. 현재 비모수 독립표본 검정에서는 Mann-Whitney U Statistic 을 제공하고 있습니다.

실행 방법



[분석] – [비모수 검정] – [독립표본]을 선택하면, **독립표본** 윈도우가 나타납니다. 분석할 필드를 선택하고 그룹변수를 선택합니다. 이때 그룹변수는 2 개의 그룹으로 구성되어야 합니다. 마지막으로 유의수준을 정합니다.

결과

독립표본 통계량과 검정결과를 보여줍니다.

비모수 검정(독립표본)

집단통계량

	이탈여부	데이터 개수	평균	표준편차	평균의 표준 오차
A1	0	3846	28.67447	12.59226	0.20305
	1	4684	31.85696	12.80715	0.18713

비모수 독립표본 검정

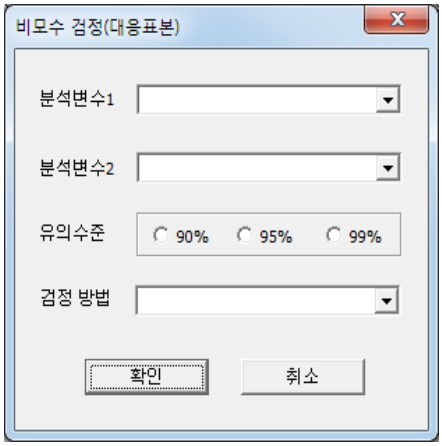
귀무가설	가설 검정 방법	유의확률(양쪽)	결론
A1의 분포는 이탈여부의 그룹에 관계없이 동일하다.	Mann-Whitney U Test	NaN	귀무가설을 기각하지 않습니다.

* 검정 통계량 연산 시, 대표본 근사를 사용하였습니다.

- 5.3.9.3 대응표본

데이터 탐색기에서는 상호 의존적인 각각의 필드에 대해 **비모수 검정 - 대응표본** 기능을 사용하여, 대응표본 검정을 실시합니다. 현재 비모수 대응표본 검정에서는 부호 검정(Sign test)와 윌콕슨 부호순위검정(Wilcoxon signed rank test)을 지원하고 있습니다.

실행 방법



[분석] – [비모수 검정] – [대응표본]을 선택하면, **대응표본** 윈도우가 나타납니다. 분석할 2 개의 필드를 선택하고 유의수준 및 검정방법을 선택합니다.

결과

대응표본 통계량과 검정결과를 보여줍니다.

비모수 검정(대응표본)

대응표본 통계량

	평균	데이터 개수	표준편차	평균의 표준 오차
A1	109.35402	7596	1.74303	0.02000
A3	115.59035	7596	2.35585	0.02703

대응표본 통계량

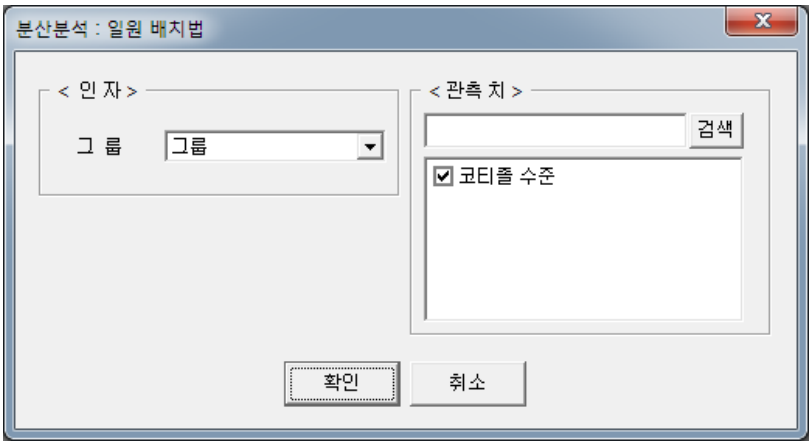
귀무가설	가설 검정 방법	유의확률(양쪽)	결론
A1와 A3 사이의 중앙값 차이는 0이다.	Wilcoxon Signed Rank Test	0	귀무가설을 기각합니다.

* 검정 통계량 연산 시, 대표본 근사를 사용하였습니다.

- **5.3.9.4 분산분석 – 일원배치법**

일원배치 분산 분석은 하나의 인자에 대한 분산 분석 방법입니다. 현재 비모수 분산분석에서는 Kruskal-Wallis 검정을 지원하고 있습니다. Kruskal-Wallis 검정은 3 개 이상의 모집단 중앙값 차이 검정에 가장 많이 사용되는 방법으로, 모집단이 2 개인 경우는 Mann-whitney 방법과 동일합니다.

실행 방법



[분석] – [비모수 분석] – [분산분석 - 일원배치법]을 선택하면, **분산분석:일원 배치법** 윈도우가 나타납니다. 분석할 2 개의 필드를 선택하고 유의수준 및 검정방법을 선택합니다.

결과

분산분석 검정결과를 보여줍니다.

비모수 검정(일원배치)

관측 변수명: 코티졸 수준

귀무가설	검정방법	검정 통계량	유의확률	결론
코티졸 수준의 분포는 그룹의 그룹에 관계없이 동일하다.	Kruskal-Wallis Test	9.2316	0.0099	귀무가설을 기각합니다.

집단 통계량

Level	N	Mean	Median	Ave Rank	StdDev
1	10	299.7000	305.5000	6.9000	77.3409
2	6	434.3333	460	15	55.4847
3	6	649.1667	729.5000	15.6667	288.8179

5.3.10 정확도 측도

분류분석, 회귀분석을 통해 어떤 변수를 예측할 때, 원래의 종속변수와 예측한 변수값을 사용하여 모델의 정확도를 계측할 수 있습니다. 이 메뉴에서는, 모델링 결과 추가된 예측변수와 원래의 변수 간의 관계를 통해 모델의 정확성 및 모델링의 타당성을 검증할 수 있습니다.

실행방법

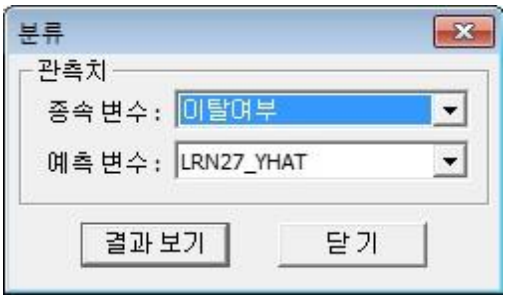
[분석] – [정확도 측도]의 하위 메뉴 중 **[분류]** 혹은 **[예측]**을 선택합니다.

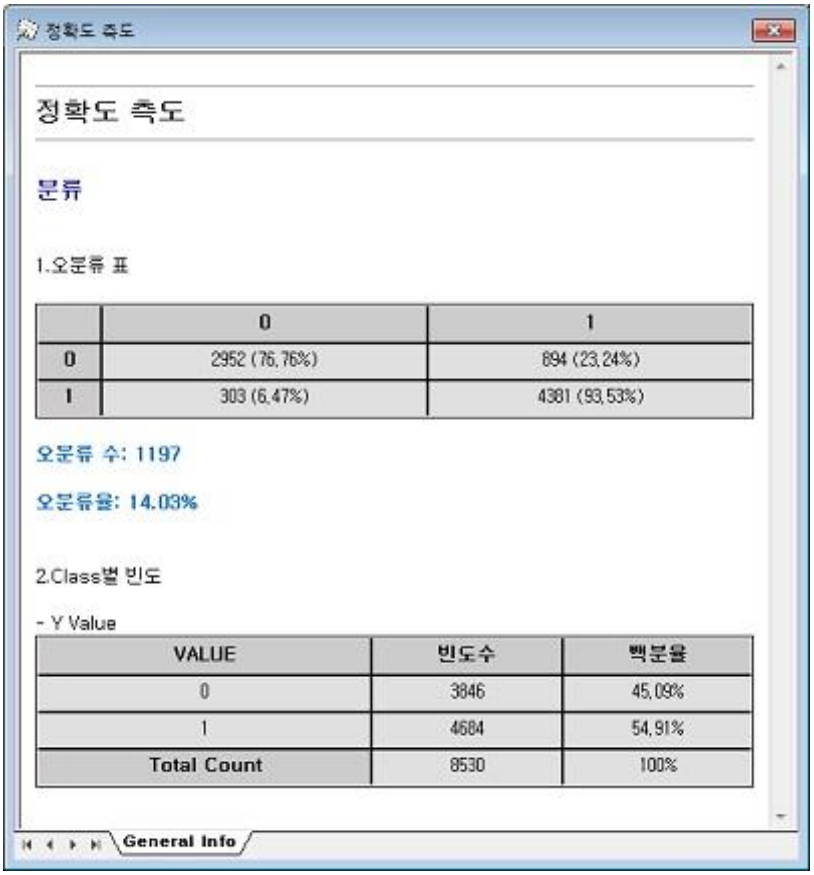
- **5.3.10.1 분류**

추정 방법

분류분석의 결과물은 독립변수를 통해 각 클래스의 label 을 예측한 값입니다. 이 값이 종속 변수와 동일하다면 올바르게 예측한 것이고, 그렇지 않다면 잘못 예측한 것입니다. 이 메뉴에서는 종속변수와 예측변수를 비교하여 오분류 표와 오분류 수 및 오분류율을 계산합니다. 또한, 종속 변수에 몇 개의 클래스가 존재하는 지와 클래스 별 데이터 빈도를 보여줍니다.

예시





- 5.3.10.2 예측

추정 방법

회귀분석 결과의 타당성을 판단하기 위해 사용되는 방법에는 여러 가지가 있습니다. 이 메뉴에서는 종속변수와 예측변수를 사용하여 R-square, MAPE(Mean Absolute Percentage Error), MAD(Mean Absolute Deviation), MSD(Mean Squared Deviation)을 제공합니다.

▪ R-square(결정계수)

도출된 회귀식이 측정값들을 대표할만한가를 확인하는 것으로, 즉, 추정된 회귀선이 관측치들을 얼마나 잘 설명하는지를 검정하는 값입니다. 1 에 가까울수록 회귀식이 데이터를 잘 설명하고 있음을 의미합니다.

▪ MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{\sum_{i=1}^{n} | (y_i - \hat{y}_i) / y_i |}{n} \times 100$$

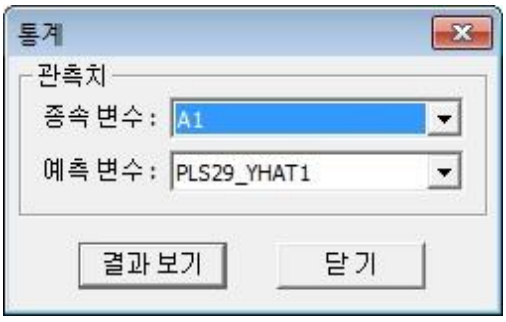
- MAE(Mean Absolute Error)

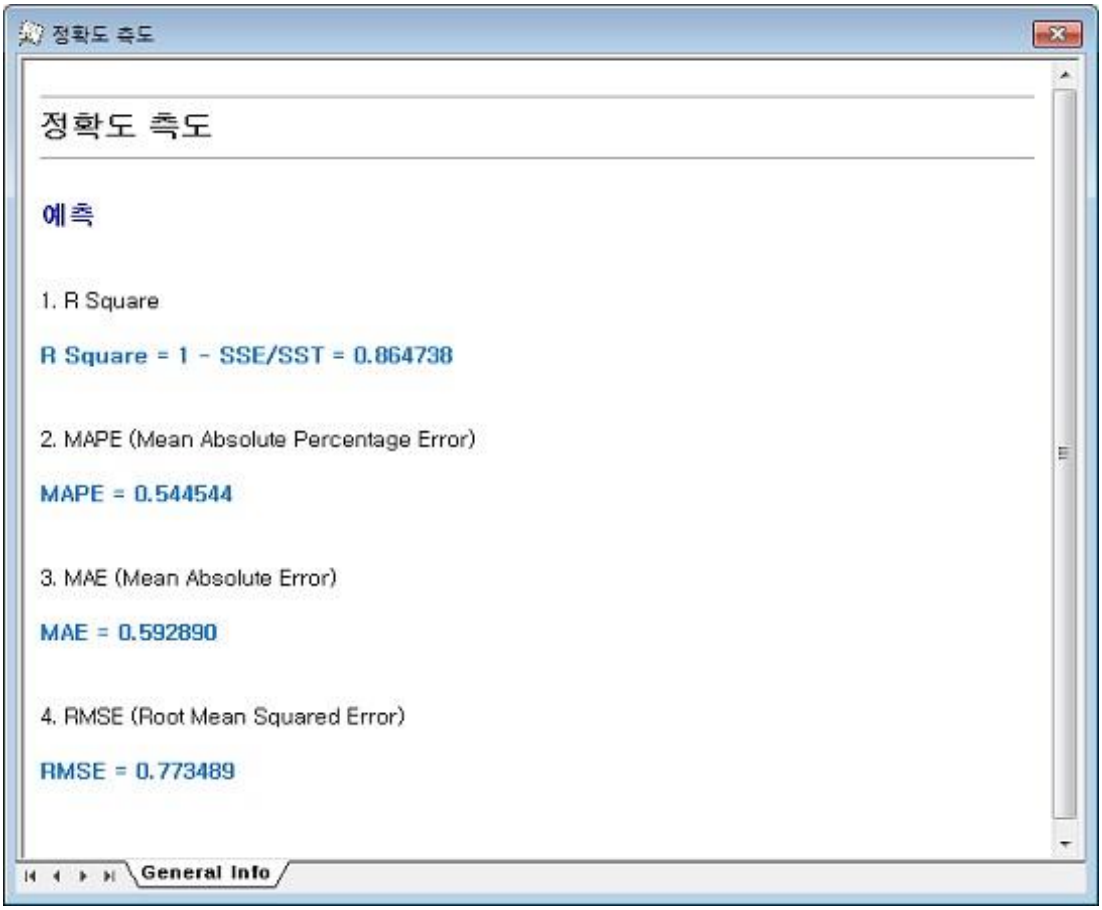
$$MAE = \frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{n}$$

- RMSE(Root Mean Squared Error)

$$RMSE = \sqrt{MSE} = \sqrt{SSE/(n-p-1)}$$

예시





Note: 상위 R-Square 값이 양수가 아닌 음수로 나오는 경우도 있습니다. ECMiner™ 에서 SSE, SST 의 계산 과정은 아래와 같습니다.

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

$$SST = \sum_i (y_i - \bar{y})^2$$

따라서 SSE 가 SST 보다 클 경우도 있어 R-Square 의 계산 값이 음수로 나올 수 있습니다.

5.3.11 Gage R&R

많은 통계 방법론이 측정된 데이터를 가지고 시작합니다. 그리고 이 측정된 데이터가 실제 상황을 대변해주는 참된 값이라는 가정에서 시작하는 통계 방법론들이 많습니다. 하지만 이는 실제로 참 값과는 다른 측정 값일 뿐이며 그렇기 때문에 정밀한 분석을 하기 위해서는 분석을 시작하기 전에 데이터가 신뢰할 만한 것인지를 확인해야 합니다. 데이터에 대한 신뢰성은 측정 시스템에 대한 신뢰성이 보장될 때 얻어질 수 있는 특성입니다. 즉 데이터에 대한 신뢰성을 평가하고자 하는 목적으로 측정 시스템에 대해 분석하는 것이 Gage R&R(Gage Repeatability & Reproducibility) 입니다.

품질 관리 방법론에서 사용하는 SPC 방법론은 부품 간의 변동을 측정하여 현재 공정에 이상이 있는지의 여부를 판단합니다. 하지만 SPC 방법론 및 다른 품질 관리 방법론을 사용하기 전에 먼저 수행해야 하는 것이 바로 Gage R&R 입니다.

5.3.11.0 Gage R&R 의 개념

용어 정리

반복성(Repeatability) : 측정 장치로 인해서 나타나는 변동이며 같은 작업자가 같은 부품을 같은 장치로 측정했을 때 나타나는 변동입니다.

재현성(Reproducibility) : 측정 시스템으로 인해서 나타나는 변동이며 다른 작업자가 같은 부품을 같은 장치로 관찰했을 때 나타나는 변동입니다.

정확성 : 정확성은 부품의 측정 값과 실제 값 사이의 차이가 나는 특성을 말합니다.

정밀성 : 정밀성은 동일한 부품을 동일한 장치로 반복해서 측정했을 때 나타나는 변동에 대해서 설명합니다.

ECMiner™ 에서 제공하는 Gage R&R 방법론은 다음과 같습니다

- Gage 런차트
- Gage 선형성 및 편향 연구
- Gage R&R 내포 설계
- Gage R&R 교차 설계

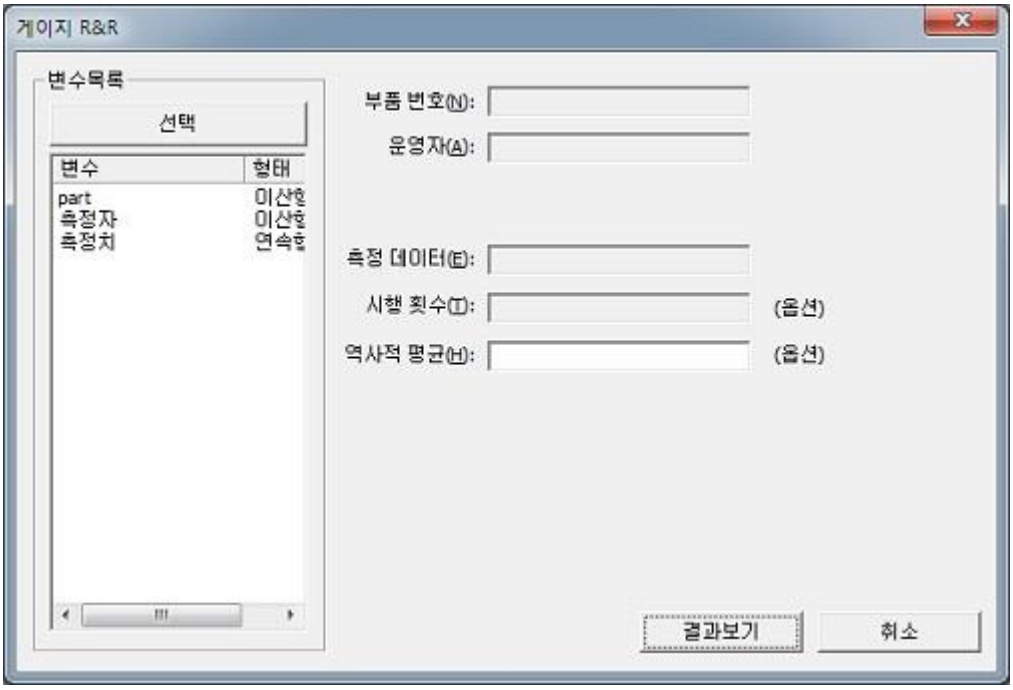
5.3.11.1 Gage 런차트

개요

Gage 런차트는 실험을 통해서 얻은 전체 관측치를 Plot 한 것입니다. 크게 부품의 종류별로 Plot 하고 각 부품당 작업자에 따라 어떻게 관측치가 달라지는 것을 보여줍니다. 본 기능을 통해서 관측치에 대한 현황을 한눈에 파악할 수 있습니다.

실행방법

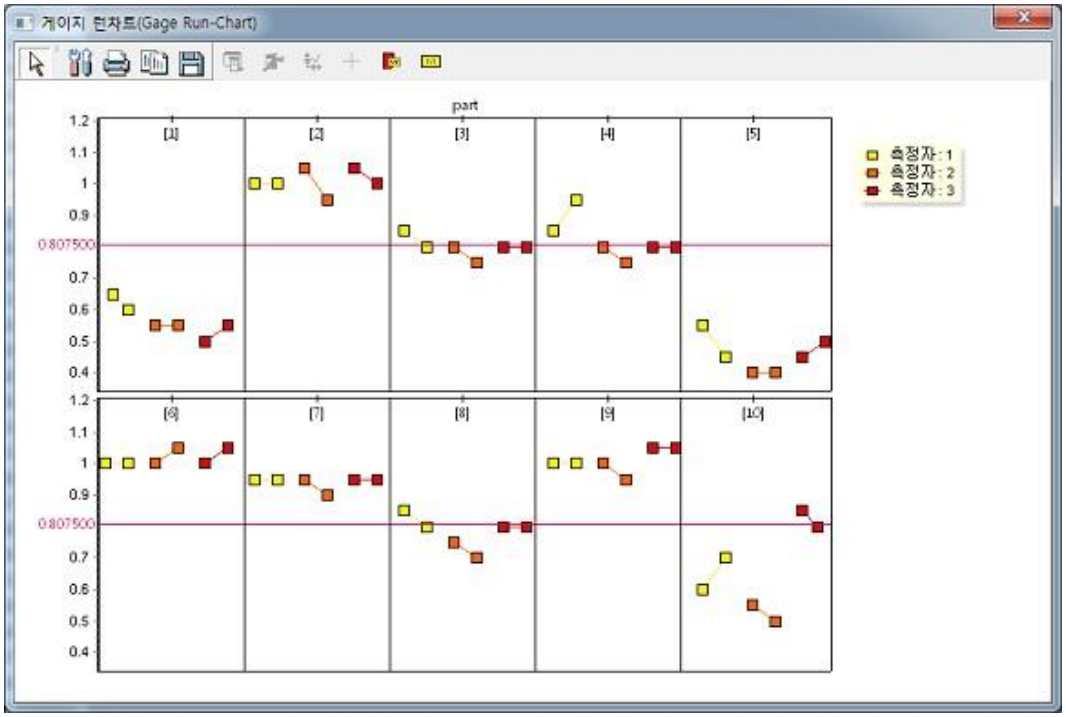
[분석] – [Gage R&R] – [Gage 런차트] 를 선택하면 [Gage 런차트] 윈도우가 나타납니다.



- 부품 번호 : 부품 번호 혹은 부품을 구별할 수 있는 데이터가 들어있는 FIELD 를 선택합니다.
- 운영자 : 부품을 측정한 사람을 구별할 수 있는 데이터가 들어있는 FIELD 를 선택합니다.
- 측정 데이터 : 실험을 통해서 측정된 데이터가 들어있는 FIELD 를 선택합니다.
- 시행 횟수 : 같은 운영자가 같은 부품을 여러 번 측정할 수 있는데 이럴 때 측정 순서가 들어가 있는 FIELD 를 선택합니다.

- 역사적 평균 : 부품의 측정치가 역사적으로 평균이 어느 정도 되는지를 입력합니다.

결과



위의 차트에서 가운데 선은 역사적 평균(사용자가 입력했을 시) 혹은 데이터의 평균입니다. 현재 데이터는 운영자는 3 명, 부품은 10 종류 그리고 반복수는 2 인 데이터입니다. 각 사각형 별로 부품의 번호가 나타나 있으며, 색으로 운영자가 구분됩니다. 이를 통해서 부품별, 운영자별 관측치 데이터에 어떠한 차이가 있는지를 확인해 볼 수 있습니다.

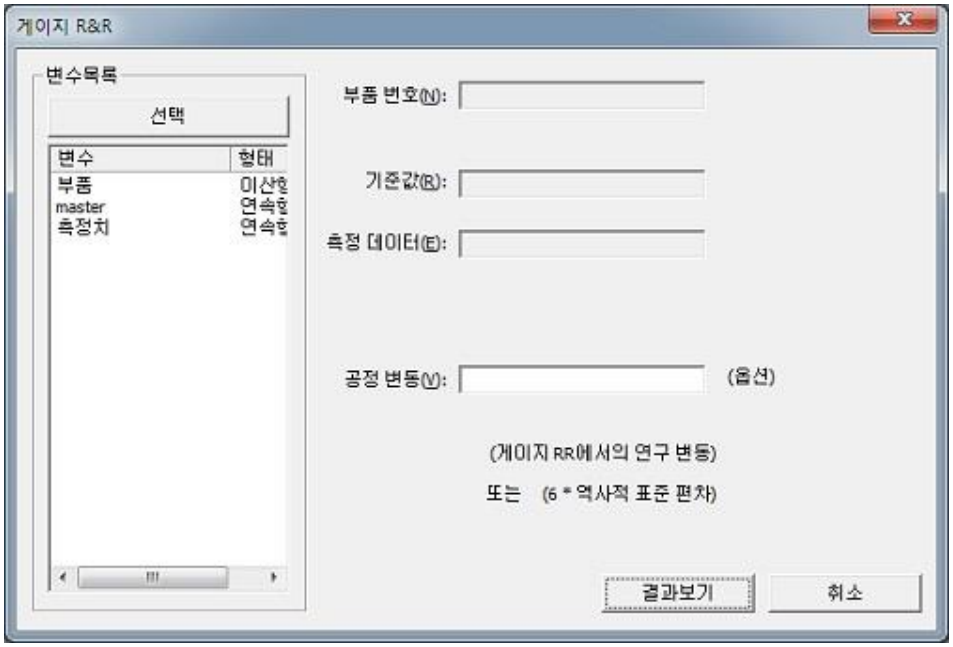
5.3.11.2 Gage 선형성 및 편향 연구

개요

Gage R&R 선형성 및 편향 연구는 부품의 규격에 따라서 측정 값과 기준 값의 차이라고 할 수 있는 편향이 선형적인 관계를 가지고 있는지를 판단하기 위해 사용합니다. 예를 들어 부품의 크기가 커질수록 편향의 정도가 커지거나 작아지는 경향이 있다고 하면 측정 시스템이 부품의 크기에 의존한다는 것이 됩니다. 이는 결코 바람직한 측정 시스템이 갖는 특성이 아니므로 이와 같은 문제가 발생하였을 때는 측정 시스템에 대한 보완이 이루어져야 할 것입니다.(바람직한 측정 시스템일수록 편향의 값이 부품의 크기와 무관해야 합니다.)

실행방법

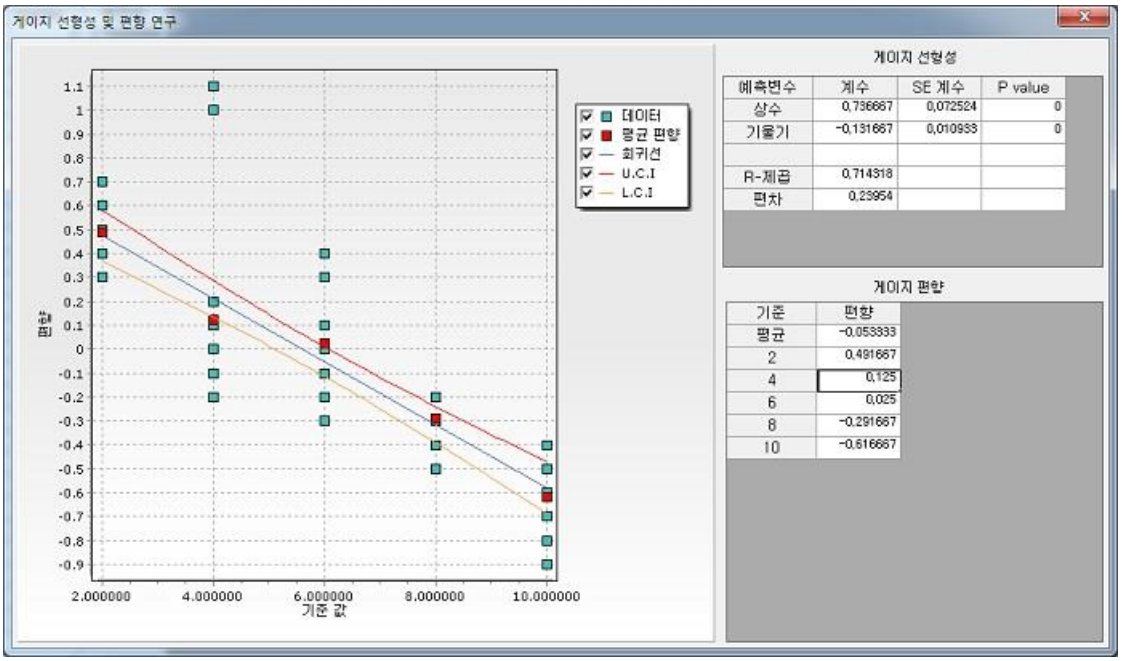
[분석] - **[Gage R&R]** - **[Gage R&R 선형성 및 편향 연구]**를 선택하면 **[Gage R&R 선형성 및 편향 연구]** 윈도우가 나타납니다.



- 부품 번호 : 부품 번호 혹은 부품을 구별할 수 있는 데이터가 들어있는 FIELD 를 선택합니다.
- 기준 값 : 부품이 실제적으로 가져야 할 값을 가지고 있는 FIELD 를 선택합니다.
- 측정 데이터 : 통해서 측정된 데이터가 들어있는 FIELD 를 선택합니다.

- 공정 변동 : 사용자가 정의한 공정 변동을 입력합니다.

결과



게이지 선형성 및 편향 연구에 대한 결과는 위와 같습니다. 왼쪽의 차트는 편향이 부품 크기에 선형적인 관계가 있는지를 시각적으로 나타내 줍니다. 그리고 이에 대한 회귀분석의 결과를 오른쪽 상단의 표에서 제시해 주고 있습니다. 현재 R 제곱의 값이 0.7 이상으로 매우 큰 것으로 보아 측정 시스템이 부품의 크기에 매우 크게 영향을 받는다는 것을 알 수 있습니다. 그리고 아래의 게이지 편향표에서도 볼 수 있듯이 편향이 매우 크게 나타나기 때문에 측정 시스템에 대한 보완이 필요함을 알 수 있습니다.

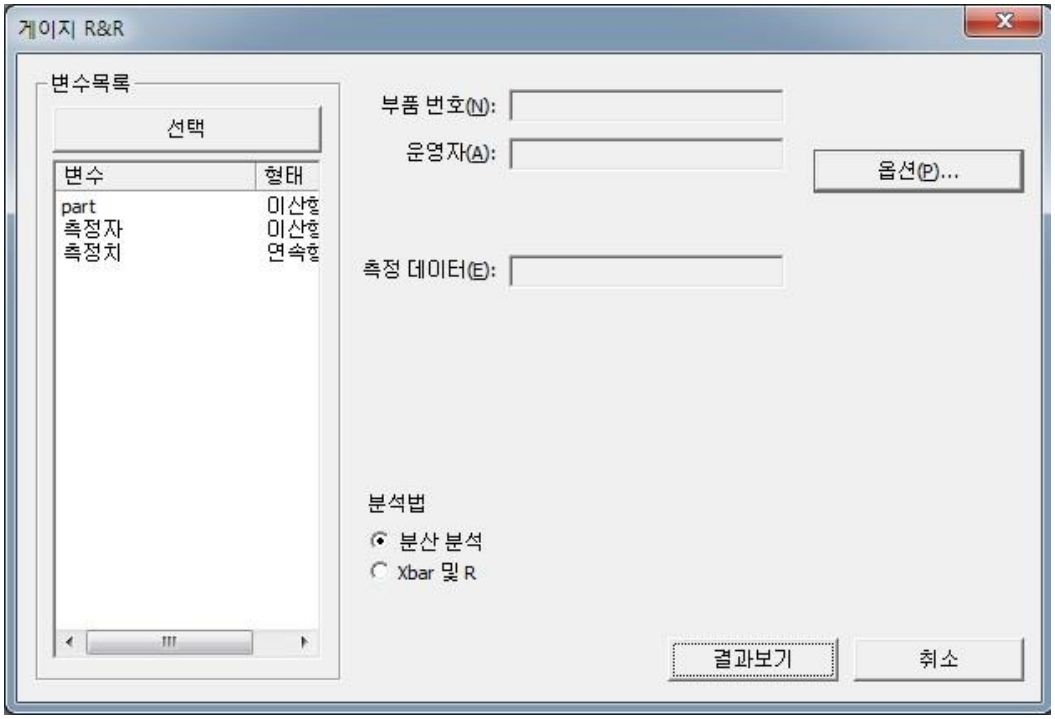
5.3.11.3 Gage R&R 교차설계

개요

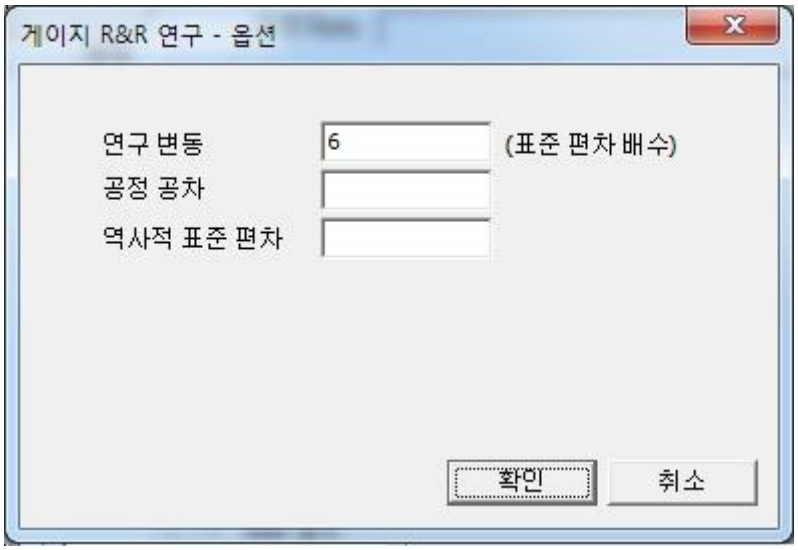
Gage R&R 교차 설계는 각 부품을 각각의 운영자가 여러 번 관측할 경우에 사용하는 방법으로 분산 분석 방법과 Xbar R 방법이 있습니다. 분산 분석의 경우 관측치의 변동을 부품 대 부품, 재현성, 반복성 성분으로 나누어 분석합니다. 그리고 분산 분석을 선택하면 재현성의 경우 운영자와 부품의 성분으로 나눕니다.(Xbar R 의 경우 재현성을 또 다시 나누지는 않습니다.) 여기에 부품과 운영자의 상호작용까지 고려하는데 ECMiner™ 에서는 부품과 운영자의 상호작용에 관련한 변동의 P value 가 0.25 보다 크면 자동적으로 상호작용을 무시하도록 하였습니다. (P value 가 0.25 보다 크면 통계적으로 유의하지 않은 것을 의미합니다.) 교차 설계를 통해서 분석자는 관측치에서 나타나는 변동이 어떤 성분에 의해서 기인하였는지를 쉽게 분석할 수 있습니다.

실행방법

[분석] - [Gage R&R] - [Gage R&R 교차설계] 를 선택하면 [Gage R&R 교차 설계] 윈도우가 나타납니다.



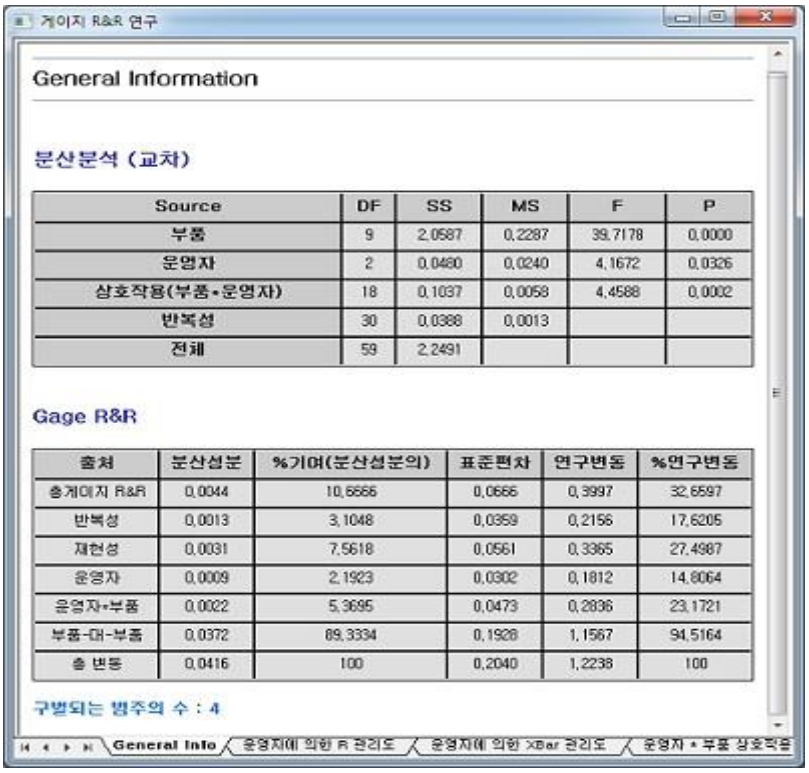
- 부품 번호 : 부품 번호 혹은 부품을 구별할 수 있는 데이터가 들어있는 FIELD 를 선택합니다.
- 운영자 : 부품을 측정한 사람을 구별할 수 있는 데이터가 들어있는 FIELD 를 선택합니다.
- 측정 데이터 : 실험을 통해서 측정된 데이터가 들어있는 FIELD 를 선택합니다.
- 분석법 : 분산 분석으로 분석할 것인지 Xbar 및 R 방법으로 분석할 것인지를 선택합니다.



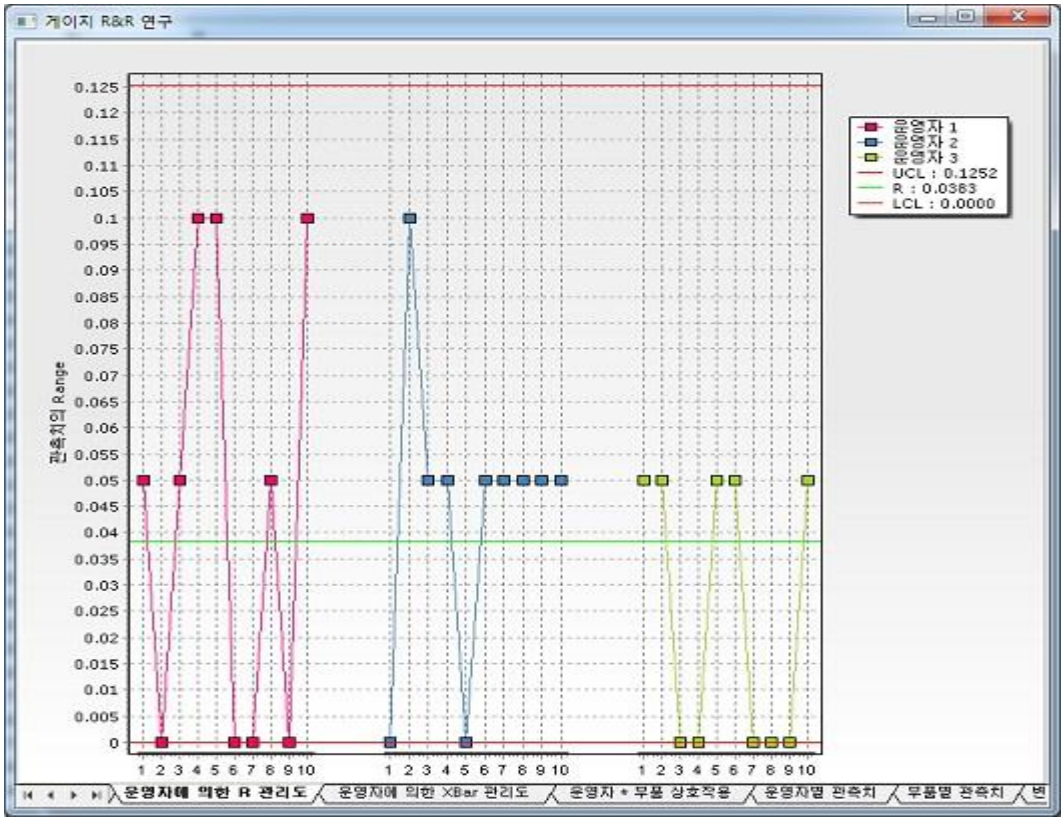
- 연구 변동 : 연구 변동을 구하기 위해 표준 편차에 곱할 계수를 입력합니다.
- 공정 공차 : 경험적으로 알려진 공차를 입력합니다. 이를 통해서 %공차를 계산하게 됩니다.
- 역사적 표준 편차 : 경험적으로 알려진 역사적 표준 편차를 입력합니다. 이를 통해서 %공정을 계산하게 됩니다.

결과

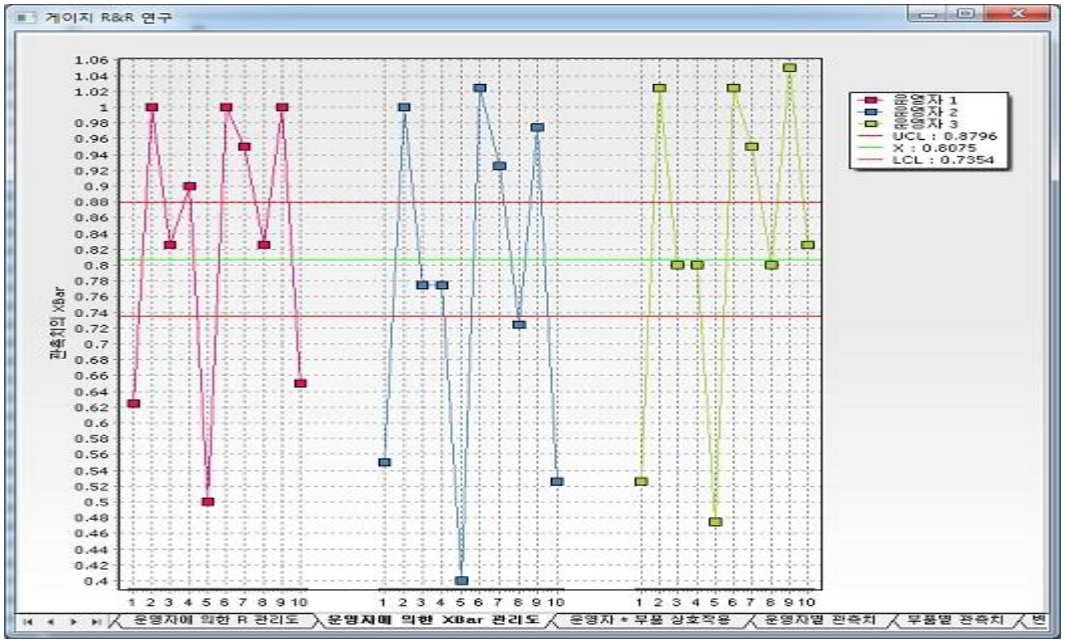
- General Information : 분산 분석 결과(분산 분석을 선택하였을 때만)와 Gage R&R 분석 결과를 보여줍니다.
  - 구별되는 범주의 수 : 측정시스템을 평가하는 척도로, 그 값이 5 이상이면 측정시스템을 인정해도 좋다고 판단합니다.



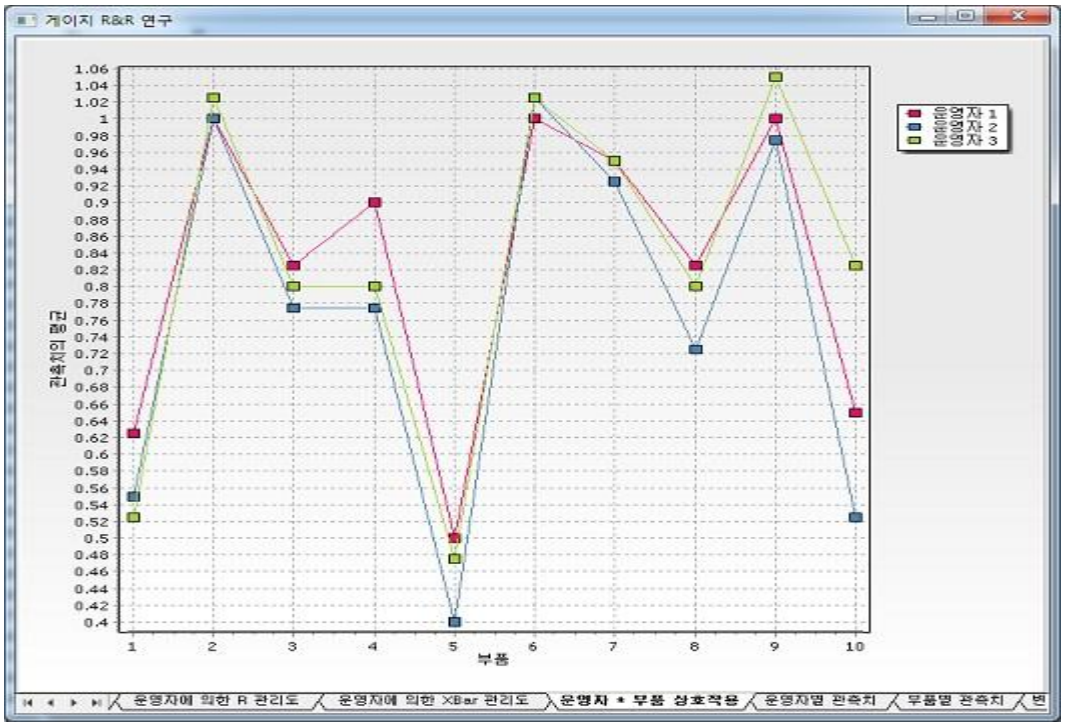
- 운영자에 의한 R 관리도 : 운영자가 각 부품을 여러 번 측정하였을 때 얼마나 차이가 났는지(Range)를 표시해 줍니다.



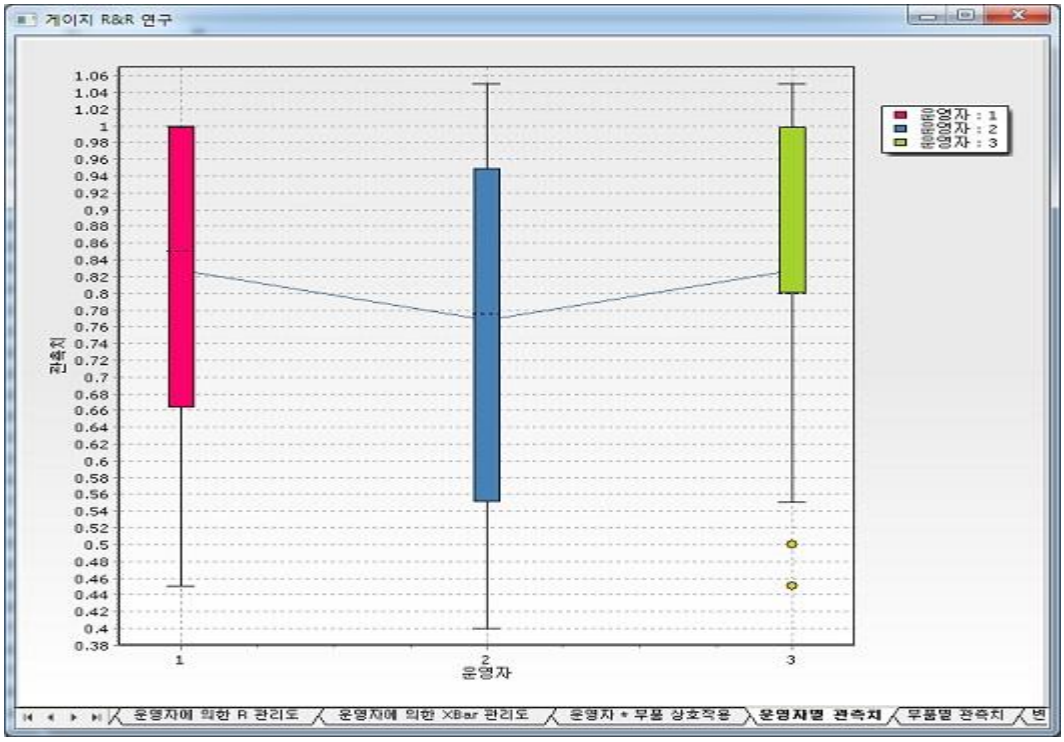
- 운영자에 의한 XBar 관리도 : 운영자가 각 부품을 여러 번 측정하였을 때 각 부품 당 관측치의 평균(XBar)을 표시해 줍니다.



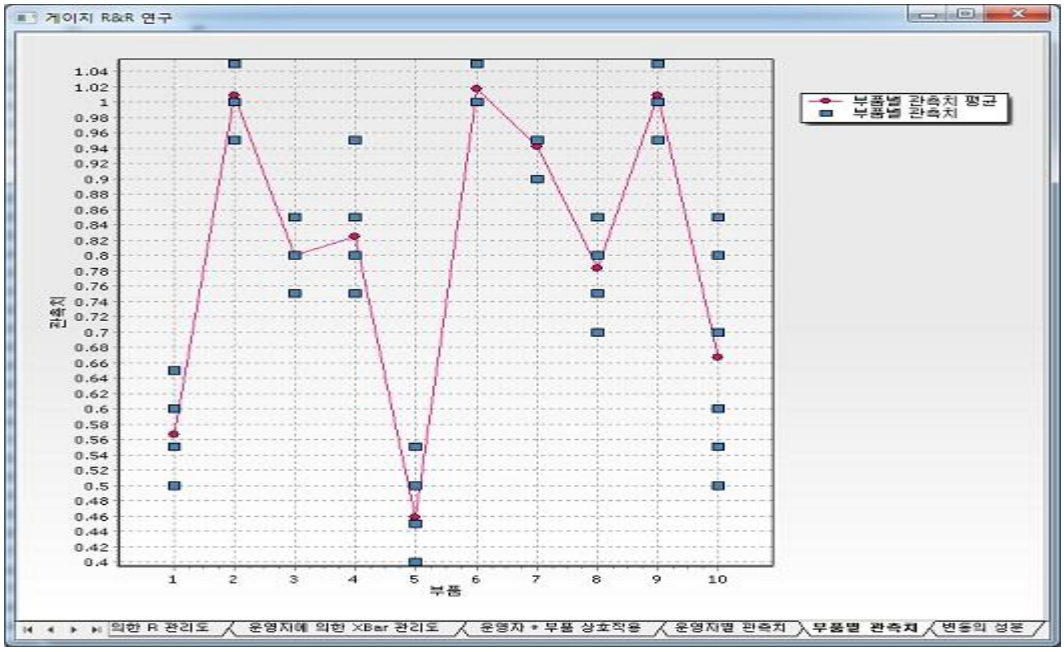
- 운영자*부품 상호 작용 : 운영자가 부품에 따라서 여러 관측치를 얻을 때 관측치의 평균을 표시해 줍니다.



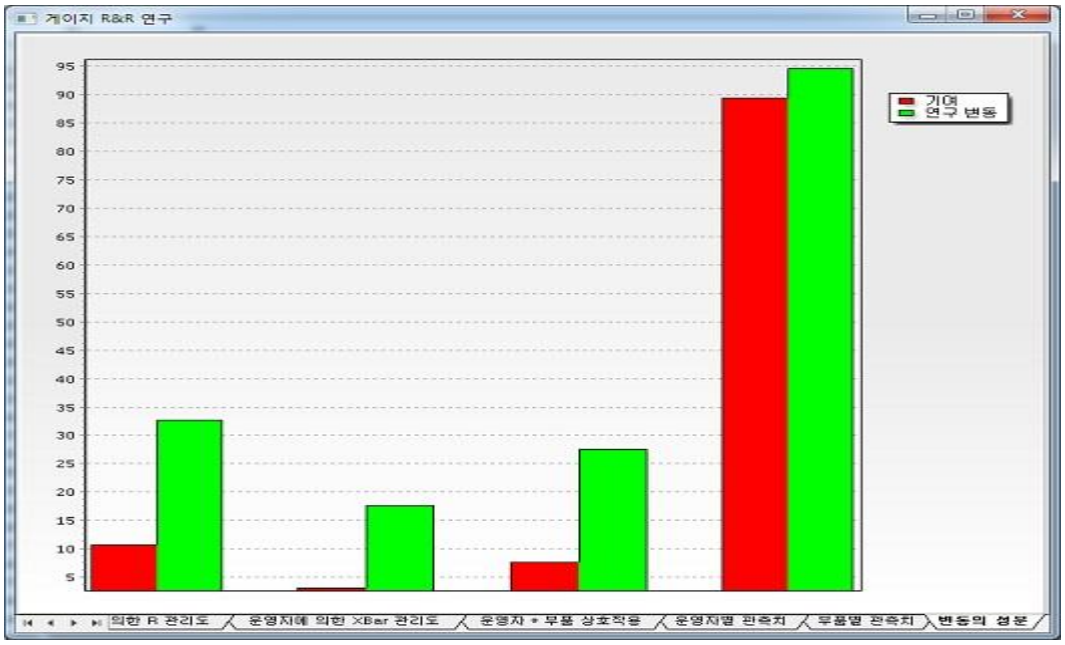
- 운영자별 관측치 : 부품이 어떠한 것인지를 고려하지 않고 오직 운영자 별로 관측치가 어떠한지를 표시해줍니다. 실선은 관측치의 평균을 나타냅니다.



- 부품별 관측치 : 운영자를 고려하지 않고 오직 부품 별로 관측치가 어떠한지를 표시해줍니다. 실선은 관측치의 평균을 나타냅니다.



- 변동의 성분 : 관측치의 변동이 어디에서 기인하였는지를 표시해줍니다.



5.3.11.4 Gage R&R 내포 설계

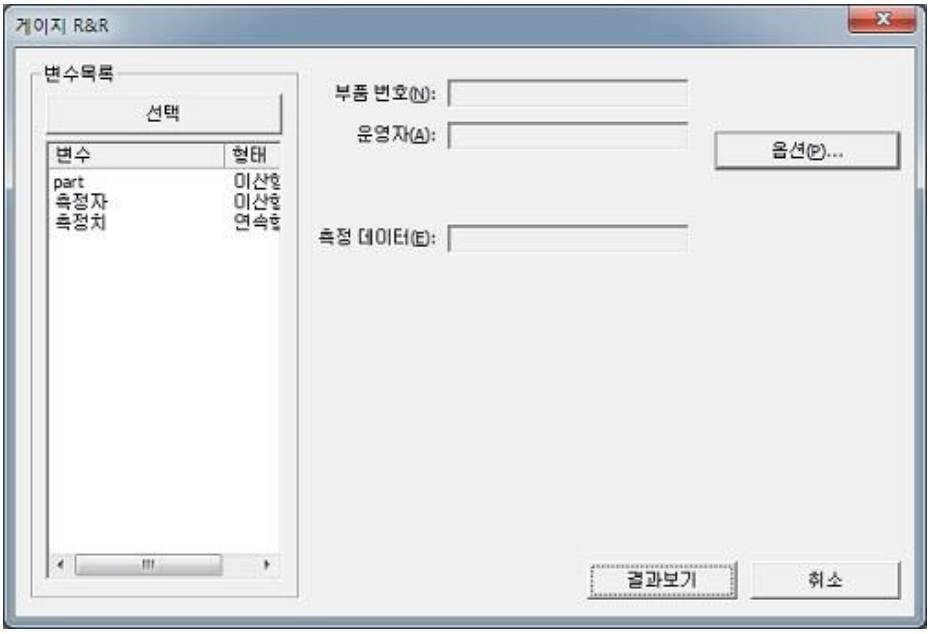
개요

Gage R&R 내포 설계는 교차 설계와 약간 다릅니다. 교차 설계의 경우 하나의 부품에 대해서 여러 운영자가 측정을 하지만 내포 설계의 경우에는 하나의 부품에 대해서는 한 운영자만이 측정을 하게 됩니다. 이는 파괴 검사에 잘 적용되는 케이스입니다. 혹은 그와 유사하게 한 운영자가 측정을 하면 더 이상 다른 운영자가 측정을 할 수 없는 시스템에서도 이와 같은 내포 설계가 사용될 수 있습니다. 이와 함께 내포 설계에서는 각 Batch 내에 있는 부품들이 거의 동일하다는 가정을 해야 합니다. 이러한 가정이 위배될 시에 서로 운영자가 완전히 다른 부품들을 측정하는 상황이 되어 결국 측정에서의 변동이 부품이 달라서 나타난 것인지 혹은 측정 시스템의 문제로 나타난 것인지 구분할 수 없기 때문입니다.

내포 설계도 역시 총 변동을 부품 대 부품, 재현성, 반복성으로 나누어 분석합니다. 이를 통해 측정 값의 변동이 어디에서 기인하였는지를 파악할 수 있습니다.

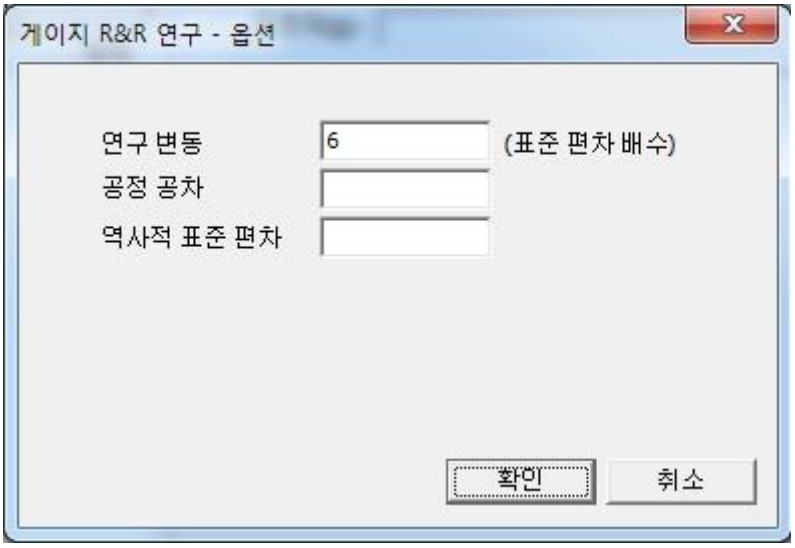
실행방법

[**분석**] - [**Gage R&R**] - [**Gage R&R 내포 설계**]를 선택하면 [**Gage R&R 내포 설계**] 윈도우가 나타납니다.



- 부품 번호 : 부품 번호 혹은 부품을 구별할 수 있는 데이터가 들어있는 FIELD 를 선택합니다.

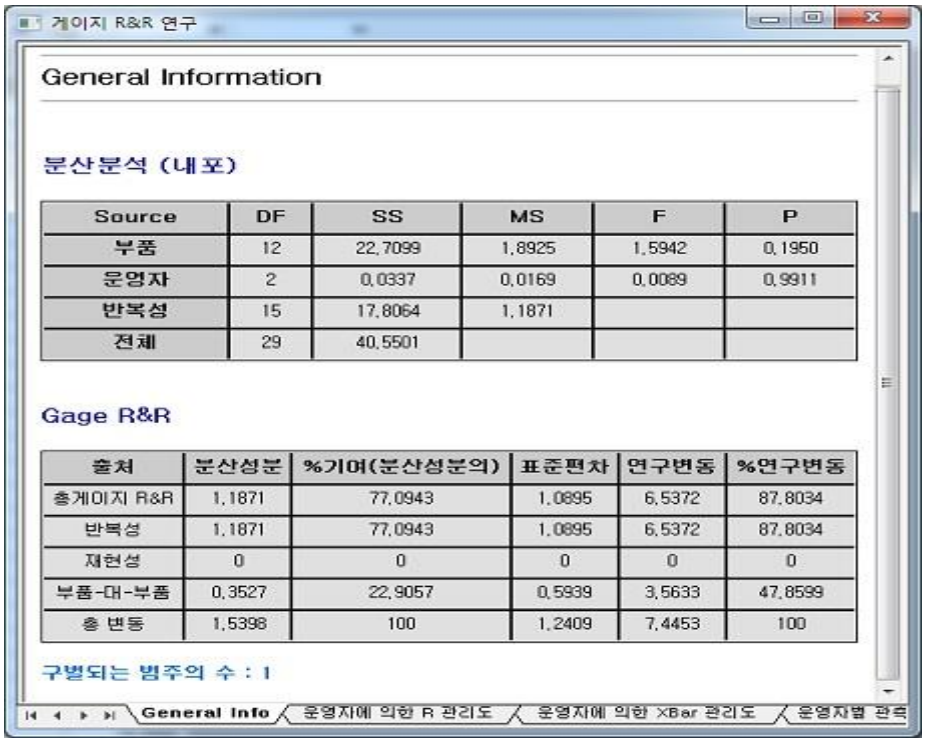
- 운영자 : 부품을 측정한 사람을 구별할 수 있는 데이터가 들어있는 FIELD 를 선택합니다.
- 측정 데이터 : 실험을 통해서 측정된 데이터가 들어있는 FIELD 를 선택합니다.
-



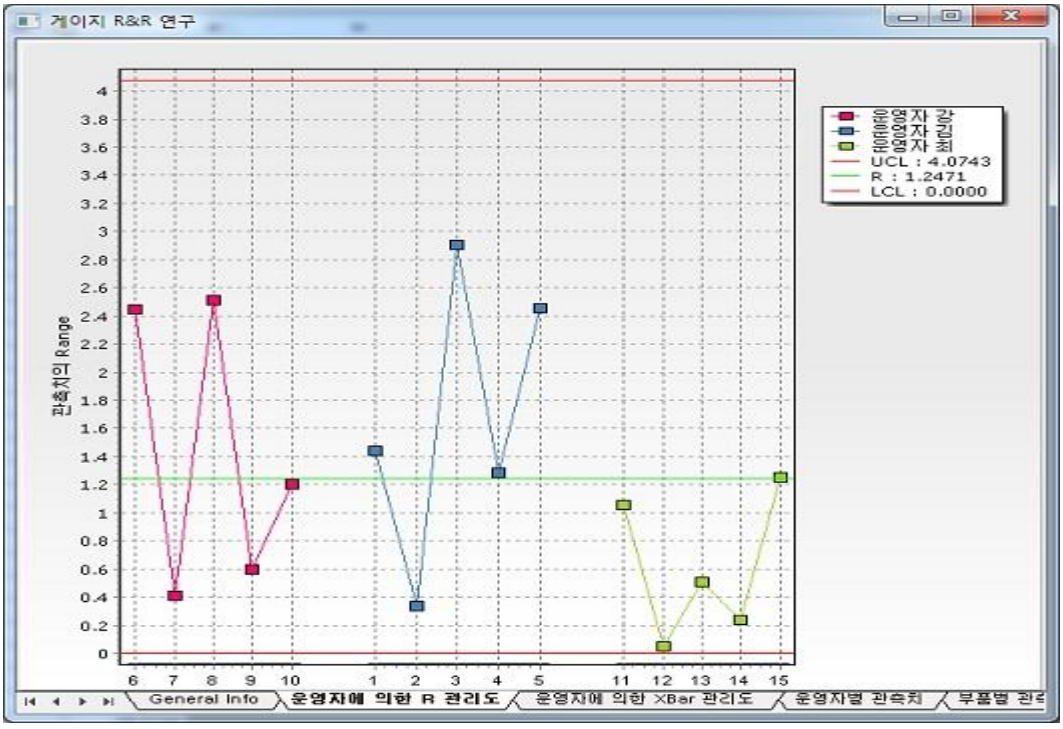
- 연구 변동 : 연구 변동을 구하기 위해 표준 편차에 곱할 계수를 입력합니다.
- 공정 공차 : 경험적으로 알려진 공차를 입력합니다. 이를 통해서 %공차를 계산하게 됩니다.
- 역사적 표준 편차 : 경험적으로 알려진 역사적 표준 편차를 입력합니다. 이를 통해서 %공정을 계산하게 됩니다.

결과

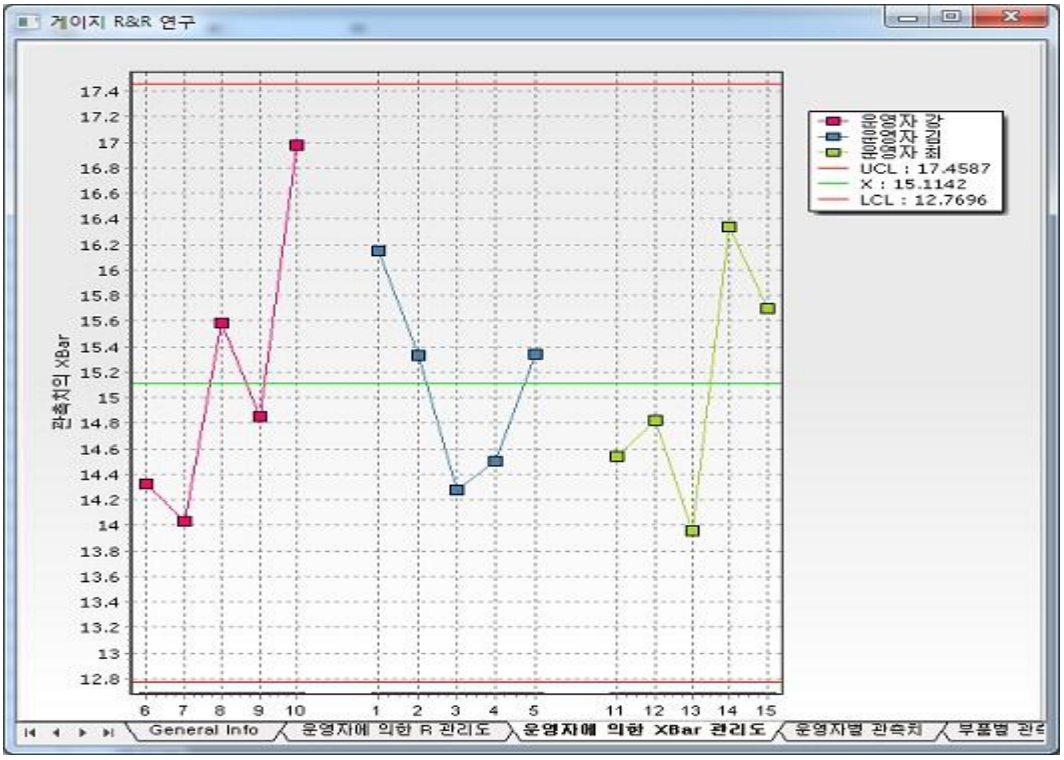
- General Information : 분산 분석 결과와 Gage R&R 분석 결과를 보여줍니다.
  - 구별되는 범주의 수 : 측정시스템을 평가하는 척도로, 그 값이 5 이상이면 측정시스템을 인정해도 좋다고 판단합니다.



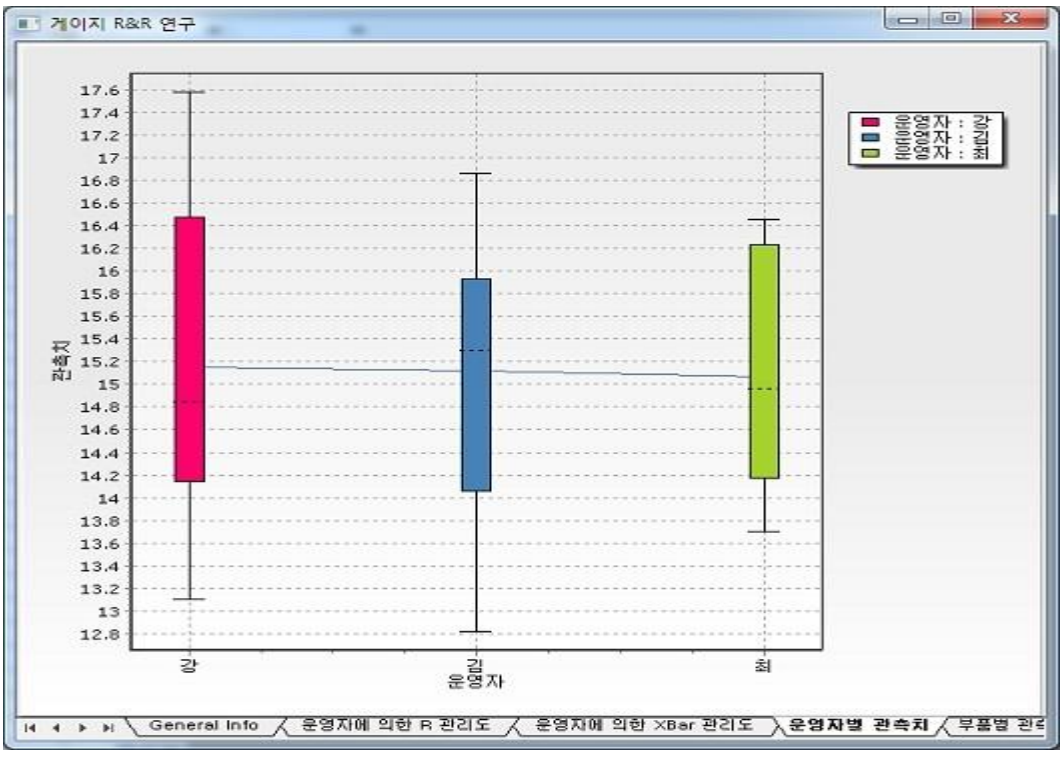
- 운영자에 의한 R 관리도 : 운영자가 각 부품을 여러 번 측정하였을 때 얼마나 차이가 났는지(Range)를 표시해 줍니다.



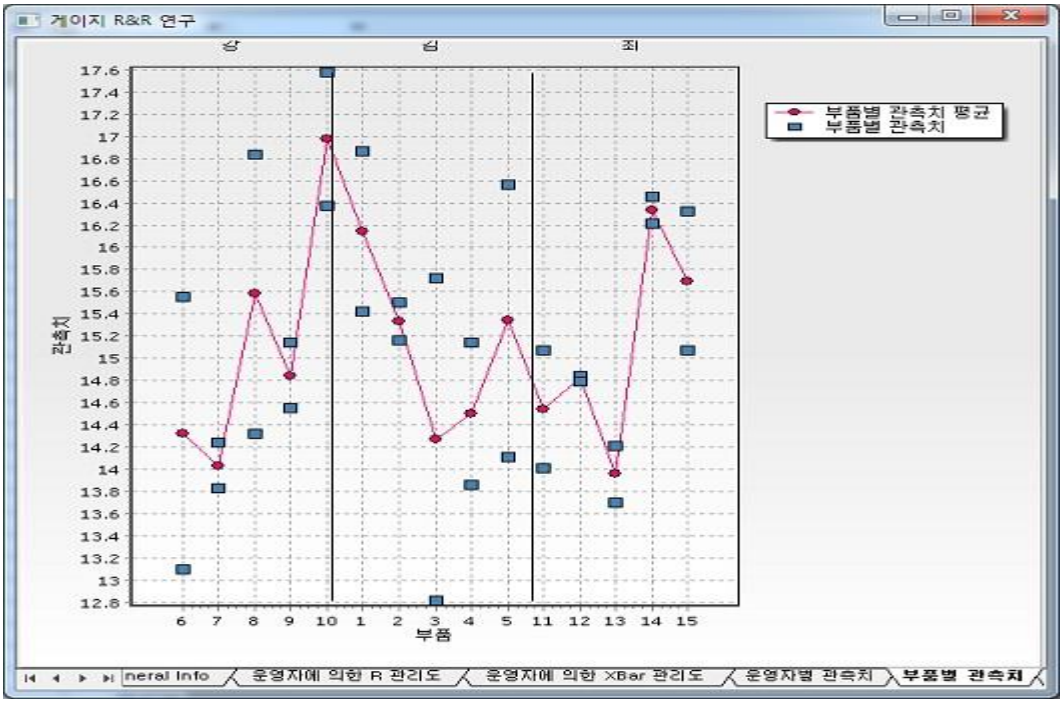
- 운영자에 의한 XBar 관리도 : 운영자가 각 부품을 여러 번 측정하였을 때 각 부품 당 관측치의 평균(XBar)을 표시해 줍니다.



- 운영자 별 관측치 : 부품이 어떠한 것인지를 고려하지 않고 오직 운영자 별로 관측치가 어떠한지를 표시해줍니다. 실선은 관측치의 평균을 나타냅니다.



- 부품별 관측치 : 운영자를 고려하지 않고 오직 부품 별로 관측치가 어떠한지를 표시해줍니다. 실선은 관측치의 평균을 나타냅니다.



- 변동의 성분 : 관측치의 변동이 어디에서 기인하였는지를 표시해줍니다.

5.4 차트 설명

여기에서는 데이터 탐색기에서 지원되는 다양한 차트 기능을 다루도록 하겠습니다.

5.4.1 지원되는 기본차트

지원되는 기본차트

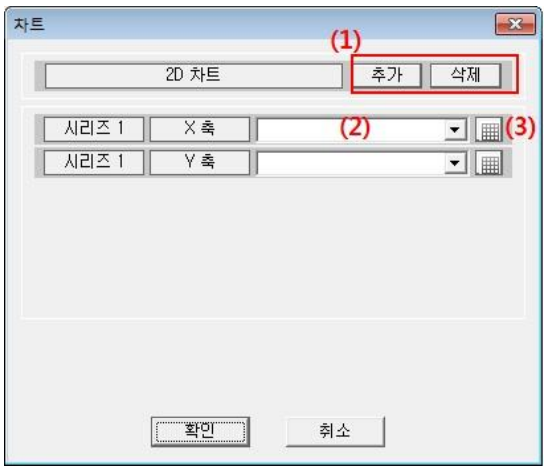
지원되는 차트로는 바차트, 2 차원차트, 3 차원차트, 파레토차트, 박스차트, 매트릭스차트, 파이차트가 있습니다.

실행 방법

[차트] 메뉴에서 원하는 차트를 선택하거나, 툴바에 있는 차트 콤보박스에서 차트 종류를 선택 하시면 됩니다.

차트 옵션 선택 방법

2 차원차트를 예제로 사용하여 차트 옵션에 대해 설명을 하겠습니다. 나머지 차트의 경우도 축의 수만 다를 뿐, 같은 방법으로 선택가능 합니다.



1: 시리즈 추가, 삭제 버튼. 시리즈를 추가하거나, 필요 없는 시리즈를 제거할 수 있습니다.

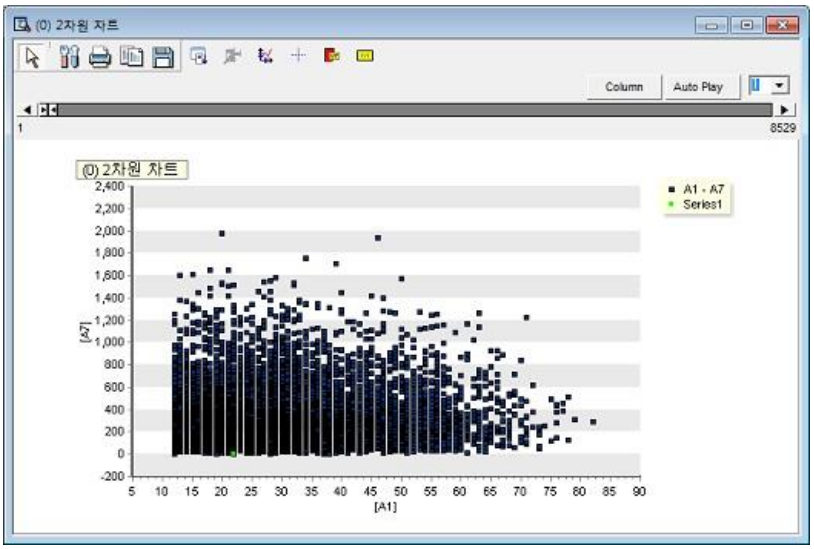
2: 필드 선택 콤보 박스. 차트로 그릴 필드를 콤보 박스 리스트 중에서 선택할 수 있습니다.

3: 필드 선택 버튼. 필드를 선택하는 다른 방법으로, 버튼을 누른 후 데이터 영역의 열 머리글을 클릭하면 필드가 선택됩니다.

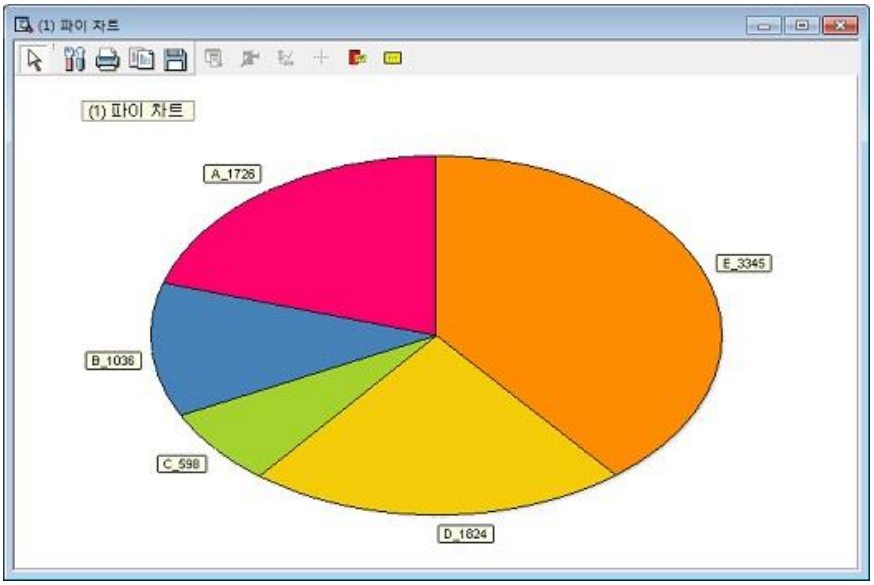
NOTE *이차원 차트의 경우 일반 필드뿐만 아니라, Series Index, Data Index 를 이용하여 차트를 그릴 수 있습니다. 이를 위해서는 X 축 선택 시에 콤보 박스에서 Series Index 또는 Data Index 를 선택하면 됩니다.*

차트 결과 예

- 이차원 차트

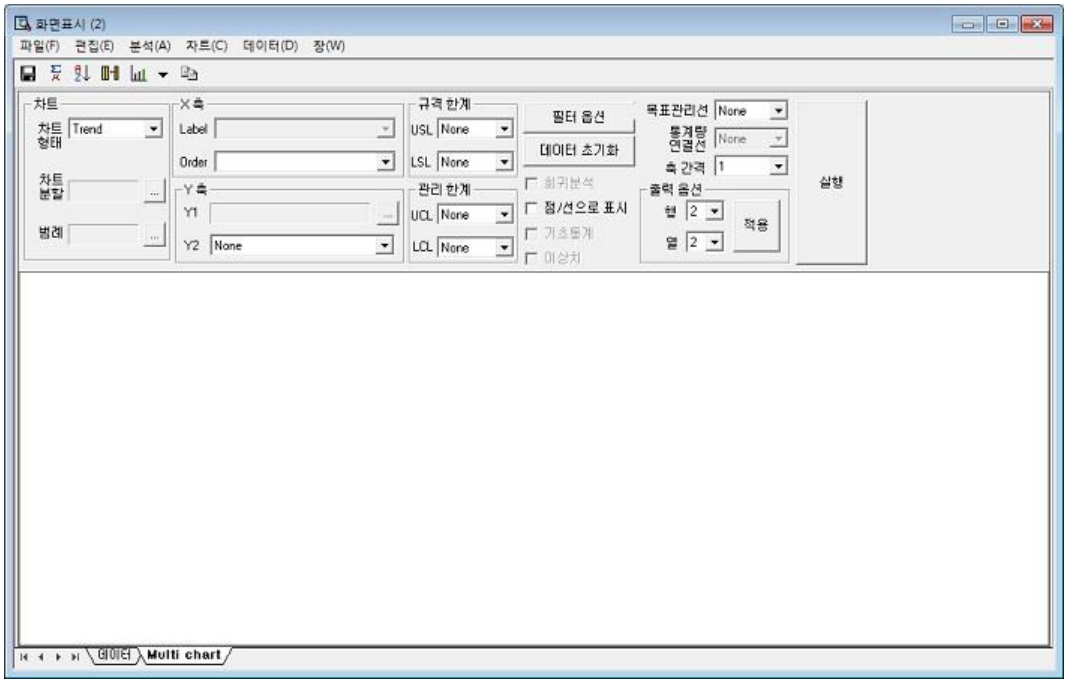


- 파이 차트



5.4.2 멀티 차트

멀티 차트는 여러 변수들의 특징을 한 눈에 보기 쉽도록 고안한 ECMiner™ 고유의 차트입니다. 데이터 탐색기의 차트에서 멀티차트를 선택하면 다음과 같은 화면이 나타납니다.



차트

- **차트형태** : 차트의 형태에는 Trend, Data, Box Plot, Distribution, Correlation 이 있습니다. Trend 는 데이터를 시간 순서대로 표현해 주는 기능이고, Data 는 데이터를 단순하게 산점도로 표현해주는 기능입니다. Box 는 데이터의 분포를 알고자 할 때 그것을 Box Plot 이라는 형태의 차트를 이용하여 표현해주는 기능입니다. Distribution 의 경우 변수의 분포를 정규 분포 확률 플롯으로 보여줍니다. Correlation 을 통해서는 변수간 관계를 쉽게 파악할 수 있습니다.
- **차트분할**: 어떠한 그룹 변수로 차트를 나누어 그릴지를 선택합니다. 만약 어떠한 변수 A 가 있어서 그 값이 1 혹은 -1 을 갖는다고 할 때 A 를 차트분할 변수로 설정하면 차트를 그릴 때 이 변수 값이 같은 것끼리 하나씩 그려주게 됩니다.
- **범례**: 어떠한 그룹 변수로 차트에 찍히는 점을 구분할지를 선택합니다. 만약 어떠한 변수 A 가 있어서 그 값이 1 혹은 -1 을 갖는다고 할 때 A 를 범례 변수로 설정하면 차트를 그릴 때 그 값이 같은 것끼리 같은 색의 점을 찍어줍니다.
- **X 축**

Label: 이는 차트형태가 Data 일 때에만 활성화됩니다. 여기에는 이산형 변수만이 들어가고 이 이산형 변수를 X 축으로 하여 차트를 그립니다.

Order: 이는 Label 에 들어가는 변수를 어떤 순서로 X 축에 나타낼지를 결정하는 변수입니다. 예를 들어 다음과 같이 변수가 나열되어 있다고 합시다.

Order 변수	Label 변수
1	A
3	B
2	C

일단 Order 변수에 의해서 정렬을 한 후 Label 변수를 순차적으로 읽는 것입니다. 즉 A->C->B 의 순서가 되는 것입니다. X 축에 이와 같은 순서로 Data Chart 가 그려집니다.

- **Y 축**

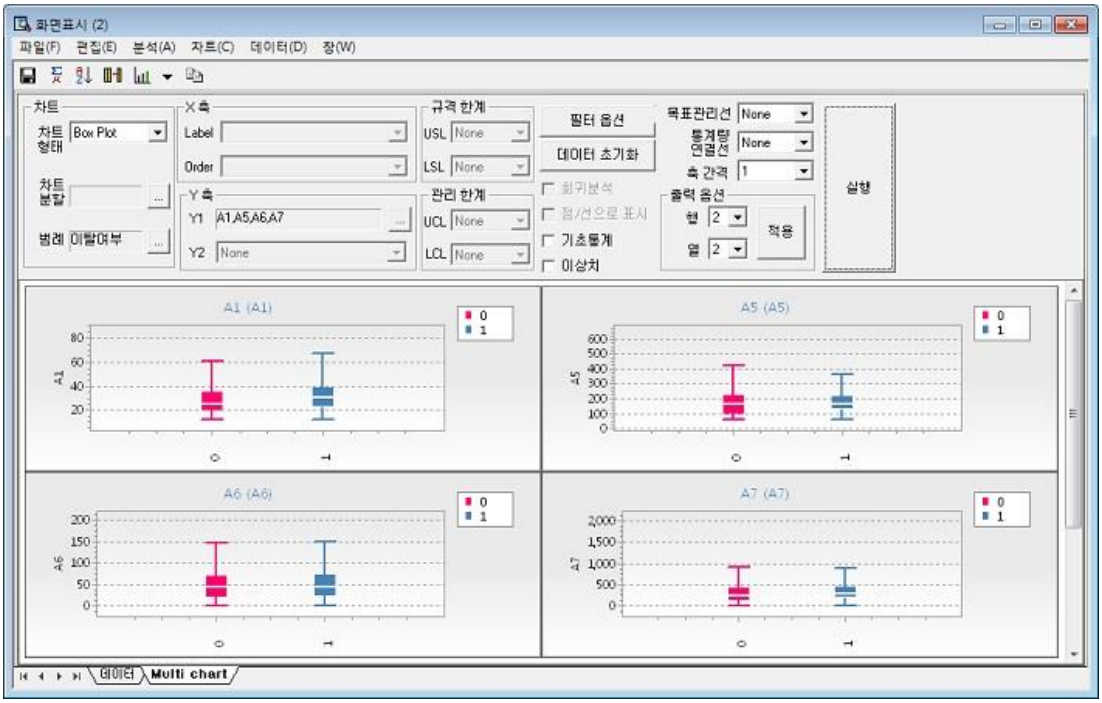
Y1: Y1 의 경우 왼쪽 Y 축에 어떤 변수가 들어갈지를 결정합니다. 복수 선택 가능이 가능합니다.

Y2: Y2 의 경우 오른쪽 Y 축에 어떤 변수가 들어갈지를 결정합니다.

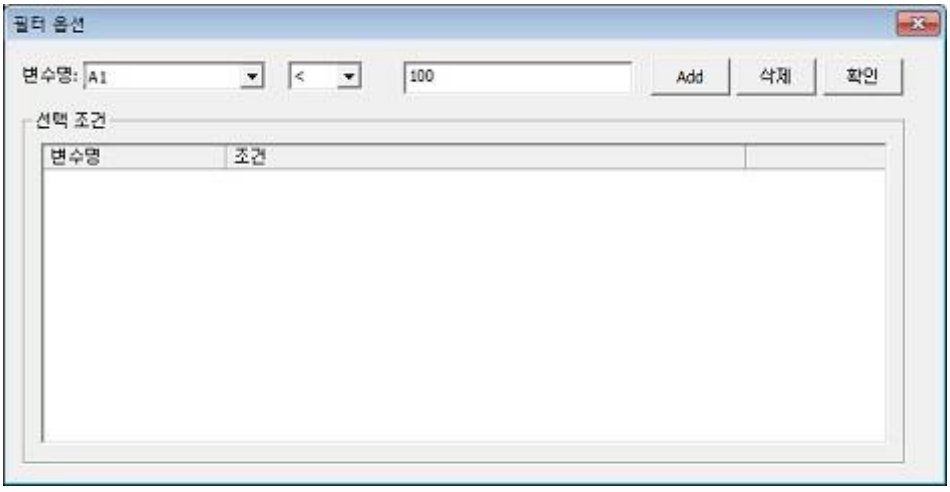
X, Y: 차트형태로 Correlation 을 선택했을 때 어떤 변수를 X 축으로 사용하고, 어떤 변수를 Y 축으로 사용할 것인지를 결정합니다. 이 때 Y 축은 복수 선택이 가능합니다.

체크 박스 관련

- **회귀분석**: 차트형태를 Correlation 으로 선택할 경우, X 축에 해당하는 변수를 독립변수로 사용하고 Y 축에 해당하는 변수를 종속변수로 사용하여 Regression 을 하는 기능 수행에 대한 여부를 결정하는 옵션입니다.
- **점/선으로 표시**: 차트형태를 Trend 혹은 Data 로 사용할 경우, 점과 점 사이를 Line 으로 연결할지에 대한 여부를 결정하는 옵션입니다.
- **기초통계**: 차트형태를 Box 로 선택할 경우, Median, Standard Deviation, Average, Maximum, Minimum, Range 와 같은 간단한 통계량을 표시할지에 대한 여부를 결정하는 옵션입니다.
- **이상치**: Outlier 를 표시할지에 대한 여부를 결정하는 옵션입니다. 다음은 '기초통계'와 '이상치'를 선택하였을 때의 Box Chart 화면입니다.



▪ **필터 옵션**



필터 옵션은 현재 데이터 테이블의 데이터 중에서 어떠한 조건을 만족하는 데이터만을 쓰고 싶을 때 이용하는 옵션입니다. 사용자는 변수를 선택하고 그 변수가 사용하고자 하는 부등호와 그에 대응하는 값을 적어줍니다. 그리고 Add 단추를 누르면 아래의 선택 조건에 작성한 목록이 등록됩니다. 등록한 여러 조건 중에서 사용하고자 하는 조건을 체크한 후 확인을 누르면 이 후부터는 원하는 데이터만을 사용할 수 있습니다.

▪ 데이터 초기화

원본 데이터로 바꾸어주는 기능입니다.

▪ 기타 옵션

출력옵션(행): 한 화면에 가로로 몇 개의 차트를 보여줄지 설정합니다.

출력옵션(열): 한 화면에 세로로 몇 개의 차트를 보여줄지 설정합니다.

축 간격: X 축에 Tick 이 몇 번 단위로 쓰여질지 설정합니다.

관리 한계: UCL, LCL 이 각각 어떤 변수를 기준으로 Control Line 을 그릴지를 결정합니다.

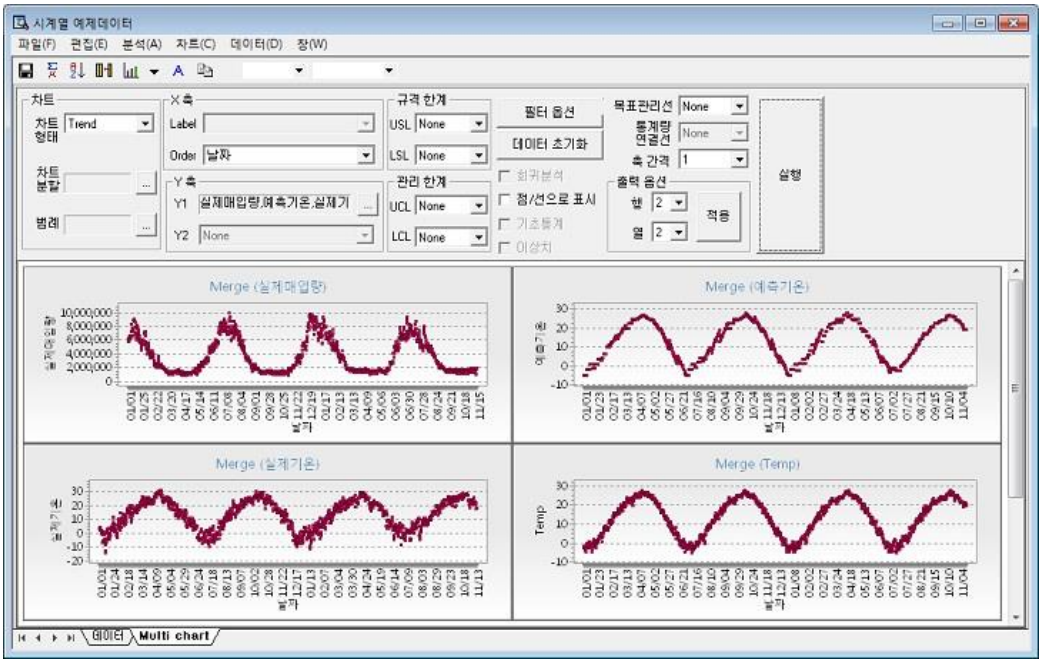
규격 한계: USL 과 LSL 이 있는데 USL 과 LSL 은 값이 같은 Field 를 사용해야 합니다.

목표 관리선: 차트형태가 Trend, Data, Box 일 때 선택한 변수에 해당하는 Target Line 을 그려지는 모든 Chart 에 그려주는 옵션입니다.

통계량 연결선: 범례에 의해서 각 차트당 두 개 이상의 Box Plot 이 그려지는 경우 Box Link 의 옵션에 해당하는 값을 이어주는 기능입니다. 예를 들어 Mean, Average, Max, Min 의 옵션 중 Average 를 선택하면 각 Box Plot 에서 평균에 해당하는 위치의 점들을 이어줍니다.

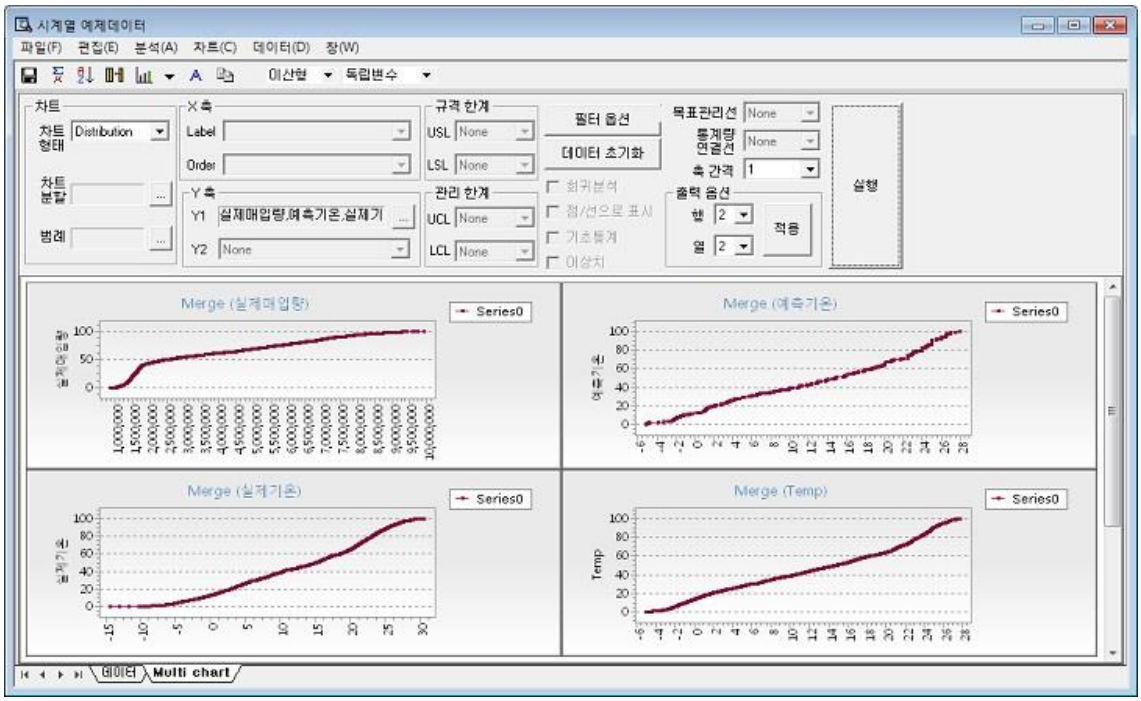
사용 예시

▪ 차트형태로 Trend 를 사용하였을 때의 화면

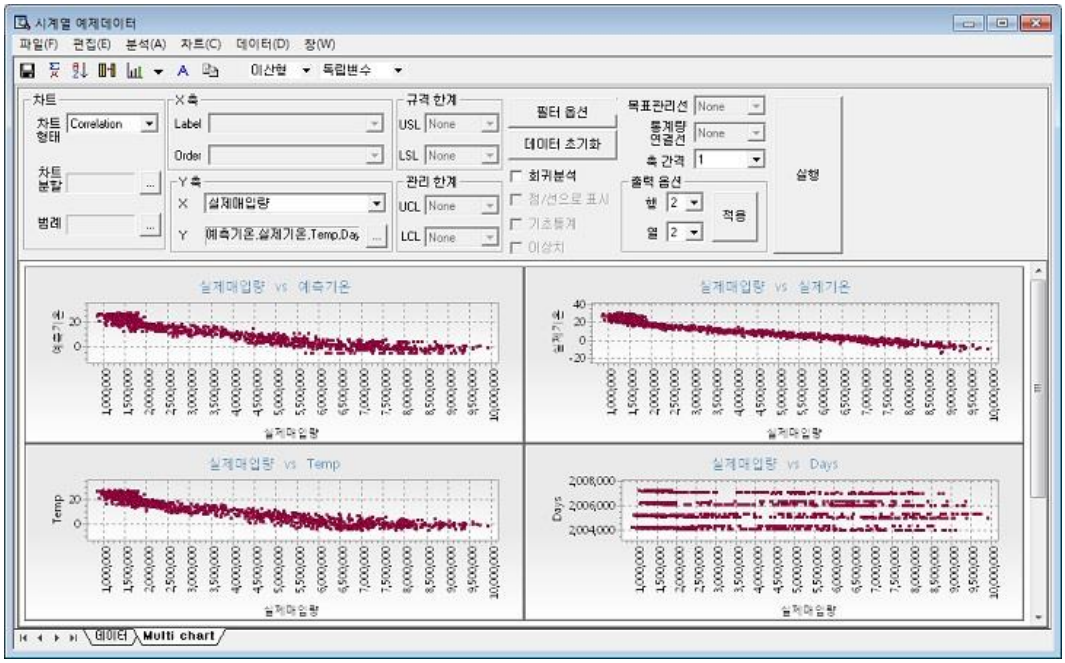


▪ 차트형태로 Box 를 사용하였을 때의 화면. 범례 변수에 의해 구별되는 데이터 별로 Box Plot 이 그려집니다.(Legend 변수는 사용하지 않을 수도 있습니다.)

▪ 차트형태로 Distribution 을 사용하였을 때의 화면



▪ 차트형태로 Correlation 을 사용하였을 때의 화면



5.5 데이터

5.5.1 데이터 정렬

데이터 탐색기에서는 다중 정렬을 지원합니다. 기능적으로 노드 창의 **SORT 노드**와 같은 기능을 수행합니다.

실행 방법

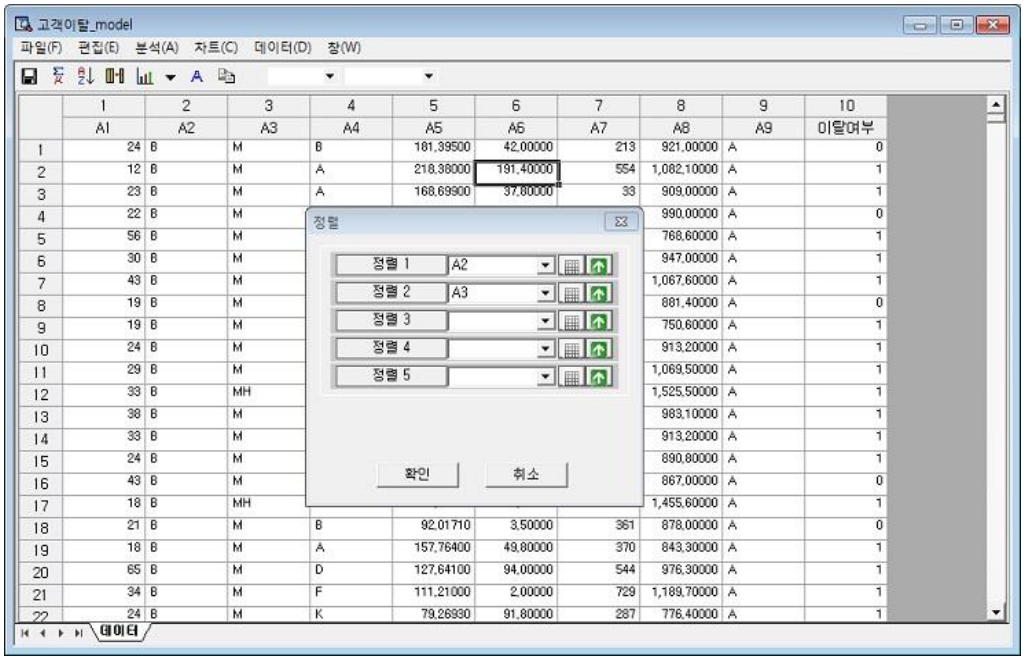
[데이터] – [정렬]를 선택하거나, 툴바에 있는 **정렬** 아이콘을 선택 하면 됩니다.

정렬 기능

정렬할 필드 선택: 콤보 박스에서 선택하거나, 필드 선택 아이콘(콤보 박스 오른쪽 아이콘)을 누른 후에, 데이터 Grid 상에서 직접 필드를 선택합니다.

정렬 방향: 방향 아이콘을 선택하여, 오름 차순 또는 내림 차순으로 정렬할 수 있습니다.

다중 정렬: 여러 필드를 선택하면, 다중 정렬이 됩니다.



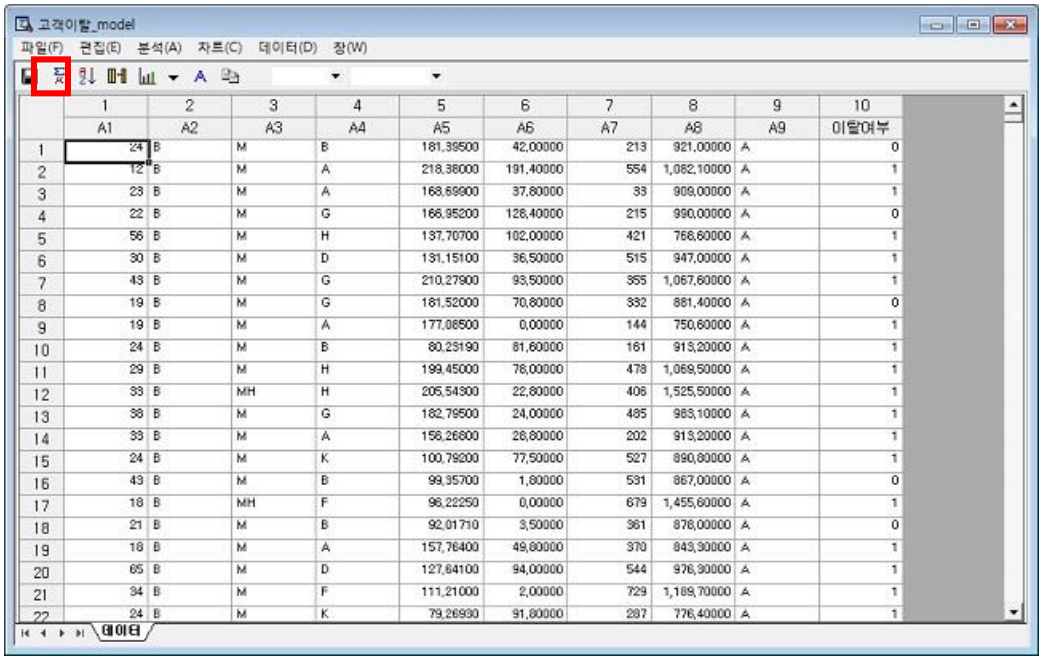
5.5.2 파생변수

데이터 탐색기에서는 파생 변수 생성을 지원합니다. 기능적으로 노드창의 **파생변수 노드**와 같은 기능을 수행합니다.

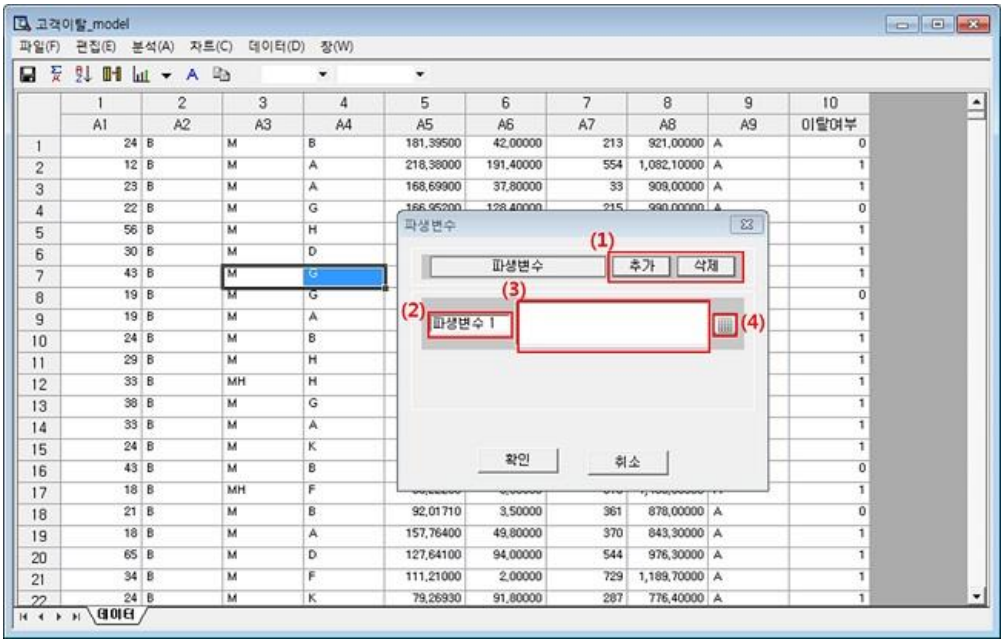
실행 방법

[데이터] – [파생변수]를 선택하거나, 또는 툴바의 **파생변수** 아이콘을 클릭하면 파생변수 다이얼로그가 나타납니다.

예제 데이터



▪ 파생변수 다이얼로그



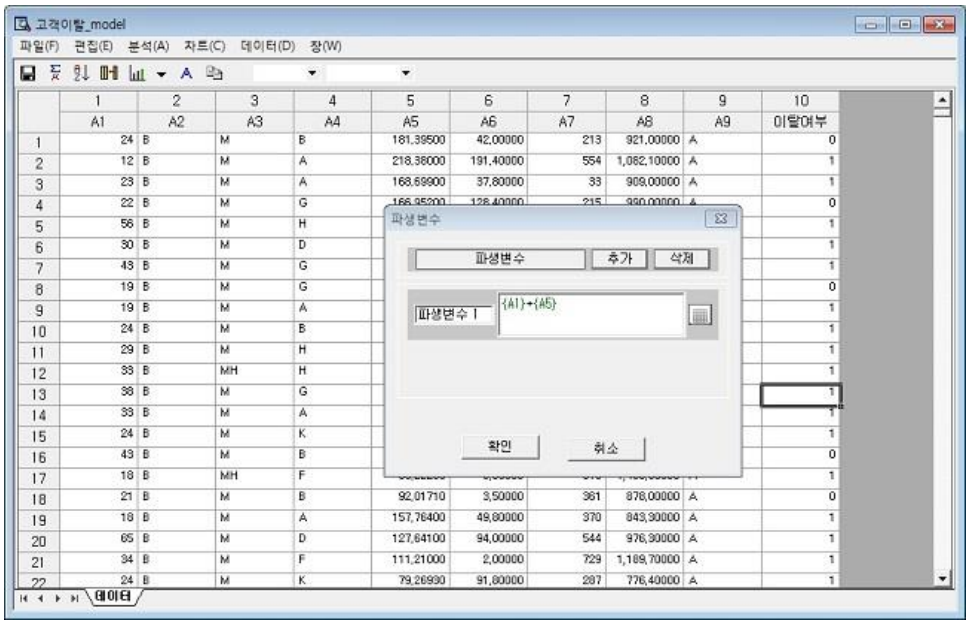
1: 파생 변수를 여러 개 만들 때, 사용합니다. 추가 버튼을 누르면 파생변수를 추가할 수 있는 리스트가 추가되고, 삭제 버튼을 누르면 마지막 파생변수 리스트가 제거됩니다. 만약 파생 변수 리스트 원소가 **1** 개일 때는 삭제 버튼을 눌러도 더 이상 리스트 원소가 줄어들지 않습니다.

2: 추가될 파생변수의 변수명을 넣을 수 있는 편집창입니다.

3: 파생 변수를 만들 규칙을 입력하는 편집창입니다. 편집 규칙은 **파생변수노드** 에서와 같습니다. 좀 더 쉽게 편집을 하기 위해서는 수식 편집기를 이용하면 됩니다.

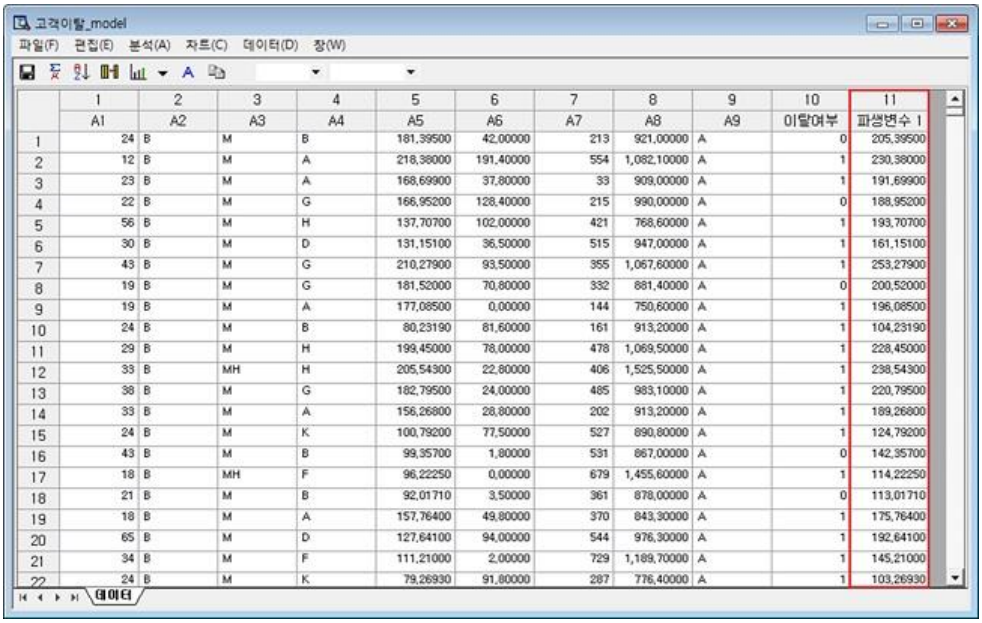
4: **수식 편집기 버튼**입니다. 버튼을 누르면, **수식 편집기**가 나타나고, 여기에서 편집 할 수 있습니다.

▪ 파생 변수 수식 편집



▪ 파생 변수 결과

앞에서 추가, 편집된 파생 변수가 데이터로 추가되었습니다.



	1	2	3	4	5	6	7	8	9	10	11
	A1	A2	A3	A4	A5	A6	A7	A8	A9	이탈여부	파생변수 1
1	24	B	M	B	181,39500	42,00000	213	921,00000	A	0	205,39500
2	12	B	M	A	218,38000	191,40000	554	1,082,10000	A	1	230,38000
3	23	B	M	A	168,69900	37,80000	33	909,00000	A	1	191,69900
4	22	B	M	G	166,95200	128,40000	215	990,00000	A	0	188,95200
5	56	B	M	H	137,70700	102,00000	421	768,60000	A	1	193,70700
6	30	B	M	D	131,15100	36,50000	515	947,00000	A	1	161,15100
7	43	B	M	G	210,27900	93,50000	355	1,067,60000	A	1	253,27900
8	19	B	M	G	181,52000	70,80000	332	881,40000	A	0	200,52000
9	19	B	M	A	177,08500	0,00000	144	750,60000	A	1	196,08500
10	24	B	M	B	80,23190	81,60000	161	913,20000	A	1	104,23190
11	29	B	M	H	199,45000	78,00000	478	1,069,50000	A	1	228,45000
12	33	B	MH	H	205,54300	22,80000	406	1,525,50000	A	1	238,54300
13	38	B	M	G	182,79500	24,00000	485	983,10000	A	1	220,79500
14	33	B	M	A	156,26800	28,80000	202	913,20000	A	1	189,26800
15	24	B	M	K	100,79200	77,50000	527	890,80000	A	1	124,79200
16	43	B	M	B	99,35700	1,80000	531	867,00000	A	0	142,35700
17	18	B	MH	F	96,22250	0,00000	679	1,455,60000	A	1	114,22250
18	21	B	M	B	92,01710	3,50000	361	878,00000	A	0	113,01710
19	18	B	M	A	157,76400	49,80000	370	843,30000	A	1	175,76400
20	65	B	M	D	127,64100	94,00000	544	976,30000	A	1	192,64100
21	34	B	M	F	111,21000	2,00000	729	1,189,70000	A	1	145,21000
22	24	B	M	K	79,26930	91,80000	287	776,40000	A	1	103,26930

5.5.3 적용

데이터 탐색기상의 여러 전처리 과정이 **프로젝트 상에 적용**되는 기능입니다. 실행된 전처리 작업에 따라, 필터 노드, 선택 노드, 파생필드 노드, 정렬 노드 등의 노드들이 생성되어 스트림이 자동으로 구성됩니다.

실행 방법

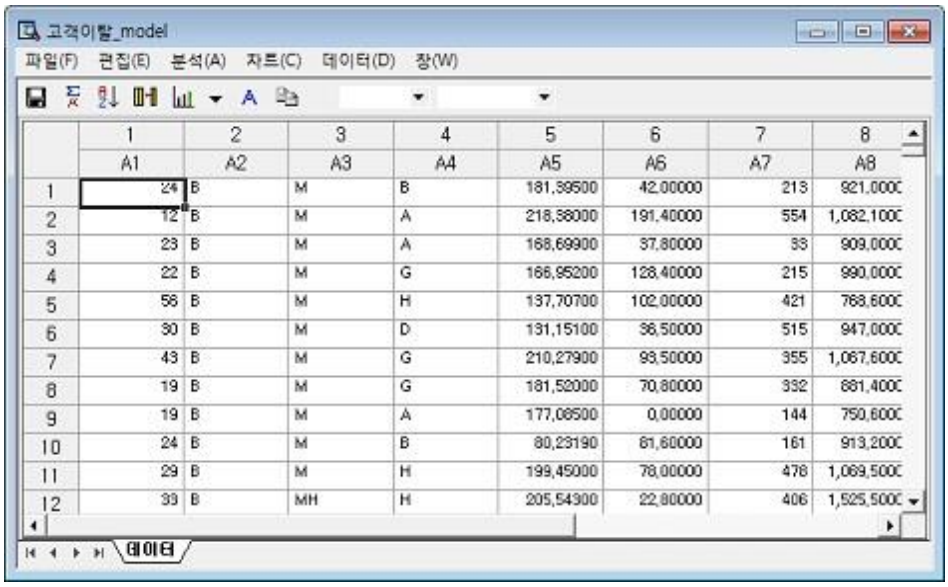
데이터 탐색기에 열 삭제, 행 삭제, 파생 필드 추가, 정렬 등의 전처리 작업을 한 후, **[데이터] – [적용]**를 선택하거나, 툴바 메뉴의 **적용**아이콘을 선택하면 됩니다.

예제

각각의 전처리 경우에 대해, 예를 들어 설명하도록 하겠습니다.

1. 예제 데이터

이후의 예제는 모두 아래 데이터를 예를 들어 설명하겠습니다.

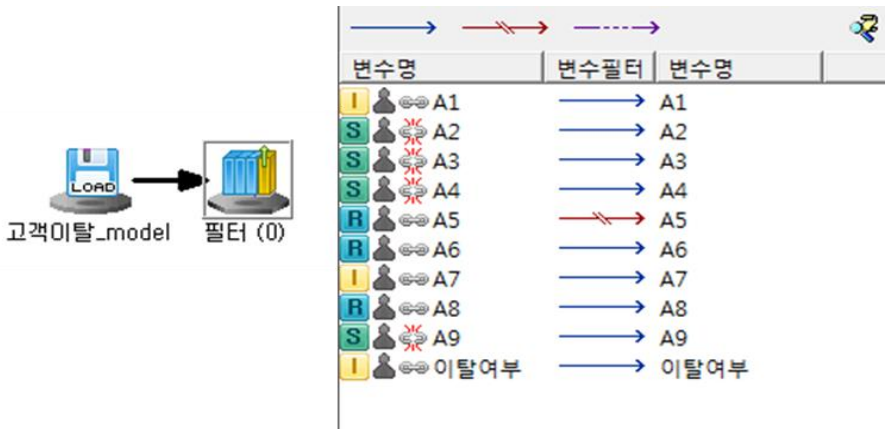


	1	2	3	4	5	6	7	8
	A1	A2	A3	A4	A5	A6	A7	A8
1	24	B	M	B	181,39500	42,00000	213	921,0000
2	12	B	M	A	218,38000	191,40000	554	1,082,1000
3	23	B	M	A	168,69900	37,80000	33	909,0000
4	22	B	M	G	166,95200	128,40000	215	990,0000
5	56	B	M	H	137,70700	102,00000	421	768,6000
6	30	B	M	D	131,15100	36,50000	515	947,0000
7	43	B	M	G	210,27900	93,50000	355	1,067,6000
8	19	B	M	G	181,52000	70,80000	332	881,4000
9	19	B	M	A	177,08500	0,00000	144	750,6000
10	24	B	M	B	80,23190	61,60000	161	913,2000
11	29	B	M	H	199,45000	78,00000	478	1,069,5000
12	33	B	MH	H	205,54300	22,80000	406	1,525,5000

2. 열삭제

열(A5) 삭제 후 프로젝트 변화

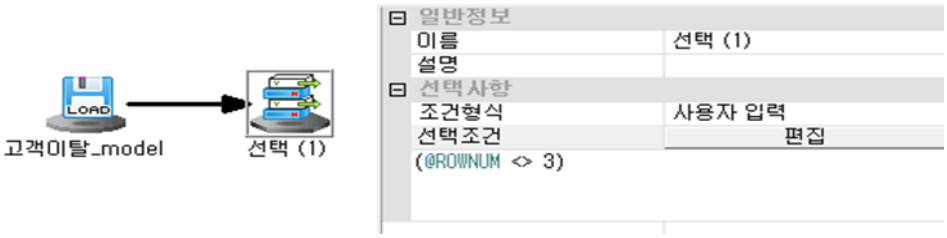
원래 단독으로 존재하던 파일입력 노드에 필터 노드가 **스트림**으로 연결된 것을 볼 수 있습니다. 또한 추가된 필터 노드의 **속성**창의 **변수필터**부분을 보면 A5 필드가 제거된 형태로 나타나는 것을 볼 수 있습니다.



3. 행삭제

행(세번째 행) 삭제 후 프로젝트 변화

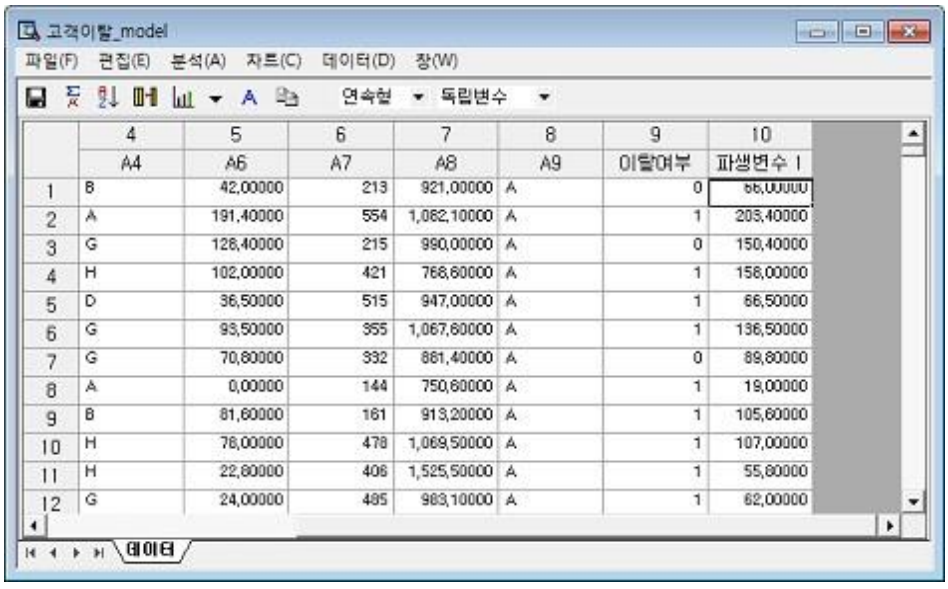
원래 단독으로 존재하던 파일입력노드에 선택 노드가 **스트림**으로 연결된 것을 볼 수 있습니다. 또한 추가된 선택 노드의 **속성**창의 **선택조건**부분을 보면 **(@ROWNUM <> 3)** 조건문이 추가된 것을 볼 수 있습니다.



NOTE *(@ROWNUM <> 3)는 "행 인덱스가 3 이 아닌 행"을 의미하는 조건문입니다.*

4. 파생필드

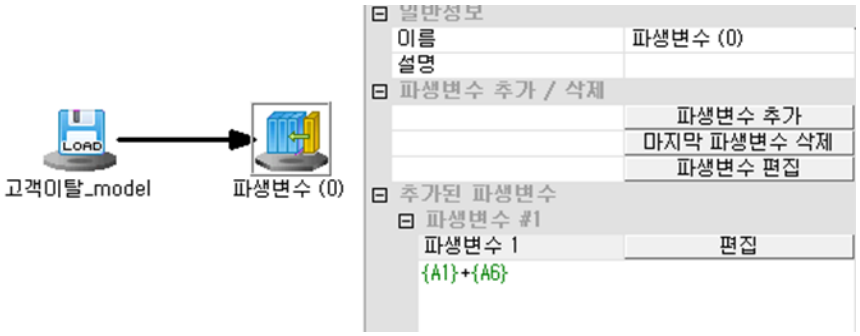
파생 필드 추가 후 데이터 변화: A1 과 A6 의 합을 새로운 값으로 가지는 필드를 새로 추가하였습니다.



	4	5	6	7	8	9	10
	A4	A6	A7	A8	A9	이탈여부	파생변수 1
1	B	42,00000	213	921,00000	A	0	66,00000
2	A	191,40000	554	1,062,10000	A	1	203,40000
3	G	128,40000	215	990,00000	A	0	150,40000
4	H	102,00000	421	768,60000	A	1	156,00000
5	D	36,50000	515	947,00000	A	1	66,50000
6	G	93,50000	355	1,067,60000	A	1	136,50000
7	G	70,80000	332	861,40000	A	0	89,80000
8	A	0,00000	144	750,60000	A	1	19,00000
9	B	81,60000	161	913,20000	A	1	105,60000
10	H	78,00000	478	1,069,50000	A	1	107,00000
11	H	22,80000	406	1,525,50000	A	1	55,80000
12	G	24,00000	485	983,10000	A	1	62,00000

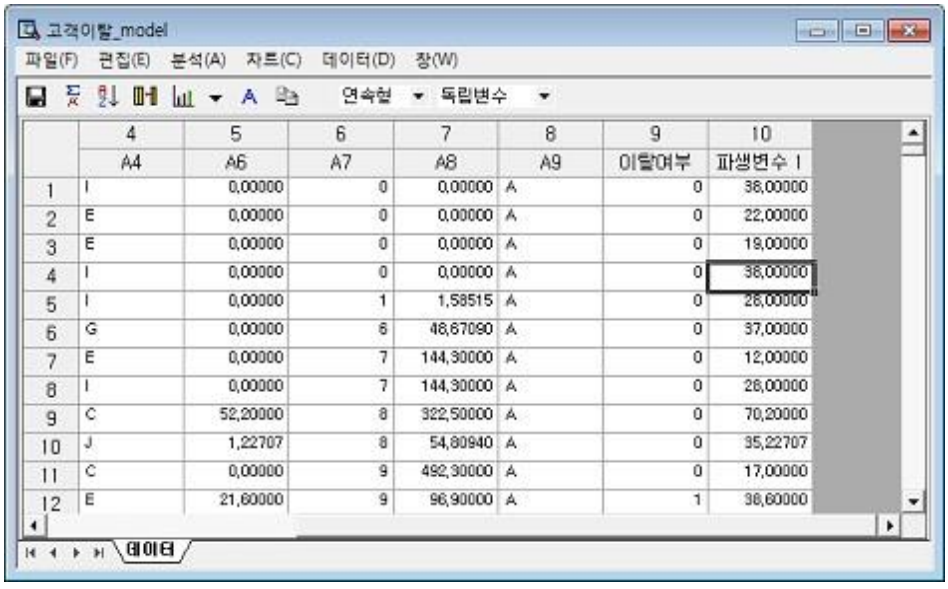
파생 필드 추가 후 프로젝트 변화

원래 단독으로 존재하던 파일입력 노드에 파생필드 노드가 **스트림**으로 연결된 것을 볼 수 있습니다. 또한 추가된 파생필드 노드의 **속성**창의 **파생변수**부분을 보면 **{A1} + {A6}** 조건문이 추가된 것을 볼 수 있습니다.



5. 정렬

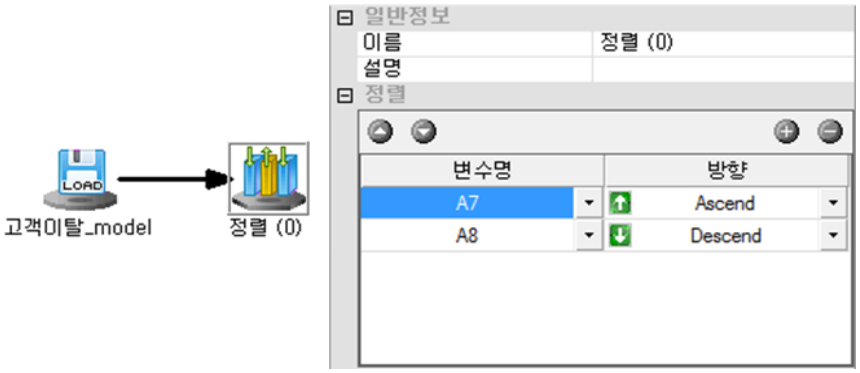
정렬 후 데이터 변화: A7 을 기준으로 오름차순, A8 을 기준으로 내림차순으로 다중 정렬한 결과입니다.



	4	5	6	7	8	9	10
	A4	A6	A7	A8	A9	이탈여부	파생변수 1
1	I	0,00000	0	0,00000	A	0	36,00000
2	E	0,00000	0	0,00000	A	0	22,00000
3	E	0,00000	0	0,00000	A	0	19,00000
4	I	0,00000	0	0,00000	A	0	36,00000
5	I	0,00000	1	1,58515	A	0	26,00000
6	G	0,00000	6	48,67090	A	0	37,00000
7	E	0,00000	7	144,30000	A	0	12,00000
8	I	0,00000	7	144,30000	A	0	26,00000
9	C	52,20000	8	322,50000	A	0	70,20000
10	J	1,22707	8	54,80940	A	0	35,22707
11	C	0,00000	9	492,30000	A	0	17,00000
12	E	21,60000	9	96,90000	A	1	38,60000

정렬 후 프로젝트 변화

원래 단독으로 존재하던 파일입력 노드에 정렬 노드가 **스트림**으로 연결된 것을 볼 수 있습니다. 또한 추가된 정렬 노드의 **속성**창의 **정렬**부분을 보면 A7 필드와 A8 필드에 대해 각각 오름 차순과 내림 차순으로 정렬하는 선택 리스트가 추가된 것을 볼 수 있습니다.



6. 여러 전처리 다중 적용

전처리 결과의 **적용**은 위의 예제처럼 하나씩 할 수도 있지만, 한번에 여러 전처리 결과를 **적용**할 수 도 있습니다.

여러 전처리 적용 후 프로젝트 변화

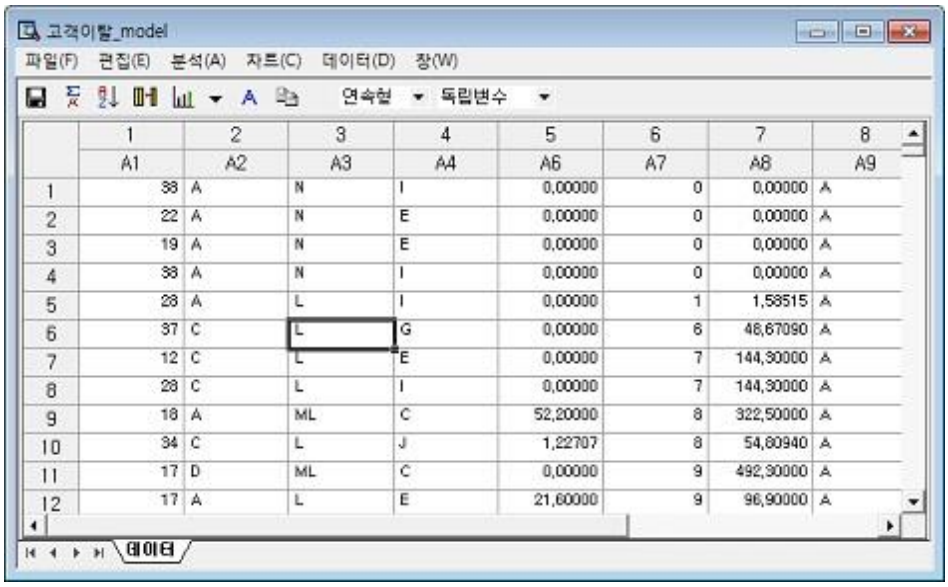
5.5.4 필터

데이터 탐색기에서는 필드 필터링을 지원합니다. 기능적으로 노드창의 **필터** 노드와 같은 기능을 수행합니다.

실행 방법

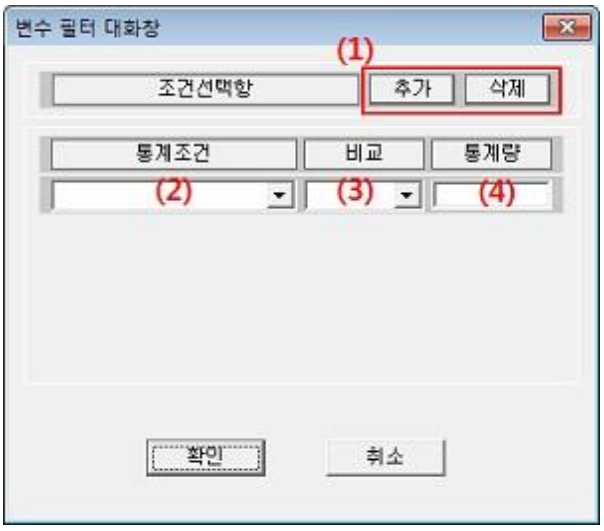
[데이터] – [필터]를 선택하면 필터 다이얼로그가 나타납니다.

예제 데이터



	1	2	3	4	5	6	7	8
	A1	A2	A3	A4	A6	A7	A8	A9
1	38	A	N	I	0,00000	0	0,00000	A
2	22	A	N	E	0,00000	0	0,00000	A
3	19	A	N	E	0,00000	0	0,00000	A
4	38	A	N	I	0,00000	0	0,00000	A
5	28	A	L	I	0,00000	1	1,58515	A
6	37	C	L	G	0,00000	6	46,67090	A
7	12	C	L	E	0,00000	7	144,30000	A
8	28	C	L	I	0,00000	7	144,30000	A
9	18	A	ML	C	52,20000	8	322,50000	A
10	34	C	L	J	1,22707	8	54,60940	A
11	17	D	ML	C	0,00000	9	492,30000	A
12	17	A	L	E	21,60000	9	96,90000	A

필터 다이얼로그



1 : 필터 조건을 여러 개 만들 때, 사용합니다. 추가 버튼을 누르면 필터링 할 수 있는 조건이 추가되고, 삭제 버튼을 누르면 마지막 조건이 제거됩니다. 만약 필터 조건 수가 1 개뿐일 때는 삭제 버튼을 눌러도 더 이상 줄어들지 않습니다.

2 : 통계 조건을 선택합니다. 통계 조건에는 표준편차가 있습니다.

3 : >, =, < 등의 조건을 선택합니다.

4 : 비교 대상이 되는 통계값을 입력합니다.

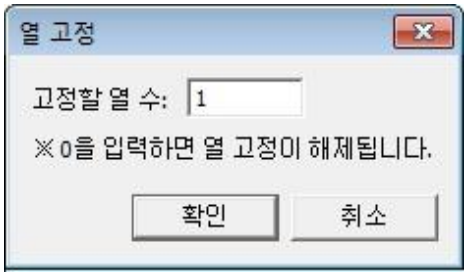
위의 과정을 거치면, 조건에 해당하는 변수는 삭제되며, 필터링이 완료됩니다.

5.5.5 관심 변수 고정

데이터의 열이 많을수록 수평 스크롤을 이용해서 변수를 살펴보는 일이 많아집니다. 이 때 특정 열을 다른 열과 비교하려고 할 경우 특정 열이 다른 열과 멀리 떨어져 있으면 눈으로 값을 비교하는 것은 불가능합니다. 이를 위해서 특정 변수를 고정할 경우 그 스크롤을 움직여도 그 변수 열은 움직이지 않도록 하는 기능이 관심 변수 고정 기능입니다.

실행 방법

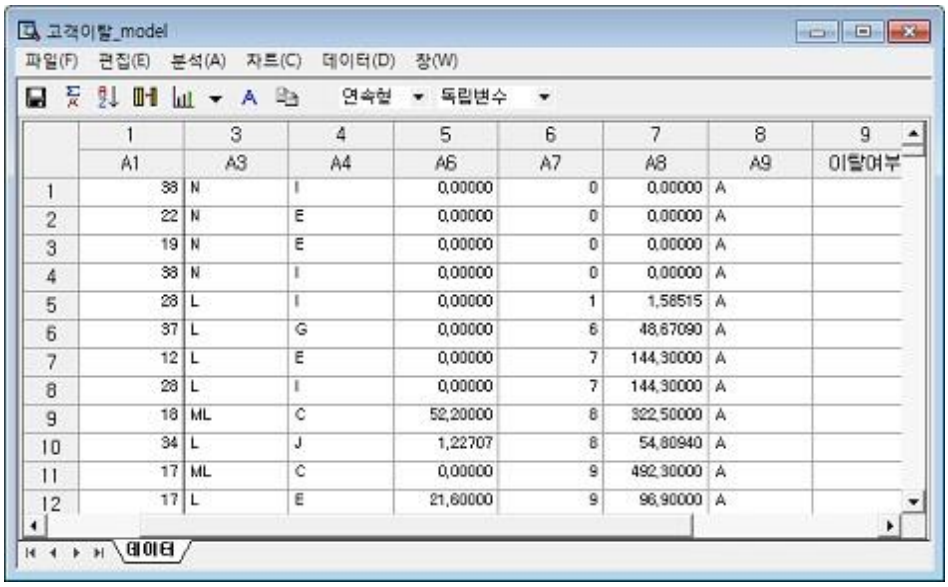
[데이터]-> [관심 변수 고정] 메뉴를 선택하면 관심 변수 고정 다이얼로그가 나타납니다.



고정할 열 개수: 첫번째 열부터 몇 번째 열까지 고정할지 선택합니다

결과

다음과 같이 FIELD 1 을 고정하면 수평 스크롤 바를 오른 쪽으로 이동했을 경우에도 FIELD1 이 왼쪽에 고정되어 있음을 확인할 수 있습니다.



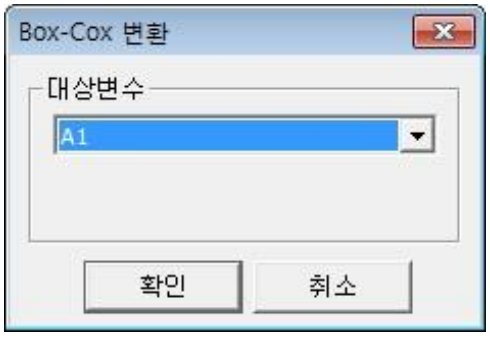
	1	3	4	5	6	7	8	9
	A1	A3	A4	A5	A7	A8	A9	이탈여부
1	38	N	I	0,00000	0	0,00000	A	
2	22	N	E	0,00000	0	0,00000	A	
3	19	N	E	0,00000	0	0,00000	A	
4	38	N	I	0,00000	0	0,00000	A	
5	28	L	I	0,00000	1	1,58515	A	
6	37	L	G	0,00000	6	48,67090	A	
7	12	L	E	0,00000	7	144,30000	A	
8	28	L	I	0,00000	7	144,30000	A	
9	18	ML	C	52,20000	8	322,50000	A	
10	34	L	J	1,22707	8	54,80940	A	
11	17	ML	C	0,00000	9	492,30000	A	
12	17	L	E	21,60000	9	96,90000	A	

5.5.6 Box-Cox 변환

관리도에 적용하기 위한 전제조건으로는 데이터가 정규분포를 따라야 합니다.. 그러나 많은 데이터가 정규분포를 따르기 어렵습니다. Box-Cox 변환은 정규분포를 따르지 않는 관측치들을 관리도에 적용하기 위하여 정규분포를 따르도록 변환 수식을 제공합니다.

실행 방법

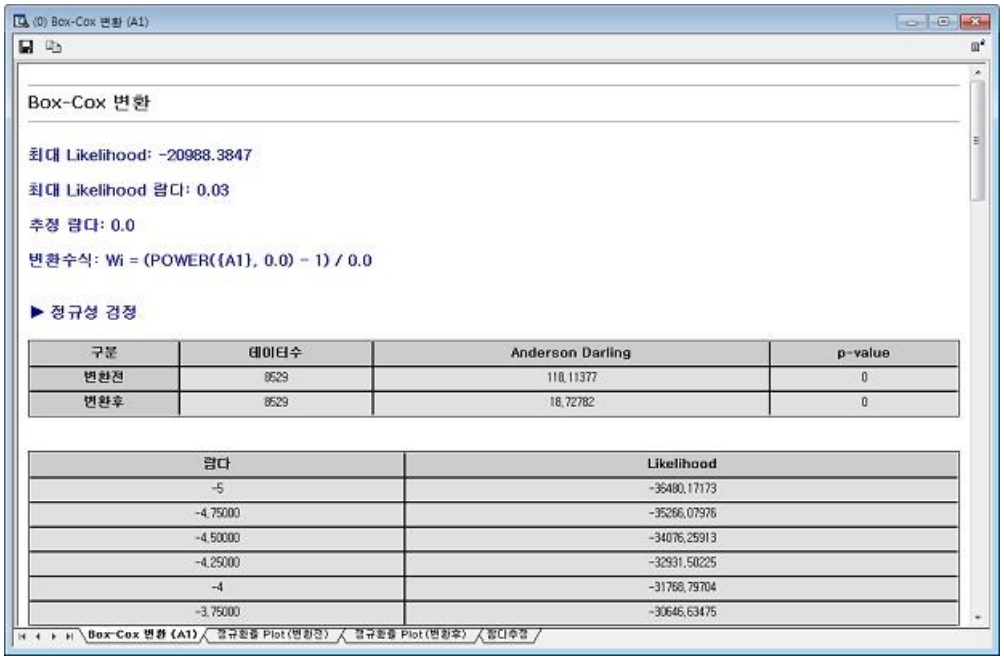
[데이터] – [Box-Cox 변환]을 선택하면 다음과 같은 Box-Cox 변환 다이얼로그가 나타납니다.



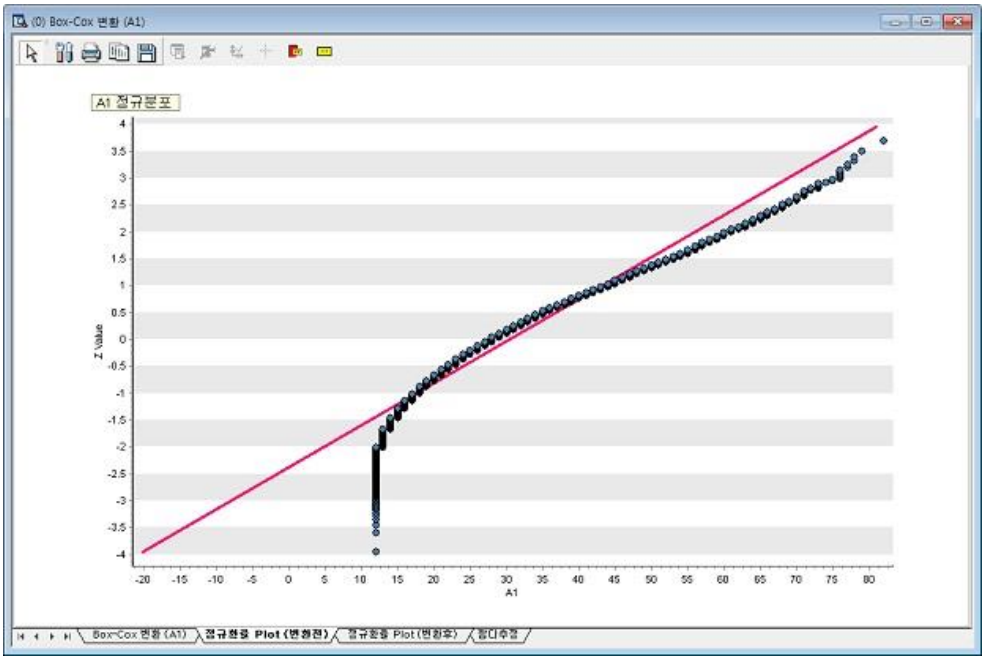
변환이 필요한 변수를 선택합니다.

결과

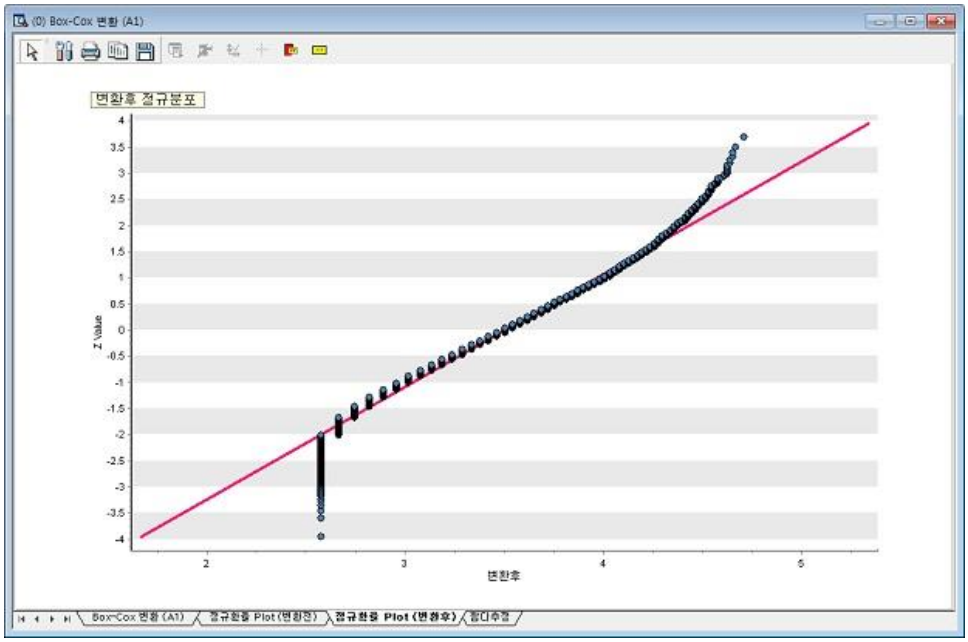
- 변환 수식을 결정하기 위한 추정 람다 정보를 제공하며, 변환 전과 후의 정규성 검정 통계량을 제공함으로써 변환이 되면서 얼마나 더 정규분포를 따르는지 알 수 있습니다. 또, 변환 수식을 파생 변수 노드와 이용하여 변환된 변수로 사용할 수 있습니다.



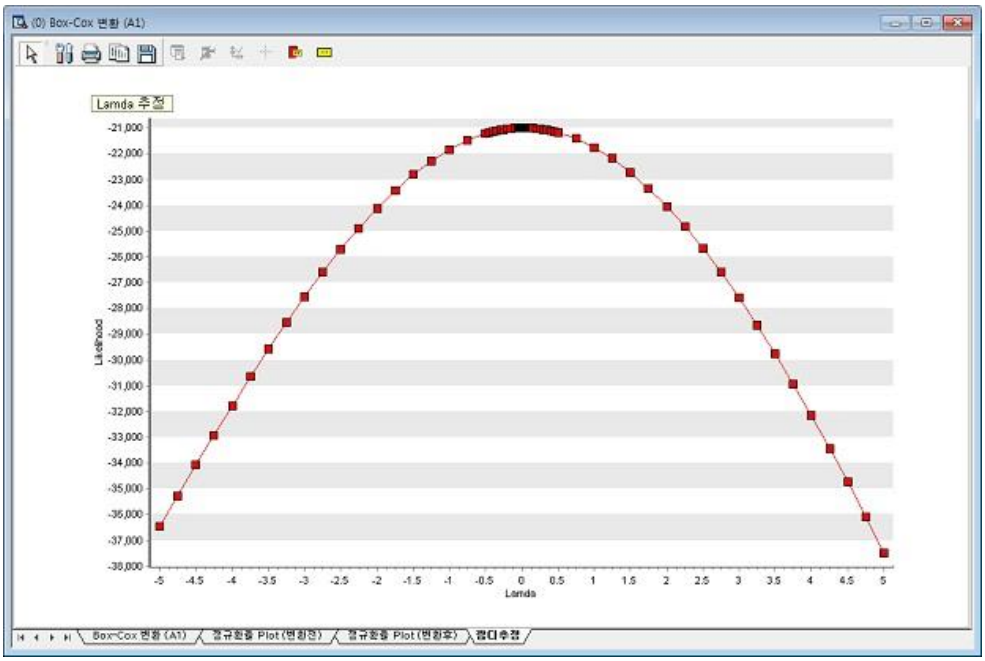
- 정규 확률 Plot(변환 전): 변환하기 전 데이터가 정규분포를 얼마나 따르는지 보여주는 그래프입니다. 빨간 선을 기준으로 빨간 선 근방에 데이터가 많이 분포하면 데이터가 정규성을 따른다고 할 수 있고, 빨간 선에 많이 벗어날수록 정규성 가정에 위배됩니다.



- 정규 확률 Plot(변환 후): 추정된 변환 수식에 의해 변환된 데이터가 정규분포를 얼마나 따르는지 보여주는 그래프입니다. 빨간 선을 기준으로 빨간 선 근방에 데이터가 많이 분포하면 데이터가 정규성을 따른다고 할 수 있고, 빨간 선에 많이 벗어날수록 정규성 가정에 위배됩니다.



- 람다 추정: Likelihood 가 최대가 되는 람다를 추정치로 사용하여 변환 수식을 결정하며, 람다에 따라 Likelihood 의 변화를 보여주는 그래프를 제공합니다.

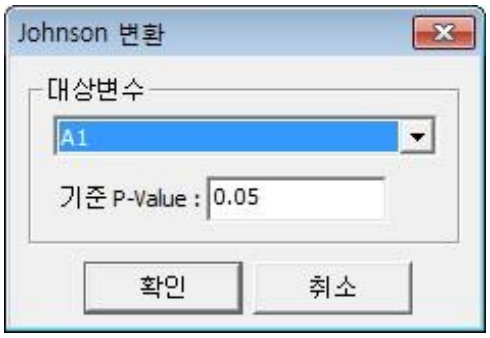


5.5.7 Johnson 변환

관리도에 적용하기 위한 전제조건으로는 데이터가 정규분포를 따라야 합니다. 그러나 많은 데이터가 정규분포를 따르기 어렵습니다. Johnson 변환은 정규분포를 따르지 않는 관측치들을 관리도에 적용하기 위하여 정규분포를 따르도록 변환 수식을 제공합니다.

실행 방법

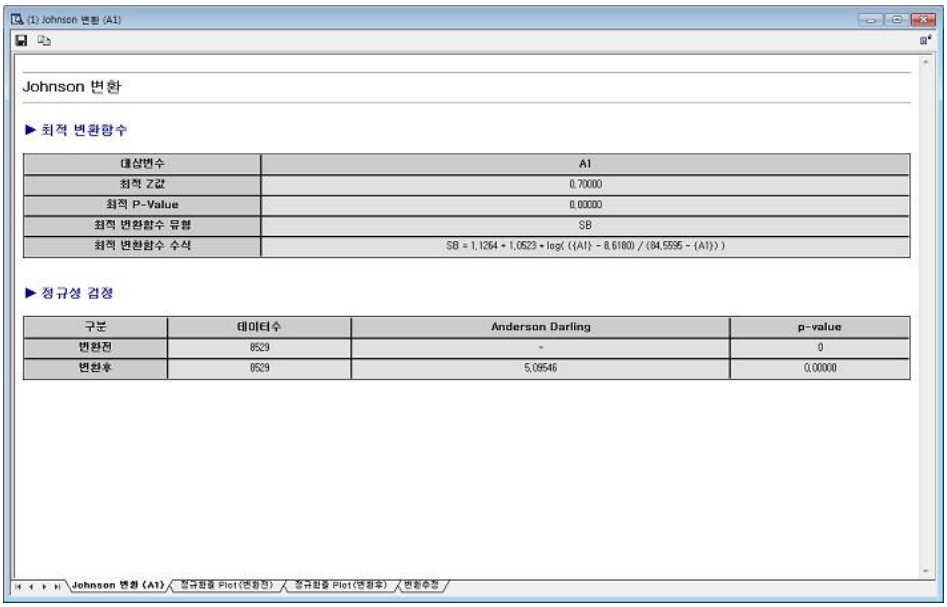
[데이터] – [Johnson 변환]을 선택하면 다음과 같은 Johnson 변환 다이얼로그가 나타납니다.



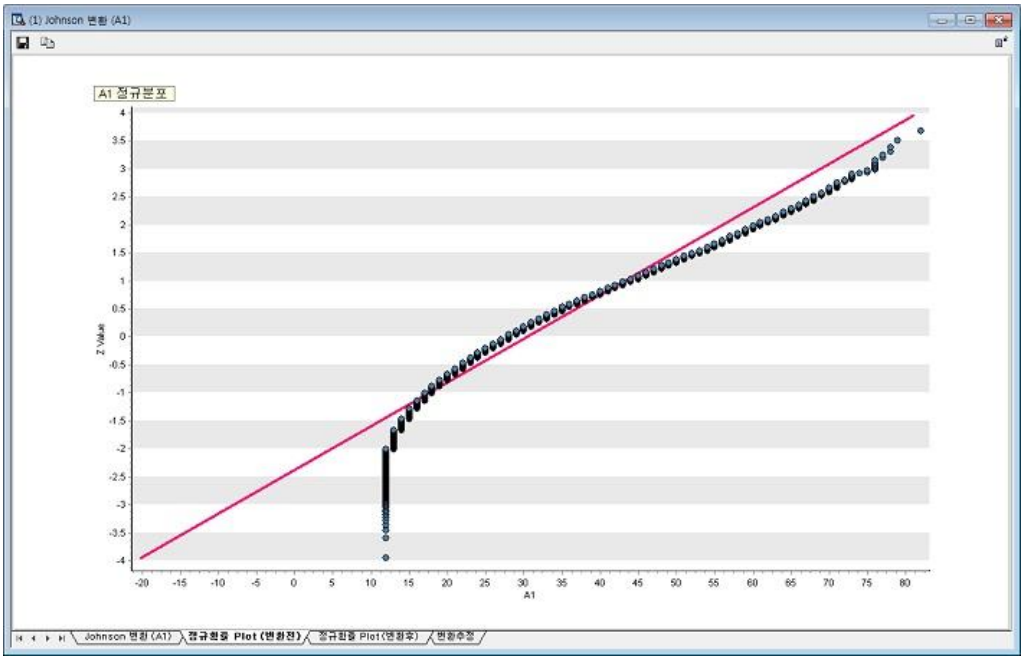
변환하고자 하는 변수를 선택하고, 허용할 수 있는 기준 P-value 를 설정합니다.

결과

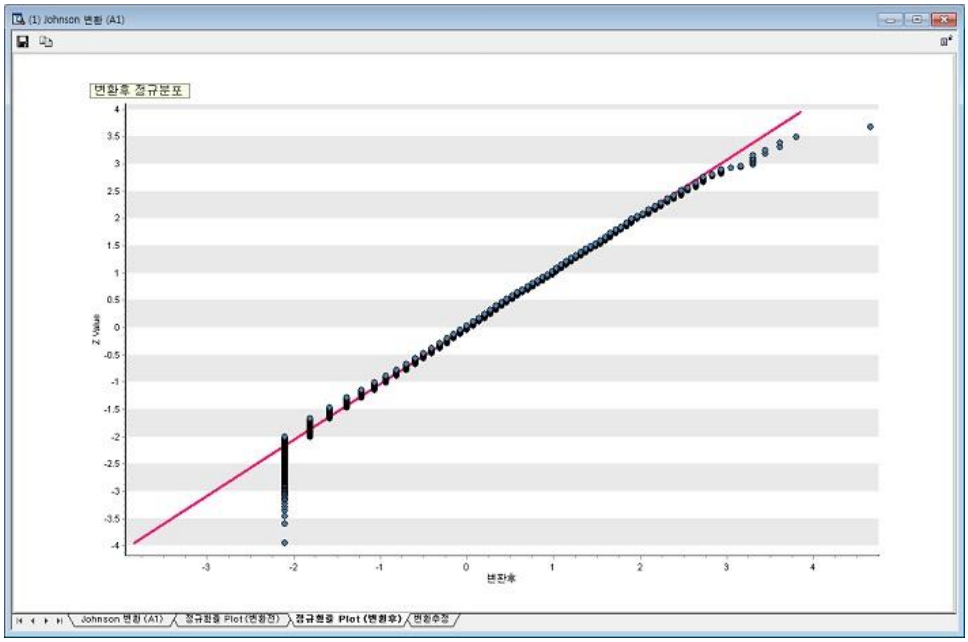
- 최적 변환함수: 정규분포를 가장 만족하는 변환함수를 추정하기 위하여 계산되는 정보들을 제공하며, 변환 함수를 파생 변수 노드와 이용하여 변환된 변수로 사용할 수 있습니다.
- 정규성 검정: 선택한 변수가 정규분포에 얼마나 따르는지 보여주는 Anderson-Darling 통계량과 P-value 를 제공합니다. P-value 가 0.05 보다 크면, 변수는 정규분포를 따른다고 할 수 있습니다.



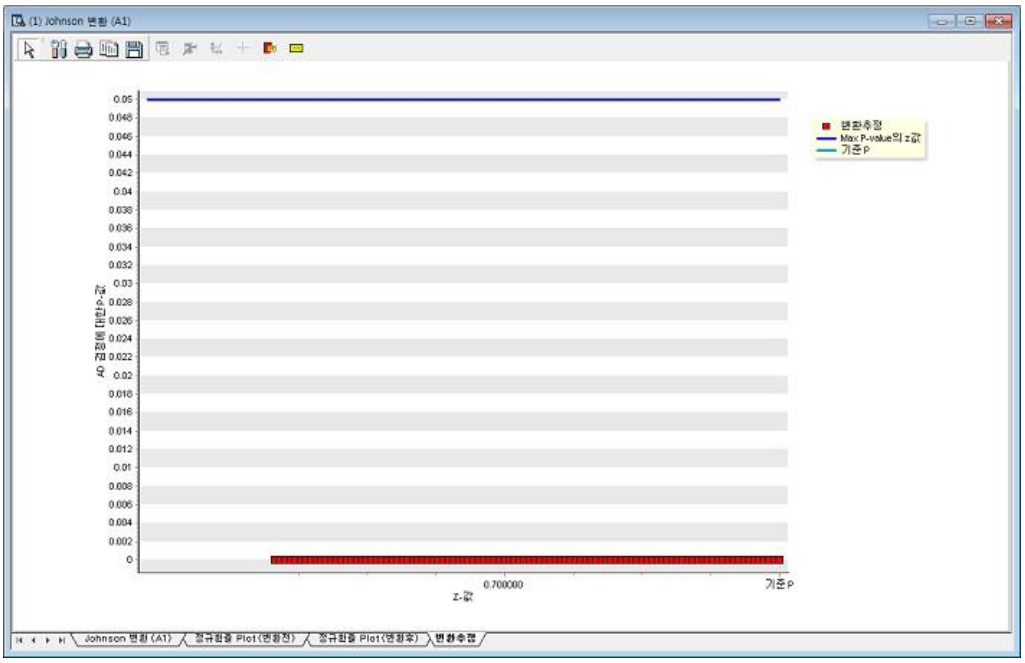
- 정규 확률 Plot(변환 전): 변환하기 전 데이터가 정규분포를 얼마나 따르는지 보여주는 그래프입니다. 빨간 선을 기준으로 빨간 선 근방에 데이터가 많이 분포하면 데이터가 정규성을 따른다고 할 수 있고, 빨간 선에 많이 벗어날수록 정규성 가정에 위배됩니다.



- 정규 확률 Plot(변환 후): 추정된 변환 수식에 의해 변환된 데이터가 정규분포를 얼마나 따르는지 보여주는 그래프입니다. 빨간 선을 기준으로 빨간 선 근방에 데이터가 많이 분포하면 데이터가 정규성을 따른다고 할 수 있고, 빨간 선에 많이 벗어날수록 정규성 가정에 위배됩니다.



- 변환추정: Z 값에 따라 달라지는 변환 데이터가 얼마나 정규분포를 따르는지 정도(P-value)를 파악할 수 있는 그래프를 제공합니다. P-value 가 크면 클수록 정규분포를 더 만족한다고 할 수 있습니다.



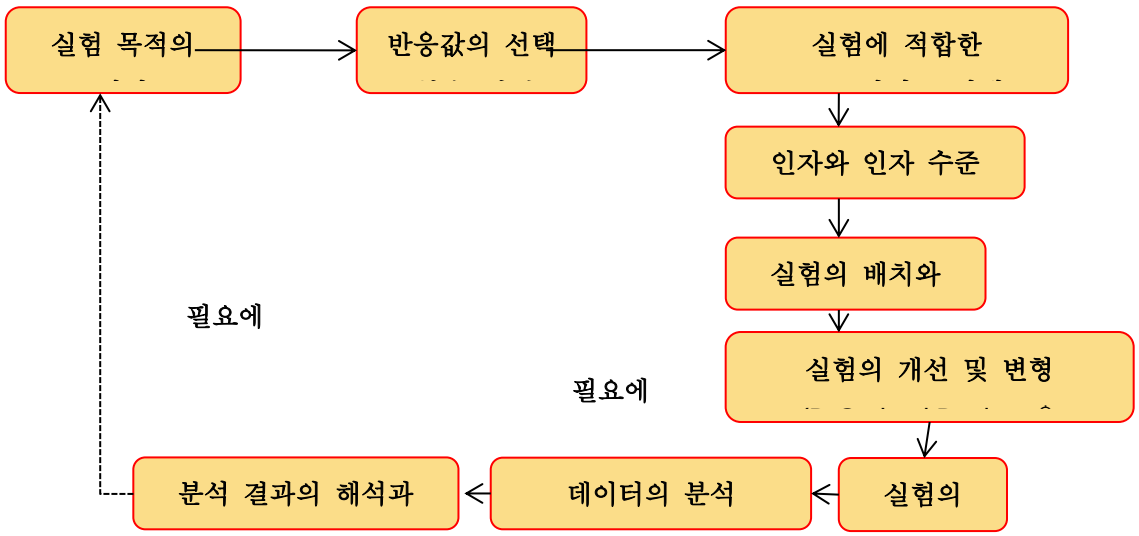
제 6 장 DOE

6.1 DOE 시작하기

6.1.1 DOE 소개

실험계획법(Design Of Experiment, 이하 DOE)는 ECMiner™ 2010 부터 추가 된 기능입니다. 실험계획법이란 실험에 대한 계획방법을 의미하는 것으로 해결하고자 하는 문제에 대하여 실험을 어떻게 행하고, 데이터를 어떻게 취하며, 어떠한 통계적 방법으로 데이터를 분석하면 최소의 실험 횟수에서 최대의 정보를 얻을 수 있는가를 계획하는 것이라고 정의할 수 있습니다. 따라서 하나의 실험계획법을 짰다고 하는 것은 해결하고자 하는 문제에 대하여 인자를 선정하고, 실험방법을 택하여 실험 순서를 정하고, 실험 후 얻어지는 데이터에 대한 최적 분석 방법을 선택하였다는 의미입니다.

이러한 실험계획법은 다음과 같은 절차에 의해서 수행됩니다.

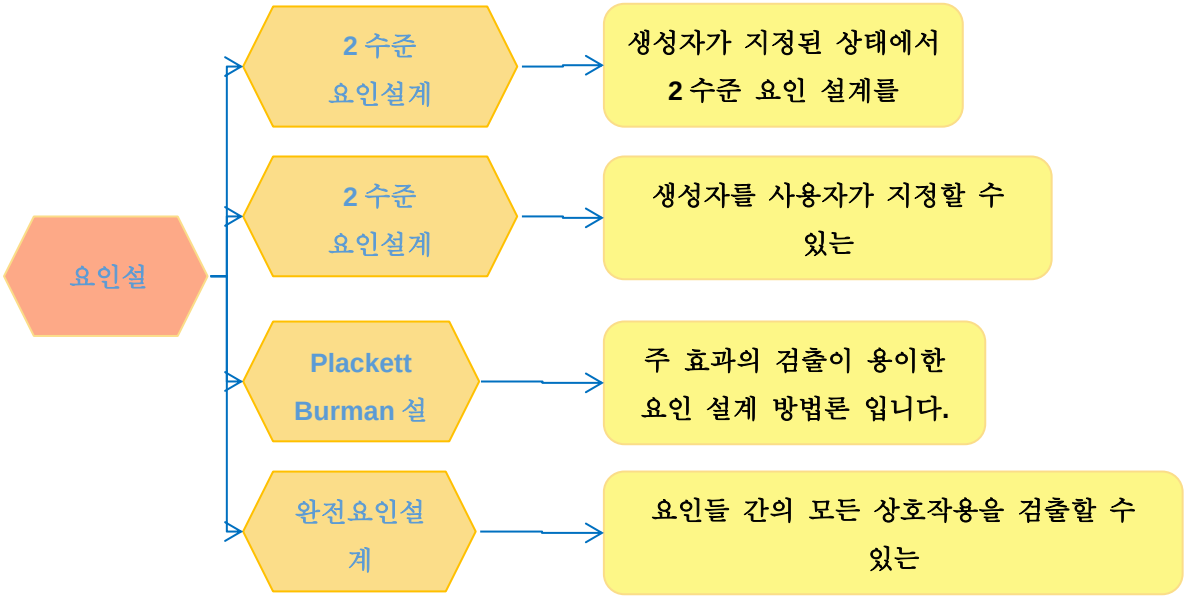


즉 먼저 실험 목적을 설정하고 인자에 영향을 받아 결정되는 반응 값을 선택합니다. (실험계획의 특성상 여러 개의 반응 값을 선택할 수 있습니다.) 그리고 실험에 적합한 DOE 방법론을 선택한 후에 인자와 수준을 선택하고 경우에 따라 D Optimal Design 을 이용하여 실험을 개선 혹은 변형하여 최종 실험표를 얻습니다. 실험표에 따라 실험을 실시하고 분산분석, 회귀 분석, 잔차 분석, 반응 최적화 방법을 통해 데이터를 분석합니다. 이를 통해 충분히 의미 있는 결과를 얻었으면 이 시점에서 종료를 하고 만약 실험을 더해야 할 것이라 판단될 경우 다시 처음으로 돌아가 실험계획을 재개하도록 합니다.

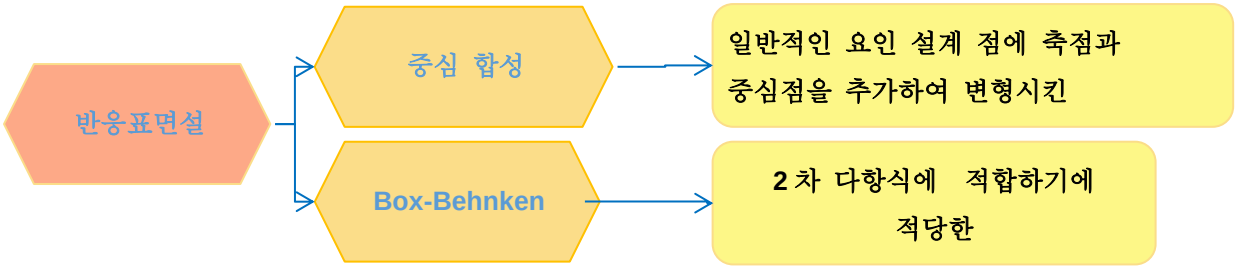
6.1.2 ECMiner™ DOE 의 구성

ECMiner™ DOE 는 요인설계, 반응 표면 설계, 혼합물 설계, 다구찌 설계 방법론을 제공하고 각 방법론 당 세부 방법론을 다음과 같이 제공합니다.

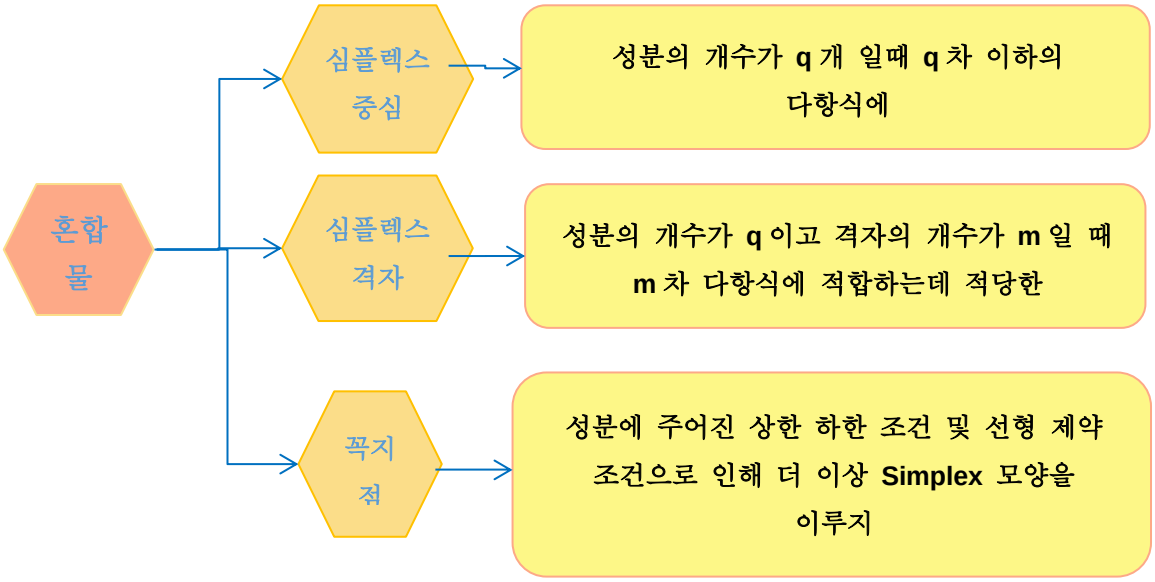
▪ 요인 설계



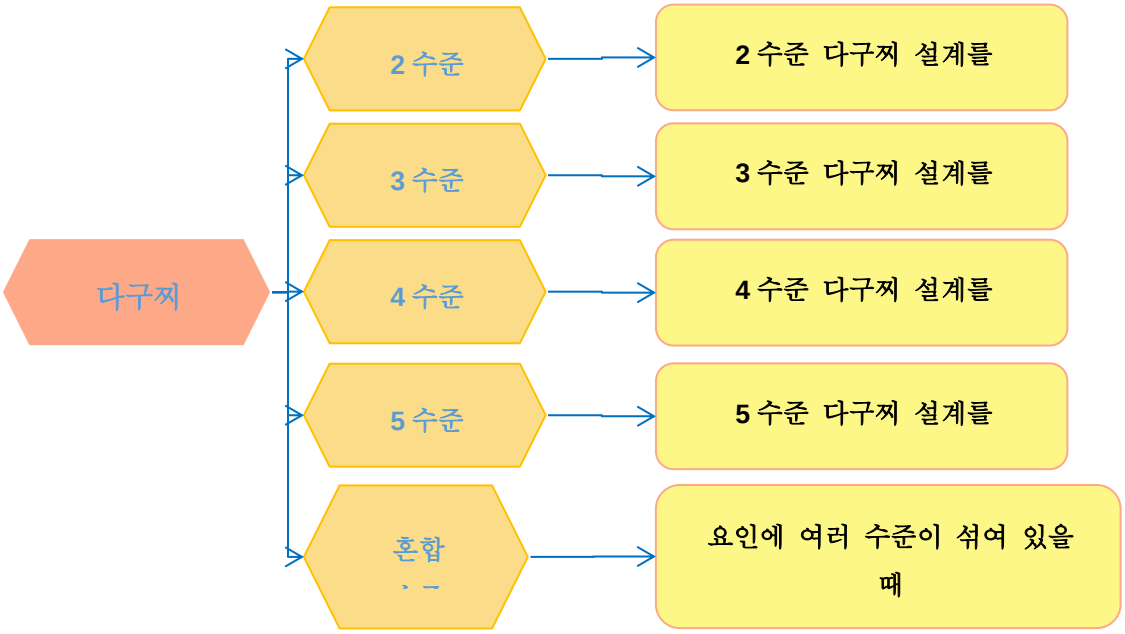
▪ 반응 표면 설계



▪ 혼합물 설계


혼합물
심플렉스 중심
성분의 개수가 q 개 일때 q 차 이하의 다항식에
심플렉스 격자
성분의 개수가 q 이고 격자의 개수가 m 일 때 m 차 다항식에 적합하는데 적당한
꼭지점
성분에 주어진 상한 하한 조건 및 선형 제약 조건으로 인해 더 이상 Simplex 모양을 이루지

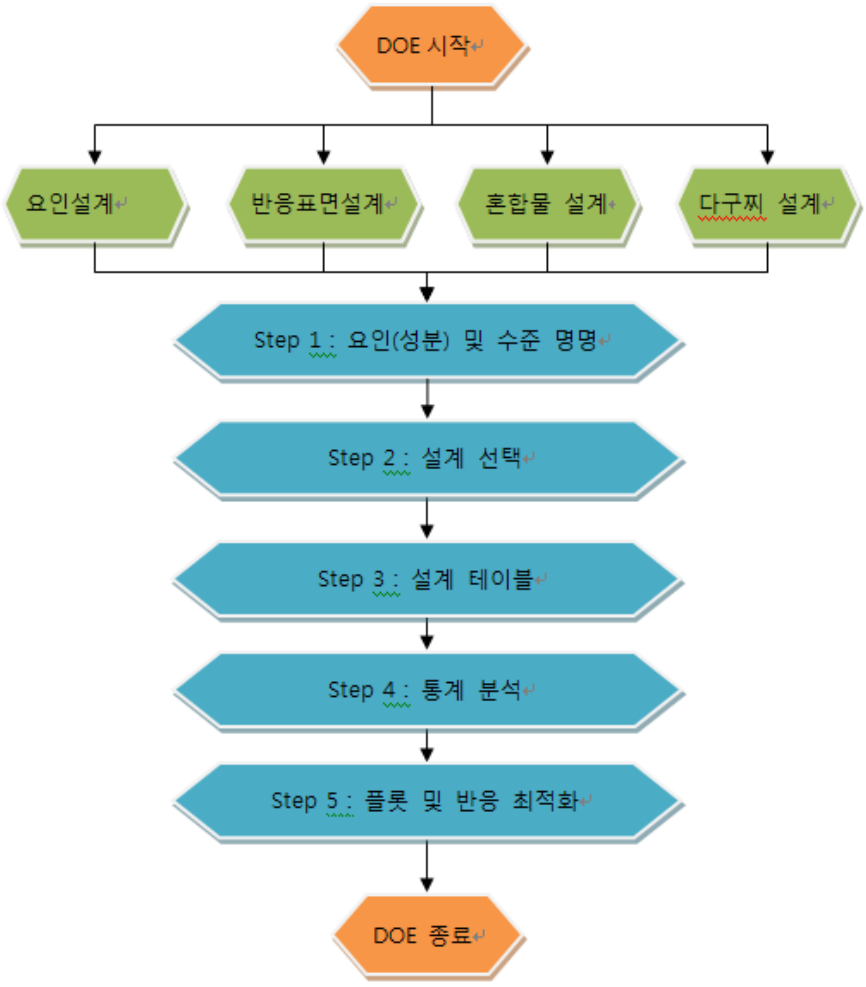
▪ 다구찌 설계


다구찌
2 수준
2 수준 다구찌 설계를
3 수준
3 수준 다구찌 설계를
4 수준
4 수준 다구찌 설계를
5 수준
5 수준 다구찌 설계를
혼합
요인에 여러 수준이 섞여 있을 때

6.2 ECMiner™ DOE 의 특징

6.2.1 일관적인 구조

ECMiner™ DOE 는 일관적인 구조를 갖습니다. DOE 의 방법론은 요인 설계, 반응 표면 설계, 혼합물 설계, 다구찌 설계가 있는데 이 설계 방법론들은 모두 매우 다른 특성을 가지고 있습니다. 따라서 이 모든 설계에 대해서 차이점 및 특성을 잘 알고 있어야 DOE 를 이용할 수 있습니다. 하지만 ECMiner™ DOE 는 이 모든 방법론들을 하나의 구조로 통합시켜 배치하였기 때문에 사용자들은 DOE 를 더욱 쉽고 편리하게 사용할 수 있습니다.

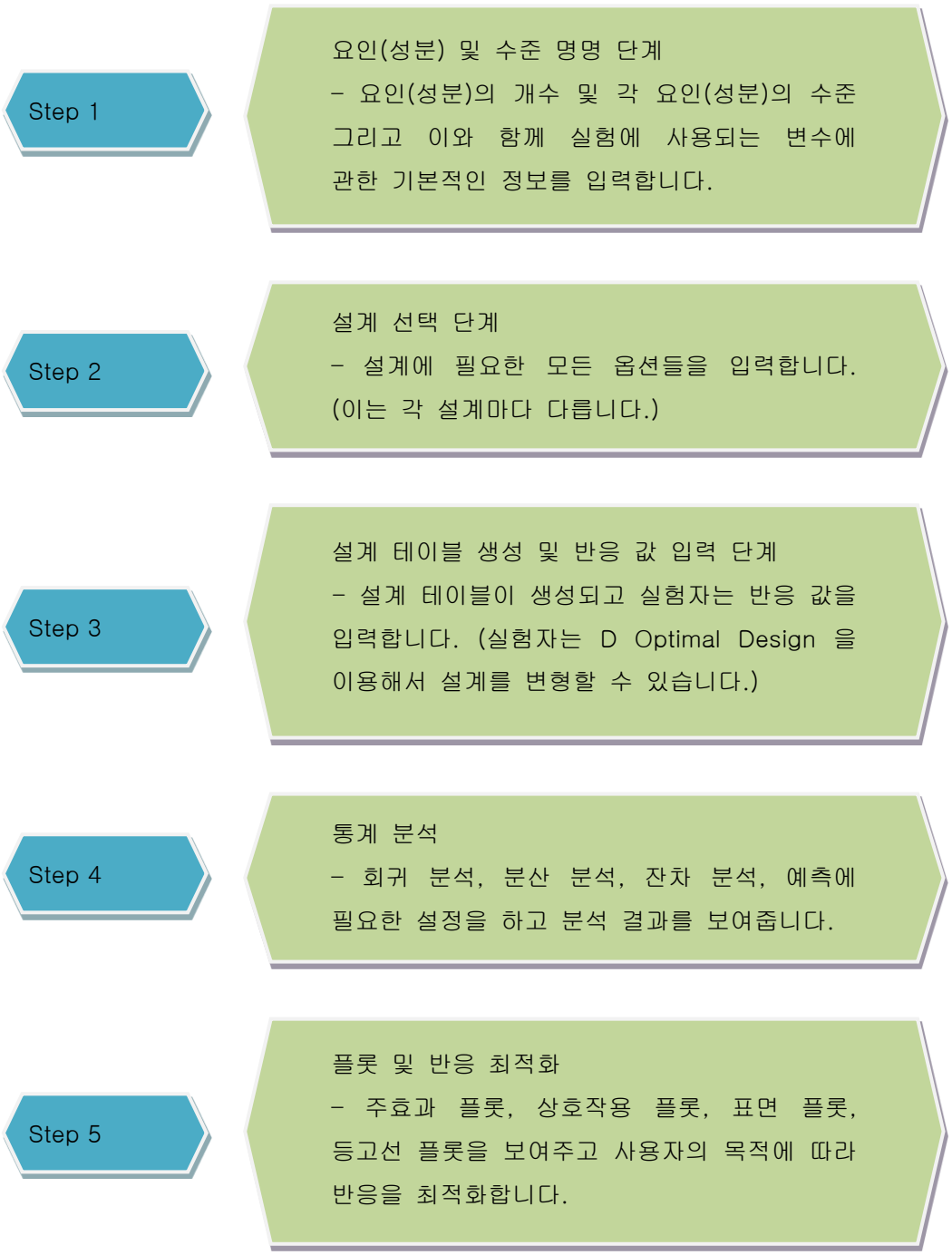


6.2.2 사용자의 편의성

사용자의 편의성은 구조의 일관성으로부터 얻어진 특징이라고 할 수 있습니다. 요인 설계, 반응 표면 설계, 혼합물 설계, 다구찌 설계를 모두 잘 알고 있어야 DOE 를 시작할 수 있는 것이 아니라 하나의 설계에 대해서 만이라도 사용법을 알고 있으면 나머지 설계의 사용은 자연스럽게 배워갈 수 있습니다. 이는 겉보기에는 서로 달라 보이는 설계 들 간의 공통적인 구조 및 특징에 기인합니다. 사용자는 이를 통해서 보다 쉽고 빠르게 DOE 를 이해하고 활용할 수 있습니다.

6.2.3 구성의 개관 및 각 step 별 특징 소개

ECMiner™ DOE 는 어떠한 설계 방법론이든 5 가지 Step 으로 나누어져 진행됩니다.



6.3 ECMiner™ DOE 방법론

6.3.1 요인 설계

6.3.1.1. 개요

ECMiner™에서 제공하는 요인설계의 종류는 크게 2 가지 입니다. 하나는 2 수준 요인 설계이고 다른 하나는 일반 완전 요인 설계입니다. 2 수준 요인 설계는 말 그대로 각 요인의 수준 값이 2 개인 실험을 말합니다. 하지만 각 요인의 수준 값이 두 개라고 하더라도 모든 격자점에서 실험을 한다고 하면 2^n(n 은 요인의 개수)번의 실험을 수행해야 합니다. 따라서 2 수준 요인 설계에서는 사용자의 목적에 맞도록 이러한 실험 횟수를 줄이는 방법을 제공합니다. ECMiner™에서 제공하는 2 수준 요인 설계는 다음과 같이 3 종류가 있습니다.

2 수준 요인 설계 (기본 생성자)

2 수준 요인 설계 (생성자 지정)

Plackett Burman 설계

이와 함께 ECMiner™에서는 일반 완전 요인 설계를 제공하는데 이는 각 요인당 수준의 개수가 다를 때 사용할 때 용이합니다. 본 실험에서는 모든 격자 점에서 실험을 하기 때문에 모든 상호 작용에 대한 유의성을 판단할 수 있습니다.

6.3.1.2. 2 수준 요인 설계(기본 생성자)

본 설계에서는 부분 요인 설계를 구할 때 기본적으로 지정해 놓은 생성자를 사용하고 블록을 나눌 때도 역시 기본적으로 지정해 놓은 블록 생성자를 사용합니다. 이 설계의 이름이 2 수준 요인 설계(기본 생성자)인 이유가 이 때문입니다.

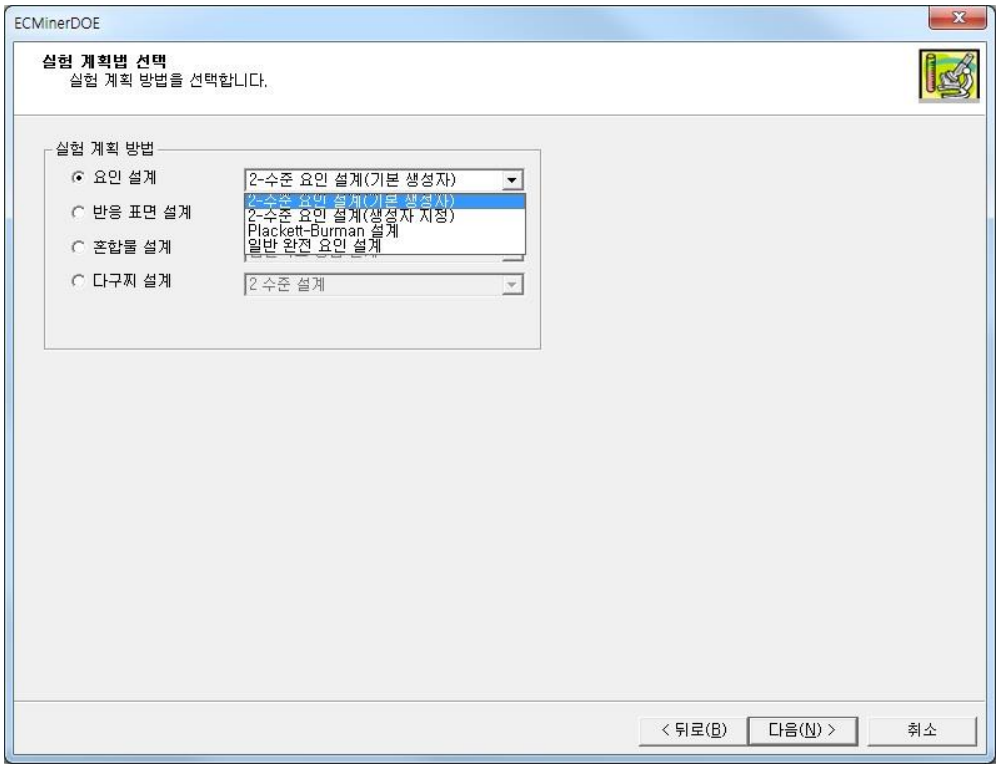
실험 소개

본 실험은 어떤 반응 공정의 수율을 향상시키기 위한 목적으로 진행되고, 요인 A 를 반응시간, 요인 B 를 반응온도, 요인 C 를 성분의 양이라고 합니다. 구체적인 실험 조건은 다음과 같습니다.

반응시간(A) : $A_0 = 4, A_1 = 5(\text{시간})$
반응온도(B) : $B_0 = 250, B_1 = 300(\text{도})$
성분의 양(C) : $C_0 = 3.0, C_1 = 3.5(g)$

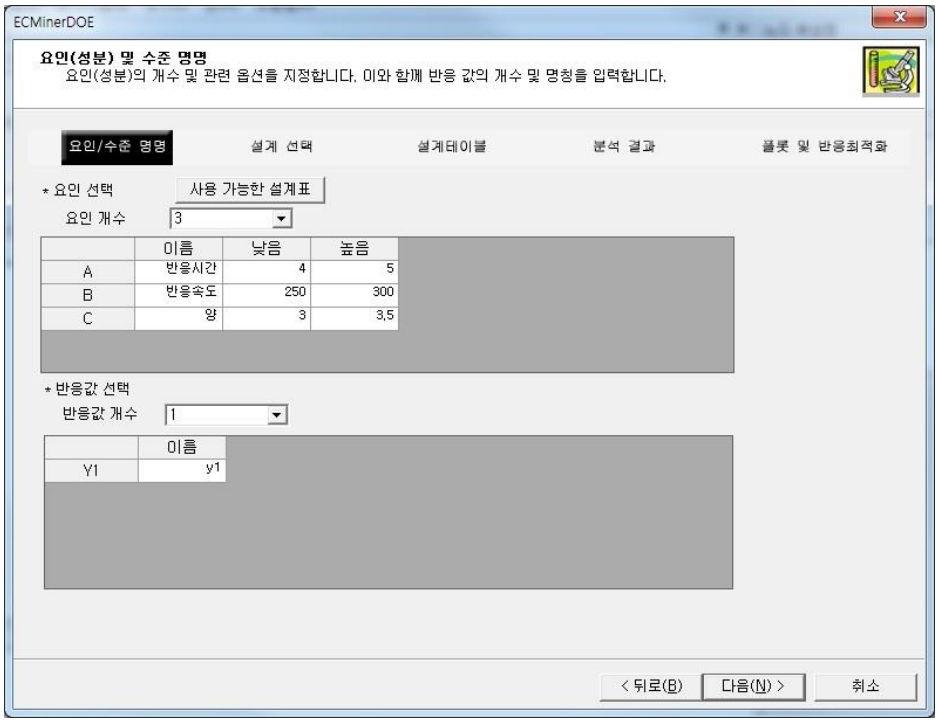
이 때 하루에 실험을 다할 수 없어서 교호작용 ABC 를 날과 교락시켜 실험하고자 합니다.

다음과 같이 2 수준 요인 설계(기본 생성자)를 선택합니다.

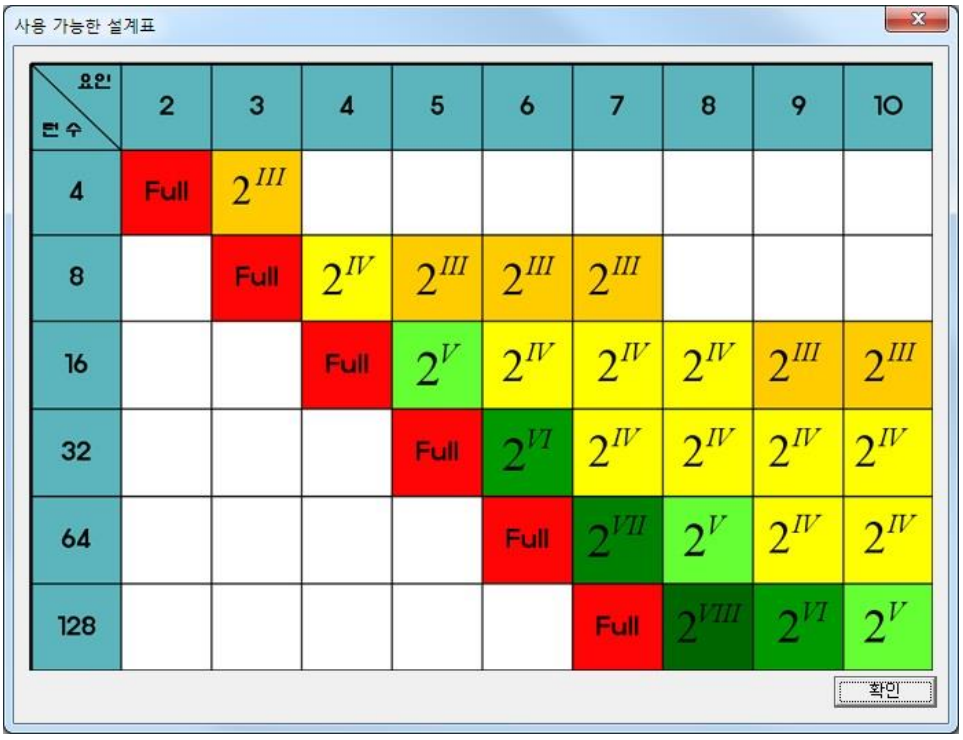


▪ **Step 1: 요인(성분) 및 수준 명명**

다음에 나타난 화면에서 각 요인당 이름을 명명하고 낮은 수준과 높은 수준에 대한 값을 입력합니다. 그리고 반응의 개수 및 각 반응 값의 명칭 또한 입력합니다.



이 때 사용 가능한 설계표를 클릭하면 다음과 같은 화면을 볼 수 있습니다.



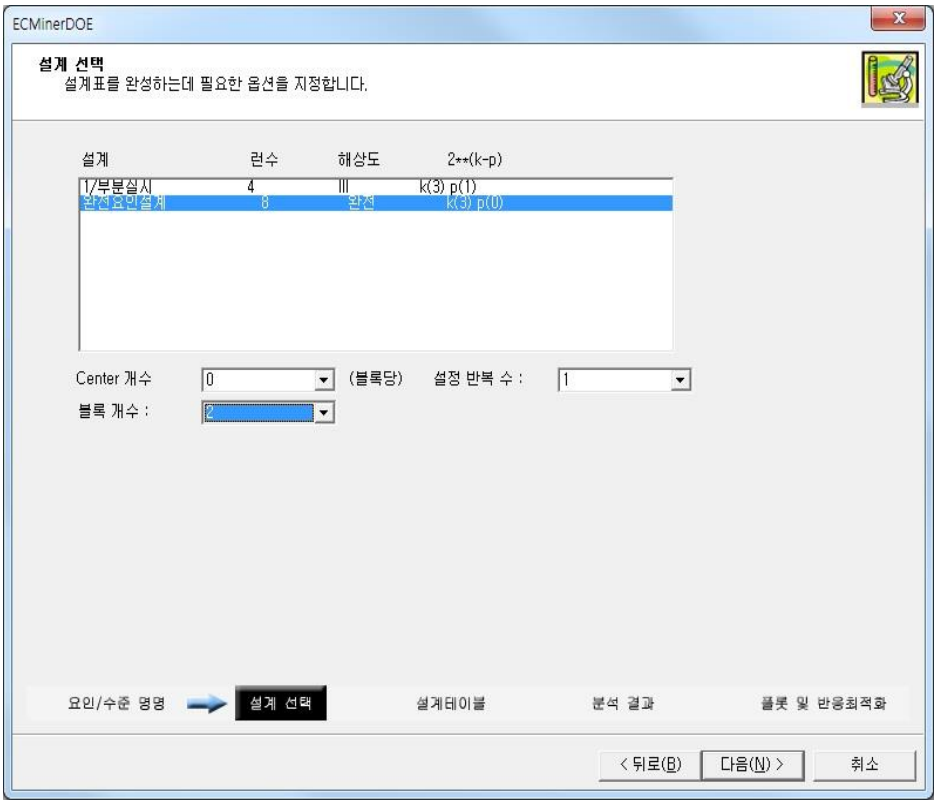
▪ **Step 2: 설계의 선택**

다음에 나타나는 화면을 통해서 세부 설계 옵션을 사용자가 지정할 수 있습니다. 요인의 개수에 따라 다르지만 요인의 개수가 3 개일 때 설계는 다음과 같이 부분실시(부분요인 설계)와 완전 요인 설계를 제공합니다. 본 과정에서 사용자는 중심점의 개수와 설정 반복 수, 블록의 개수를 조절할 수 있습니다.

중심점: 요인들의 수준 값이 낮은 값과 높은 값의 중간에 해당할 때의 실험 점을 의미합니다. 이러한 중심점을 설정함으로써 반응 값을 종속변수로 하고, 요인을 독립변수로 하여 Regression Model 을 만들 때 이 Regression 모델에 곡면성이 있는지를 체크해 볼 수 있습니다.(만일 곡면성이 있다는 판단이 되면, 반응표면 설계와 같이 2 차 모형을 적합할 수 있는 실험 설계법을 사용해야 합니다.)

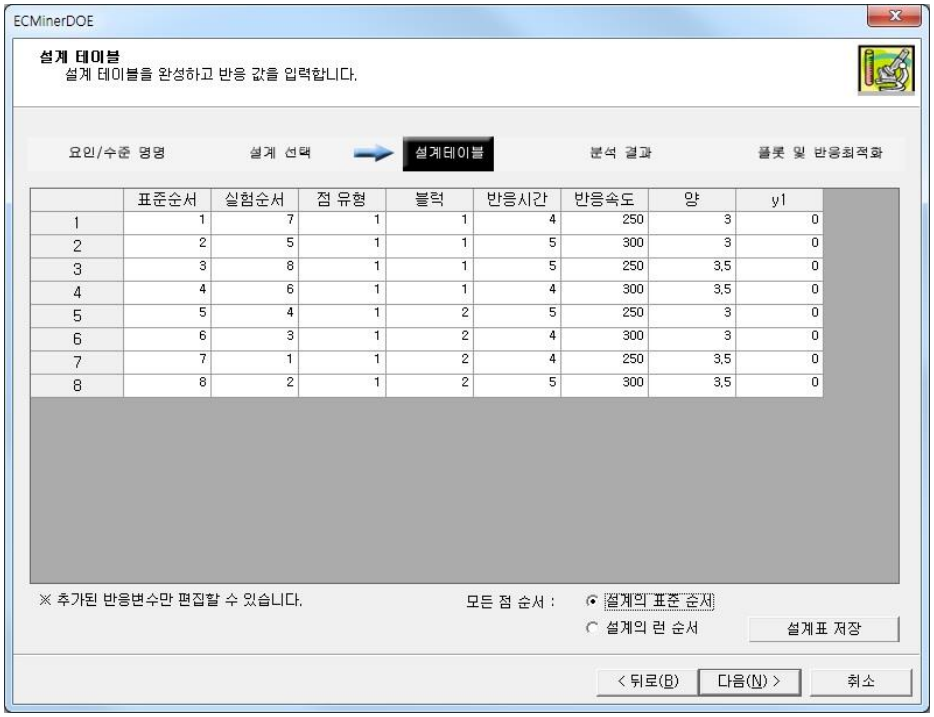
설정 반복 수: 같은 실험을 몇 번 반복할 수 있는지를 설정할 수 있습니다. 같은 실험을 반복함을 통해서 순수 오차가 어느 정도 되는지(즉 같은 점에서 측정할 때 값이 얼마나 차이가 나는지를 나타내 주는 지표)를 계산할 수 있습니다.

블록 개수: 기본 생성자 및 반복에 블록을 할당하도록 합니다. 현재 상황에서 블록 개수 2 를 선택하면 ABC 라는 생성자를 사용하여 블록을 할당합니다.



▪ **Step 3: 설계 테이블**

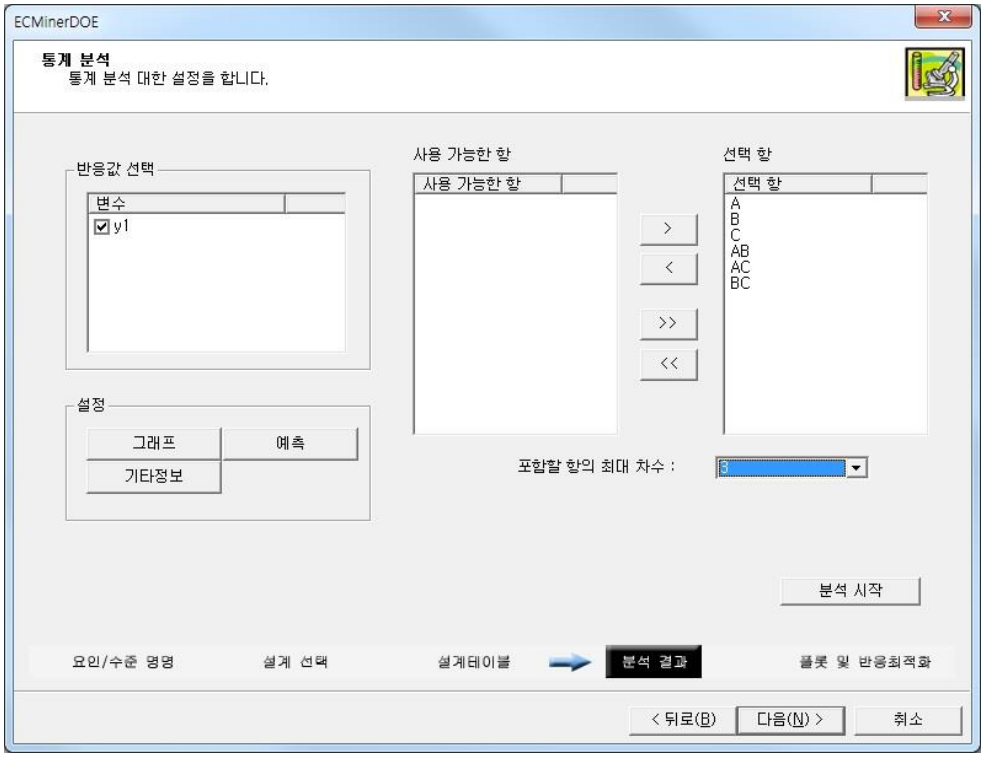
본 화면에서는 앞의 여러 설정들로부터 만들어지는 설계표가 얻어지고, 사용자는 이 설계표에 반응 값을 입력할 수 있습니다. 본 화면에서는 설계점을 실행순서(랜덤으로 정해진 순서)로 정렬하거나 표준순서로 정렬할 수 있습니다.



	표준순서	실험순서	점 유형	블럭	반응시간	반응속도	양	y1
1	1	7	1	1	4	250	3	0
2	2	5	1	1	5	300	3	0
3	3	8	1	1	5	250	3,5	0
4	4	6	1	1	4	300	3,5	0
5	5	4	1	2	5	250	3	0
6	6	3	1	2	4	300	3	0
7	7	1	1	2	4	250	3,5	0
8	8	2	1	2	5	300	3,5	0

▪ **Step 4: 통계 분석**

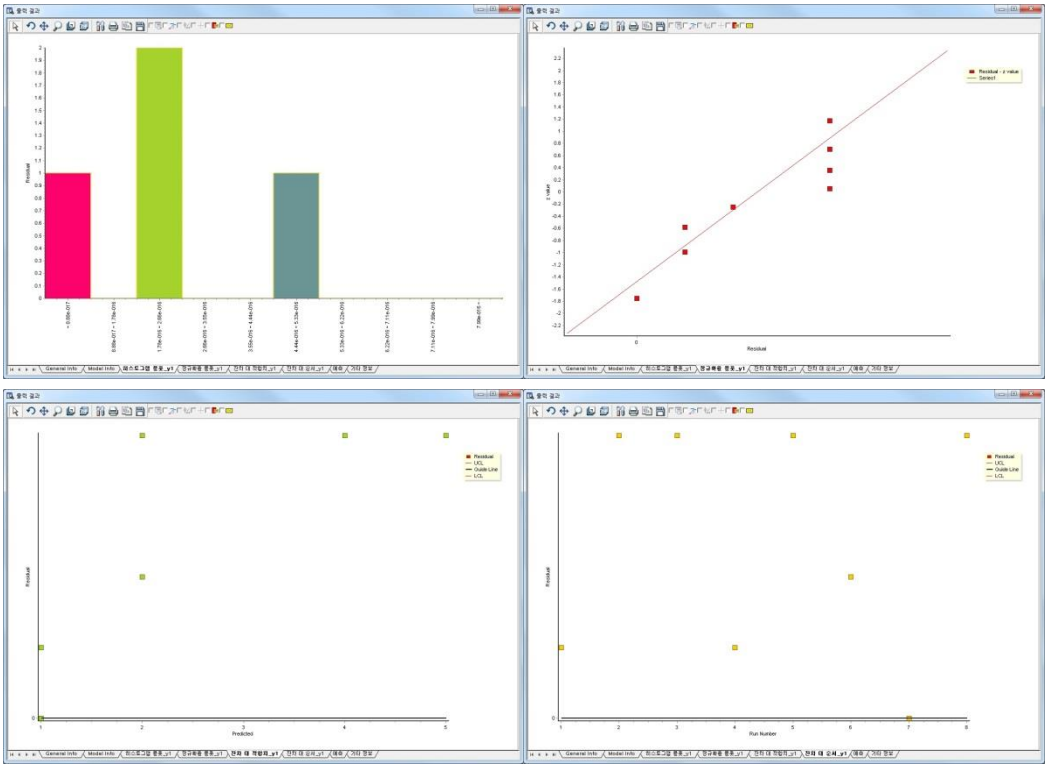
본 단계에서는 회귀 분석, 잔차 분석, 분산 분석에 필요한 설정들을 제공합니다. 먼저 그래프, 예측, 기타정보 단추를 클릭하여 이에 대한 설정을 완료합니다. Step 4 의 Main 화면에서는 포함할 항의 최대 차수를 선택합니다. 선택한 요인의 개수(3)만큼 최대 차수를 선택할 수 있는데 현재 I = ABC 라는 블록 생성자를 사용하여 블록을 생성하였으므로 ABC 라는 항은 자동으로 선택 불가능합니다. 분석시작 단추를 클릭하면 분석 결과를 볼 수 있습니다.



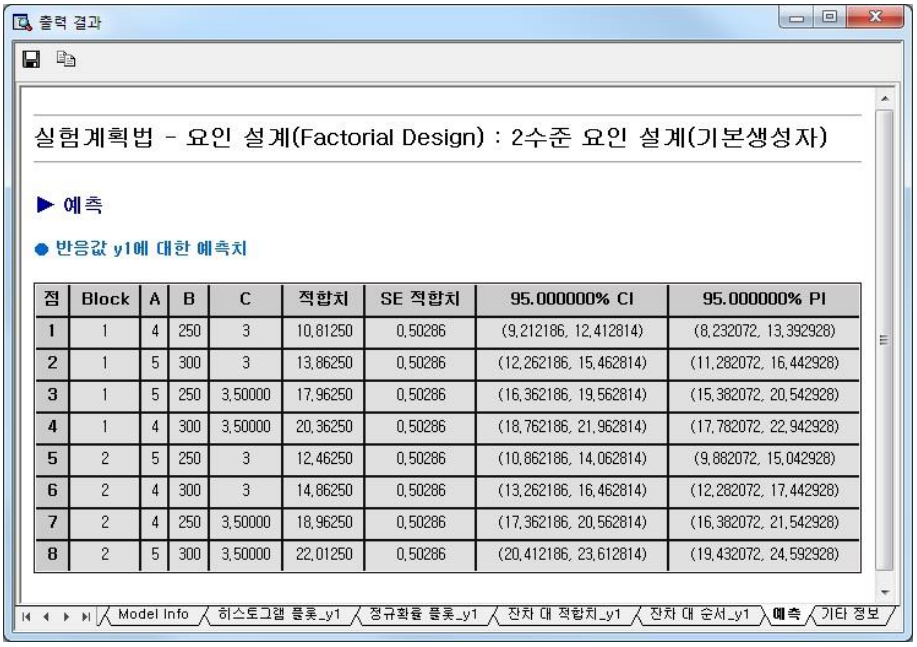
General Info: 각 요인에 대한 기본 정보와 설계 정보, 그리고 별칭 구조를 볼 수 있습니다.

Model Info: 회귀분석 결과, 분산분석 결과, 비정상적 관측치(극단 레버리지, 표준화 잔차)을 볼 수 있습니다.

잔차 관련 플롯: 잔차 히스토그램 플롯, 정규 확률 플롯, 잔차 대 적합치, 잔차 대 순서 을 볼 수 있습니다.

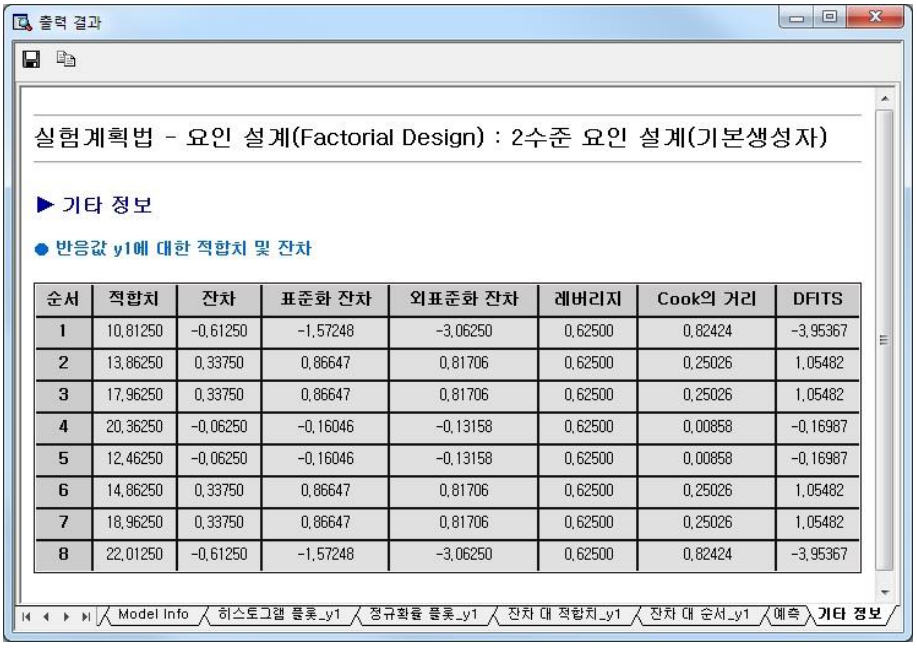


예측: 예측에서 설정한 값에 대한 예측 값을 얻을 수 있습니다.



점	Block	A	B	C	적합치	SE 적합치	95.000000% CI	95.000000% PI
1	1	4	250	3	10,81250	0,50286	(9,212186, 12,412814)	(8,232072, 13,392928)
2	1	5	300	3	13,86250	0,50286	(12,262186, 15,462814)	(11,282072, 16,442928)
3	1	5	250	3,50000	17,96250	0,50286	(16,362186, 19,562814)	(15,382072, 20,542928)
4	1	4	300	3,50000	20,36250	0,50286	(18,762186, 21,962814)	(17,782072, 22,942928)
5	2	5	250	3	12,46250	0,50286	(10,862186, 14,062814)	(9,882072, 15,042928)
6	2	4	300	3	14,86250	0,50286	(13,262186, 16,462814)	(12,282072, 17,442928)
7	2	4	250	3,50000	18,96250	0,50286	(17,362186, 20,562814)	(16,382072, 21,542928)
8	2	5	300	3,50000	22,01250	0,50286	(20,412186, 23,612814)	(19,432072, 24,592928)

기타 정보: 잔차 관련 통계량 및 여러 통계량을 얻을 수 있습니다.



순서	적합치	잔차	표준화 잔차	외표준화 잔차	레버리지	Cook의 거리	DFITS
1	10,81250	-0,61250	-1,57248	-3,06250	0,62500	0,82424	-3,95367
2	13,86250	0,33750	0,86647	0,81706	0,62500	0,25026	1,05482
3	17,96250	0,33750	0,86647	0,81706	0,62500	0,25026	1,05482
4	20,36250	-0,06250	-0,16046	-0,13158	0,62500	0,00858	-0,16987
5	12,46250	-0,06250	-0,16046	-0,13158	0,62500	0,00858	-0,16987
6	14,86250	0,33750	0,86647	0,81706	0,62500	0,25026	1,05482
7	18,96250	0,33750	0,86647	0,81706	0,62500	0,25026	1,05482
8	22,01250	-0,61250	-1,57248	-3,06250	0,62500	0,82424	-3,95367

자세한 설명은 6.4. 설정 및 분석을 참고하세요.

- **Step 5: 플롯 및 반응 최적화**

Step 5 에서는 여러 종류의 차트를 볼 수 있고 반응 최적화를 통해서는 사용자가 원하는 반응 값을 얻기 위해서 최적화를 해주는 알고리즘을 수행합니다. 2 수준 요인 설계(기본 생성자)에서 제공하는 차트는 다음과 같습니다.

주효과 플롯

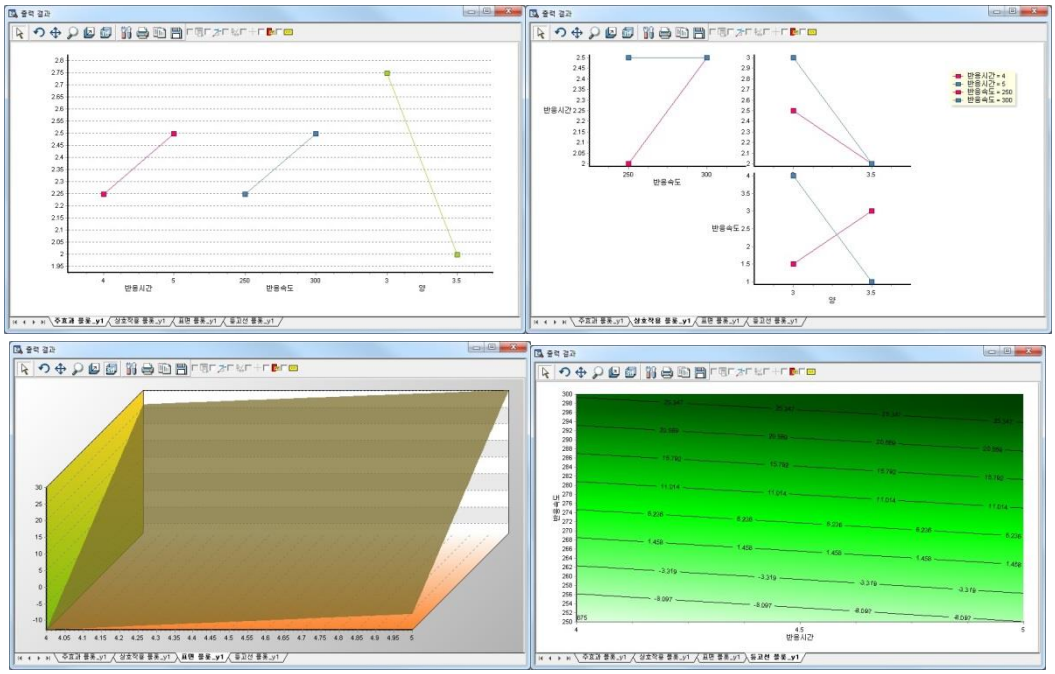
상호작용 플롯

표면 플롯

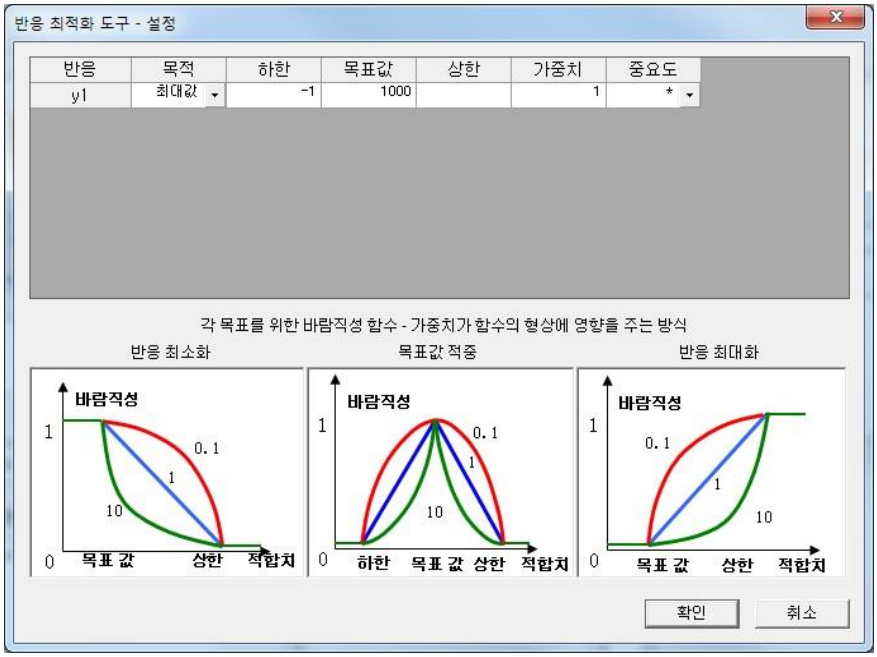
등고선 플롯

주효과 플롯 및 상호 작용 플롯을 그리기 위해서는 먼저 데이터 평균을 사용할 것인지 적합 평균을 사용할 것인지를 선택합니다. 그리고 설정 단추를 통해서는 어떠한 요인들에 대해서 플롯을 그릴지를 선택합니다.

표면 플롯 및 등고선 플롯을 그리기 위해서는 플롯의 특성상 요인이 3 개 이상일 때 하나 이상의 요인에 대해서 고정 값을 입력해야 합니다. 이러한 설정을 마친 후 얻게 되는 예시 화면은 다음과 같습니다.



반응 최적화는 사용자가 반응 값을 최대화 시킬지 최소화 시킬지 혹은 목표 값에 적중하고자 하는 목적이 있을 때 이를 수행하기 위한 방법입니다. 일단 최적화 할 반응 변수를 선택하고 다음과 같은 설정 창에서 목적, 하한, 목표 값, 상한, 가중치, 중요도를 입력합니다.



반응 최적화 도구의 옵션 창에서는 초기값을 설정하는 방법 및 최적화에 사용되는 여러 설정들을 입력한 후에 Step 5 의 Main 화면에서 결과 보기 버튼을 클릭하면 다음과 같은 화면을 얻을 수 있습니다.

Number	반응시간	반응속도	양	y1	종합 바람직성
1	1	-1	-1	23,75000	0,02473
2	0,99977	-0,99993	-0,99977	23,74592	0,02472
3	0,99915	-0,99911	-0,99932	23,73316	0,02471
4	0,99857	-0,99766	-0,99995	23,72922	0,02470
5	0,99349	-0,99959	-0,99950	23,71114	0,02469
6	0,99849	-0,99991	-0,99431	23,67839	0,02465
7	0,99978	-0,99984	-0,98679	23,59942	0,02457
8	-0,87921	-0,99992	-0,99986	14,82235	0,01581
9	-1	-1	-1	14,25000	0,01523

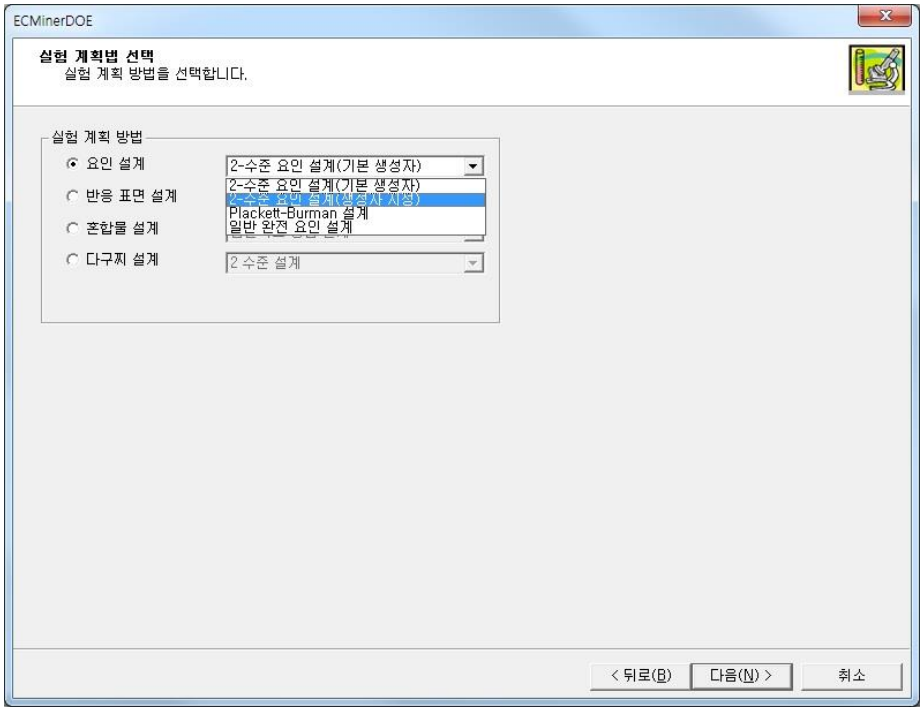
위의 과정을 통해서 반응 값(수율)을 최대화 하기 위해서 각 요인(반응시간, 반응온도, 양)의 수준 값을 어떻게 결정해야 하는지를 알 수 있습니다.(이 때 각 요인의 수준 값은 코드화된 단위로 나타나게 됩니다.)

6.3.1.3. 2 수준 요인 설계(생성자 지정)

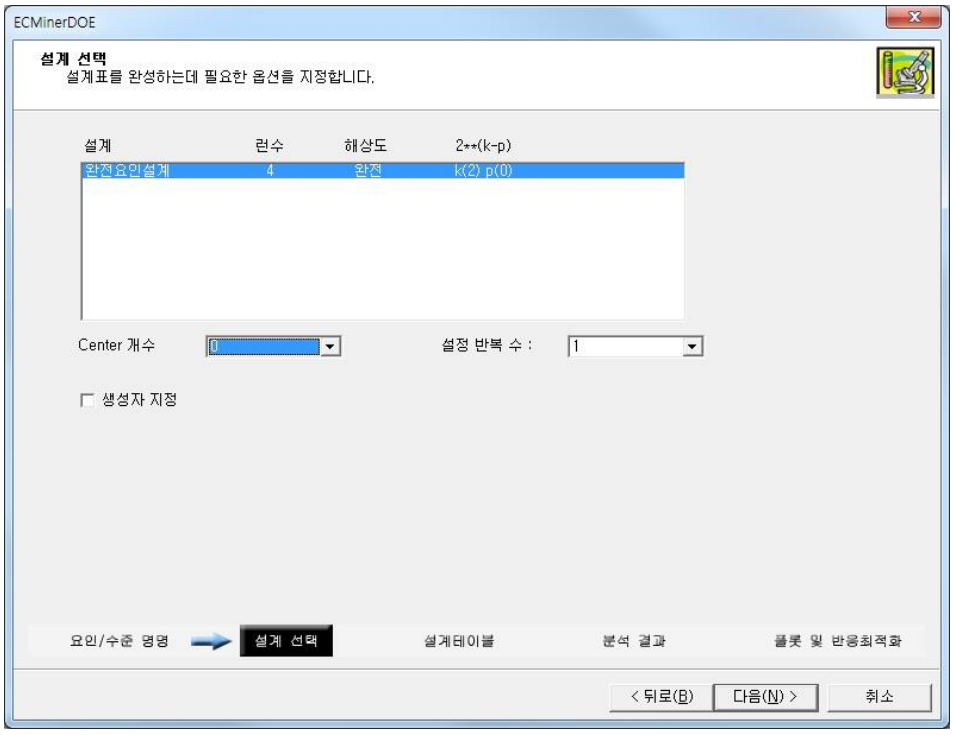
본 설계에서는 부분 요인 설계를 구할 때 사용자가 지정하는 생성자를 사용하고 블록을 나눌 때도 역시 사용자가 지정하는 블록 생성자를 사용합니다. 이 설계의 이름이 생성자 지정인 이유가 이 때문입니다.

2 수준 요인 설계 (생성자 지정)는 2 수준 요인 설계(기본 생성자) 분석 이후의 단계가 동일합니다. 따라서 이를 위해서 차이가 있는 부분을 중점적으로 설명하도록 하겠습니다.

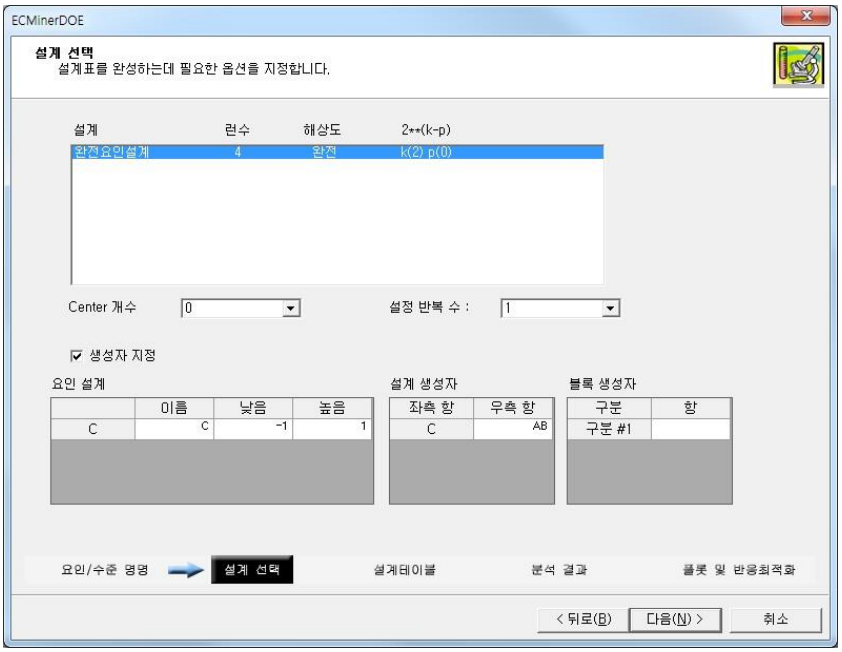
2 수준 요인 설계 (생성자 지정)을 선택합니다.

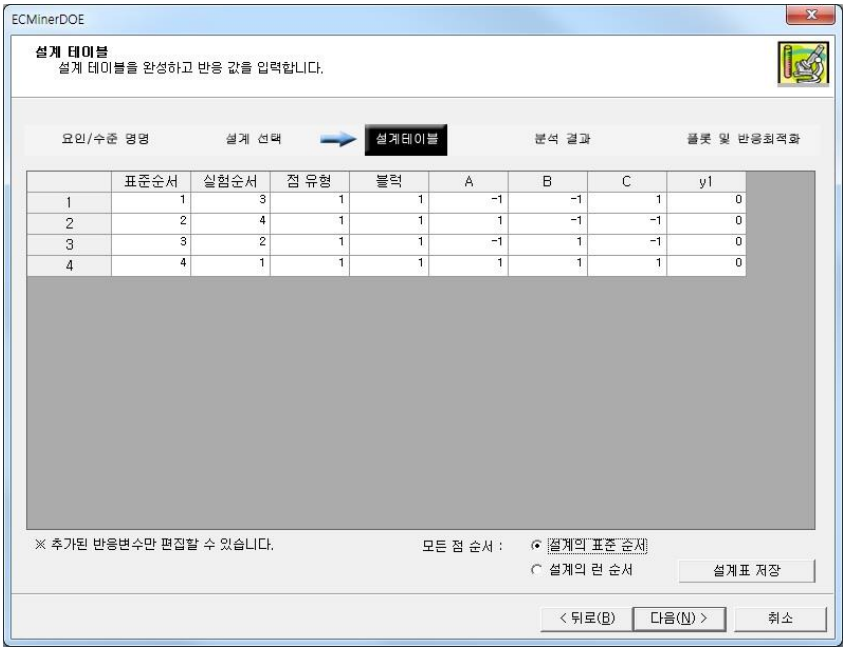


Step 1 화면에서는 설계 생성자를 이용하여 설계를 증대하기 이전의 요인 수를 입력합니다. 현재 상황에서는 요인의 수를 2 개로 합니다. Step 2 화면은 다음과 같습니다.



본 화면에서 생성자를 지정하여 요인의 개수를 늘리고자 하면 생성자 지정 체크 박스를 클릭하고 설계 생성자에 해당하는 부분을 완성합니다. 사용자의 필요에 따라 블록 생성자를 입력할 수도 있습니다. 예를 들어 C=AB 라는 생성자를 이용하여 요인의 수를 늘리고 싶으면 위와 같이 좌측 항을 C 로 하고, 우측 항을 AB 라고 입력하여 다음 단추를 클릭합니다. 다음 단추를 클릭한 후에는 다음과 같은 화면이 나타납니다.





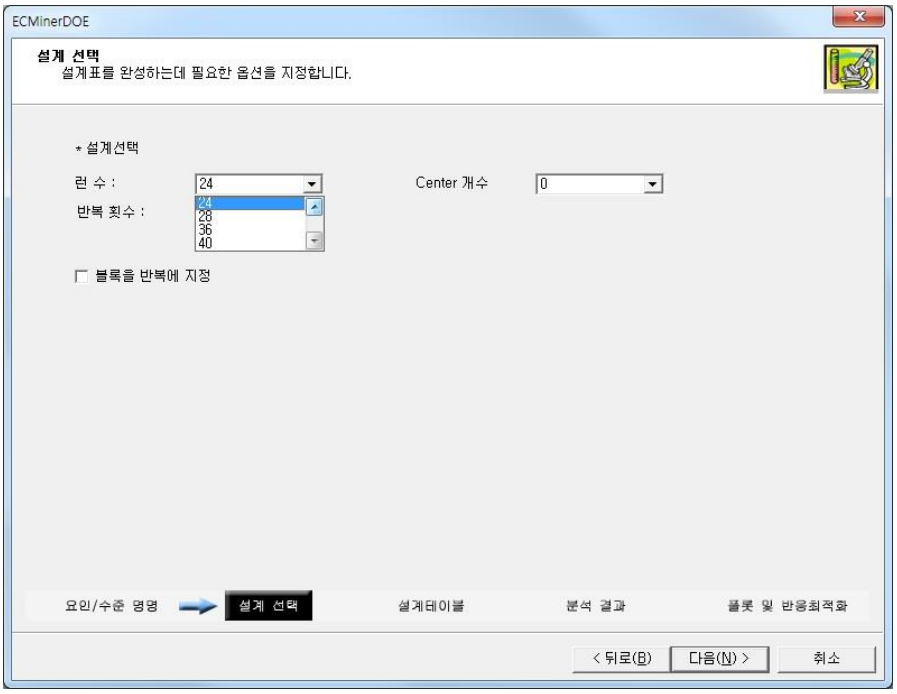
C 라는 요인에 해당하는 Column 이 하나 더 생긴 것을 볼 수 있는데 이 Column 의 값은 A Column 에 있는 값과 B Column 에 있는 값을 서로 곱한 것과 같은 값입니다. 이런 식으로 사용자의 필요에 따라 서로 다른 생성자를 이용하여 설계표를 만들 수 있습니다. 이 후의 분석 과정은 2 수준 요인 설계(기본 생성자)의 과정과 동일합니다.

6.3.1.4. Plackett Burman 설계

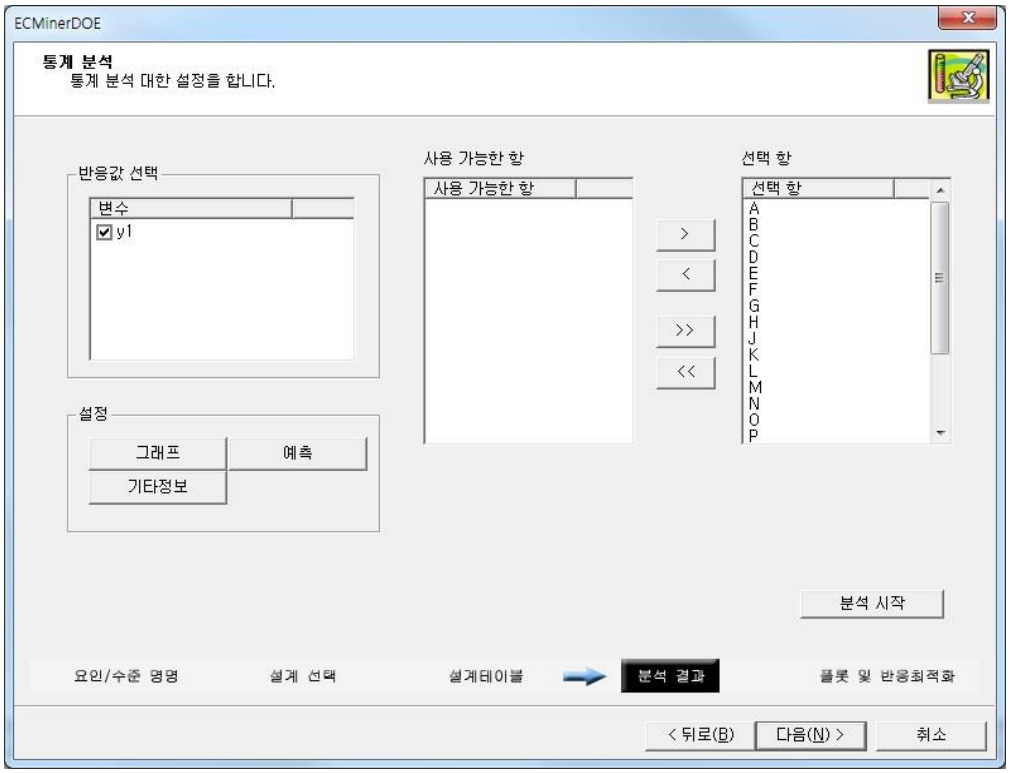
본 설계는 요인의 개수가 매우 많을 때 사용하는 설계입니다. 본 설계는 사용자가 요인이 반응값에 줄 수 있는 효과는 오직 주효과뿐이라는 사전지식이 있을 때 사용 가능합니다. 이러한 사전지식은 실험의 수를 매우 획기적으로 줄이는데 도움이 됩니다.
Plackett Burman 설계는 2 수준 요인 설계와 분석 이후의 단계가 비슷합니다. 따라서 2 수준 요인 설계와 비교하여 차이가 있는 것을 위주로 설명하도록 하겠습니다.

Plackett Burman 의 특징은 실험자가 사전에 요인과 반응 값 사이에 선형적인 관계만이 있다는 것을 알 때 실험의 수를 획기적으로 줄인다는 것에 있습니다. 예를 들어 47 개의 요인에 대해서 48 번의 실험만으로도 주효과를 검출해 낼 수가 있습니다. 이렇게 실험을 획기적으로 줄일 수 있다는 사실이 바로 Plackett Burman 설계를 사용하는 이유입니다.

Plackett Burman 설계를 선택하고 Step 1 화면에서 요인을 20 개를 선택합니다. Step 2 그 후에 Step 2 화면으로 넘어가면 다음과 같은 실험 수를 선택할 수 있습니다.



만약 실험자가 한번의 실험을 최대한 적게 하고 싶으면 런 수가 24 번인 설계를 선택하면 됩니다. 최대 런수가 48 까지 선택가능 하지만 20 개의 요인에 48 번의 실험은 그리 부담스러운 수는 아닙니다. 사용자의 목적에 맞게 런 수를 설정하고 중심점의 개수와 반복 회수를 설정합니다. 중심점은 곡면성을 체크하기 위한 것이며 반복회수를 늘림으로써 같은 실험점에서 반응 값이 얼마나 큰 차이가 나는 것을 통해 순수 오차(Pure Error)를 체크할 수 있습니다.
Step 3 에서는 설계 테이블이 완성되고 Step 4 에서는 다음과 같은 화면이 나타납니다.



다른 설계와는 다르게 포함할 수 있는 항의 최대 차수라는 항목을 선택할 수 없습니다. 왜냐하면 Plackett Burman 설계의 특징상 주효과 외의 상호작용은 검출할 수가 없기 때문입니다. 위의 선택항은 따라서 주효과를 검출하고자 하는 요인을 말합니다. 요인을 선택하여 분석을 진행하면 그 이후의 과정은 2 수준 요인 설계(기본 생성자)와 동일합니다.

6.3.1.5. 일반 완전 요인 설계

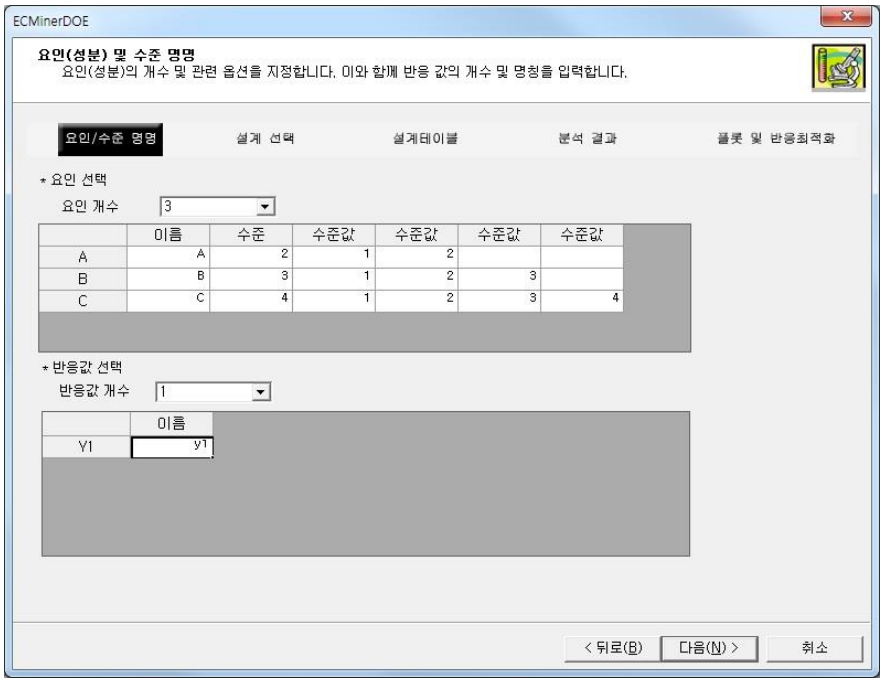
일반 완전 요인 설계는 앞의 3 가지 설계 (2 수준 요인 설계(기본 생성자), 2 수준 요인 설계(생성자 지정), Plackett Burman 설계) 가 오직 2 가지 수준만의 실험을 할 수 있다는 한계를 극복할 수 있는 실험 설계입니다. 각 요인의 수준이 어떠한 값도 가질 수 있습니다. 예를 들어 요인의 개수가 3 개일 때 수준 수를 각각 2, 3, 4 와 같이 가질 수 있는 것입니다.

일반완전요인설계에서는 가능한 모든 조합에 대해서 실험을 하기 때문에 수준 수가 2, 3, 4 라면 다음과 같은 실험 설계 표를 얻게 됩니다.

A	B	C
1	1	1
2	1	1
1	2	1
2	2	1
1	3	1
2	3	1
1	1	2
2	1	2
1	2	2
2	2	2
1	3	2
2	3	2
1	1	3
2	1	3
1	2	3
2	2	3
1	3	3
2	3	3
1	1	4
2	1	4
1	2	4
2	2	4
1	3	4
2	3	4

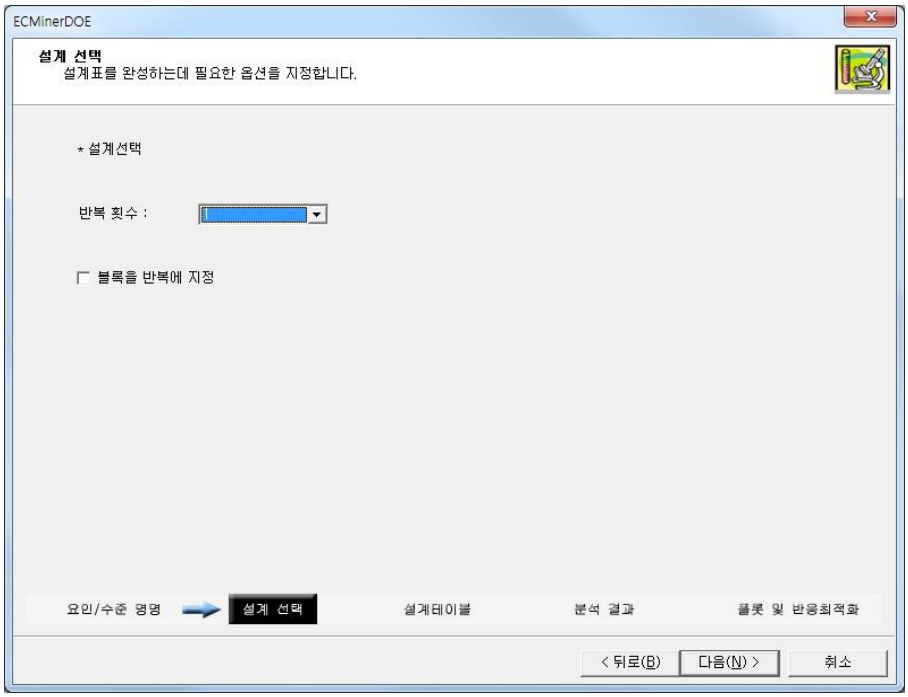
일반 완전 요인 설계를 설명하기 위해서 다음의 과정을 따릅니다.

- **Step 1: 요인(성분) 및 수준 명명**



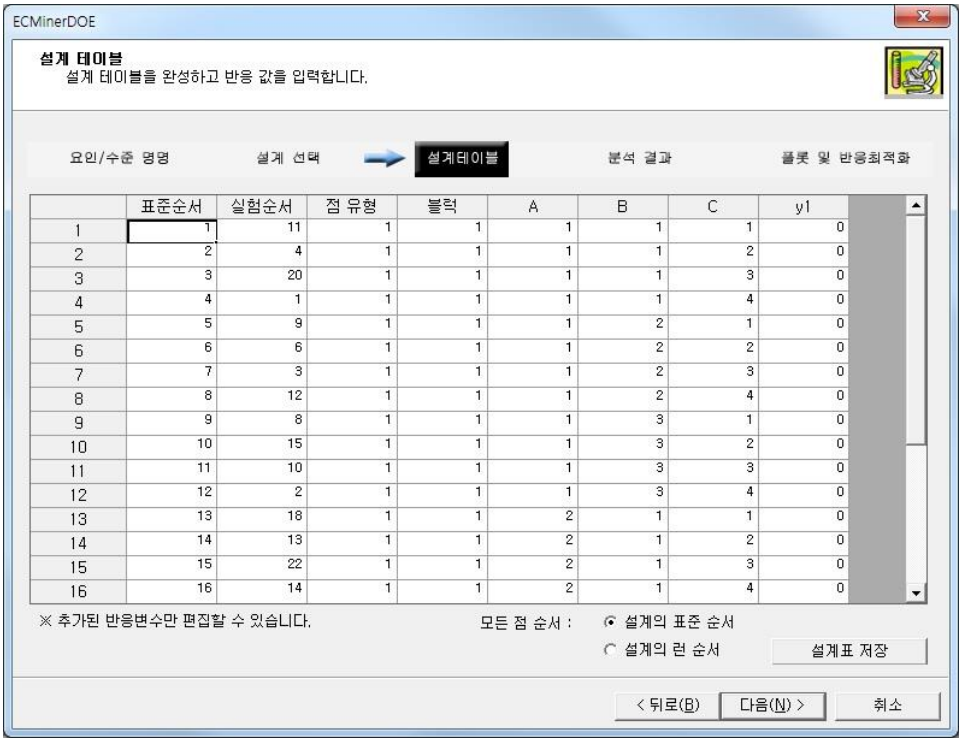
어떠한 3 개의 요인이 있어서(A, B, C) 각각 2 수준, 3 수준, 4 수준의 값을 가질 수 있으면 위와 같이 설정 값을 입력해 주면 됩니다.

▪ **Step 2: 설계 선택**



설계 선택 창에서는 반복횟수와 블록을 반복에 지정할 것인지의 여부를 결정합니다.

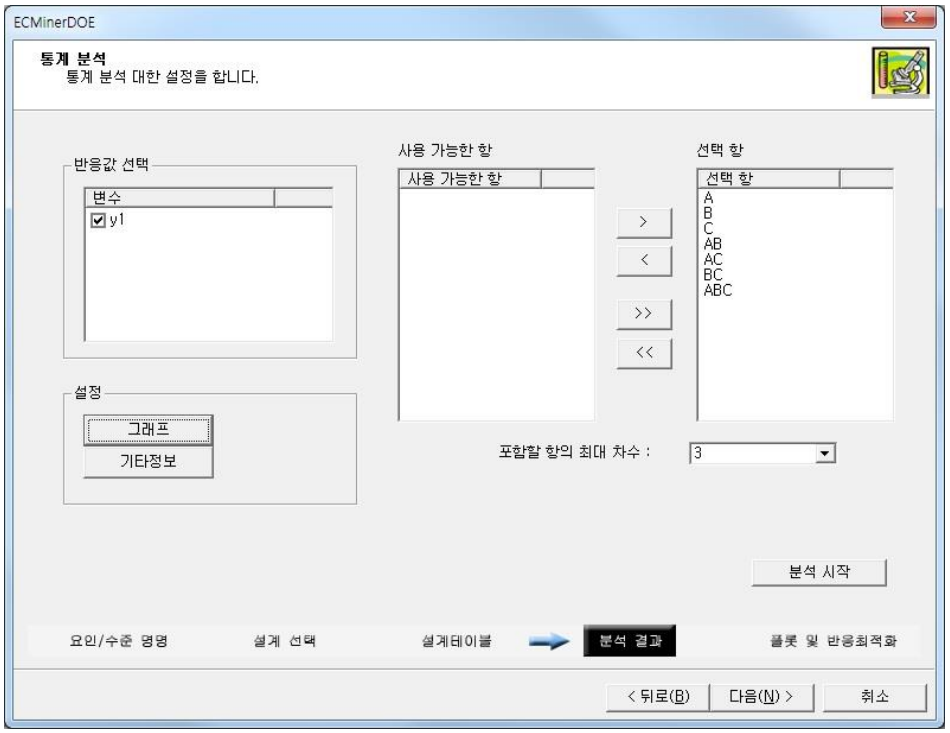
▪ **Step 3: 설계 테이블**



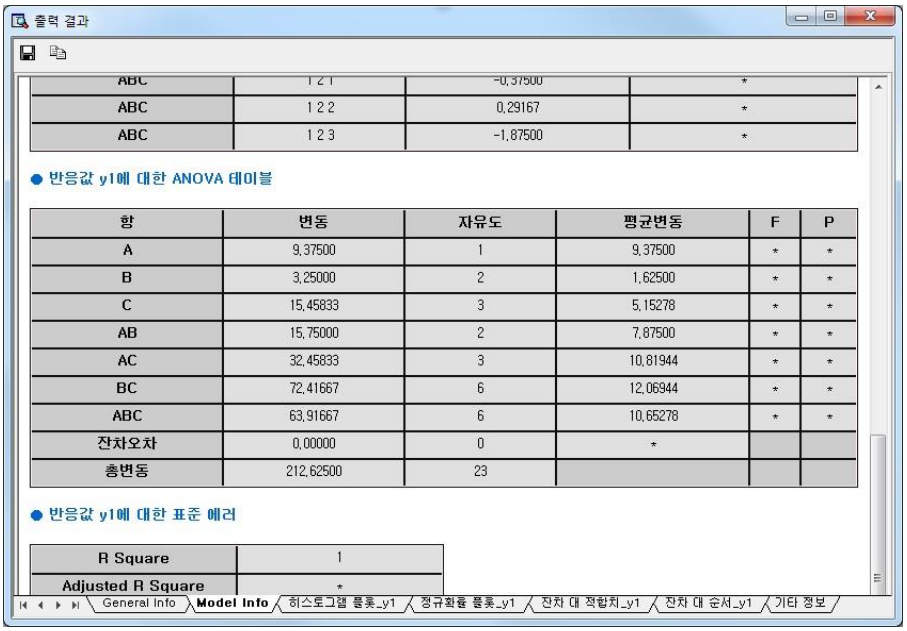
	표준순서	실험순서	점 유형	블럭	A	B	C	y1
1	1	11	1	1	1	1	1	0
2	2	4	1	1	1	1	2	0
3	3	20	1	1	1	1	3	0
4	4	1	1	1	1	1	4	0
5	5	9	1	1	1	2	1	0
6	6	6	1	1	1	2	2	0
7	7	3	1	1	1	2	3	0
8	8	12	1	1	1	2	4	0
9	9	8	1	1	1	3	1	0
10	10	15	1	1	1	3	2	0
11	11	10	1	1	1	3	3	0
12	12	2	1	1	1	3	4	0
13	13	18	1	1	2	1	1	0
14	14	13	1	1	2	1	2	0
15	15	22	1	1	2	1	3	0
16	16	14	1	1	2	1	4	0

설계 테이블이 생성된 후에는 반응값 y1 을 입력하고 다음 단추를 클릭하여 Step 4 로 넘어갑니다.

▪ **Step 4: 통계 분석**



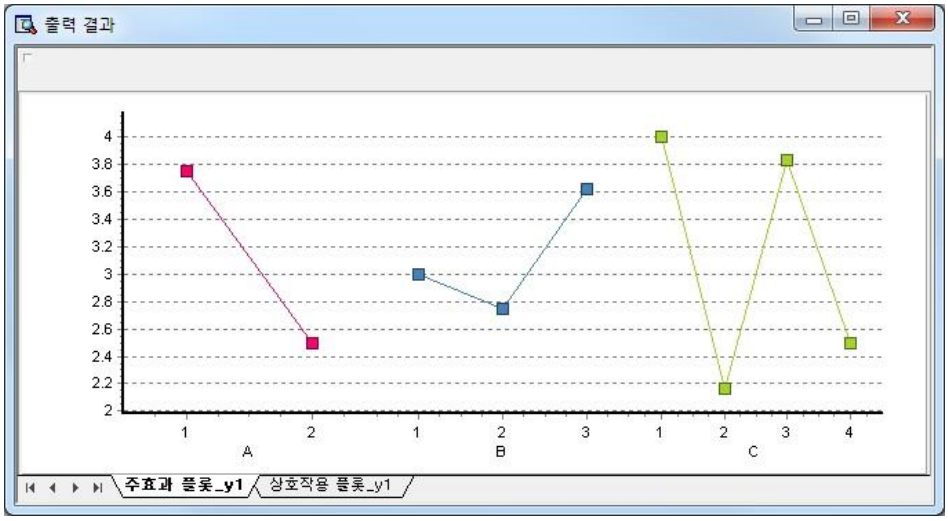
포함할 항의 최대 차수를 3 으로 하고 분석을 시작합니다. 포함할 항의 최대 차수를 요인의 개수만큼 지정함으로써 모든 상호작용에 대해서 반응값 y1 이 어느 정도의 영향을 받는지 체크해 볼 수 있습니다. 그래프, 기타정보에 대한 설정을 마친 후 분석 시작을 하면 다음과 같은 결과 화면이 나타납니다.

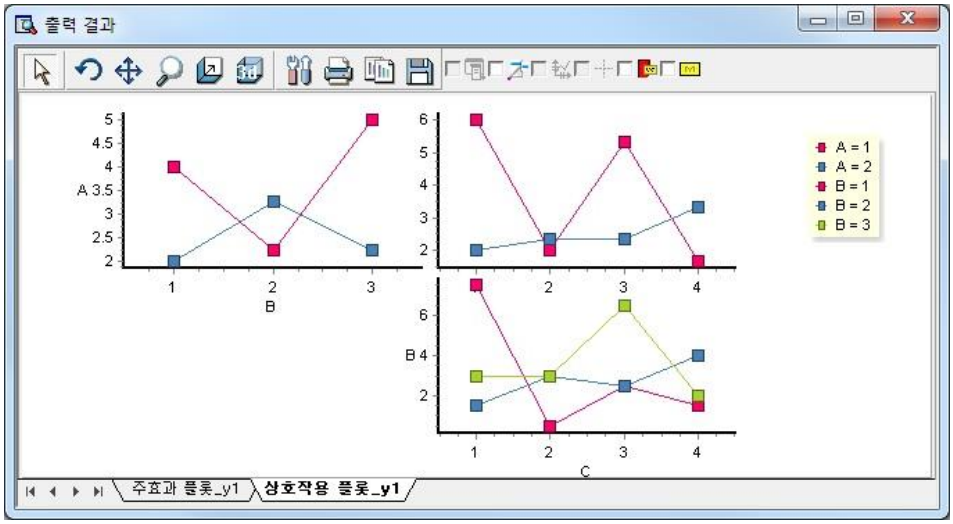


General Info, 잔차 분석 및 기타 정보에 대한 해석은 2 수준 요인 설계(기본 생성자)와 동일합니다. 분산 분석의 결과인 ANOVA 테이블을 볼 때 상호작용의 차수가 높을 때 변동이 매우 커짐을 알 수 있습니다. 이 같은 경우는 단지 A, B, C 각각의 효과로는 반응 값을 설명하기 힘들고 복합적인 작용에 의하여 반응값 y1 이 결정됩니다.

▪ **Step 5: 플롯 및 반응 최적화**

본 단계에서는 주효과 플롯과 상호작용 플롯을 그릴 수 있습니다. (설계의 특성상 표면 플롯, 등고선 플롯, 반응 최적화 기능은 제공하지 않습니다.) 아래와 같은 주효과 플롯과 상호작용 플롯을 통해 요인의 각 수준에 따라 반응값 y1 이 어떠한 움직임을 보이는지 시각적으로 이해할 수 있습니다.





6.3.2 반응 표면 설계

6.3.2.1. 개요

반응 표면 설계는 여러 개의 설명 변수 (요인)이 복합적인 작용을 함으로써 어떤 반응 값에 영향을 주고 있을 때, 설명 변수와 반응 값 사이의 관계를 규명하는 통계적인 분석 방법입니다. 예를 들어 어떤 화학 반응에 있어서 반응량이 온도와 시간에 따라 변화한다고 합니다. 이 때 온도와 시간을 x_1, x_2라고 하고 반응량을 y 라고 하면

$$y = f(x_1, x_2)$$

의 관계를 가질 것입니다. 이렇게 와 y 사이의 관계를 방정식으로 표현하여 주고자 하는 것이 반응 표면 설계에서 다루는 내용입니다. 실험자가 인자와 반응 사이에 대해서 어느 정도의 사전지식을 가지고 있을 경우 적은 실험으로 인자와 반응 사이의 관계를 규명할 수 있도록 방법론을 제시하는 것이 바로 반응 표면 설계라고 할 수 있습니다. ECMiner™ DOE 의 반응 표면 설계는 다음의 두 가지 방법론을 제공합니다.

- 중심 합성 설계
- Box Behnken 설계

6.3.2.2. 중심 합성 설계

독립변수의 수가 k 인 2 차 회귀모형은

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i \le j}^{k} \beta_{ij} x_i x_j + \epsilon \quad \epsilon \sim N(0, \sigma^2)$$

으로 2 수준 요인 배치법으로 회귀계수를 추정할 수 없습니다. 왜냐하면, 2 수준 요인실험에서는 각 변수의 두 수준에서만 실험이 되므로, 변수의 수준 변화에 따라서 발생되는 반응량의 곡면적인 변화를 감지할 수 없으며, 2 차 회귀 모형에서 제곱항의 계수 등을 추정할 수 없습니다. 이런 단점을 보완하고 적은 횟수의 실험으로 곡면을 추정하기 위해서 다음과 같이 중심점과 축점을 2 수준 요인 실험에 추가시킨 실험계획을 중심합성계획이라고 부릅니다. 중심합성계획에서는 중심점의 수는 제한이 없이 하나 이상이면 되며, 축점의 수는 2k 개가 됩니다. 여기서 k 의 값은 0 이 아닌 양수면 됩니다. 예를 들어 요인이 3 이고 중심점이 두 개이고 완전요인 설계에서의 중심합성 설계는 다음과 같습니다.

$$\begin{bmatrix} -1 & -1 & -1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \\ 1 & -1 & 1 \\ -1 & 1 & 1 \\ 1 & 1 & 1 \\ \alpha & 0 & 0 \\ -\alpha & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & -\alpha & 0 \\ 0 & 0 & \alpha \\ 0 & 0 & -\alpha \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

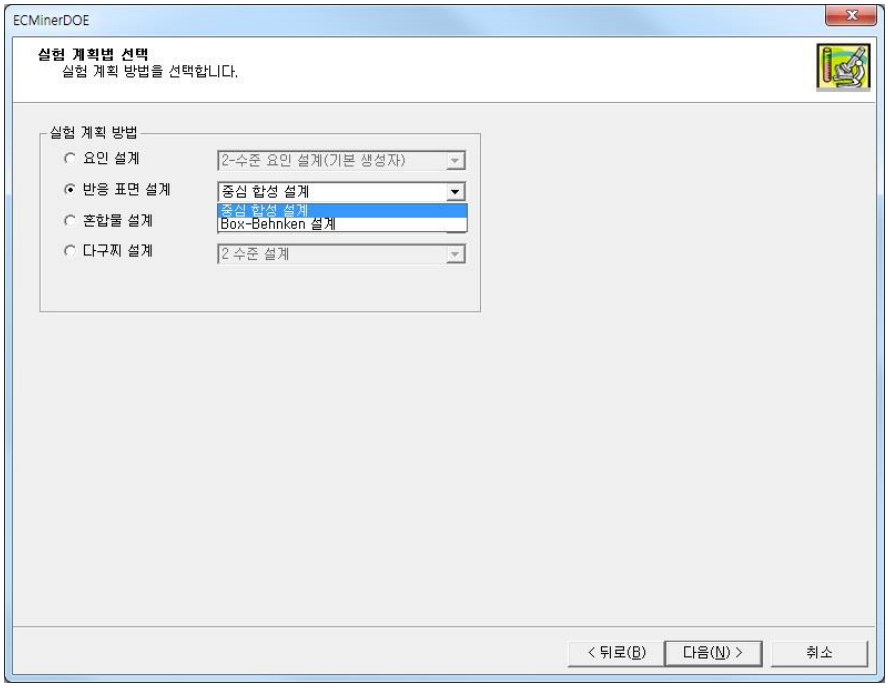
1 행부터 8 행까지는 격자 점에서의 실험입니다. 그리고 9 행부터 14 행까지는 축점입니다. 그리고 마지막 두 행은 중심점입니다. 이렇게 하면 2 차 곡선에 대한 회귀식을 충분히 잘 추정할 수 있는 설계가 됩니다. 이와 같은 실험 설계는 3 수준 요인 설계보다 훨씬 적은 수의 실험으로 2 차 곡선에 대한 회귀식을 찾을 수 있다는 장점이 있습니다.
이 실험의 실험 횟수는 모두

$$n = 2^k + 2k + n_0, \quad n_0 \text{ is a number of center points.}$$

입니다.
이 중심합성계획이 갖는 또 하나의 장점은 축차 실험이 가능하다는 것입니다. 예를 들면, 실험자가 2 수준 요인 실험 계획에 의하여 실험을 한 후에 반응표면에 관한 연구를 1 차 회귀모형을 사용하여 진행하다가 나중에 이 모형이 적당하지 않은 것을 발견하였습니다. 그러면 처음부터 새로이 2 차 회귀모형을 추정하기 위하여 다른 실험계획법을 사용할 필요 없이 이제까지 얻은 2 수준 요인 실험에 추가하여 중심과 축에 새로운 설계점을 증가시키면 간단히 우리가 쉽게 이용할 수 있는 중심합성계획이 됩니다.

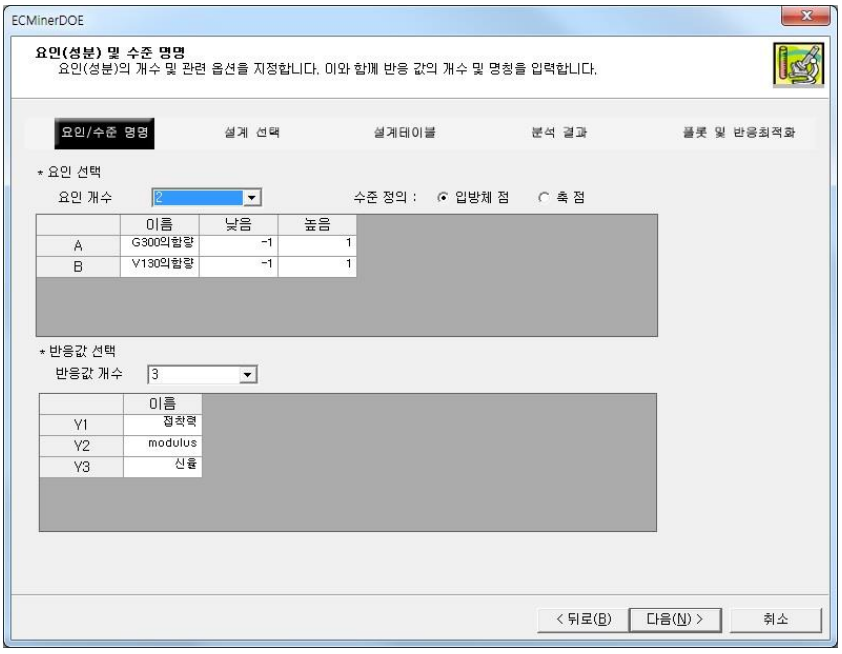
실험 소개
본 실험은 1979 년 모 타이어 회사에서 Monopoly radial 타이어의 주행성능을 향상시키기 위하여 실시된 반응 표면 실험 계획법 적용 사례입니다. G300 의 함량, V130 의 함량이 접착력, modulus, 신율이라는 반응 값에 영향을 미치는 실험입니다.

반응 표면 설계의 중심 합성 설계를 선택합니다.



▪ **Step 1: 요인(성분) 및 수준 명명**

먼저 다음과 같이 요인과 각 수준을 명명합니다. 이 때 낮음, 높음 단위는 바꾸지 않고 코드화된 단위를 사용하도록 합니다. 본 실험에서 반응 값은 접착력, modulus, 신율이 될 수 있으므로 반응 값의 개수는 3 으로 하고 다음 버튼을 클릭하여 진행합니다.



▪ **Step 2: 설계 선택**

다음의 화면에서 설계 선택, 중심점 개수, 반복 수, 알파 값을 지정합니다.

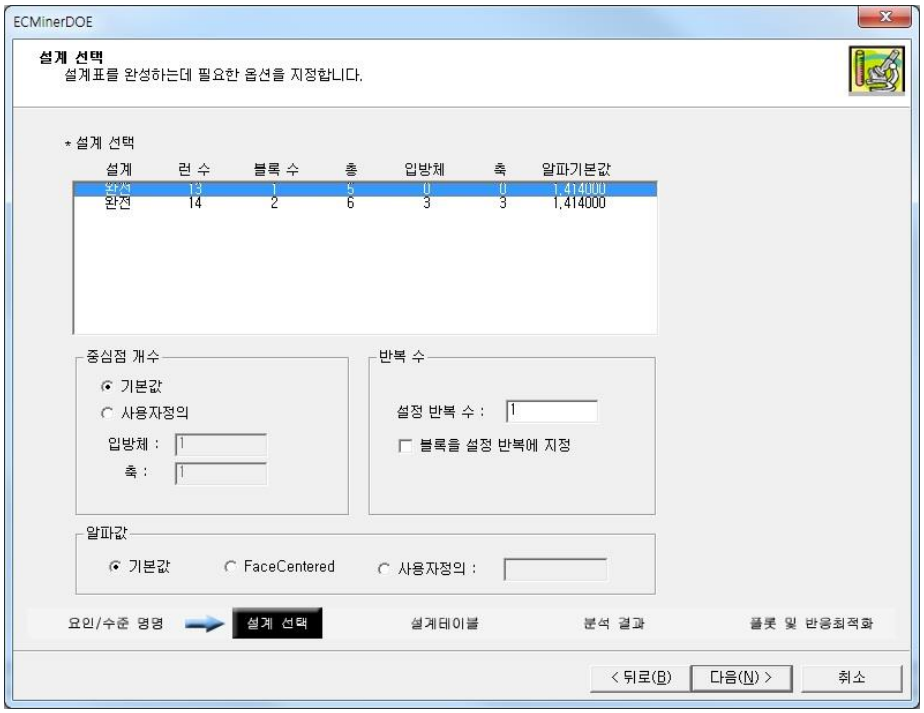
설계 선택: 사용할 수 있는 설계 유형 중에 하나를 선택합니다. 사용 가능한 설계에서 '완전'이 의미하는 것은 격자 점이 2 수준 완전 요인 설계의 모든 점을 포함한다는 것이고 '부분'이 의미하는 것은 격자 점이 2 수준 부분 요인 설계의 점을 포함한다는 것입니다.

중심점 개수: 기본 값 혹은 사용자 지정 값을 입력합니다.

반복 수: 기본 반복수는 1 이지만 사용자 임의로 반복수를 입력할 수 있습니다. 필요에 따라서 블록을 반복에 지정할 수도 있습니다.

알파 값: 기본값을 사용하거나 Face Centered(알파 값이 1), 사용자 정의 값을 입력할 수 있습니다.

각 요인 점과 축 점, 그리고 중심점에서 실험을 세 번씩 하려고 하므로 다음 화면에서와 같이 중심점의 개수를 1 개로 하고 알파의 값은 1(Face Centered), 그리고 설정 반복 수는 3 으로 하고 다음 버튼을 클릭하여 진행합니다.



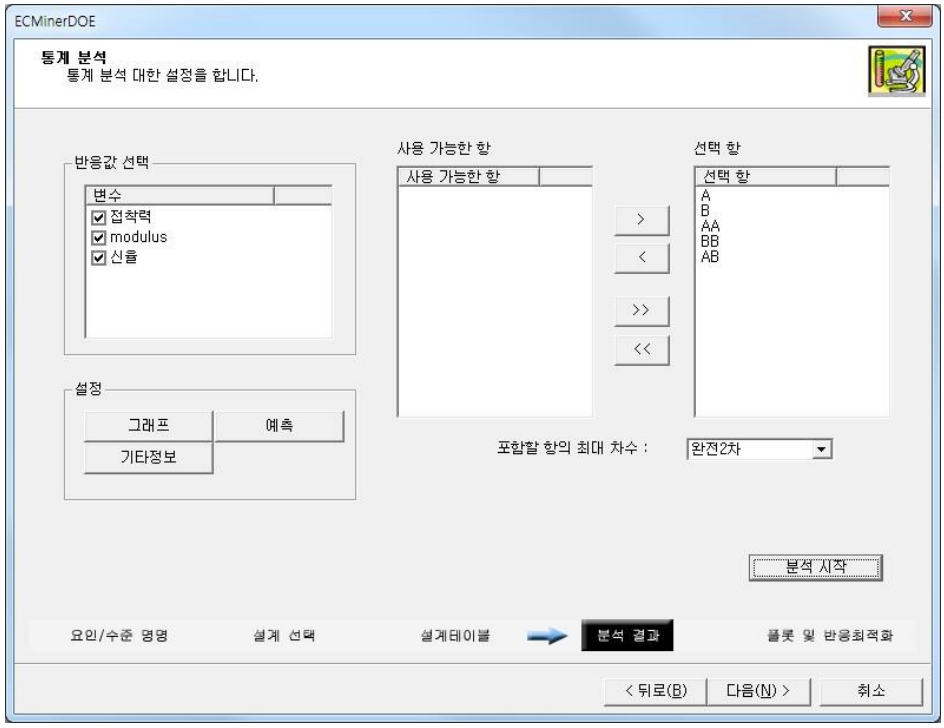
▪ **Step 3: 설계 테이블**

본 단계에서는 Step 1, Step 2 의 설정으로부터 설계표가 완성됩니다. 완성된 설계표에서 접착력, modulus, 신율에 해당하는 반응 값을 실험을 통해 입력하도록 합니다. 이 때 D Optimal Design 옵션의 경우 만들어진 설계표를 변형하는 방법론으로 이에 대해서는 6.3.2.4. 반응 표면 설계 D Optimal Design 을 참고하세요.

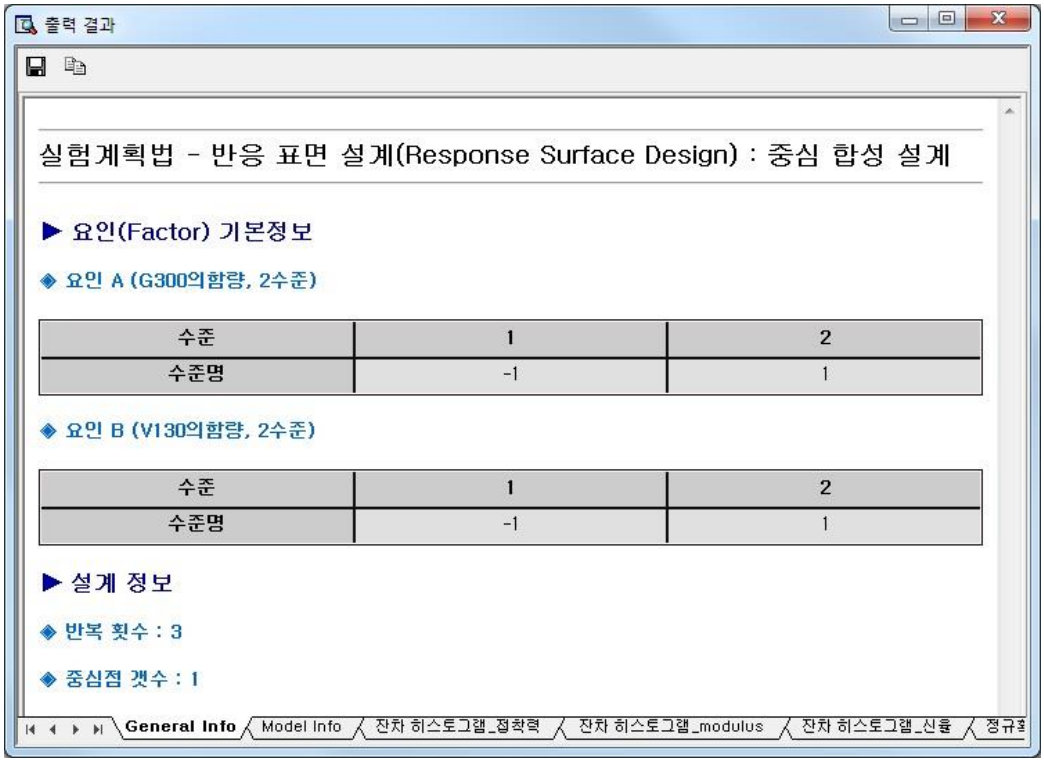
	표준순서	실험순서	점 유형	블럭	G300의함량	V130의함량	접착력	modulus	신율
1	1	29	1	1	-1,000000	-1,000000	0	0	0
2	2	30	1	1	1,000000	-1,000000	0	0	0
3	3	17	1	1	-1,000000	1,000000	0	0	0
4	4	16	1	1	1,000000	1,000000	0	0	0
5	5	6	-1	1	-1,414000	0,000000	0	0	0
6	6	27	-1	1	1,414000	0,000000	0	0	0
7	7	2	-1	1	0,000000	-1,414000	0	0	0
8	8	8	-1	1	0,000000	1,414000	0	0	0
9	9	12	0	1	0,000000	0,000000	0	0	0
10	10	32	0	1	0,000000	0,000000	0	0	0
11	11	13	0	1	0,000000	0,000000	0	0	0
12	12	31	0	1	0,000000	0,000000	0	0	0
13	13	26	0	1	0,000000	0,000000	0	0	0
14	14	25	1	1	-1,000000	-1,000000	0	0	0
15	15	19	1	1	1,000000	-1,000000	0	0	0
16	16	36	1	1	-1,000000	1,000000	0	0	0

▪ **Step 4: 통계 결과**

본 단계는 실험 결과를 분석하기 위한 단계입니다, 세 반응 값에 대해 모두 분석을 할 것이므로 반응 값 선택에서 세 반응 값을 모두 선택하고 그래프, 예측, 기타정보에서 필요한 설정을 하도록 합니다. 현재 완전 2 차로 회귀 분석을 하기를 원하므로 다음과 같이 선택을 하고 결과보기 단추를 클릭합니다.



General Info: 설계에 대한 일반적인 정보를 보여줍니다.



Model Info: 회귀분석 결과, 분산 분석 결과, 비정상적 관측치(극단화 레버리지, 표준화 잔차) 에 대한 정보를 보여줍니다.

회귀분석 결과

● 반응값 접착력에 대한 추정된 계수(코드화 된 단위)

항	계수	계수SE	T	P
Const	75	1,89267	39,62652	0
A	3,88889	1,03666	3,75137	0,00118
B	-1,44444	1,03666	-1,39337	0,17808
AA	-4,33333	1,79555	-2,41338	0,02502
BB	4,66667	1,79555	2,59902	0,01675
AB	2,66667	1,26964	2,10033	0,04796

● 반응값 modulus에 대한 추정된 계수(코드화 된 단위)

항	계수	계수SE	T	P
Const	144,14815	2,57564	55,96594	0
A	3,88889	1,41074	2,75664	0,01182
B	7,44444	1,41074	5,27699	0,00003
AA	-8,55556	2,44347	-3,5014	0,00213
BB	-3,55556	2,44347	-1,45513	0,16042
AB	0,25	1,72779	0,14469	0,88633

● 반응값 신율에 대한 추정된 계수(코드화 된 단위)

항	계수	계수SE	T	P
Const	402,40741	8,44079	47,67415	0
A	-9,44444	4,62321	-2,04283	0,05382
B	-20,	4,62321	-4,326	0,0003
AA	10,55556	8,00763	1,31819	0,20164
BB	2,22222	8,00763	0,27751	0,7841
AB	-3,75	5,66225	-0,66228	0,51499

● 표준 에러

R Square	0,61115
Adjusted R Square	0,51856
RMSE	4,39817

R Square	0,70357
Adjusted R Square	0,633
RMSE	5,98525

R Square	0,54487
Adjusted R Square	0,43651
RMSE	19,61462

분산 분석 결과

● 반응값 접착력에 대한 ANOVA 테이블

항	변동	자유도	평균변동	F	P
선형	309,77778	2	154,88889	8,00711	0,0026
제곱	243,33333	2	121,66667	6,28966	0,00724
상호작용	85,33333	1	85,33333	4,41138	0,04796
잔차오차	406,22222	21	19,34392		
Lack of Fit	56,88889	3	18,96296	0,9771	0,42539
PureError	349,33333	18	19,40741		
총변동	1044,66667	26			

● 반응값 modulus에 대한 ANOVA 테이블

항	변동	자유도	평균변동	F	P
선형	1269,77778	2	634,88889	17,72285	0,00003
제곱	515,03704	2	257,51852	7,1886	0,00419
상호작용	0,75	1	0,75	0,02094	0,88633
잔차오차	752,28704	21	35,82319		
Lack of Fit	351,62037	3	117,20679	5,26553	0,00876
PureError	400,66667	18	22,25926		
총변동	2537,85185	26			

● 반응값 신율에 대한 ANOVA 테이블

항	변동	자유도	평균변동	F	P
선형	8805,55556	2	4402,77778	11,44372	0,00044
제곱	698,14815	2	349,07407	0,90731	0,41885
상호작용	168,75	1	168,75	0,43862	0,51499
잔차오차	8079,39815	21	384,73325		
Lack of Fit	2446,06481	3	815,35494	2,60528	0,08349
PureError	5633,33333	18	312,96296		
총변동	17751,85185	26			

비정상적 관측치

● 반응값 접착력에 대한 비정상적 관측치 - 극단 레버리지

순서	실제 관측값	예측치	잔차	표준화 잔차

● 반응값 접착력에 대한 비정상적 관측치 - 표준화 잔차

순서	실제 관측값	예측치	잔차	표준화 잔차
6	85	74.55556	10.44444	2.63077
25	73	81.11111	-8.11111	-2.04305

● 반응값 modulus에 대한 비정상적 관측치 - 극단 레버리지

순서	실제 관측값	예측치	잔차	표준화 잔차

● 반응값 modulus에 대한 비정상적 관측치 - 표준화 잔차

순서	실제 관측값	예측치	잔차	표준화 잔차
7	144	133.14815	10.85185	2.00859

● 반응값 신율에 대한 비정상적 관측치 - 극단 레버리지

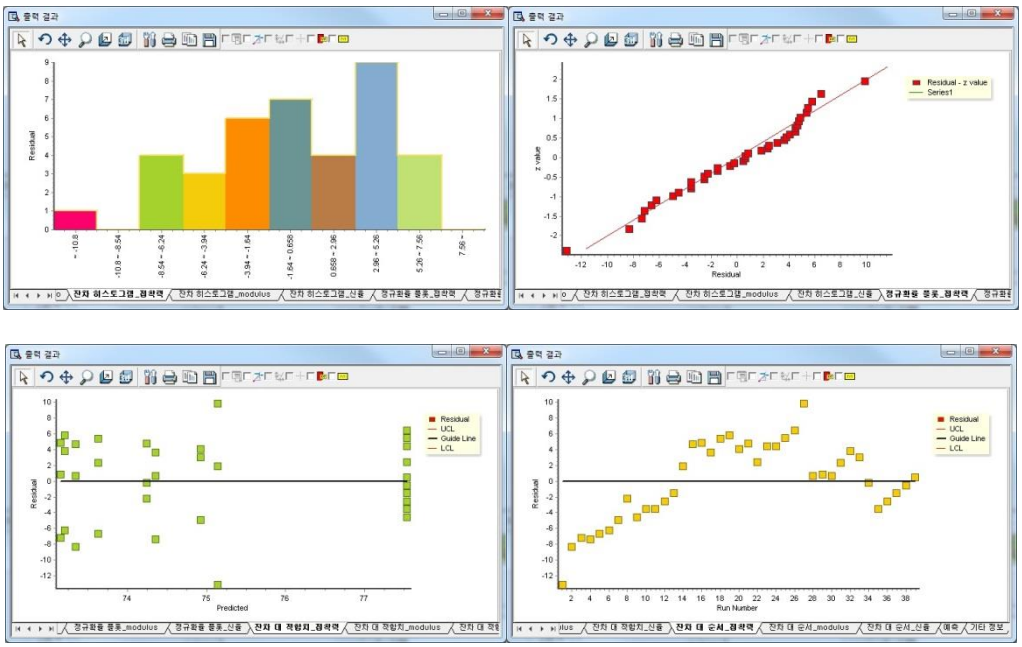
순서	실제 관측값	예측치	잔차	표준화 잔차

● 반응값 신율에 대한 비정상적 관측치 - 표준화 잔차

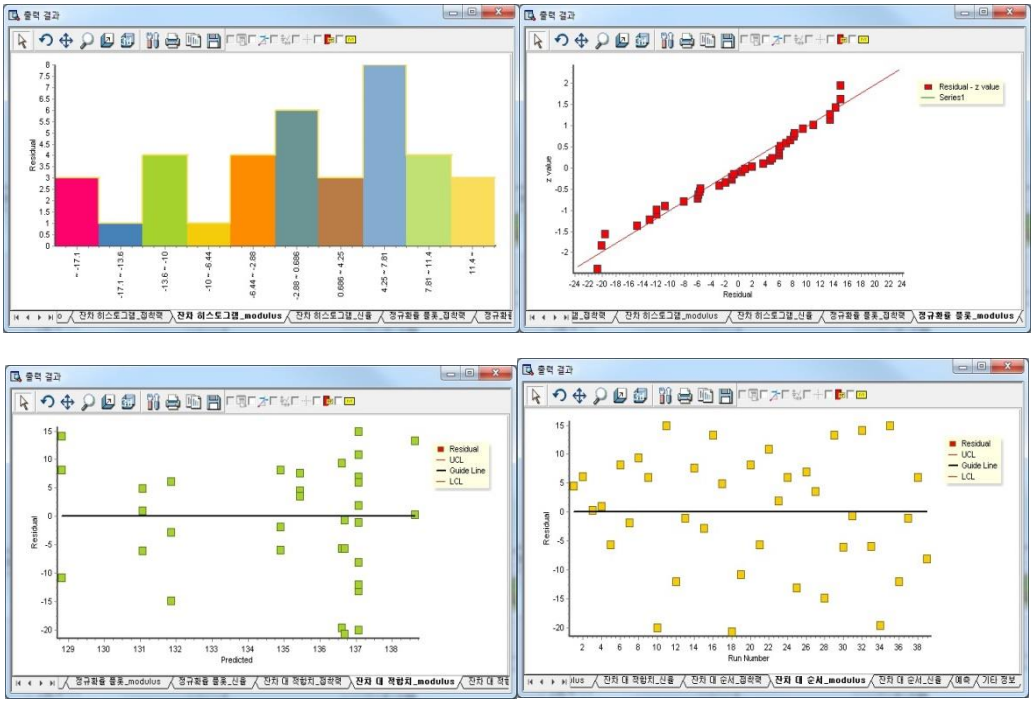
순서	실제 관측값	예측치	잔차	표준화 잔차
13	420	381.99074	38.00926	2.26573
17	340	384.62963	-44.62963	-2.52066

잔차 관련 플롯: 잔차 히스토그램, 잔차 정규확률 플롯, 잔차 대 순서, 잔차 대 적합치를 볼 수 있습니다.

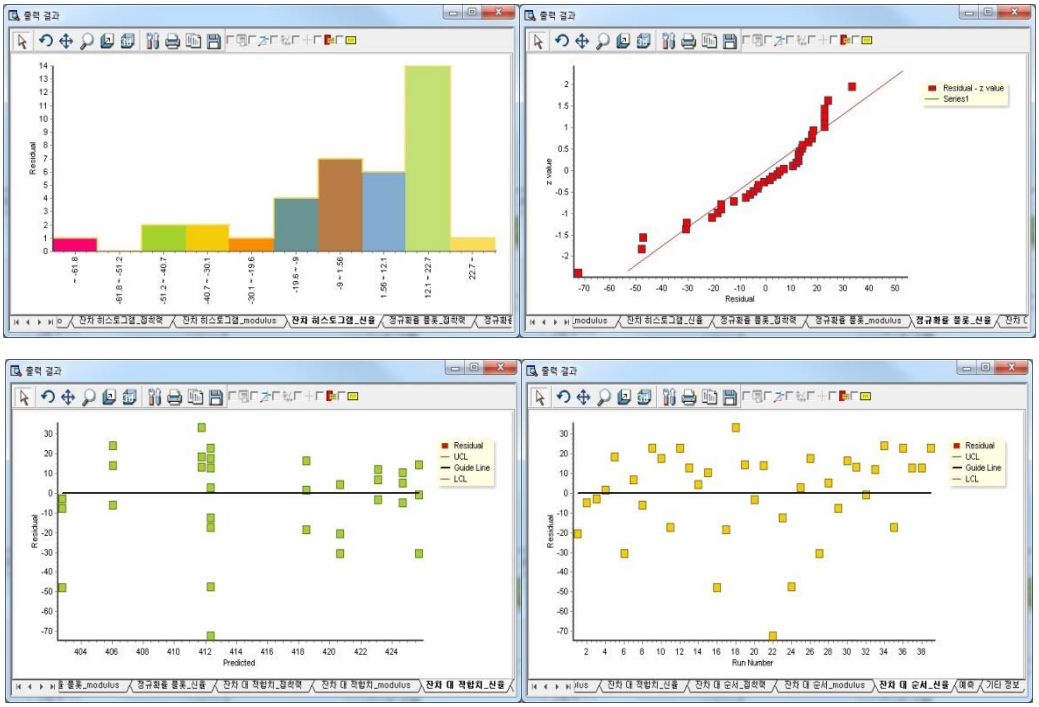
반응 값 접착력에 대한 잔차 관련 플롯



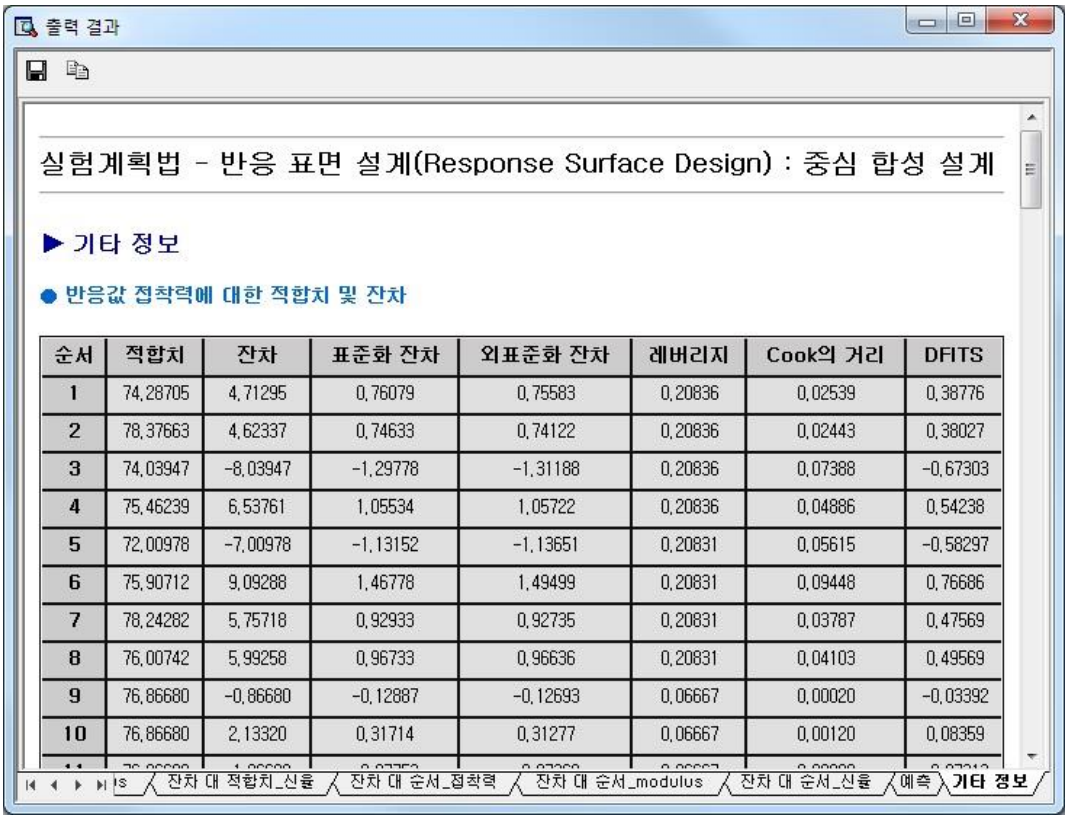
반응 값 **modulus** 에 대한 잔차 관련 플롯



반응 값 신율에 대한 잔차 관련 플롯



기타 정보: 잔차 관련 통계량을 보여줍니다.



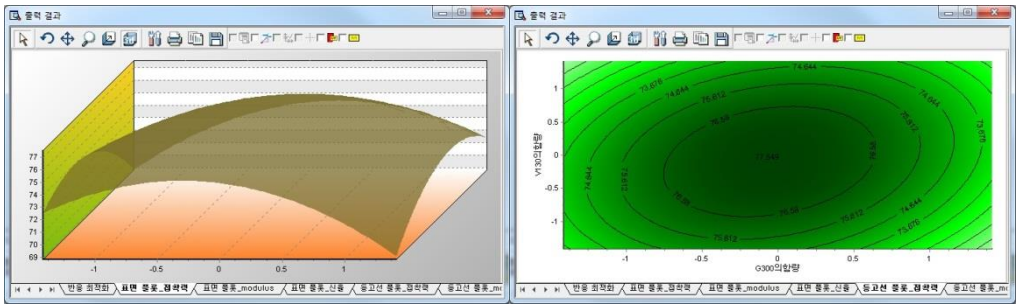
순서	적합치	잔차	표준화 잔차	외표준화 잔차	레버리지	Cook의 거리	DFITS
1	74,28705	4,71295	0,76079	0,75583	0,20836	0,02539	0,38776
2	78,37663	4,62337	0,74633	0,74122	0,20836	0,02443	0,38027
3	74,03947	-8,03947	-1,29778	-1,31188	0,20836	0,07388	-0,67303
4	75,46239	6,53761	1,05534	1,05722	0,20836	0,04886	0,54238
5	72,00978	-7,00978	-1,13152	-1,13651	0,20831	0,05615	-0,58297
6	75,90712	9,09288	1,46778	1,49499	0,20831	0,09448	0,76686
7	78,24282	5,75718	0,92933	0,92735	0,20831	0,03787	0,47569
8	76,00742	5,99258	0,96733	0,96636	0,20831	0,04103	0,49569
9	76,86680	-0,86680	-0,12887	-0,12693	0,06667	0,00020	-0,03392
10	76,86680	2,13320	0,31714	0,31277	0,06667	0,00120	0,08359

자세한 설명은 6.4. 설정 및 분석을 참고하세요.

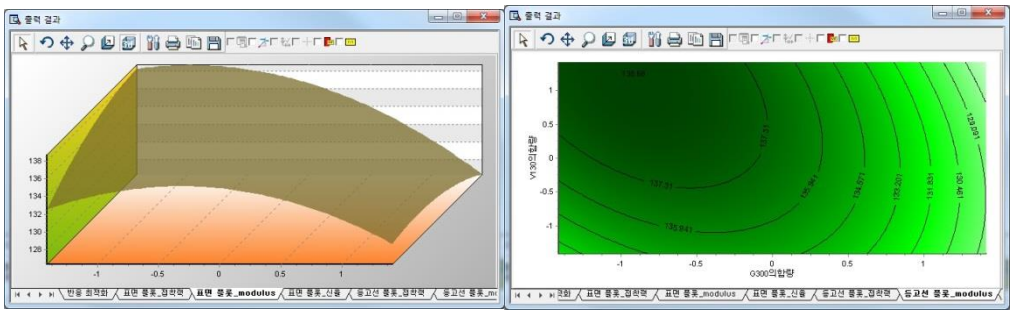
▪ **Step 5: 플롯 및 반응 최적화**

본 단계에서는 Step 4 에서 만든 Regression Model 로 표면 플롯, 등고선 플롯을 그리고 이와 함께 사용자의 목적에 따라 반응최적화를 할 수 있습니다. 현재 요인은 2 개이기 때문에 따로 고정 값을 입력할 필요가 없습니다. 각 반응 값에 따른 표면 플롯과 등고선 플롯을 그리면 다음과 같습니다.

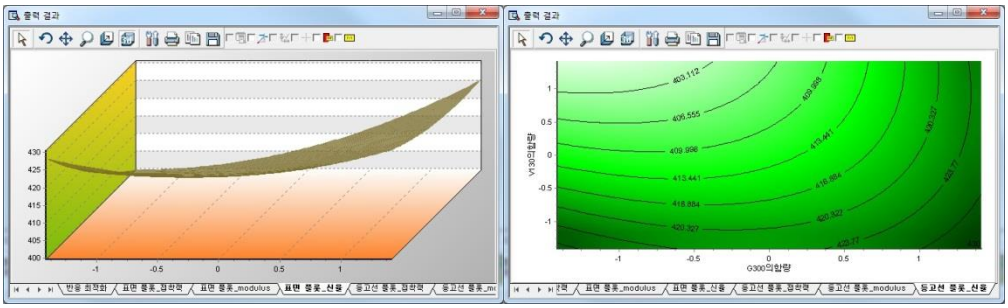
반응 값 접착력에 대한 표면 플롯, 등고선 플롯



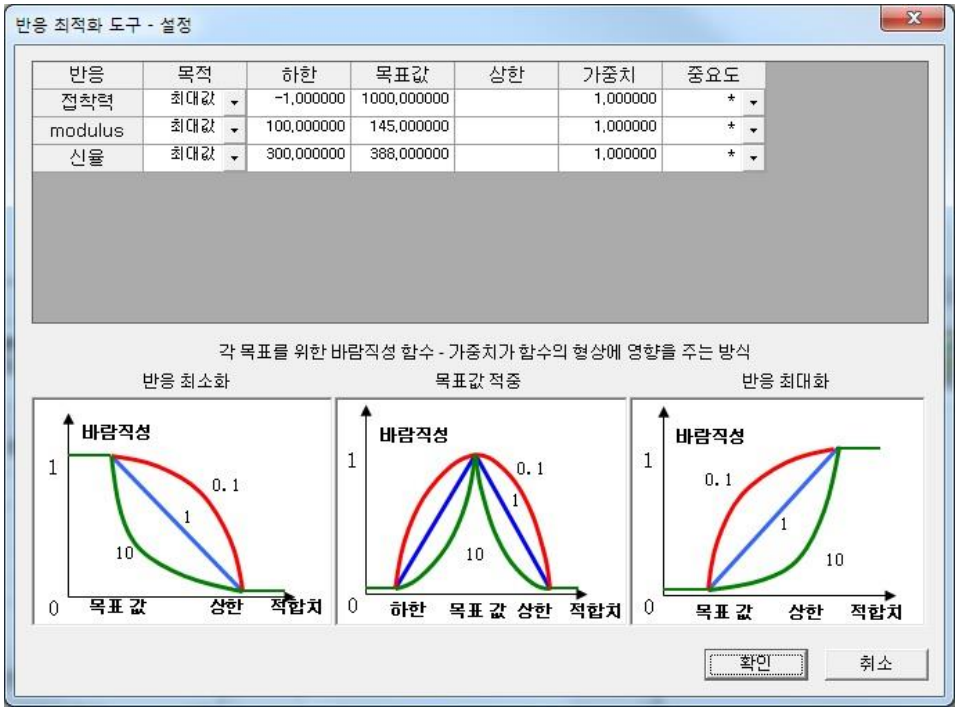
반응 값 modulus 에 대한 표면 플롯, 등고선 플롯



반응 값 신율에 대한 표면 플롯, 등고선 플롯

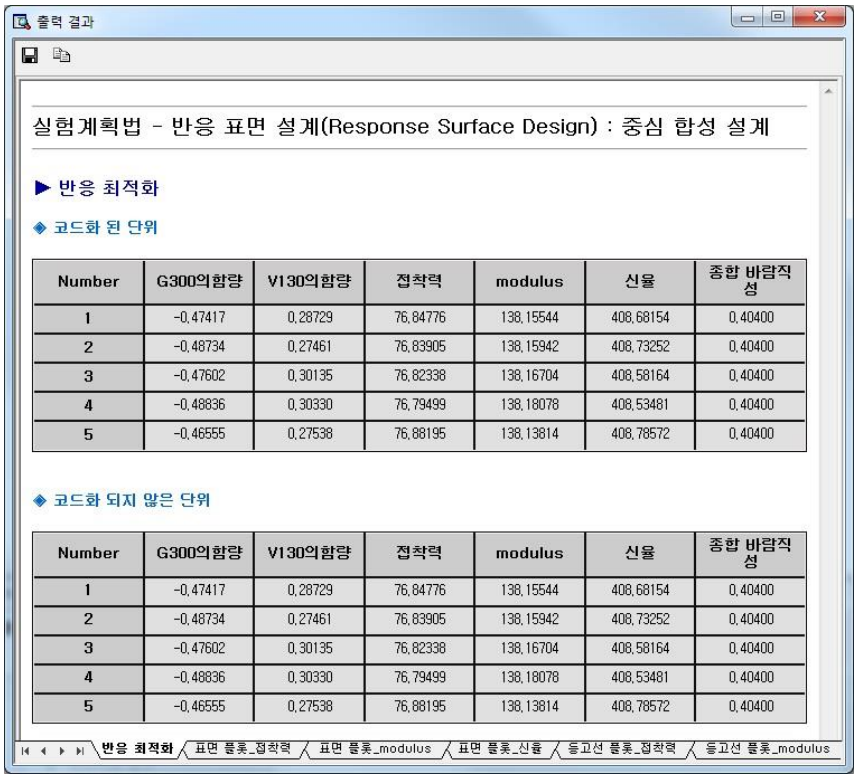


본 실험에서 실험자는 modulus 의 값은 145 이상, 신율의 값은 388 이상이면서 접착력의 값은 최대화하고자 하는 것이 목적입니다. 이를 위해서 다음과 같은 설정을 합니다.



위의 창에서 modulus 의 목표값, 신율의 목표값을 145, 388 로 한 것은 이 값에 도달하면 더 이상 값을 증가시킬 필요가 없다는 것입니다. 그리고 접착력의 목표 값을 최대한 크게 잡은 것은 접착력의 경우 크면 클수록 좋기 때문입니다. (사실 여기서 modulus 의 하한과 신율의 하한을 145, 388 로 잡는 것이 합리적일 것 같아 보입니다. 하지만 이 경우 modulus 값이 145 보다 작거나 신율의 값이 388 보다 작으면 바람직성 함수가 0 이 됩니다. 실제로 이러한 영역이 매우 넓어서 최적화를 수행할 때 최적 값을 잘 찾지 못하는 경우가 대부분입니다. 이를 위하여 하한을 좀 더 작게 잡아서 최적화 수행의 성능을 향상시킨 것입니다.)

이렇게 설정을 마치고 최적화를 수행하면 다음과 같은 결과를 얻을 수 있습니다.



Number	G300의함량	V130의함량	접착력	modulus	신율	종합 바람직성
1	-0.47417	0.28729	76.84776	138.15544	408.68154	0.40400
2	-0.48734	0.27461	76.83905	138.15942	408.73252	0.40400
3	-0.47602	0.30135	76.82338	138.16704	408.58164	0.40400
4	-0.48836	0.30330	76.79499	138.18078	408.53481	0.40400
5	-0.46555	0.27538	76.88195	138.13814	408.78572	0.40400

Number	G300의함량	V130의함량	접착력	modulus	신율	종합 바람직성
1	-0.47417	0.28729	76.84776	138.15544	408.68154	0.40400
2	-0.48734	0.27461	76.83905	138.15942	408.73252	0.40400
3	-0.47602	0.30135	76.82338	138.16704	408.58164	0.40400
4	-0.48836	0.30330	76.79499	138.18078	408.53481	0.40400
5	-0.46555	0.27538	76.88195	138.13814	408.78572	0.40400

이를 통해 볼 때 G300 의 함량이 -0.47417, V130 의 함량이 0.28729 이면 modulus 와 신율의 조건을 만족시키면서 접착력 76.84776 로 최대화시킬 수 있음을 알 수 있습니다.

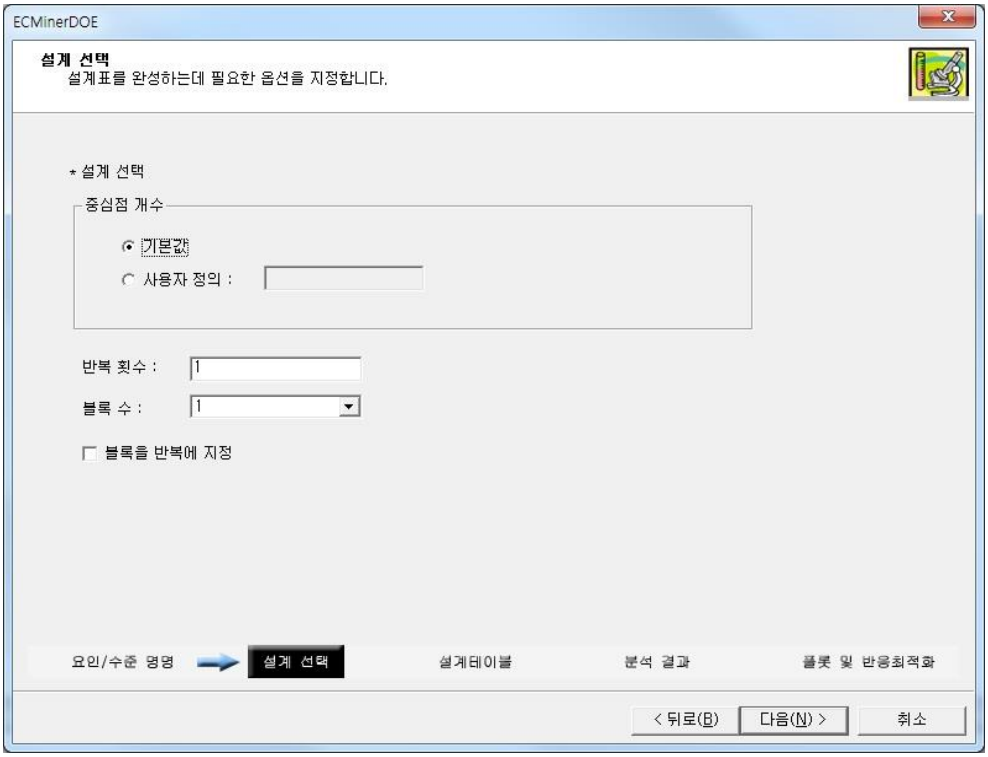
6.3.2.3. Box-Behnken 설계(내용 검토)

Box Behnken 의 설계는 2 수준 설계 및 3 수준 설계의 단점을 보완하기 위해서 만들어진 설계입니다. 일반적으로 p 수준 설계는 p-1 차의 polynomial 적합에 적당합니다. 따라서 2 수준 설계로는 second order polynomial 적합이 어렵고 3 수준 설계는 가능하긴 하지만 실험수가 많다는 단점이 있습니다. Box Behnken 의 설계는 3 수준 설계의 일부를 택하여 하는 실험으로서 더욱 효율적인 second order polynomial 적합이 가능합니다.

Box Behnken 설계 또한 중심 합성 설계와 같은 목적에서 만들어진 설계입니다. 따라서 이에 대해서는 자세히 예를 들어 설명하기 보다는 중심 합성 설계에 대한 차이점에 대해서 설명하도록 하겠습니다. (반응 표면 설계의 기본적인 방법에 대해서는 6.3.2.2. 중심 합성 설계를 참고하세요.)

반응 표면 설계의 Box Behnken 설계를 선택합니다. Step1 화면에서는 요인의 개수, 요인 이름, 그리고 상한과 하한을 입력하고, 반응 값의 개수와 반응 값에 따른 이름을 입력합니다.

Step 2 의 화면은 다음과 같습니다.



선택하는 옵션은 보는 바와 같이 매우 단순합니다. 이러한 옵션을 지정하면 그 이후의 과정은 중심 합성 설계와 동일합니다.

6.3.2.4. 반응 표면 설계 D Optimal Design

D Optimal Design 은 사용자의 필요에 따라 설계를 변형하고 싶을 때 향후 통계 분석 과정에 가장 적합하게 설계를 변형하는 방법론을 말합니다. 설계의 우수성을 판단하는 지표로 다음과 같은 지표가 있습니다.

D Optimality(Determinant)

D Optimality 는 가장 흔히 사용되는 기준으로 X'X 역행렬의 Determinant 를 가장 크게 만드는 설계를 찾을 때 사용됩니다. 여러 후보점들의 집합에서 필요한 후보점들을 모아서 만든 Design Matrix X 구할 때 이 중에서 X'X 역행렬의 Determinant 를 가장 크게 만드는 설계표를 D-Optimal Design 이라고 합니다.

A Optimality(Trace)

여러 후보점들의 집합에서 필요한 후보점들을 모아서 만든 Design Matrix X 구할 때 이 중에서 X'X 역행렬의 TRACE 를 가장 크게 만드는 설계표를 A-Optimal Design 이라고 합니다.

하지만 A-Optimality 의 경우 계산상의 어려움 때문에 잘 사용하진 않습니다.

G Optimality(평균 레버리지 / 최대 레버리지)

G Optimality 는 평균 레버리지를 최대 레버리지로 나눈 값을 의미합니다. 여기서 레버리지는 Generalized Linear Model Matrix 를 X 라고 할 때, H Matrix 를

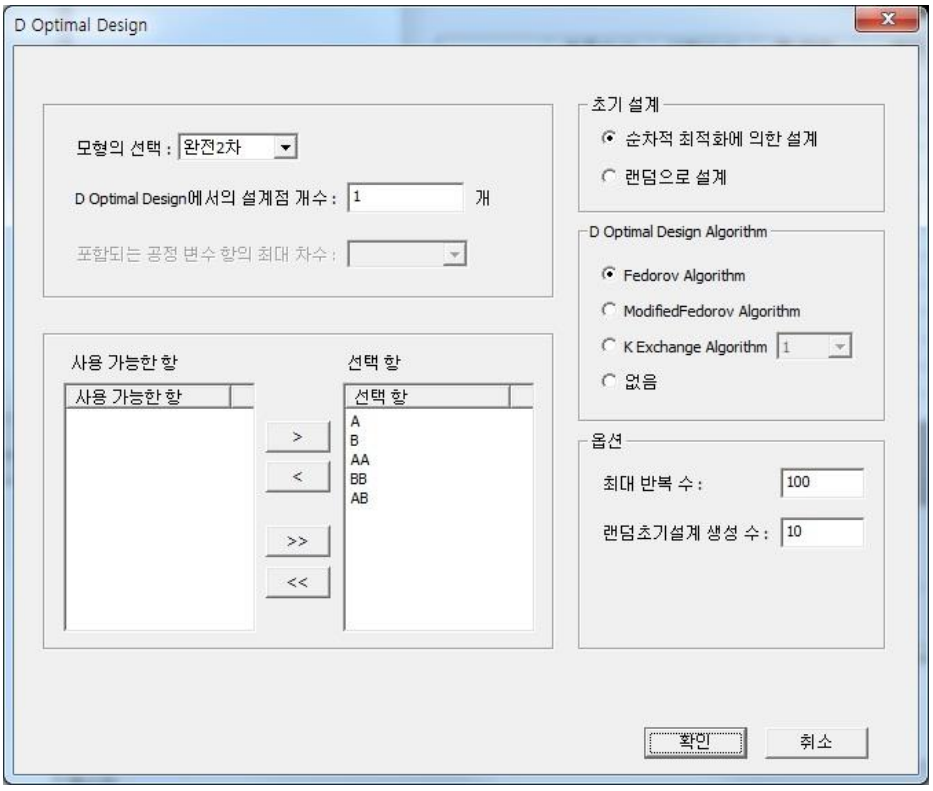
$$H = X(X'X)^{(-1)}X'$$

라고 하면 그것들의 diagonal component 를 말합니다. 이 레버리지들의 평균을 최대값으로 나누면 그것이 바로 G-Optimality 입니다.

V Optimality

레버리지의 평균이 V Optimality 입니다.

이런 여러 지표 중에서 D Optimality 를 최대화하기 위한 방법론으로 D Optimal Design 이라는 방법론이 사용됩니다. 반응 표면 설계의 D Optimal Design 화면은 다음과 같이 구성되어 있습니다.



- 모형의 선택: 먼저 모형을 선택합니다. 이렇게 선택한 모형에 따라 아래의 선택 항 창의 선택 항이 바뀌게 되는데 이 중에서 일부만 사용해서 쓸 수도 있습니다.
- D Optimal Design 에서의 설계 점 개수: 이는 변형된 설계표에서 과연 설계 점이 몇 개가 있을지를 정하는 것입니다.
- 초기 설계: D Optimal Design 을 수행하기 위해서는 초기 설계를 선택해야 합니다. D Optimal Design 에서의 설계 점 개수와 같은 수의 설계 점을 갖는 초기 설계를 어떻게 선택할지를 설정합니다.
- D Optimal Design Algorithm: 어떤 알고리즘을 사용하여 초기설계를 개선하여 D Optimal Design 을 얻을 것인지를 결정합니다. ECMiner™ DOE 에서는 Fedorov Algorithm, Modified Fedorov Algorithm, K Exchange Algorithm 을 제공하는데 경험적으로 이 중에서 Modified Fedorov Algorithm, K Exchange Algorithm 이 좋은 성능을 발휘한다고 알려져 있습니다.
- 옵션: 최대 반복 수란 D Optimal Design Algorithm 에서 최대 몇 번의 반복을 수행할 것인지를 의미합니다. 랜덤 초기 설계의 생성 수는 초기 설계를 랜덤으로 설계할 때 몇 개의 랜덤 설계를 만들지를 의미합니다. 이 때 만들어진 여러 랜덤 설계 중에서 D Optimality 가 가장 큰 것을 초기 설계로 사용하여 Algorithm 을 수행합니다.

6.3.3 혼합물 설계

6.3.3.1. 개요

대부분의 실험 계획은 하나 또는 두 개 이상의 인자 $(x_1, x_2, ..., x_k)$ 가 관심 있는 반응 값 y 에 유의한 영향을 미치는가를 발견하거나, 더 나아가서 y 를 최대 또는 최소화시키는 x_i 들의 최적조건을 찾는데 그 목적이 있습니다. 요인 설계, 반응 표면 설계 등이 여기에 속하며 이러한 실험 계획법들은 요인들이 취할 수 있는 상호간의 비율이나 그 합에 제약 조건이 없습니다. 하지만 혼합물 설계 방법론은 인자(혼합물 설계에서는 인자를 성분이라고 합니다.)들의 합이 일정한 경우의 실험 설계를 제공합니다.

$$\sum_{i=1}^{k} x_i = \text{일정}\ x_1 \geq 0, x_2 \geq 0, \ldots x_k \geq 0$$

ECMiner™ DOE 에서 제공하는 혼합물 설계 방법론은 다음과 같이 세가지 입니다.

- 심플렉스 중심 설계
- 심플렉스 격자 설계
- 꼭지점 설계

뿐만 아니라 혼합물 설계는 실험의 종류에 따라 세 가지로 나누어집니다.

- 혼합물 성분만을 사용하는 실험 설계
- 공정변수를 추가한 실험 설계
- 혼합물 양 실험 설계

결국 심플렉스 중심 설계에도 위와 같은 세 종류의 실험이 있고, 심플렉스 격자 설계에도 위와 같은 세 종류의 실험이 있고, 꼭지점 설계에도 위와 같은 세 종류의 실험이 있어서 결국 총 9 가지에 해당하는 방법을 제공하는 것입니다. 이를 설명하기 위해서 혼합물 성분만을 사용하는 실험 설계를 심플렉스 중심 설계를 이용하여 설명하고, 혼합물 양 설계를 심플렉스 격자 설계를 이용하여 설명하고, 공정변수를 추가한 실험을 꼭지점 설계를 이용해서 설명하겠습니다. 이를 통해서 혼합물 설계를 전반적으로 이해할 수 있습니다.

6.3.3.2. 심플렉스 중심 설계

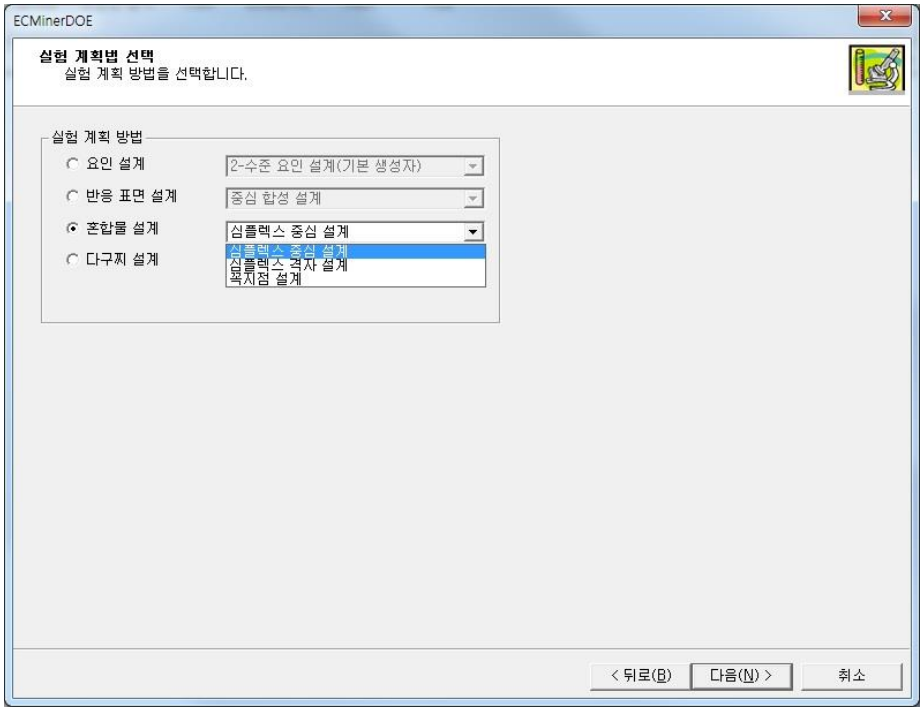
q 개의 Component 로 이루어진 심플렉스 중심 설계는 $2^q - 1$ 개의 서로 다른 실험점으로 이루어집니다. 이 점들은 가능한 모든 permutation 에 대응됩니다. 본 설계는 다음과 같은 Polynomial 에 적합을 하는데 적당한 설계라고 할 수 있습니다.

$$\eta = \sum_{1}^{q} \beta_i x_i + \sum_{i<j}^{q} \beta_{ij} x_i x_j + \sum_{i<j<k}^{q} \beta_{ijk} x_i x_j x_k + \cdots + \beta_{12\ldots q} x_1 x_2 \ldots x_q$$

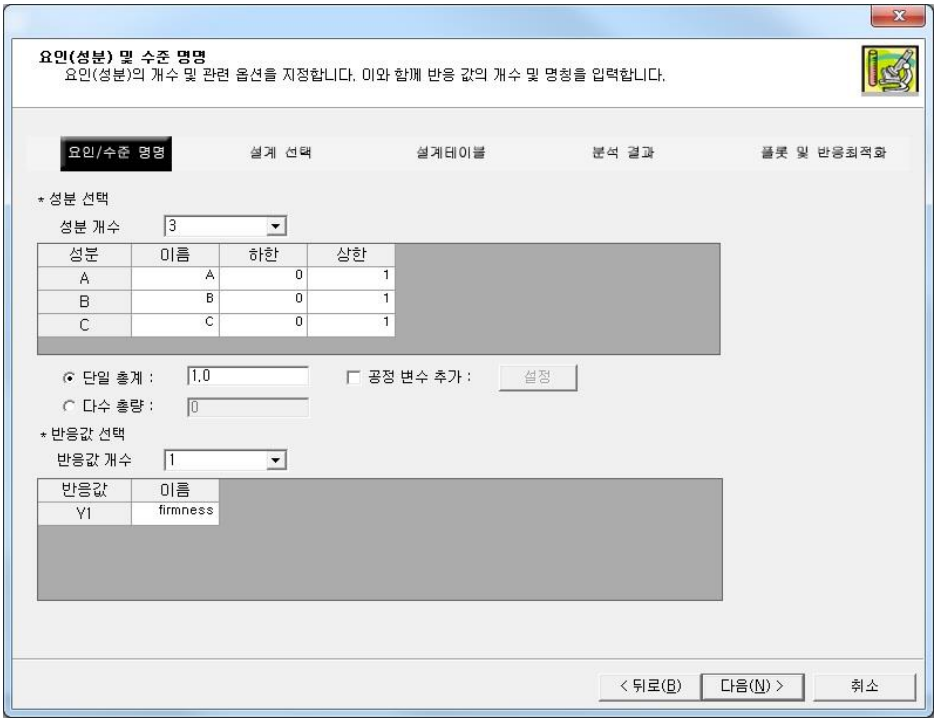
이를 통해 볼 때 성분의 개수가 많아질수록 적합할 수 있는 Polynomial 의 차수도 높아지는 것을 알 수 있습니다.

실험 소개
본 실험에서는 3 개의 성분이 반응 값 product 의 firmness 에 영향을 미치는 정도를 알아보려고 합니다. 성분에는 특별히 상한, 하한 조건이 없이 기본값(0,1)이고 실험자는 각 성분이 어떻게 반응 값에 영향을 미치는지를 알아보려고 합니다.

심플렉스 중심 설계를 수행하기 위해 다음의 화면에서 혼합물 설계 -> 심플렉스 중심 설계를 선택합니다.



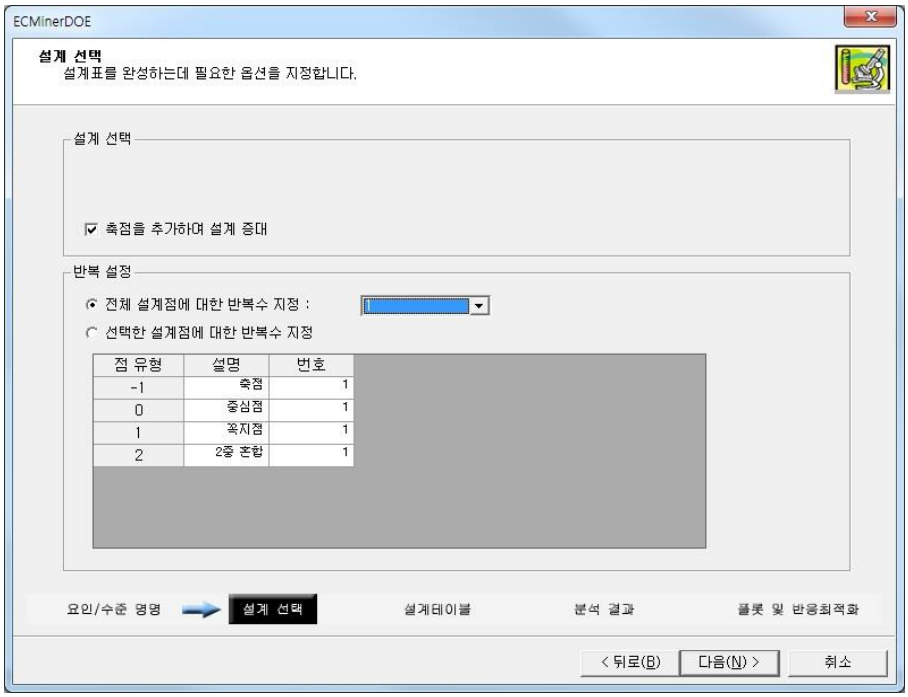
Step 1: 요인 및 수준 명명



본 화면에서 성분의 개수를 3 개로 하고 반응 값의 개수는 1 개, 그리고 반응 값의 이름은 firmness 라고 합니다. 본 화면에서 나타나는 옵션에 대해서 설명하면 다음과 같습니다.

- 단일 총계: 각 성분의 합을 지정합니다. 기본 값은 1 입니다.
- 다수 총량: 혼합물 양 설계일 때, 혼합물 양의 값을 입력합니다. 양의 값은 모두 양수만 가능하고, 구분은 space 로 합니다.
- 공정변수 추가: 공정변수를 추가한 혼합물 설계를 하고자 할 때 사용합니다.

Step 2: 설계 선택



본 화면에서 설계를 확정하기 위해서 여러 옵션을 설정합니다.

- 축 점을 추가하여 설계 증대: 본 옵션은 중심점에서 각 꼭지점에 해당하는 점 사이에 축 점을 추가하는 옵션입니다.
- 반복 설정

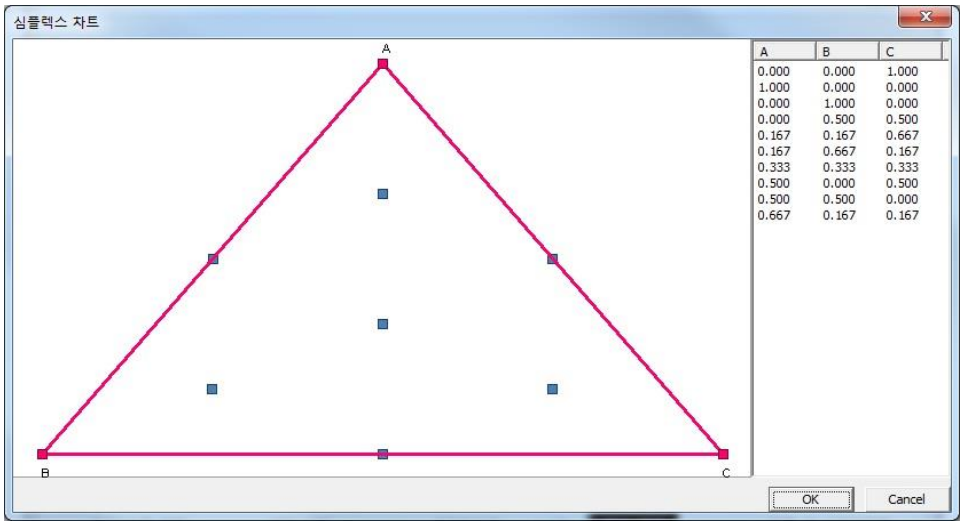
  전체 설계 점에 대한 반복 수 지정: 이 옵션을 통해서 만들어진 설계표를 몇 번이나 반복할지를 지정합니다.

  선택한 설계 점에 대한 반복 수 지정: 이 옵션은 점 유형에 따라 사용자가 반복할 횟수를 정해줄 수 있습니다.

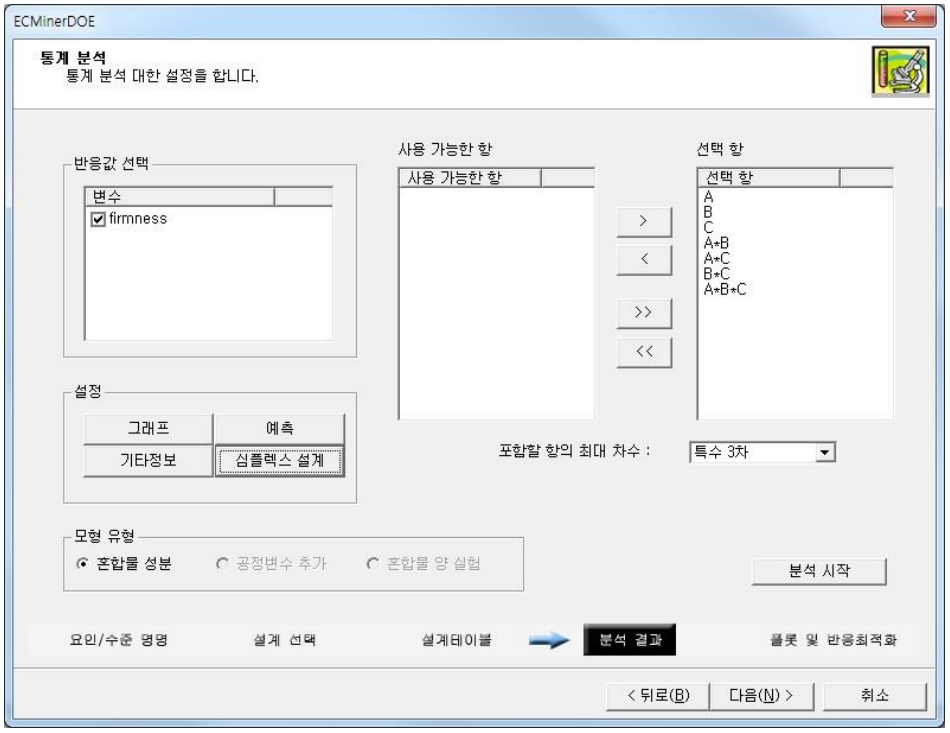
Step 3: 설계 테이블 생성

본 단계에서는 설계표가 완성되고, 실험자는 생성된 실험표에 따라 실험을 진행하고 결과 반응 값을 입력합니다. 이렇게 하면 분석에 필요한 모든 준비가 끝난 것입니다. 다음 단추를 클릭하여 Step 4 로 진행합니다.

Step 4: 통계 분석

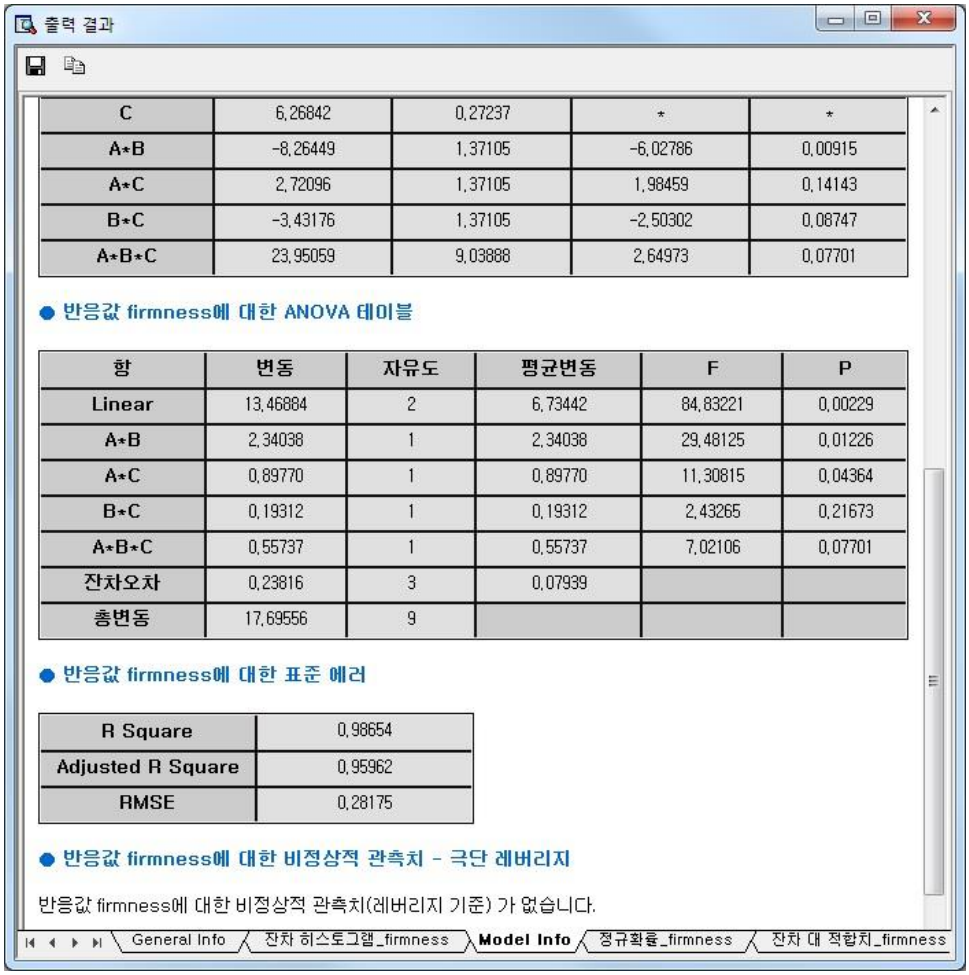


분석에 관련된 설정을 하기 전에 설계 표에서 만들어진 설계 점들이 어떻게 배치되어 있는지를 심플렉스 설계 플롯을 이용하여 볼 수 있습니다. 위의 심플렉스 설계 플롯을 통해서 점들이 매우 균일하게 배치되어 있음을 알 수 있습니다.



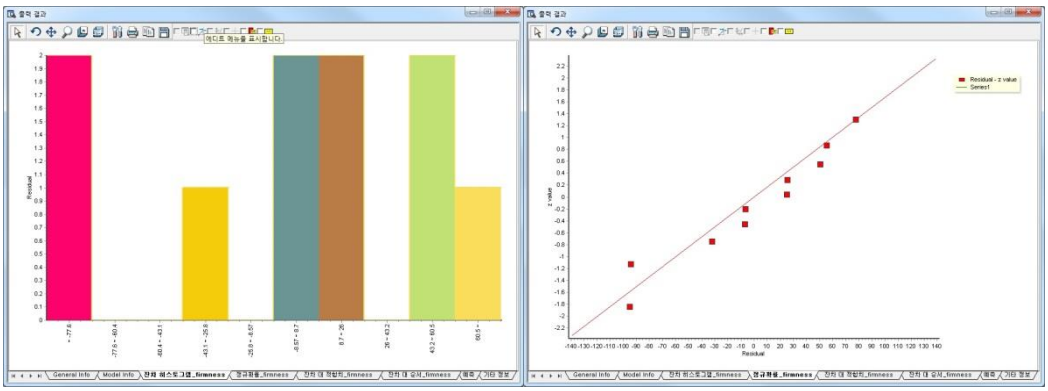
Step 4 의 Main 화면에서 모델을 설정하고 분석을 시작합니다. 이를 통해 회귀분석, 분산분석, 잔차 분석을 할 수 있습니다. Step 4 의 Main 화면에서 예측을 했을 경우에는 예측 값 또한 알 수 있습니다.

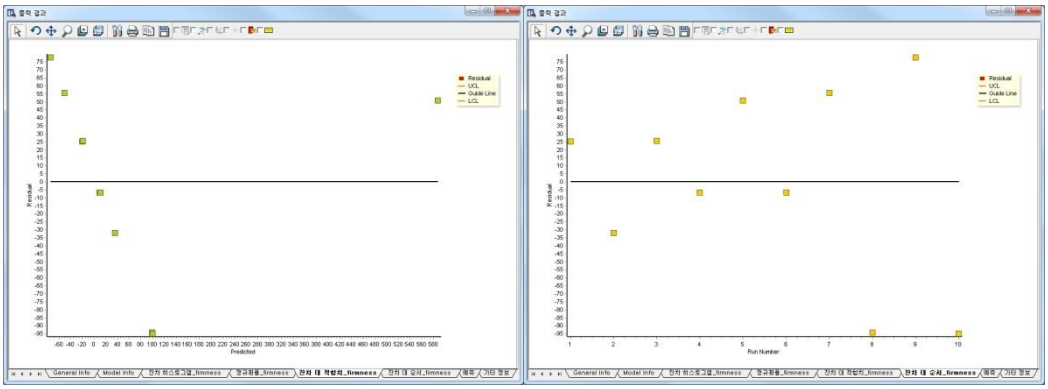
- General Info: 설계에 대한 일반적인 정보를 보여줍니다.
- Model Info: 회귀분석 결과, 분산분석 결과, 비정상적 관측치(극단 레버리지, 표준화 잔차)을 볼 수 있습니다.



회귀분석 및 분산분석을 통해서 모델의 적합성이 매우 높음을 알 수 있고, 아래의 잔차 분석을 통해서 정규성 가정이 합리적임을 알 수 있습니다.

- 잔차 관련 플롯: 잔차 히스토그램, 잔차 정규확률 플롯, 잔차 대 순서, 잔차 대 적합치를 볼 수 있습니다.





▪ 기타 정보: 잔차 관련 통계량을 보여줍니다.

출력 결과

실험계획법 - 혼합물 설계(Mixture Design) : 심플렉스 중심 설계

▶ 기타 정보

● 반응값 firmness에 대한 적합치 및 잔차

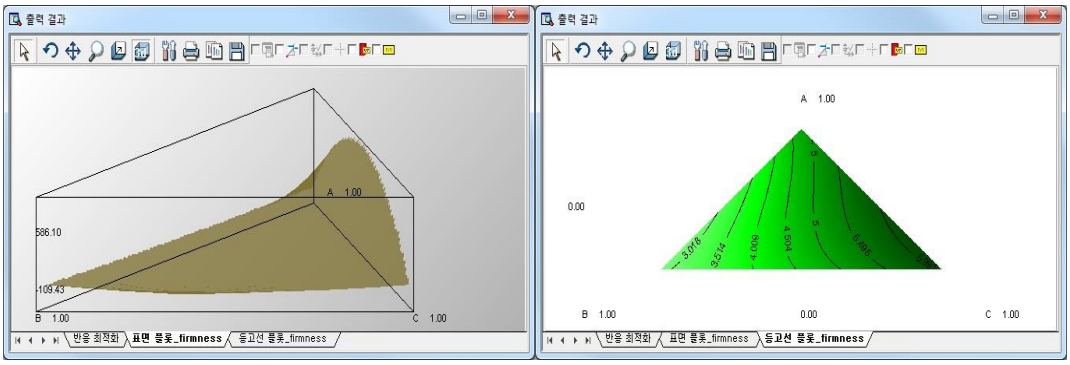
순서	적합치	잔차	표준화 잔차	외표준화 잔차	레버리지	Cook의 거리	DFITS
1	5,09206	-0,09206	-1,27658	-1,54223	0,93449	3,32108	-5,82492
2	3,27570	-0,02570	-0,35632	-0,29729	0,93449	0,25873	-1,12284
3	6,26842	0,11158	1,54725	2,81080	0,93449	4,87868	10,61622
4	2,11775	-0,11775	-1,47564	-2,30109	0,91979	3,56699	-7,79206
5	6,36048	0,01952	0,24460	0,20174	0,91979	0,09801	0,68313
6	3,91412	0,08588	1,07624	1,12154	0,91979	1,89740	3,79781
7	4,76853	-0,01853	-0,10437	-0,08537	0,60294	0,00236	-0,10520
8	4,71765	0,28235	1,17944	1,31500	0,27807	0,07655	0,81613
9	3,29674	0,08326	0,34780	0,28988	0,27807	0,00666	0,17991
10	5,70856	-0,32856	-1,37244	-1,83695	0,27807	0,10365	-1,14007

General Info | Model Info | 잔차 히스토그램_firmness | 정규확률_firmness | 잔차 대 적합치_firmness | 잔차 대 순서_firmness | 예측 | 기타 정보

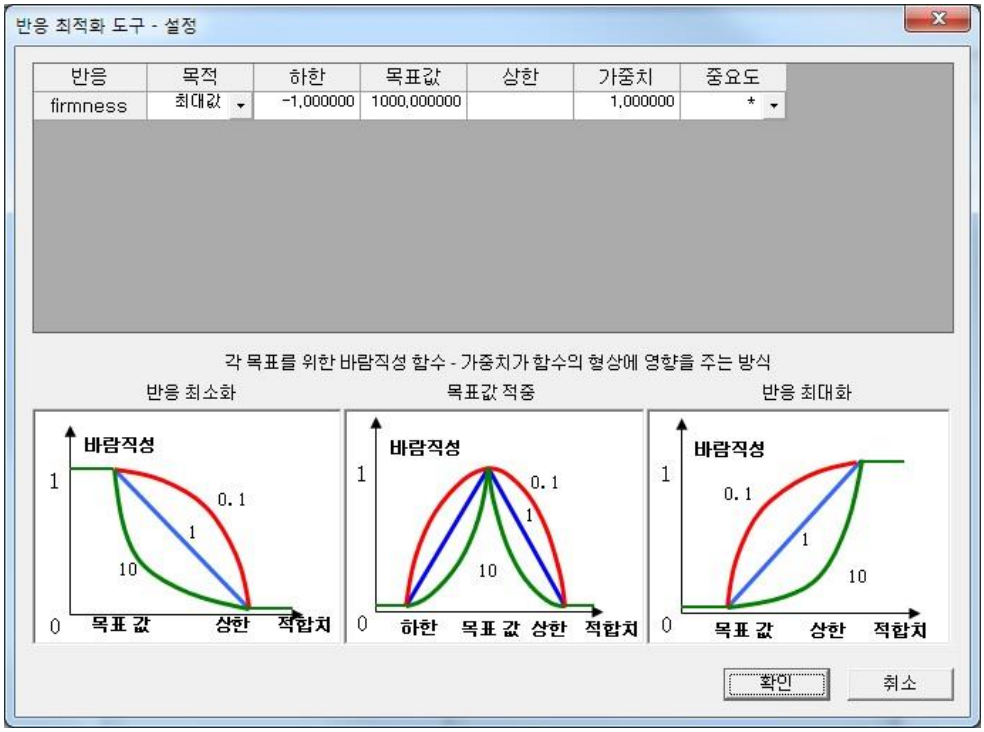
자세한 설명은 6.4. 설정 및 분석을 참고하세요.

Step 5: 플롯 및 반응 최적화

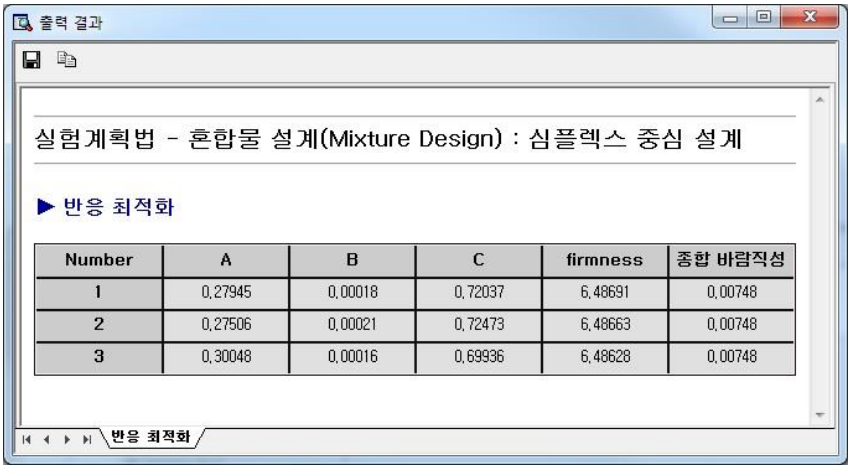
혼합물 설계의 Step 5 를 통해서는 표면 플롯, 등고선 플롯을 볼 수 있고 이와 함께 사용자의 목적에 맞는 반응최적화를 할 수 있습니다. 아래는 Step 4 에서 만든 Regression Model 에 대한 표면 플롯과 등고선 플롯입니다.



본 실험에 있어서의 목적이 어떠한 성분 조합에서 반응 값(Firmness)이 최대화되는지를 알고 싶은 것이라고 할 때 설정 창에서 다음과 같이 입력을 합니다. 하한을 -1 이라고 둔 것은 반응 값이 -1 이하이면 모두 똑같이 '좋지 않다'라는 것이고 목표 값을 1000 이라는 매우 큰 값으로 둔 것은 반응 값을 증가시킬 수 있는 한 계속 증가시키고자 하기 때문입니다.



위와 같이 설정을 하고 간단한 옵션 설정을 마친 후에 결과 보기를 클릭하면 다음과 같은 결과를 얻게 됩니다.



Number	A	B	C	firmness	종합 바람직성
1	0,27945	0,00018	0,72037	6,48691	0,00748
2	0,27506	0,00021	0,72473	6,48663	0,00748
3	0,30048	0,00016	0,69936	6,48628	0,00748

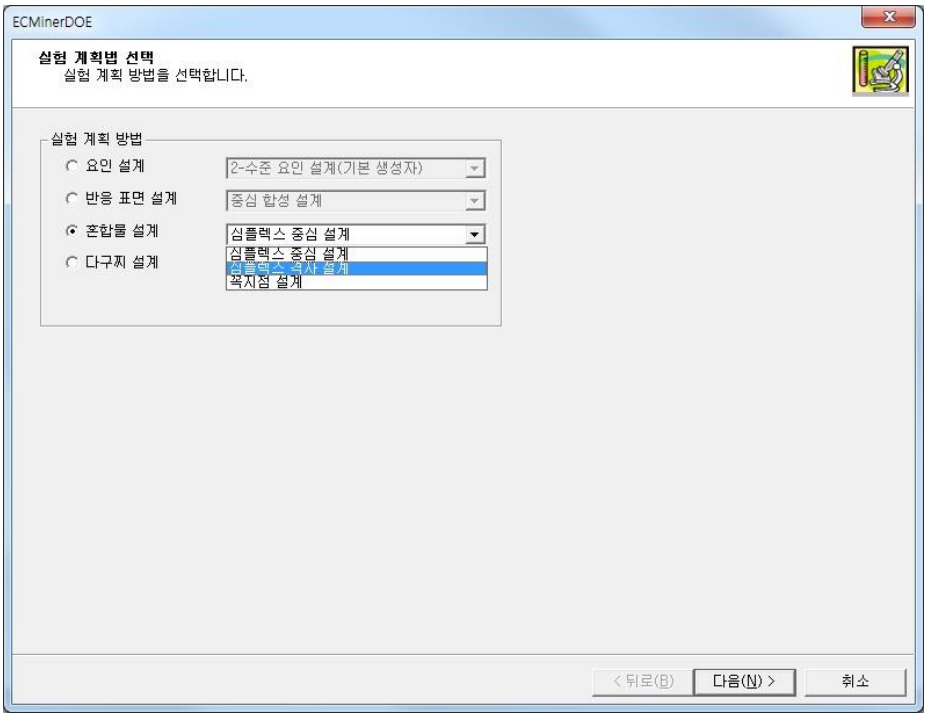
즉 반응 최적화를 통해서 A 성분이 0.27945, B 성분이 0.00018, C 성분이 0.72037 일 때 firmness 의 값이 6.48691 로 가장 커짐을 알 수 있고 실험자는 이 성분 조합에서 firmness 의 값이 만족할 수준으로 커지는지를 확인함으로써 실험을 종료 혹은 실험을 재개할 수 있습니다.

6.3.3.3. 심플렉스 격자 설계

심플렉스 Region 에서 실험을 하고 그리고 그 실험 결과로 그 위에서의 Regression Model 을 만드는 것이 목적이라면 되도록 이면 실험 점이 심플렉스 Region 전역에 골고루 분포되는 것이 바람직할 것입니다. 이러한 목적에 부합하는 설계가 바로 심플렉스 격자 설계입니다. 심플렉스 격자 설계는 성분의 개수가 n 개 있을 때, m 차 다항식에 적합을 하는데 유용한 설계입니다.

실험 소개
어떤 금속의 강도가 어떠한 성분들의 비율뿐 아니라 양에도 영향을 받는다고 합니다. 그렇다면 단순히 성분들의 비율만 가지고는 성분과 금속 강도 사이의 관계를 알 수 없습니다. 따라서 이러한 상황에서 사용할 수 있는 혼합물 양 설계를 사용할 수 있습니다.

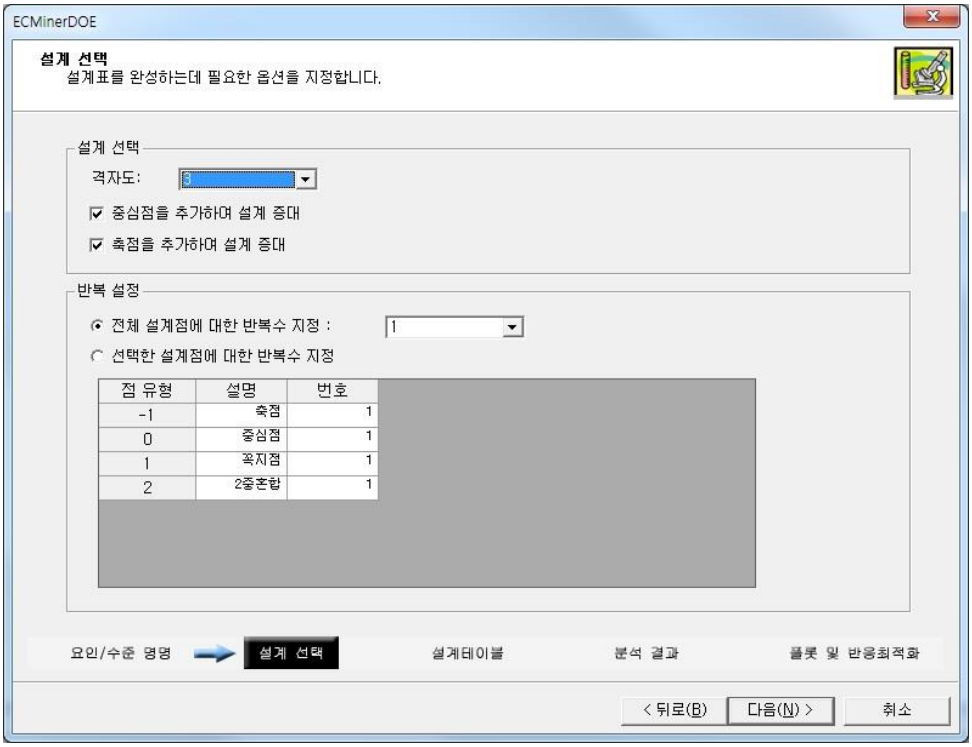
심플렉스 격자 설계를 통해서는 혼합물 양 설계를 설명하도록 하겠습니다. 먼저 심플렉스 격자 설계를 선택합니다.



Step 1: 요인(성분) 및 수준 명명

본 단계에서는 실험에 사용되는 성분에 대한 기본적인 설정을 합니다. 이번 실험은 혼합물 양 설계를 하므로 단일 총계가 아닌 다수 총량을 선택합니다. 만약 양이 반응 값에 2 차 이상의 영향을 줄 것이라 생각되면 양의 값을 3 개 이상으로 하여 실험을 합니다. 양이 반응 값에 선형적인 영향을 줄 것이 확실하면 2 개의 양 값으로도 충분합니다. 양의 개수가 적을수록 실험 수가 줄어들기 때문에 이에 대해서는 신중한 검토가 필요합니다. 하지만 이렇게 진행하는 첫 번째 실험의 분석 결과 혼합물의 양이 주는 영향이 2 차 이상이라는 것이 확인될 경우 실험을 완전히 새로 하는 것이 아니라 수행한 실험에서 추가적으로 실험을 몇 번 더하면 되기 때문에 처음부터 완벽히 실험을 수행할 필요는 없습니다. 설정을 마친 후에는 다음 단추를 클릭하여 다음 화면으로 넘어갑니다.

Step 2: 설계 선택



본 단계에서는 설계를 완성하기 위해서 여러 세부적인 옵션을 설정합니다.

격자도: 격자도에 따라서 적합할 수 있는 Model 의 차수가 결정되기 때문에 이 격자도를 어떻게 선택하느냐는 매우 중요합니다. 1 차로 적합을 하고 싶으면 격자도를 1 이상으로, 2 차로 적합하고 싶으면 격자도를 2 이상으로 해야 합니다. 현재 성분이 금속의 강도에 2 차적인 영향을 줄 것이라 예상되므로 여유를 두어서 격자도를 3 으로 합니다. (격자도의 수가 커질수록 실험 수가 증가하므로 이를 염두에 두고 옵션을 선택하도록 해야 합니다.)

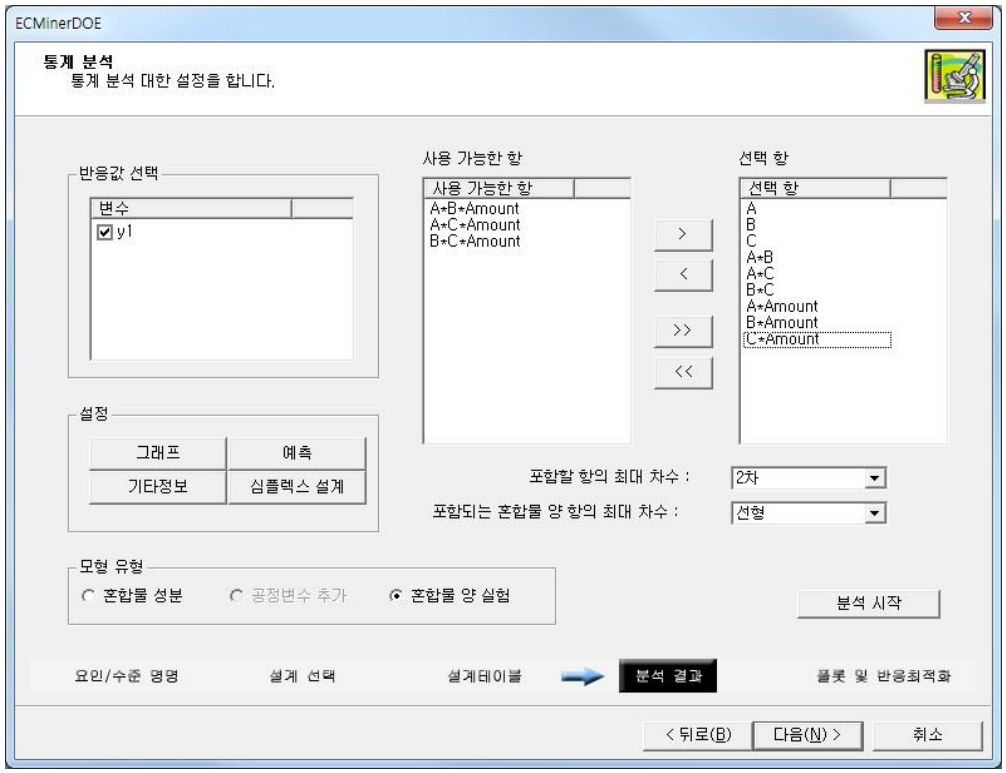
반복 설정의 경우 심플렉스 중심 설계와 동일합니다.

Step 3: 설계 테이블

	표준순서	실험순서	점 유형	블럭	A	B	C	양	강도
1	1	12	1	1	1,000000	0,000000	0,000000	1,000000	0
2	2	11	1	1	0,000000	1,000000	0,000000	1,000000	0
3	3	14	1	1	0,000000	0,000000	1,000000	1,000000	0
4	4	10	0	1	0,333333	0,333333	0,333333	1,000000	0
5	5	5	-1	1	0,666667	0,166667	0,166667	1,000000	0
6	6	1	-1	1	0,166667	0,666667	0,166667	1,000000	0
7	7	8	-1	1	0,166667	0,166667	0,666667	1,000000	0
8	8	4	1	1	3,000000	0,000000	0,000000	3,000000	0
9	9	2	1	1	0,000000	3,000000	0,000000	3,000000	0
10	10	3	1	1	0,000000	0,000000	3,000000	3,000000	0
11	11	9	0	1	1,000000	1,000000	1,000000	3,000000	0
12	12	7	-1	1	2,000000	0,500000	0,500000	3,000000	0
13	13	6	-1	1	0,500000	2,000000	0,500000	3,000000	0
14	14	13	-1	1	0,500000	0,500000	2,000000	3,000000	0

Step 3 를 통해서 설계 테이블이 완성되고 실험자는 주어진 설계 테이블에 따라 실험을 한 후 반응 값(강도)를 입력합니다. 입력을 마친 후에는 다음 단추를 클릭하여 분석 단계로 넘어갑니다.

Step 4: 통계 분석



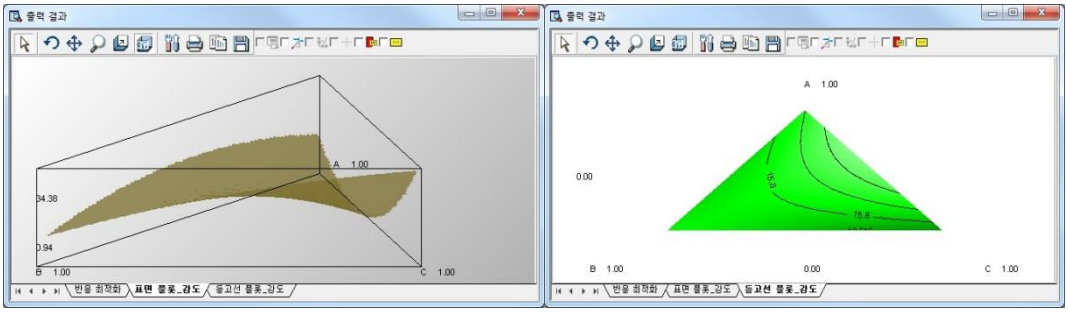
Step 4 의 Main 에서는 모형 유형에서 혼합물 양 실험을 선택할 수 있습니다. 만약 혼합물 성분을 선택하면 양이 달라도 비율만 같으면 같은 것으로 생각하여 분석을 합니다. 하지만 현재 주어진 실제 실험이 혼합물의 양을 고려해야 하는 실험임이 분명하므로 혼합물 양 실험을 선택합니다. 포함할 항의 최대차수, 포함되는 혼합물 양항의 최대 차수를 각각 2 차, 선형으로 하고 영향을 주지 않을 것으로 사전적으로 알고 있는 항을 위와 같이 제거합니다. (혹은 가능한 모든 항을 넣어서 분석을 한 후 분석 결과를 보고 의미가 없다고 생각하는 항을 빼는 것도 좋은 방법입니다.) 분석 내용은 6.3.3.1. 심플렉스 중심 설계를 참고하세요.

다만 위에서 만들어지는 Regression Model 은 다음과 같은 형태를 갖습니다.

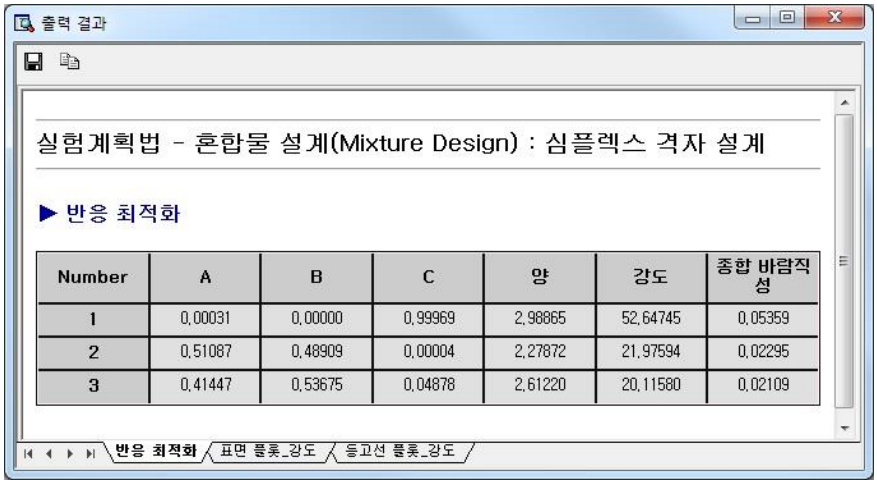
$$y = b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{1,\mathrm{Amount}}x_1(\mathrm{Amount} - \bar{A}) + b_{2,\mathrm{Amount}}(\mathrm{Amount} - \bar{A}) + b_{3,\mathrm{Amount}}(\mathrm{Amount} - \bar{A})$$

Step 5: 플롯 및 반응 최적화

본 단계에서는 Step 4 에서 만들어진 Regression Model 을 표면 플롯, 등고선 플롯으로 표현해 주고 실험자의 목적에 맞게 반응 최적화 기능을 제공합니다. 혼합물의 양이 2 일 때 플롯을 그리면 다음과 같습니다.



실험자의 목적은 반응 값(강도)를 최대화하는 것이므로 목적은 최대화이고 하한 및 목표 값은 적당한 값을 입력합니다. 이번에도 역시 대부분의 예제에서와 같은 이유로 하한을 -1, 상한을 1000 으로 합니다.



Number	A	B	C	양	강도	종합 바람직성
1	0,00031	0,00000	0,99969	2,98865	52,64745	0,05359
2	0,51087	0,48909	0,00004	2,27872	21,97594	0,02295
3	0,41447	0,53675	0,04878	2,61220	20,11580	0,02109

위의 결과에서 볼 때 (A, B, C) = 3*(0, 0, 1) 에서 금속의 최대 강도를 얻을 수 있음을 알 수 있습니다.

6.3.3.4. 꼭지점 설계

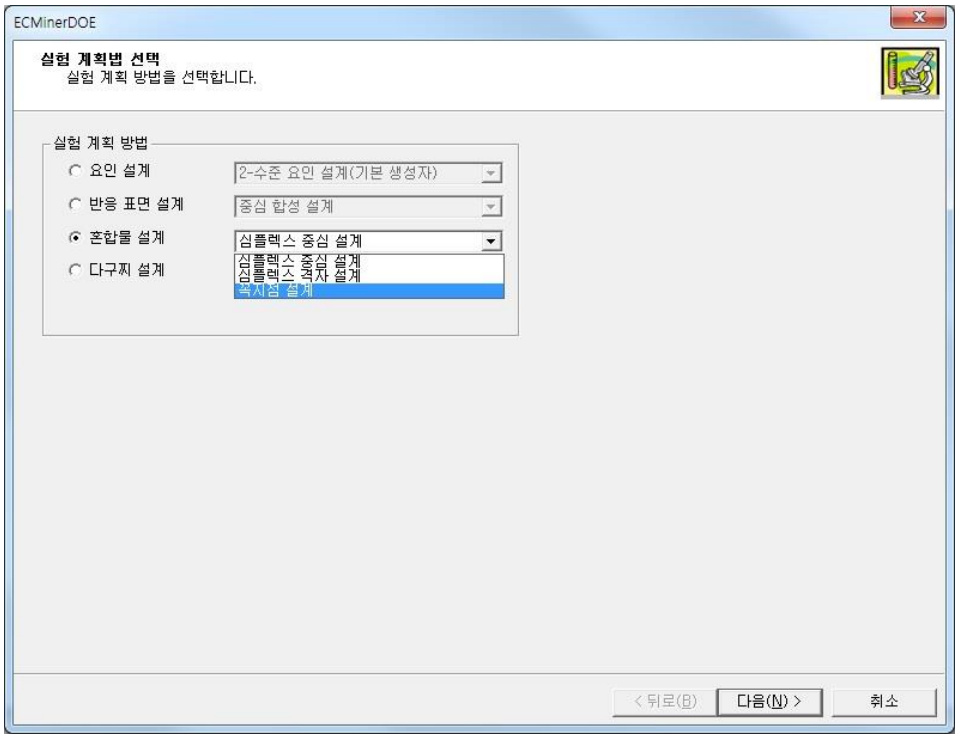
꼭지점 설계는 상한 및 하한 조건 그리고 여기에 선형 제약 조건까지 더해졌을 때 이러한 조건이 만족되는 Convex Set 의 꼭지점을 구해서 이 꼭지점을 가지고 실험을 증대한 후 이 점들에게 실험을 하는 것을 말합니다. n 차원 공간에서 오직 상한과 하한 조건, 그리고 선형 제약 조건만이 있을 때 이러한 조건들로 만들어지는 공간의 꼭지점을 구하는 것은 상당히 쉽지 않은 과정이라는 것을 직감적으로 알 수 있습니다. 따라서 이는 단순한 과정을 통해서는 구할 수 없는데 이러한 꼭지점을 구하는 알고리즘을 Piepel(1988)이 제시 하였습니다.
꼭지점 설계를 통해서는 혼합물 설계에 공정 변수를 추가할 경우에 대해서 설명하겠습니다.

실험 소개

본 실험은 3 가지의 성분과 함께 어떠한 공정변수(작업온도)가 product 의 quality 에 영향을 미쳐서 그 관계를 알아보고자 하는 것입니다. 이 때 3 가지 성분의 제한 조건은 다음과 같습니다.

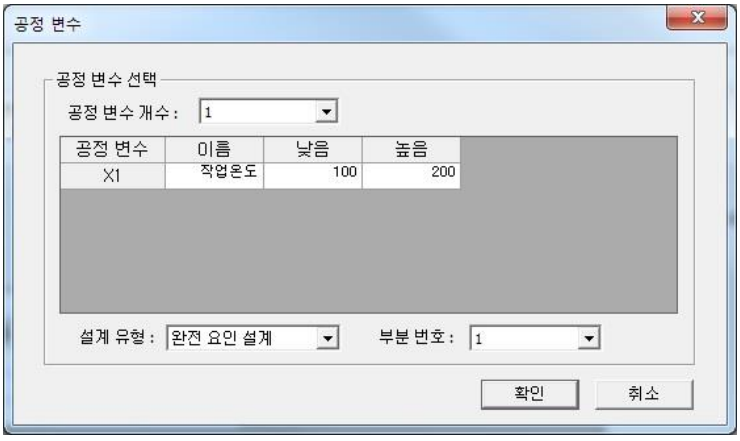
$$0.1 \le x_1 \le 0.8, \qquad 0.0 \le x_2 \le 0.6, \qquad 0.0 \le x_3 \le 0.5$$

먼저 꼭지점 설계를 선택합니다.

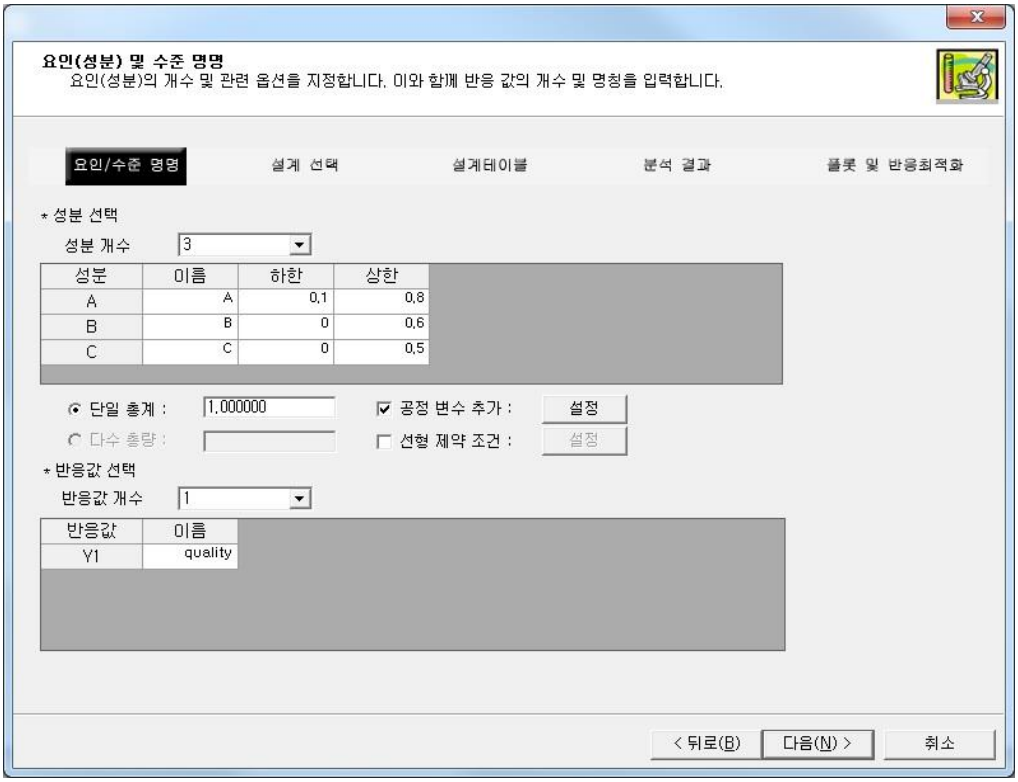


Step 1: 요인(성분) 및 수준 명명

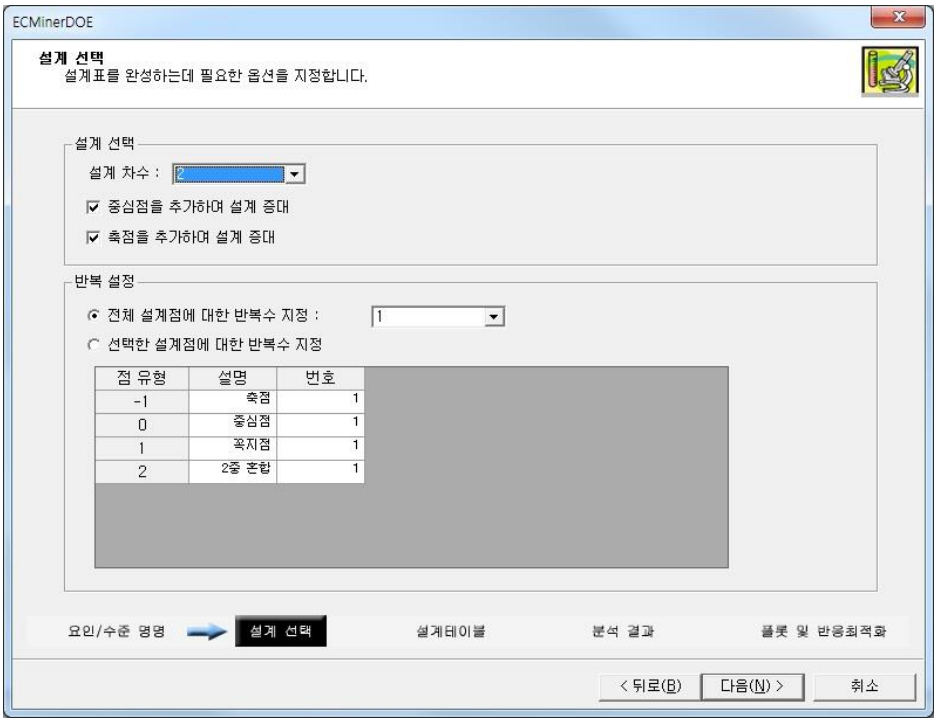
먼저 다음과 같은 공정변수 추가를 선택하여 공정변수를 추가합니다.



그리고 Step1 의 Main 화면에서 성분들의 제한 조건을 입력하고 다음 단추를 클릭합니다.



Step 2: 설계 선택



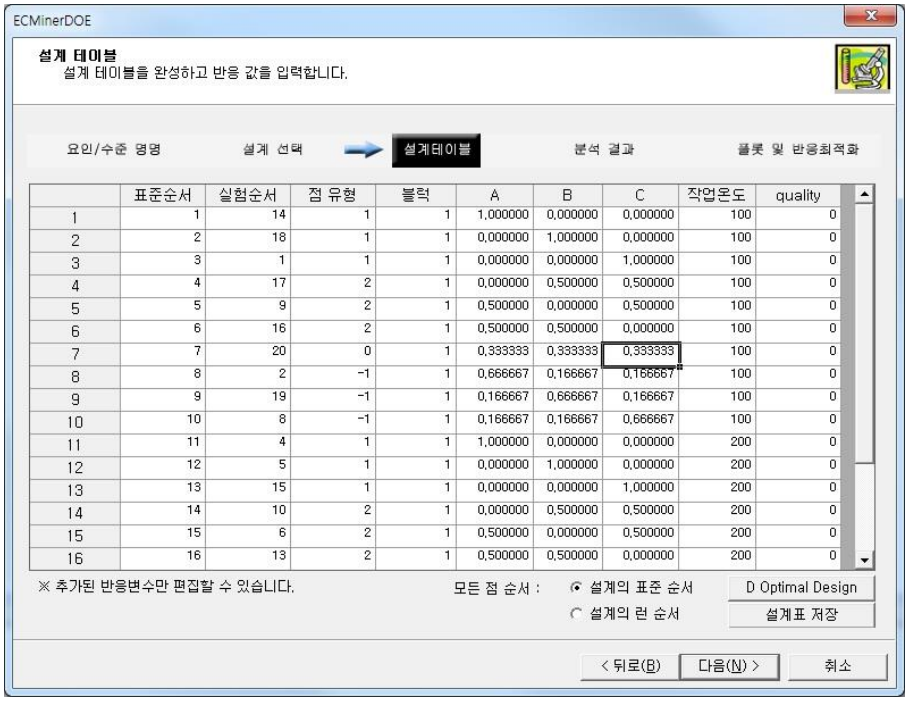
설계 선택 단계에서는 여러 가지 옵션을 지정할 수 있습니다.

- 설계 차수: 현재 꼭지점을 가지고 몇 차까지 점을 증대할지를 결정합니다.
- 중심점을 추가하여 설계 증대: 중심점을 추가할 지의 여부
- 축점을 추가하여 설계 증대: 중심점과 각 꼭지점에 이르는 중간점인 축점을 추가할 것인지의 여부
- 반복 설정

  전체 설계점에 대한 반복 수 지정: 같은 실험 전체를 몇 번이나 반복할지를 지정

  선택한 설계점에 대한 반복 수 지정: 점의 유형에 따라 반복 수를 서로 다르게 합니다.

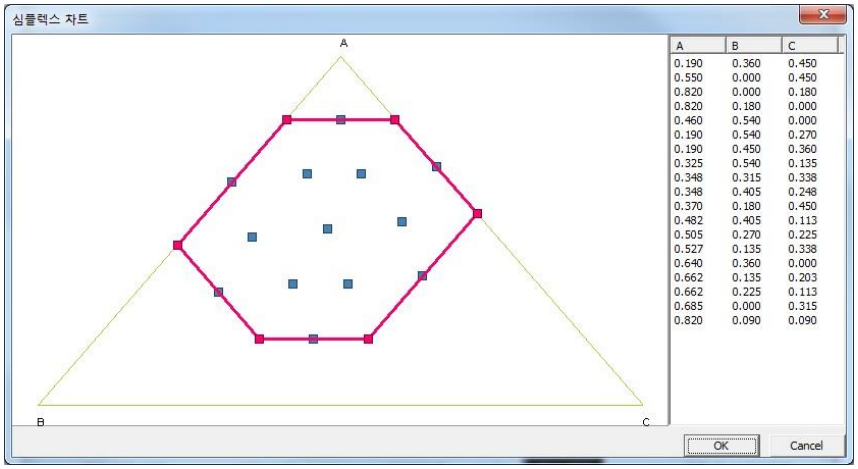
Step 3: 설계 테이블



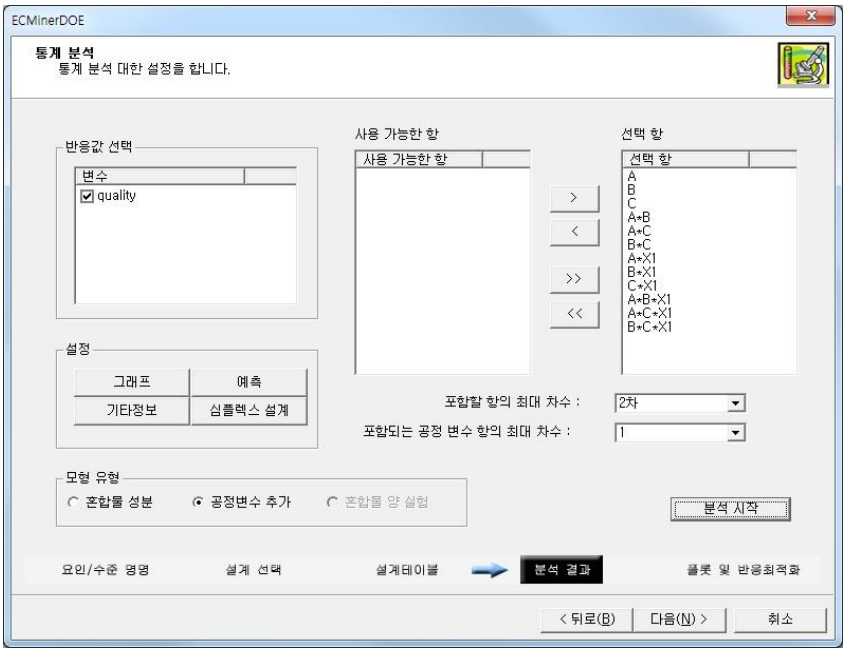
설계 테이블 단계에서는 설계 테이블이 생성되고 실험자는 생성된 설계에 따라 실험을 한 후 반응 값을 입력합니다. 입력을 완료한 다음에는 다음 단추를 클릭하여 Step 4 로 넘어갑니다.

Step 4: 통계 분석

통계 분석을 시작하기 전에 현재 알고리즘을 통해서 만들어진 설계점들이 공간상에 어떻게 분포되어 있는지를 심플렉스 설계 플롯을 통해 확인해 봅니다.



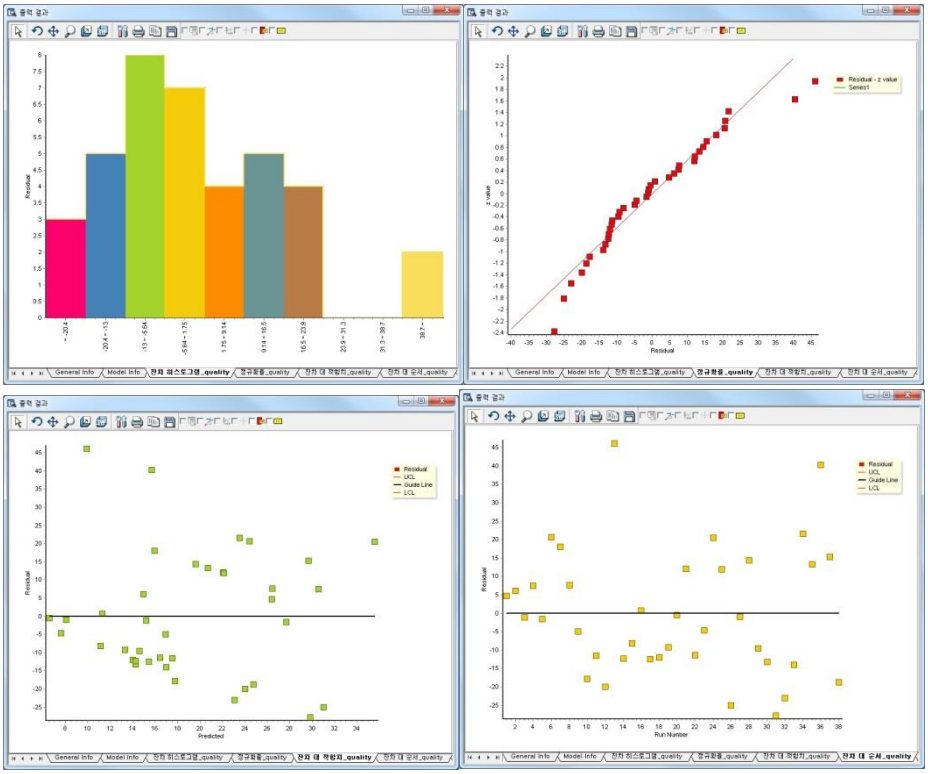
제한 조건으로 인하여 설계 점들의 집합 모양이 삼각형은 아니지만 제한된 공간 상에서 매우 균일하게 설계 점들이 배치되어 있는 것을 확인할 수 있습니다.



모형 유형에서는 공정 변수 추가를 선택하고 포함할 항의 최대 차수와 포함되는 공정 변수 항의 최대 차수를 사용자의 필요에 따라 입력합니다. 그럼 다음과 같은 결과를 얻을 수 있습니다.

- General Info: 설계에 대한 기본적인 정보를 제공합니다.
- Model Info: 회귀분석, 분산 분석, 비정상적 관측치(극단 레버리지, 표준화 잔차)에 대한 정보를 제공합니다.

▪ 잔차 관련 (잔차 히스토그램, 잔차 정규 확률 플롯, 잔차 대 순서, 잔차 대 적합치)



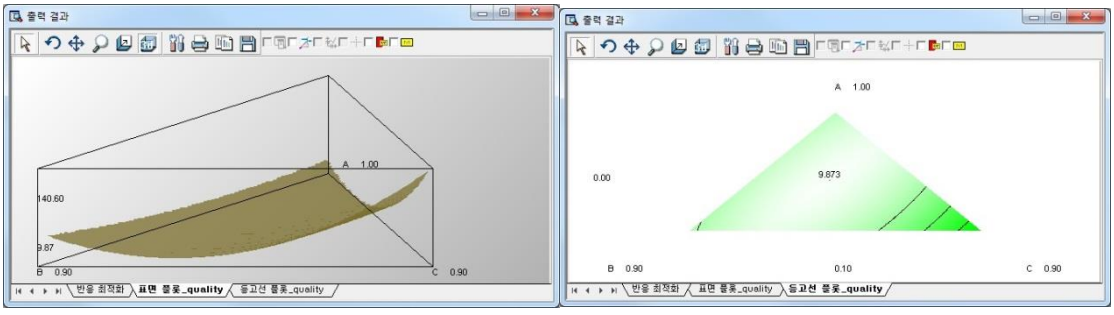
▪ 기타 정보: 기타 잔차 관련 정보들을 정리하여 보여줍니다.

순서	적합치	잔차	표준화 잔차	외표준화 잔차	레버리지	Cook의 거리	DFITS
1	26,37045	4,62955	0,31207	0,30659	0,47187	0,00725	0,28980
2	14,91813	6,08187	0,39035	0,38389	0,41744	0,00910	0,32496
3	15,19003	-1,19003	-0,07638	-0,07490	0,41744	0,00035	-0,06341
4	30,57041	7,42959	0,54973	0,54222	0,56168	0,03227	0,61380
5	27,66037	-1,66037	-0,12285	-0,12050	0,56168	0,00161	-0,13641
6	24,40920	20,59080	1,38800	1,41445	0,47187	0,14344	1,33700
7	15,91950	18,08050	1,06290	1,06567	0,30562	0,04144	0,70699
8	26,46431	7,53569	0,44837	0,44137	0,32213	0,00796	0,30426
9	16,94020	-4,94020	-0,28677	-0,28164	0,28780	0,00277	-0,17904
10	17,78488	-17,78488	-1,03237	-1,03373	0,28780	0,03589	-0,65713
11	17,51048	-11,51048	-0,67667	-0,66945	0,30562	0,01679	-0,44413
12	24,02866	-20,02866	-1,19169	-1,20183	0,32213	0,05624	-0,82848
13	9,90068	46,09932	2,57781	2,92971	0,23254	0,16779	1,61266
14	14,26328	-12,26328	-0,65948	-0,65215	0,17020	0,00743	-0,29535
15	11,12105	-8,12105	-0,43659	-0,42969	0,16969	0,00325	-0,19425

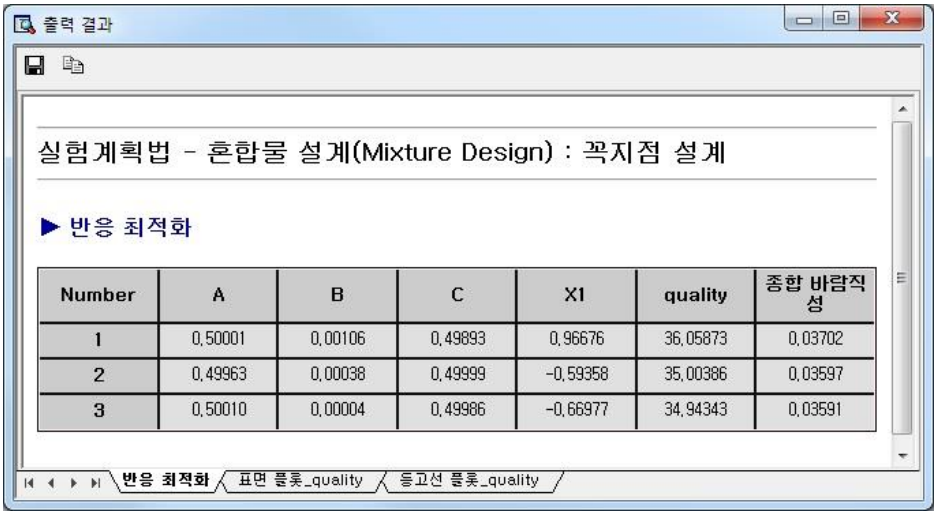
자세한 내용은 6.4. 설정 및 분석을 참고하세요.

Step 5: 플롯 및 반응 최적화

Step 5 에서는 Step 4 에서 만든 Regression Model 이 어떠한 모양인지를 표면 플롯, 등고선 플롯을 통해서 보여줍니다.



실험자의 목적에 따라 반응 최적화를 실행합니다. 현재 실험은 Quality 를 최대화시키는 것이 목적일 것이므로 목적을 최대화로 하고 하한과 목표값을 각각 -1, 1000 으로 해서 최적화를 수행하면 다음과 같은 결과를 얻을 수 있습니다.



Number	A	B	C	X1	quality	종합 바람직성
1	0,50001	0,00106	0,49893	0,96676	36,05873	0,03702
2	0,49963	0,00038	0,49999	-0,59358	35,00386	0,03597
3	0,50010	0,00004	0,49986	-0,66977	34,94343	0,03591

즉 성분 (A, B, C) = (0.5, 0, 0.5), 공정변수=1 (작업온도=100 도) 일 때 최대의 Quality 36.05873 를 얻을 수 있음을 알 수 있습니다.

6.3.3.5. 혼합물 설계 D Optimal Design

D Optimal Design 은 사용자의 필요에 따라 설계를 변형하고 싶을 때 향후 통계 분석 과정에 가장 적합하게 설계를 변형하는 방법론을 말합니다. 설계의 우수성을 판단하는 지표로 다음과 같은 지표가 있습니다.

D Optimality(Determinant)

D Optimality 는 가장 흔히 사용되는 기준으로 X^TX 역행렬의 Determinant 를 가장 크게 만드는 설계를 찾을 때 사용됩니다. 여러 후보점들의 집합에서 필요한 후보점들을 모아서 만든 Design Matrix X 구할 때 이 중에서 X^TX 역행렬의 Determinant 를 가장 크게 만드는 설계표를 D-Optimal Design 이라고 합니다.

A Optimality(Trace)

여러 후보점들의 집합에서 필요한 후보점들을 모아서 만든 Design Matrix X 구할 때 이 중에서 X^TX 역행렬의 TRACE 를 가장 크게 만드는 설계표를 A-Optimal Design 이라고 합니다.

하지만 A-Optimality 의 경우 계산상의 어려움 때문에 잘 사용하진 않습니다.

G Optimality(평균 레버리지 / 최대 레버리지)

G Optimality 는 평균 레버리지를 최대 레버리지로 나눈 값을 의미합니다. 여기서 레버리지는 Generalized Linear Model Matrix 를 X 라고 할 때, H Matrix 를

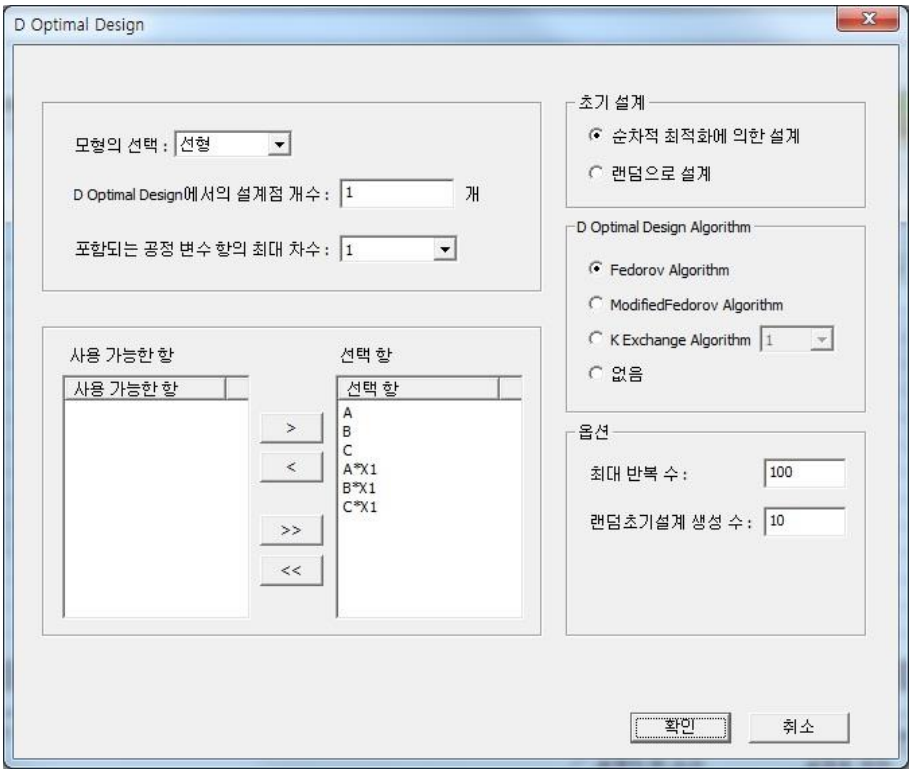
$$H = X(X^TX)^{-1}X^T$$

라고 하면 그것들의 diagonal component 를 말합니다. 이 레버리지들의 평균을 최대값으로 나누면 그것이 바로 G-Optimality 입니다.

V Optimality

레버리지의 평균이 V Optimality 입니다.

이런 여러 지표 중에서 D Optimality 를 최대화하기 위한 방법론으로 D Optimal Design 이라는 방법론이 사용됩니다. 반응 표면 설계의 D Optimal Design 화면은 다음과 같이 구성되어 있습니다.



모형의 선택: 먼저 모형을 선택합니다. 이렇게 선택한 모형에 따라 아래의 선택 항 창의 선택 항이 바뀌게 되는데 이 중에서 일부만 사용해서 쓸 수도 있습니다.

D Optimal Design 에서의 설계 점 개수: 이는 변형된 설계표에서 과연 설계 점이 몇 개가 있을지를 정하는 것입니다.

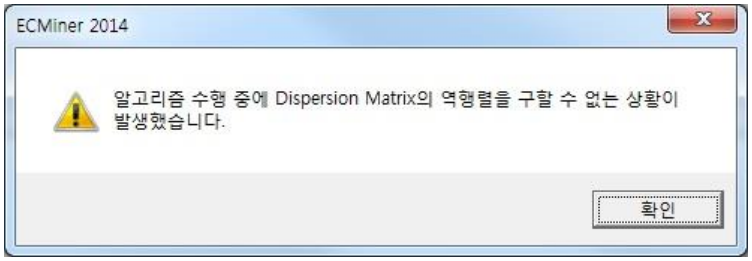
포함되는 공정 변수 항의 최대 차수: 이는 혼합물 설계에서 공정변수를 추가한 실험을 하고자 할 때 나타나는 옵션입니다. 공정변수 개수가 2 개 이상이면 최대 2 차까지 선택할 수 있으며 공정 변수가 1 개인 경우에는 최대 1 차까지만 선택할 수 있습니다.

초기 설계: D Optimal Design 을 수행하기 위해서는 초기 설계를 선택해야 합니다. D Optimal Design 에서의 설계 점 개수와 같은 수의 설계 점을 갖는 초기 설계를 어떻게 선택할지를 설정합니다.

D Optimal Design Algorithm: 어떤 알고리즘을 사용하여 초기설계를 개선하여 D Optimal Design 을 얻을 것인지를 결정합니다. ECMiner™ DOE 에서는 Fedorov Algorithm, Modified Fedorov Algorithm, K Exchange Algorithm 을 제공하는데 경험적으로 이 중에서 Modified Fedorov Algorithm, K Exchange Algorithm 이 좋은 성능을 발휘한다고 알려져 있습니다.

옵션: 최대 반복 수란 D Optimal Design Algorithm 에서 최대 몇 번의 반복을 수행할 것인지를 의미합니다. 랜덤 초기 설계의 생성 수는 초기 설계를 랜덤으로 설계할 때 몇 개의 랜덤 설계를 만들지를 의미합니다. 이 때 만들어진 여러 랜덤 설계 중에서 D Optimality 가 가장 큰 것을 초기 설계로 사용하여 Algorithm 을 수행합니다.

주의: 반응 표면 설계의 D Optimal Design 과는 달리 혼합물 설계의 D Optimal Design 에서는 알고리즘 수행도중 다음과 같은 Message 가 종종 나타나는데 이는 알고리즘을 수행하는 과정 중 Dispersion Matrix 의 역행렬을 구하는 상황에서 Determinant 가 0 에 매우 가까울 때 나타나는 현상입니다. 이와 같은 상황에서는 설계 모형의 선택에서 모형의 차수를 낮추어서 진행하면 이러한 에러가 나타나지 않습니다.



6.3.4 다구찌 설계

6.3.4.1. 개요

다구찌 실험 계획법은 종래의 실험 계획법에 비하여 특히 확장된 역할로서 다음의 몇 가지 역할을 가지고 있습니다.

종래에는 제어 불가능한 환경조건이나 제어하기 어려운 생산조건, 공정 조건 등의 원인(이를 잡음인자라고 통칭합니다.)들이 데이터에 주는 영향의 정도를 평가하기 어려웠지만, 차츰 객관적으로, 그리고 계량적으로 이를 평가할 수 있는 방법이 제시되었는데 다구찌 실험계획에서는 이를 SN 비를 이용하여 평가합니다. 즉 다구찌 실험계획은 잡음을 포함한 실험 환경에서 작업자가 원하는 조건에 도달할 수 있는 최적의 실험 조합을 찾아줍니다. 또한 직교 배열표 사용을 통해 실험 회수를 획기적으로 줄입니다.

위의 특징뿐 아니라 다구찌 실험계획은 많은 이점을 가지고 있습니다. ECMiner™ DOE 에서 제공하는 다구찌 방법론은 다음과 같습니다.

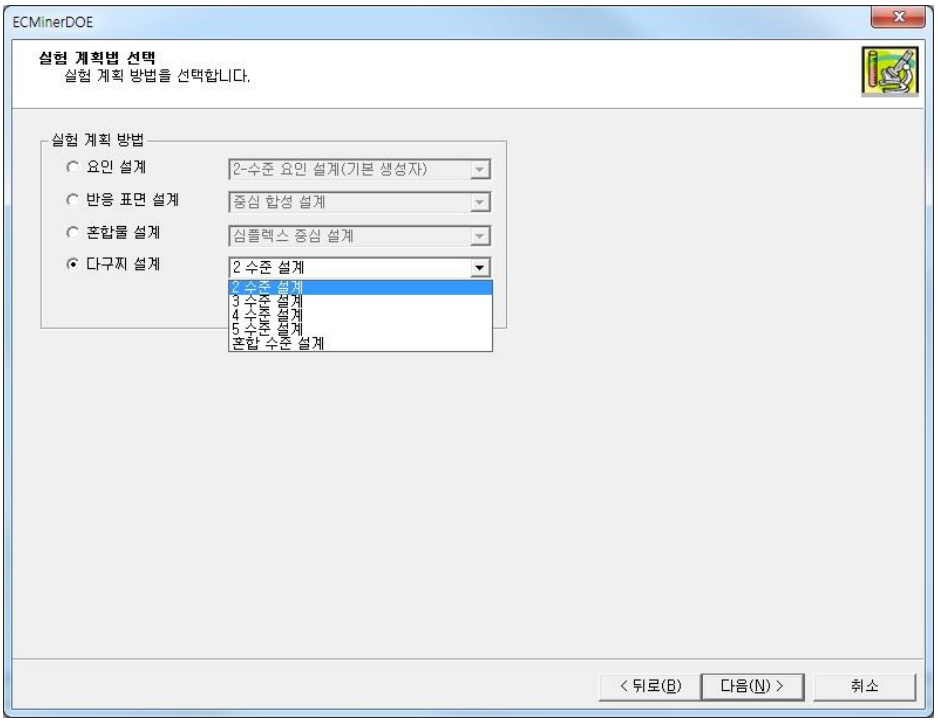
- 2 수준 설계
- 3 수준 설계
- 4 수준 설계
- 5 수준 설계
- 혼합 수준 설계

6.3.4.2. 2 수준 설계

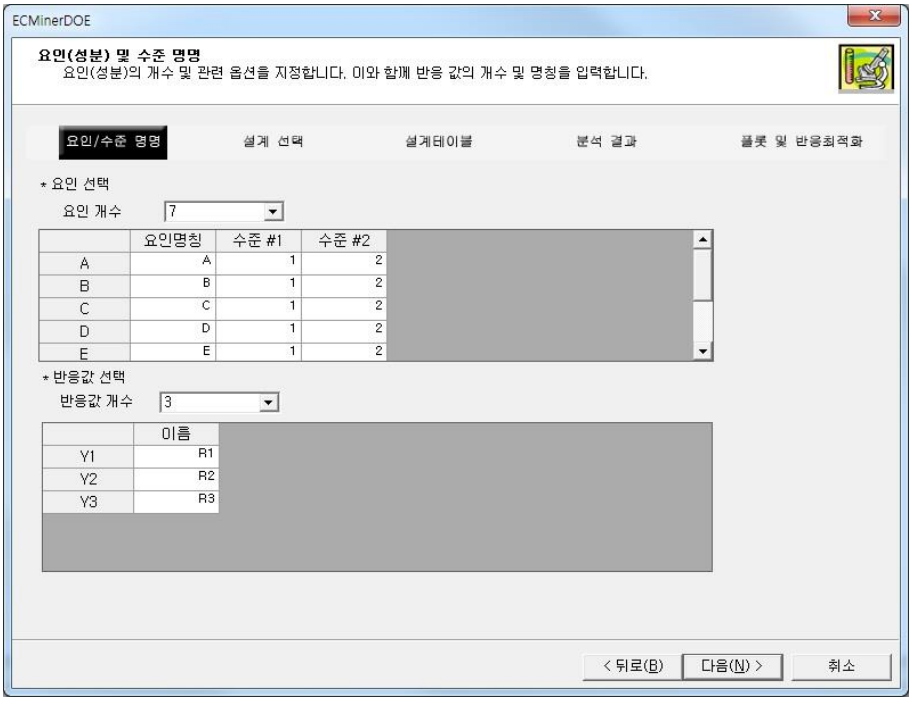
2 수준 설계는 요인당 수준이 2 개일 때 사용하는 설계입니다. 다음과 같은 실험이 있다고 합니다.

실험 소개
배기가스 가운데 CO 의 양을 줄이기 위하여 관련이 있는 인자 A-H 를 직교 배열표에 배치하고, 잡음인자로 세종류의 도로 (R1,R2,R3)에서 어떤 정해진 방법으로 주행했을 때 최소의 CO 를 발생 시키는 조건을 구하여라.

다구찌 2 수준 설계를 선택합니다.

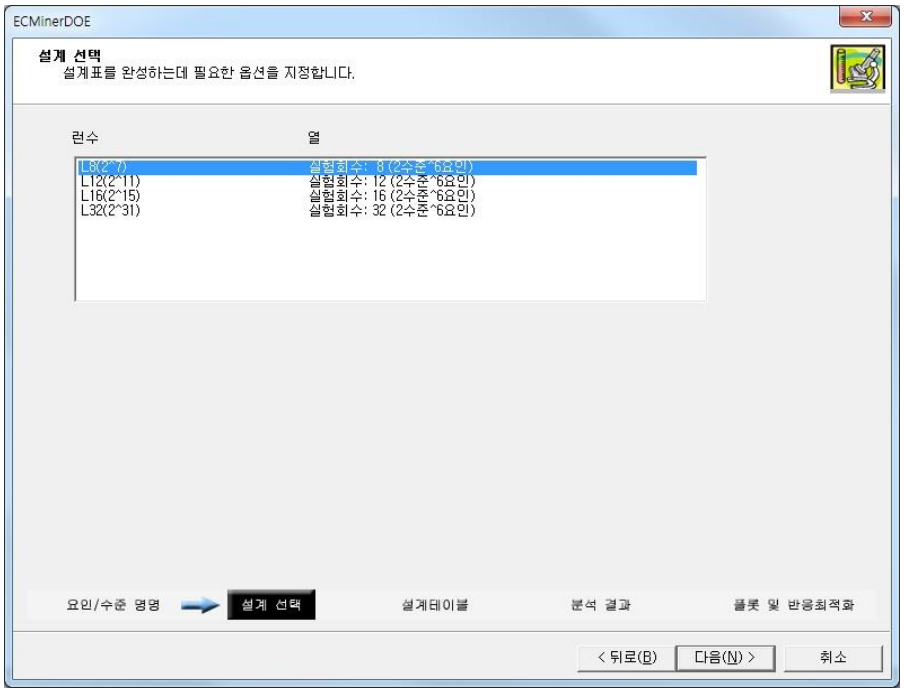


Step 1: 요인(성분) 및 수준 명명



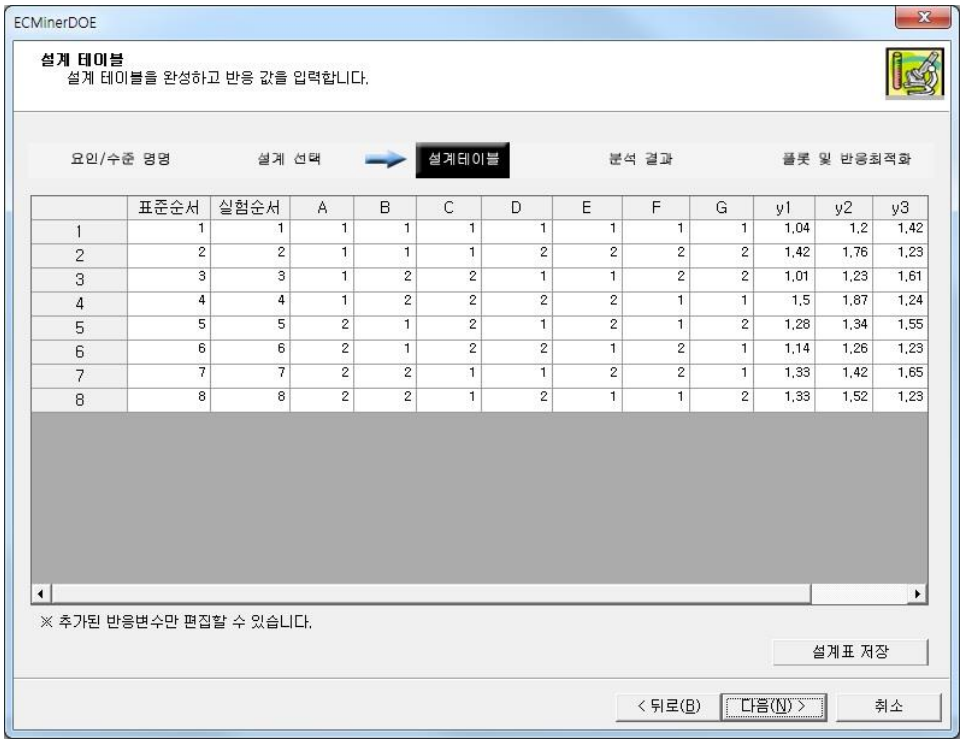
7 가지의 요인을 선택하고 반응 값의 개수로 3 을 선택합니다. 이 때 반응 값의 개수 3 은 세 종류의 도로를 의미합니다.

Step 2: 설계 선택



선택할 수 있는 설계가 많이 있는데 이 중에서 실험 수가 가장 적은 L8 설계를 선택합니다.

Step 3: 설계 테이블



	표준순서	실험순서	A	B	C	D	E	F	G	y1	y2	y3
1	1	1	1	1	1	1	1	1	1	1,04	1,2	1,42
2	2	2	1	1	1	2	2	2	2	1,42	1,76	1,23
3	3	3	1	2	2	1	1	2	2	1,01	1,23	1,61
4	4	4	1	2	2	2	2	1	1	1,5	1,87	1,24
5	5	5	2	1	2	1	2	1	2	1,28	1,34	1,55
6	6	6	2	1	2	2	1	2	1	1,14	1,26	1,23
7	7	7	2	2	1	1	2	2	1	1,33	1,42	1,65
8	8	8	2	2	1	2	1	1	2	1,33	1,52	1,23

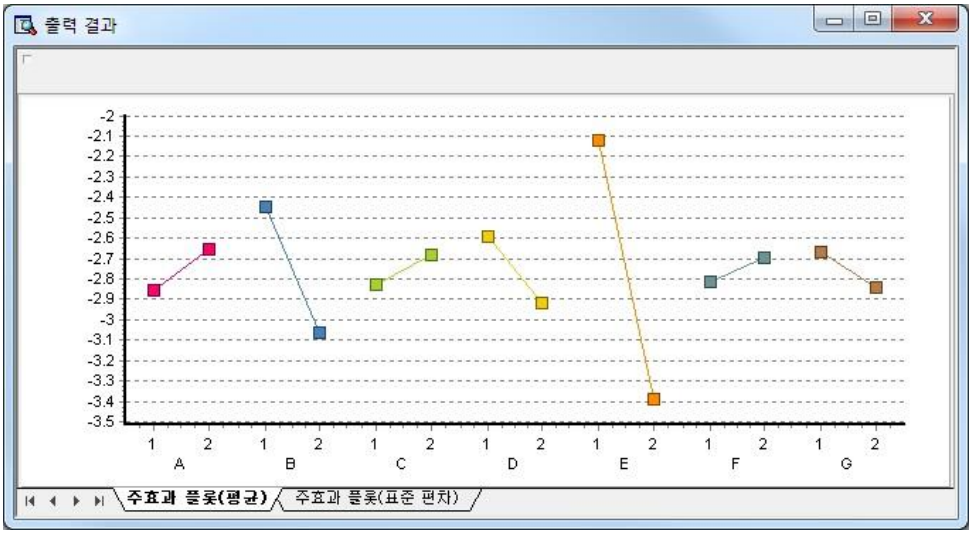
완성된 설계 테이블에 세 종류의 도로에서의 CO 의 양을 입력합니다. 이렇게 입력된 값은 추후 통계분석에 사용될 것입니다.

Step 4: 통계 분석

현재 CO 값은 작으면 작을수록 좋은 값이므로 망소 특성을 선택하고 분석을 시작합니다. 분석 결과로 SN 비와 함께 분산분석표를 얻을 수 있습니다. 현재 SS 가 매우 작은 요인의 경우 처음에 실험할 때 아예 제외를 하고 실험을 해도 좋습니다. E 의 SS 가 매우 큰 것으로 보아 이 요인이 CO 의 SN 비에 매우 결정적인 요인들임을 알 수 있습니다.

이에 대한 자세한 설명은 6.4. 설정 및 분석을 참고하세요.

Step 5: 플롯 및 반응 최적화



본 단계에서는 SN 비에 대한 주효과 플롯을 통해서 최적 조건을 찾을 수 있습니다. SN 비는 크면 클수록 좋습니다. 따라서 A2, B1, C2, D1, E1, F2, G1 조건을 최적조건이라고 할 수 있습니다. 하지만 이 중에서 몇 개의 요인의 수준 만을 결정할 수 있다면 우선적으로 B,E 를 각각 1,1 수준으로 맞추어야 합니다. 그리고 남은 요인에 대해서는 경제적, 비용적 효과를 고려하여 수준을 결정하도록 합니다. 재실험을 통해 최적으로 결정된 수준에서 재현성을 판단해 본 후 실험계획을 종료합니다.

6.3.4.3. 3 수준 설계

3 수준 설계는 요인당 수준이 3 개일 때 사용하는 설계입니다. 다음과 같은 실험이 있다고 가정합니다.

실험 소개

어떤 수지를 생산하는 한 화학업체에서는 이 수지에 포함되는 불순물의 함량률을 줄이기 위한 실험을 실시하고자 한다. 규격 상한은 4.0%이고 , 이 규격이 만족되지 않으면 10kg 당 50000 원의 손실이 발생한다. 불순물에 영향을 주리라고 예상되는 4 가지 제어인자를 다음과 같이 취하였다.

A : 본드의 배합비 3 수준$(\mathrm{A}_0, \mathrm{A}_1, \mathrm{A}_2)$

B : 본딩 방법 3 수준$(\mathrm{B}_0, \mathrm{B}_1, \mathrm{B}_2)$

C : 표면처리방법 3 수준$(\mathrm{C}_0, \mathrm{C}_1, \mathrm{C}_2)$

D : 열처리 방법 3 수준 $(\mathrm{D}_0, \mathrm{D}_1, \mathrm{D}_2)$

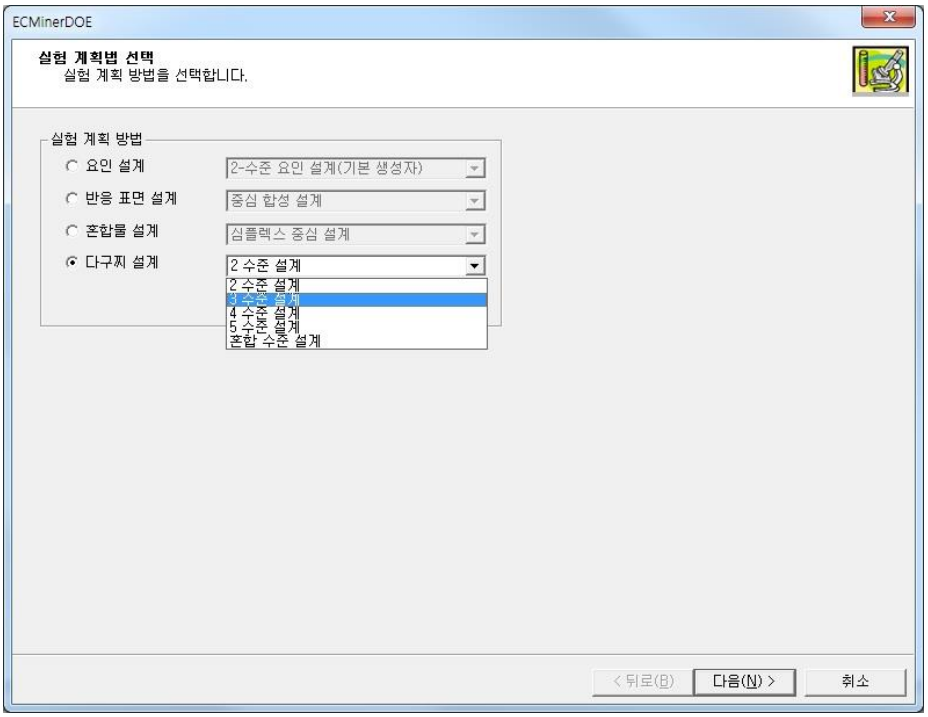
비 제어 인자로서

U : 작업자의 2 수준 $(\text{비숙련공}, \text{숙련공})$

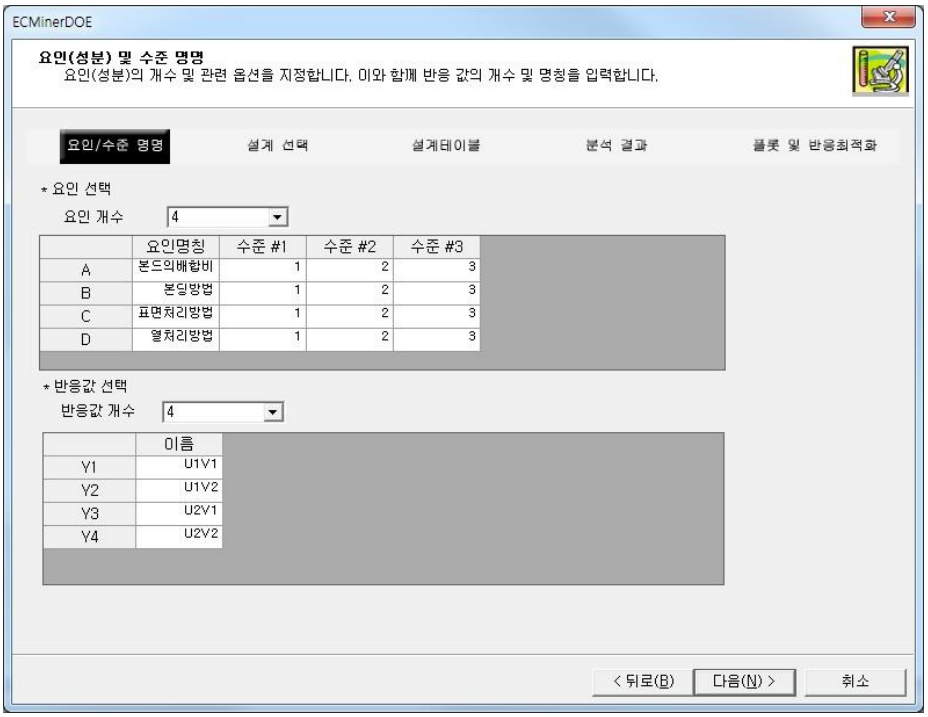
V : 수지 생산라인 2 수준

을 선택하고 생산된 수지를 실험실에서 분석하여 불순물의 함량을 얻었다.

다구찌 설계 3 수준 설계를 선택합니다.

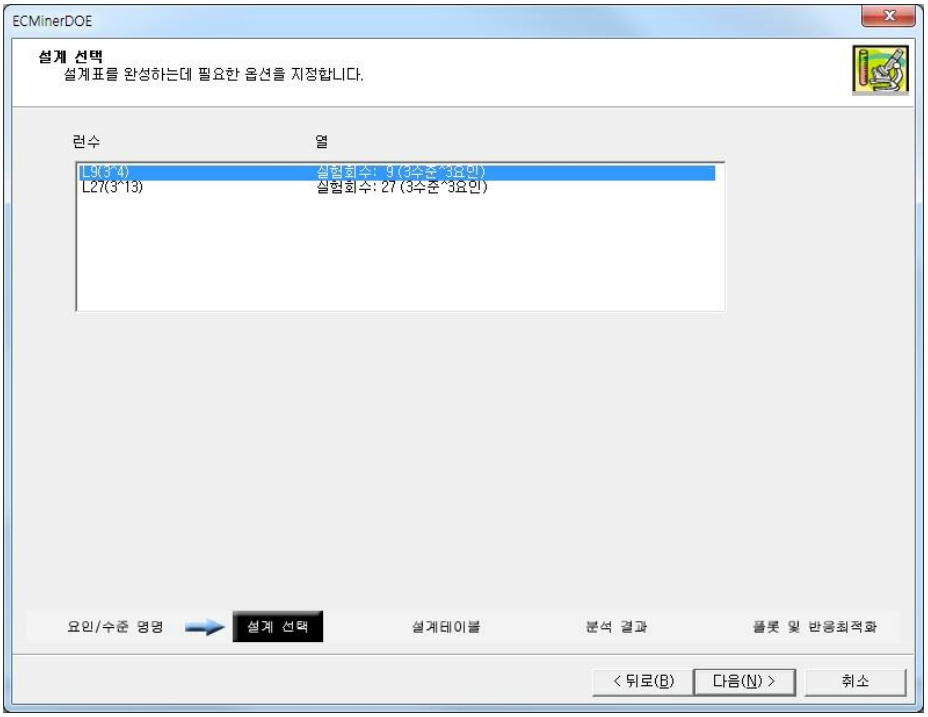


Step 1: 요인(성분) 및 수준 명명



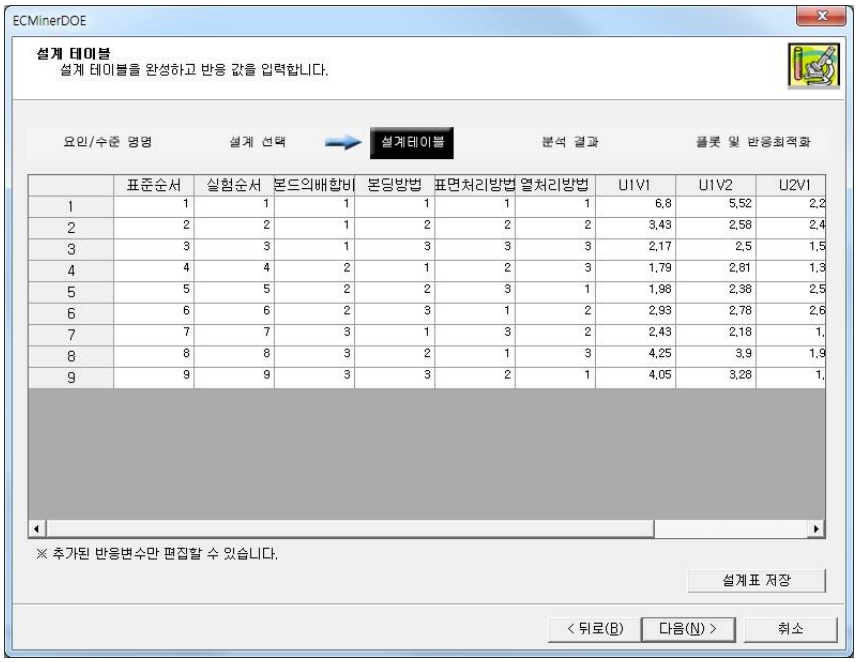
위와 같이 요인의 이름을 붙이고 반응 값의 개수를 4 개를 선택해서 각 잡음 요인의 수준에 맞는 이름을 붙입니다.

Step 2: 설계 선택



현재 선택할 수 있는 설계는 2 개인데 이 중에서 L9 를 선택합니다.

Step 3: 설계 테이블



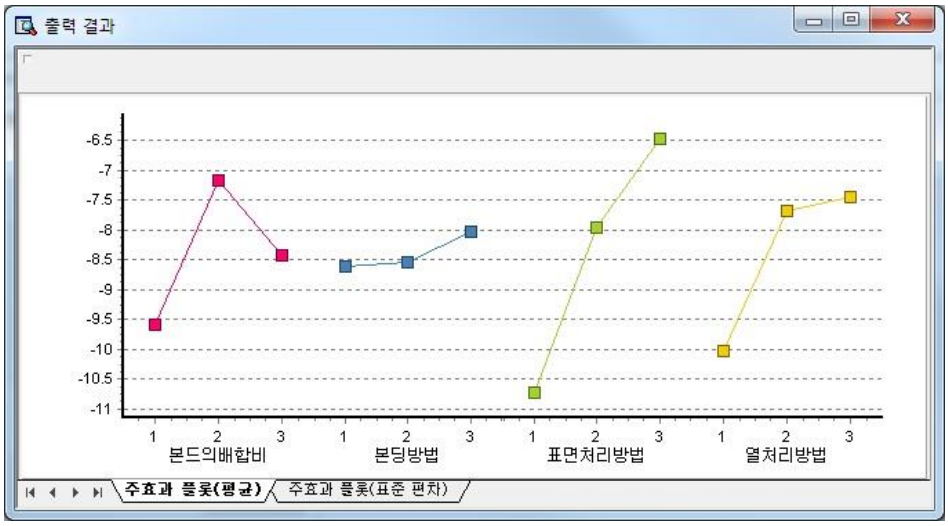
설계 테이블이 완성되면 비제어인자들의 각 수준에서 실험한 특성치를 입력합니다.

Step 4: 통계 분석

현재 특성치가 불순물의 함량이므로 이것은 작으면 작을수록 좋은 값입니다. 따라서 망소 특성을 선택하고 분석을 시작합니다. SN 비를 기준으로 한 분산분석표는 B 의 SS(Sum of Square)가 매우 작음을 보여줍니다. 이를 통해 볼 때 A, C, D 요인이 모두 SN 비의 변동을 잘 설명하는 것을 알 수 있습니다.

자세한 설명은 6.4. 설정 및 분석을 참고하세요.

Step 5: 플롯 및 반응 최적화



본 단계에서는 SN 비에 대한 주효과 플롯을 통해 최적의 실험 조건을 찾을 수 있습니다. 현재 A 는 2 수준, B 는 3 수준, C 는 3 수준, D 는 3 수준에서 SN 비가 가장 크므로 A, B, C 의 상호 작용이 없다는 전제하에 A2, B3, C3, D3 이 가장 좋은 조건임을 알 수 있습니다. 이에 대해서 확인 실험을 하여 재현성을 확인하고 실험을 마칩니다.

6.3.4.4. 4 수준 설계, 5 수준 설계, 혼합 수준 설계

모든 다구찌 설계의 분석 방법은 2, 3 수준 설계와 동일합니다. 따라서 4 수준 설계, 5 수준 설계, 혼합 수준 설계의 특징만을 언급하도록 하겠습니다.

4 수준 설계

4 수준 설계는 요인의 수준 수가 4 인 설계입니다.

5 수준 설계

5 수준 설계는 요인의 수준 수가 5 인 설계입니다.

혼합 수준 설계

혼합 수준 설계는 요인의 수준 수가 요인에 따라 다른 설계입니다. ECMiner™ DOE 에서는 $2^1 \times 3^7$ 설계를 제공합니다.

자세한 분석 방법은 6.3.4.1. 2 수준 설계, 6.3.4.2. 3 수준 설계, 6.4. 설정 및 분석을 참고하세요.

6.4 설정 및 분석

6.4.1 설정

6.4.1.1. 예측 설정

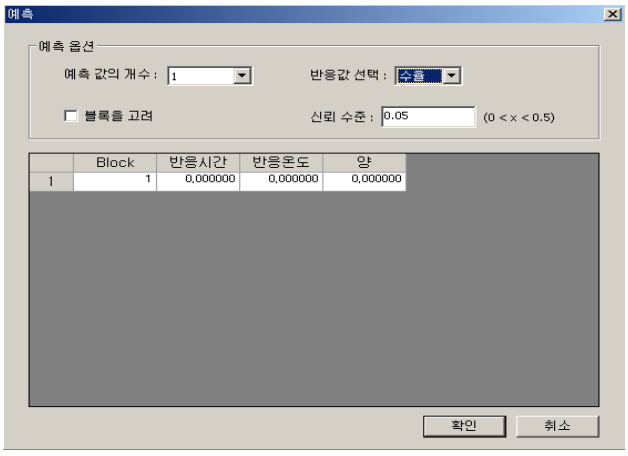
예측은 Step 4 에서 만들어지는 Regression Model 을 이용하여 특정 요인(성분) 수준에서 반응값을 예측하고자 할 때 사용합니다. 예측에서 설정해야 하는 입력 값은 다음과 같습니다.

예측 값의 개수: 몇 개의 점에서 예측을 하고 싶은지를 입력합니다.

반응 값의 선택: 복수의 반응 값이 있을 경우 예측을 하고자 하는 반응 값을 선택합니다.

신뢰 수준: 반응 값의 신뢰구간을 구할 때 사용됩니다.

블록을 고려: 블록을 고려할지 여부를 결정합니다.



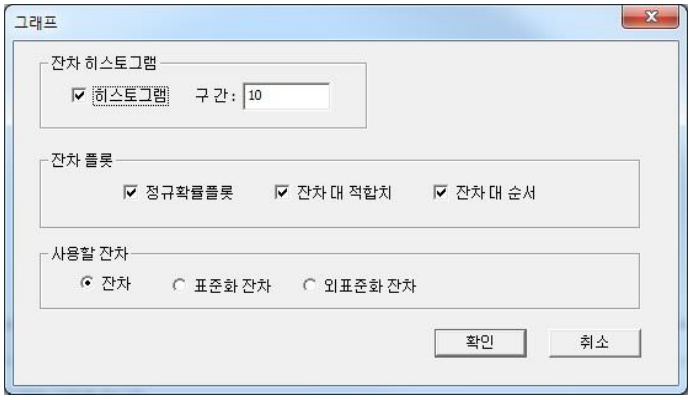
6.4.1.2. 그래프 설정

그래프 설정을 통해서는 Step 4 에서 Regression Model 을 통해서 얻어진 여러 통계량과 관련된 그래프에 관한 설정을 합니다.

잔차 히스토그램: 잔차 히스토그램을 그릴 지의 여부와 그릴 경우 구간을 몇 개로 나눌 것인지를 체크합니다.

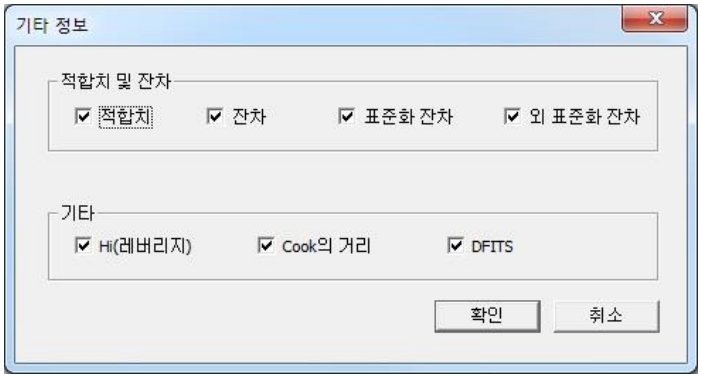
잔차 플롯: 어떠한 잔차 관련 플롯을 그릴지를 선택합니다.

사용할 잔차: 잔차 관련 플롯을 그릴 때 어떠한 잔차를 사용할지를 선택합니다.



6.4.1.3. 기타 정보

기타 정보를 통해서는 분석 결과에 어떠한 통계량을 보여줄지를 선택합니다. 선택 가능한 통계량으로는 적합치, 잔차, 표준화 잔차, 외 표준화 잔차, 레버리지, Cook 의 거리, DFFITS 가 있습니다.



6.4.2 분석

6.4.2.1. 회귀 분석(Regression Analysis)

설계 표가 생성되고 생성된 설계에 따라 반응 값을 얻게 되면 본격적인 분석을 시작할 수 있습니다. ECMiner™ DOE 에서는 회귀분석, 분산분석, 잔차 분석, 예측, 그래프 분석, 반응최적화 등의 기능을 제공합니다. 이 중에서 가장 기본이 되는 것이 바로 회귀분석입니다. 회귀분석 모델을 수학적으로 표현하면 다음과 같습니다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \quad \epsilon \sim N(0, \sigma^2)$$

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I})$$

$$\text{where } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1k} \\ x_{21} & \dots & x_{2k} \\ \dots & \dots & \dots \\ x_{N1} & \dots & x_{Nk} \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_N \end{bmatrix}$$

이 때 최소 제곱법(Least Squares)를 이용한 $\boldsymbol{\beta}$의 estimator 를 구하면 다음과 같습니다.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \quad \mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$$

이러한 사실을 이용하여 데이터를 가장 잘 설명하는 회귀 분석 모델을 세우면 다음과 같습니다.

$$\hat{\mathbf{y}} = \boldsymbol{X}\,\hat{\boldsymbol{\beta}}$$

주의: 혼합물 설계에서는 위의 Regression Model 에서 상수 항이 제외됩니다. 혼합물 성분의 합이 항상 일정해야 한다는 제약 조건으로 인하여 상수 항은 사라지게 되는 것입니다.

요인 설계 회귀분석 결과

● 반응값 y1에 대한 추정된 효과 및 계수

항	효과	계수	계수SE	T	P
Const	NaN	4,5	0,20412	22,04541	0,0002
Block1	NaN	-2	0,20412	-9,79796	0,00226
A	1,5	0,75	0,20412	3,67423	0,0349
B	-1,5	-0,75	0,20412	-3,67423	0,0349
C	0	0	0,20412	0	1

● 표준 에러

R Square	0,97619
Adjusted R Square	0,94444
RMSE	0,57735

반응 표면 설계 회귀 분석 결과

● 반응값 y1에 대한 추정된 계수(코드화 된 단위)

항	계수	계수SE	T	P
Const	9,33333	2,82154	3,30788	0,02129
A	-0,375	1,72784	-0,21703	0,83676
B	-1,375	1,72784	-0,79579	0,46224
C	-3,5	1,72784	-2,02566	0,09866
AA	-0,41667	2,54331	-0,16383	0,87628
BB	-2,41667	2,54331	-0,95021	0,38564
CC	0,33333	2,54331	0,13106	0,90084
AB	-1,5	2,44353	-0,61387	0,56615
AC	-0,75	2,44353	-0,30693	0,77127
BC	1,75	2,44353	0,71618	0,50593

● 반응값 y1에 대한 추정된 계수(코드화 되지 않은 단위)

항	계수	계수SE	T	P
Const	-13,91667	42,82752	-0,32495	0,75838
A	0,155	0,34815	0,44521	0,67478
B	0,6	0,6963	0,8617	0,42823
C	-0,375	0,64938	-0,57748	0,58865
AA	-0,00017	0,00102	-0,16383	0,87628
BB	-0,00387	0,00407	-0,95021	0,38564
CC	0,00083	0,00636	0,13106	0,90084
AB	-0,0012	0,00195	-0,61387	0,56615
AC	-0,00075	0,00244	-0,30693	0,77127
BC	0,0035	0,00489	0,71618	0,50593

● 표준 에러

R Square	0,57351
Adjusted R Square	0
RMSE	4,88706

혼합물 설계 회귀분석 결과

● 반응값 y1에 대한 추정된 계수(비율)

항	계수	계수SE	T	P
A	8,45875	3,388	NaN	NaN
B	6,45875	3,388	NaN	NaN
C	1,27694	3,388	NaN	NaN
A*B	-12,53535	15,61481	-0,80279	0,46709
A*C	1,10101	15,61481	0,07051	0,94717
B*C	13,10101	15,61481	0,83901	0,44868

● 반응값 y1에 대한 추정된 계수(양)

항	계수	계수SE	T	P
A	8,45875	3,388	NaN	NaN
B	6,45875	3,388	NaN	NaN
C	1,27694	3,388	NaN	NaN
A*B	-12,53535	15,61481	-0,80279	0,46709
A*C	1,10101	15,61481	0,07051	0,94717
B*C	13,10101	15,61481	0,83901	0,44868

● 표준 에러

R Square	0,40166
Adjusted R Square	0
RMSE	3,51293

6.4.2.2. 분산 분석(Analysis of Variance, ANOVA)

분산 분석은 전체 변동에서 각 항에 의해 생기는 변동의 정도가 어느 정도인지를 파악하여 각 항이 반응 값에 영향을 미치는지에 대한 여부를 판단할 수 있도록 해 줍니다. 분산 분석은 회귀 분석에 사용된 Regressor Matrix 를 통해서 쉽게 수행될 수 있습니다.

먼저 블록이 있는 경우 블록에 대한 변동을 구해야 합니다. Regressor Matrix X 의 첫 번째부터 i 번째 Column 으로 이루어진 Submatrix 를 X_i라고 하고, 블록의 총 개수를 bn(>=2)이라고 할 때

$$\text{Sum of squares of block} = y^{T}\left(X_{bn}\left(X_{bn}^{T}X_{bn}\right)^{-1}X_{bn}^{T} - X_1(X_1^{T}X_1)^{-1}X_1^{T}\right)y$$

입니다. 만약 블록의 개수가 1 이면 위의 과정을 실행하지 않습니다. 그리고 이 때의 자유도는 bn-1 입니다.

이제 각 항에 대한 Sum of Squares 를 구해야 합니다. 하나의 항은 X 의 하나의 Column 에 해당합니다. 어떤 항이 j 번 째 Column 에 해당한다고 하면 그 항의 Sum of square 는

$$y^T\left(X_i(X_i^TX_i)^{-1}X_i - X_{i-1}(X_{i-1}^TX_{i-1})^{-1}X_{i-1}\right)y$$

입니다. 이 때 이 항의 자유도는 1 입니다. 하지만 ECMiner™ DOE 에서는 각 항에 대해서 Sum of Square 를 구하지 않고 주 효과와 2 원 상호작용, 3 원 상호작용 혹은 다른 특징에 의해 묶어서 SS 를 구하도록 합니다. 먼저는 주 효과에 의한 열이 i 부터 j 까지라고 한다면

$$y^T\left(X_j(X_j^TX_j)^{-1}X_j - X_{i-1}(X_{i-1}^TX_{i-1})^{-1}X_{i-1}\right)y$$

의 공식을 써서 Sum of Square 를 구합니다. 이는 결국 주효과를 구성하는 열에 대한 각각의 Sum of Square 를 모두 더한 것과 같은 값입니다. 이것의 자유도는 i-j 가 됩니다. 마찬가지 방법으로 k 원 상호작용도 묶어서 Sum of Square 를 구하도록 합니다.
만약 설계에 중심점이 있다면 가장 마지막 열이 그에 해당합니다. 중심점에 대한 변동을 Curvature 라고 합니다. X 의 column 의 개수를 Ncols 라고 할 때

$$\text{Sum of Squares of Curvature} = y^T\left(X_{Ncols}(X_{Ncols}^TX_{Ncols})^{-1}X_{Ncols}^T - X_{Ncols-1}(X_{Ncols-1}^TX_{Ncols-1})^{-1}X_{Ncols-1}\right)y$$

이고 이에 대한 자유도는 1 이 됩니다.
SSE 와, SST 의 경우는 다음과 같이 구합니다.

$$\text{SSE} = \text{SST} - \text{모든 항목의 Sum of Squares.}$$

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

로 구합니다. 이 때의 자유도는 '전체 데이터 개수 – 1' 이 됩니다. 블록, Curvature, 주효과, 2 원 상호작용…. 의 F value 및 P value 는

$$\text{Fvalue} = \frac{SS/df}{SST/(n-1)} \qquad p-value = P(F > Fvalue)\ \ where\ F \sim F(df, n-1)$$

을 통해서 구합니다. 이에 덧붙여 고려해야 할 것은 적합성 결여입니다.(Lack of Fit) SSE 를 구할 때 SSE 는 Pure Error 와 Lack of Fit 으로 구성된다고 할 수 있습니다. Pure Error 는 같은 점에서 실험을 여러 번 한 경우(2 번 이상)에 구할 수 있습니다.

(단 여기서 같은 점이라는 것은 실제의 같은 점을 의미하지 않습니다. 여기서는 Regressor Matrix 의 행이 같으면 같은 점이라고 말한 것입니다. 실제 실험 점이 다르더라도 선택 항을 어떻게 선택하느냐에 따라서 Regressor Matrix 의 행은 같을 수 있습니다.)
만약 ak(1<=k<=m)실험 점에서 (Nk 번의 실험을 하였을 경우) Pure Error 는

$$\text{Pure Error} = \sum_{k=1}^{m} \text{SST at } a_k$$

$$\text{where SST at } a_k = \sum_{i \in A} y_i^2 - N_{a_k} * \left(\frac{\sum_{i \in A} y_i}{N_{a_k}} \right)^2 \quad \textit{The set A consists of} \text{ the rows of } a_k$$

그리고 Pure Error 의 자유도는 다음과 같습니다.

$$\text{DF of Pure Error} = \sum_{k=1}^{m} (N_{a_k} - 1)$$

이와 같이 Pure Error 와 그것의 자유도를 구하면 Lack of Fit 의 SS 와 자유도 또한 다음과 같이 구할 수 있습니다.

$$\text{Sum of Squres of } \quad \text{'Lack of Fit'} = \text{SSE} - \text{Pure Error}$$

$$\text{DF of Lack of Fit} = \text{DF of SSE} - \text{DF of Pure Error}$$

이 때 구한 Lack of Fit 의 F value 및 P value 는

$$\text{Fvalue} = \frac{\dfrac{\text{SS of Lack of Fit}}{DF\ of\ Lack\ of\ Fit}}{\dfrac{}{Pure\ Error\ \ DF\ of\ Pure\ Error}} \text{ and}$$

$$p - \text{value} = P(F > Fvalue) \text{ where } F \sim F(\text{DF of Lack of Fit, DF of Pure Error})$$

이를 통해 본 모델의 적합성에 결여가 있는지를 통계적으로 확인해 볼 수 있습니다.

요인 설계 분산 분석 결과

● 반응값 y1에 대한 ANOVA 테이블

항	변동	자유도	평균변동	F	P
Block	32	1	32	96	0,00226
주효과	9	3	3	9	0,05204
잔차오차	1	3	0,33333		
총변동	42	7			

반응 표면 설계 분산 분석 결과

● 반응값 y1에 대한 ANOVA 테이블

항	변동	자유도	평균변동	F	P
선형	114,25	3	38,08333	1,59456	0,30202
제곱	22,83333	3	7,61111	0,31868	0,81227
상호작용	23,5	3	7,83333	0,32798	0,80611
잔차오차	119,41667	5	23,88333		
Lack of Fit	36,75	3	12,25	0,29637	0,82928
PureError	82,66667	2	41,33333		
총변동	280	14			

혼합물 설계 분산 분석 결과

● 반응값 y1에 대한 ANOVA 테이블

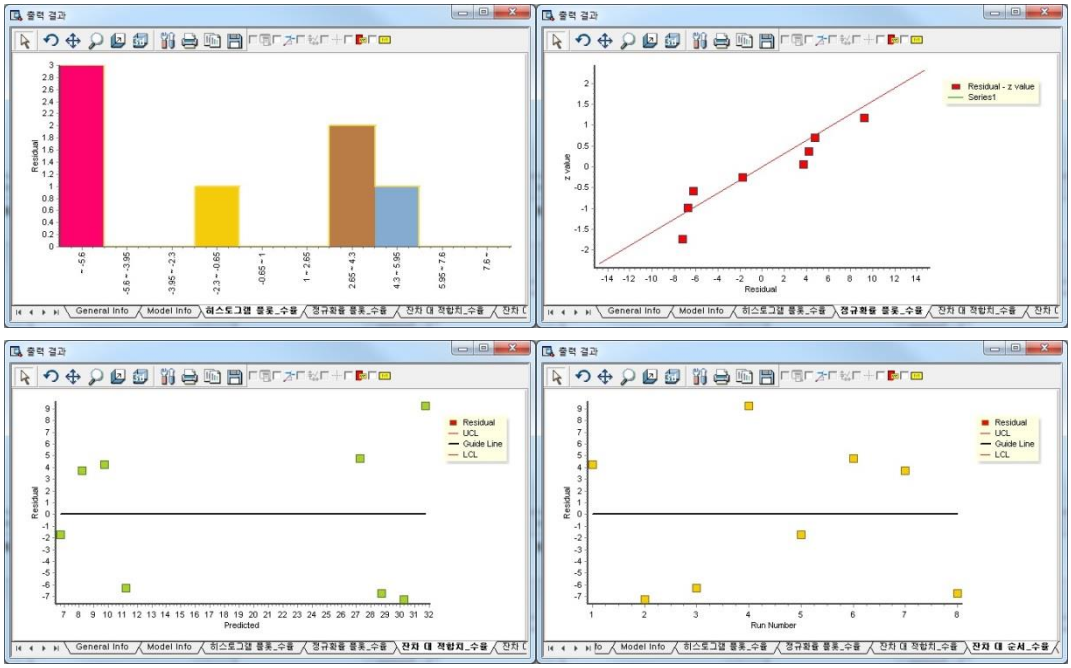
항	변동	자유도	평균변동	F	P
Linear	16,33333	2	8,16667	0,66177	0,56457
A*B	8,06402	1	8,06402	0,65345	0,46422
A*C	0,05273	1	0,05273	0,00427	0,95102
B*C	8,68712	1	8,68712	0,70394	0,44868
잔차오차	49,36279	4	12,3407		
총변동	82,5	9			

6.4.2.3. 잔차 분석(Residual Analysis)

잔차 분석을 위해서는 여러 가지 통계량이 사용됩니다.(잔차 뿐 아니라 잔차와 관련된 모든 통계량을 말합니다. ECMiner™ DOE 에서는 다음과 같은 통계량을 제공합니다.

- 잔차
- 표준화잔차
- 외표준화잔차
- 레버리지
- Cook 의 거리
- DFITTS

그리고 잔차와 관련된 그래프로서 잔차 히스토그램, 잔차 정규확률 플롯, 잔차 대 순서, 잔차 대 적합치 그래프를 제공합니다.



6.4.2.4. 다구찌 통계 분석

6.4.2.4.1. 손실함수와 SN 비

손실함수

품질의 특성치를 다음과 같이 세가지로 분류하여 각각에 대한 손실함수를 별도로 정의합니다.

- **망목 특성**

특정한 목표치가 주어져 있는 경우로 길이, 무게, 두께 등과 같이 지정된 목표가 있는 경우입니다. 측정치가 y 이고 목표치가 m 인 경우에 손실함수

$$L(y) = k(y - m)^2$$

로 정의합니다. 특성치의 소비자 허용 한계점 $m+\Delta, m-\Delta$ 에서의 소비자의 손실이 A 원이면 이는 다음과 같은 식에 의해 결정됩니다.

$$A = k\Delta^2$$

즉 망목 특성인 경우에 손실함수는

$$L(y) = \frac{A}{\Delta^2}(y - m)^2$$

이 됩니다.

▪ **망소 특성**

특성치가 작으면 작을수록 좋은 경우로 마모, 진동, 불량률 등의 특성치입니다. 이 경우는 망목 특성의 m=0 이라고 볼 수 있으므로 손실함수는 다음과 같이 됩니다.

$$L(y) = ky^2 = \frac{A}{\Delta^2}y^2$$

▪ **망대 특성**

특성치가 클수록 좋은 경우로 강도, 수명, 연료 효율 등의 특성치입니다. 이 때의 손실함수는 다음과 같습니다.

$$L(y) = \frac{A}{\Delta^2}\left(\frac{1}{y^2}\right)$$

SN 비

$$\text{SN 비} = \frac{\text{신호의 힘(power of signal)}}{\text{잡음의 힘}(power\ of\ noise)}$$

로 표현되는데 이는 종류별로 다르게 정의됩니다.

▪ **망목 특성인 경우**

$$\frac{\text{신호의 힘}}{\textit{잡음의 힘}} = \frac{\text{모평균 } \mu \text{ 의 제곱}(\mu^2)\text{의 추정값}}{\textit{분산 } \sigma^2 \textit{의 추정값}}$$

이 때 분산의 추정 값은

$$\widehat{\sigma^2} = V = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}$$

입니다. $S_m = \frac{(y_1 + y_2 + \ldots + y_n)}{n}$ 인 경우에 $E(S_m) = \sigma^2 + n\mu^2$이 성립이 되므로,

$$S_m = \widehat{\sigma^2} = n\,\widehat{\mu^2}$$

$$\hat{\mu}^2 = \frac{1}{n}(S_m - V)$$

이 됩니다.

$$SN\text{ 비} = \frac{\frac{1}{n}(S_m - V)}{V}$$

에서 상용로그를 위하면 다음의 값을 얻을 수 있습니다.

$$SN\text{ 비} = 10\log\left[\frac{\frac{1}{n}(S_m - V)}{V}\right]$$

이 값이 크면 클수록 신호의 힘이 크고 잡음의 힘이 작아지는 것으로, 이 SN 값을 가장 크게 하는 조건이 최적 조건이 됩니다. 그런데 $S_m = n(\bar{y})^2$ 이므로 SN 비는 다음과 같이 되고

$$SN = 10\log\left[\frac{(\bar{y})^2 - \frac{V}{n}}{V}\right]$$

n 이 충분히 크면 V/n 가 무시될 수 있을 정도로 작아지므로 $V = s^2$ 으로 나타나는 경우에

$$SN = 10\log\left[\frac{(\bar{y})^2}{V}\right] = 10\log\left[\frac{(\bar{y})^2}{s^2}\right] = 20\log\left(\frac{\bar{y}}{s}\right)$$

가 됩니다.

▪ 망소 특성인 경우

망소 특성인 경우에는 손실함수의 기대값을 최소화시키는 SN 비를 생각해줍니다. 반복 측정 데이터 $y_1, y_2, \ldots, y_n$이 얻어진 경우에 E(y)의 추정 값은

$$MSD = \frac{1}{n}\sum_{i=1}^{n}(y_i - 0)^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2$$

으로 볼 수 있습니다. 이를 이용하여 다음과 같이 SN 비를 구해줍니다.

$$\mathrm{SN} = -10\log\left[\frac{1}{n}\sum_{i=1}^{n} y_i^2\right]$$

▪ **망대 특성인 경우**

망소 특성인 경우와 같이 기대손실을 L 작게 해주기 위하여 E(1/y)의 추정값을

$$\mathrm{MSD} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\frac{1}{\mathrm{y_i}} - 0\right)^2 = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{1}{\mathrm{y_i^2}}$$

으로 사용하여 SN 비를

$$\mathrm{SN\ 비} = -10\log\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{y_i^2}\right]$$

으로 정해 주도록 합니다.

정리하면 다음과 같습니다.

특성치의 종류	1 개 데이터 y 의 손실함수	n 개 데이터가 얻어진 경우 평균 손실함수	SN 비
망목 특성	$\frac{\mathrm{A}}{\Delta^2}(\mathrm{y}-\mathrm{m})^2$	$\frac{\mathrm{A}}{\Delta^2}\left[\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i}-\mathrm{m})^2\right]$	$10\log\left[\frac{\frac{1}{n}(S_m - V)}{V}\right] = 20\log\left(\frac{\bar{y}}{s}\right)$
망소 특성	$\frac{\mathrm{A}}{\Delta^2}\mathrm{y}^2$	$\frac{\mathrm{A}}{\Delta^2}\left[\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{y_i^2}\right]$	$-10\log\left[\frac{1}{n}\sum_{i=1}^{n} y_i^2\right]$
망대 특성	$\mathrm{A}\Delta^2\left(\frac{1}{\mathrm{y}^2}\right)$	$\mathrm{A}\Delta^2\left[\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{1}{\mathrm{y_i^2}}\right]$	$-10\log\left[\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{1}{\mathrm{y_i^2}}\right]$

6.4.2.4.2. 계량치의 파라미터 설계

파라미터 설계는 제품 설계와 공정설계에서 유용하게 사용되는 다구찌 실험계획법의 핵심입니다. 파라미터는 제품 성능의 특성치에 영향을 주는 제어 가능한 인자를 의미하며, 파라미터 설계는 이들 인자들의 최적수준을 정하여 주는 것을 말합니다. 이러한 파라미터를 설계 변수라고도 부르며, 파라미터 설계에서는 제품의 품질 변동이 잡음에 둔감하면서 목표 품질을 가질 수 있도록 설계변수들의 최적조건을 구해 줍니다.

파라미터 설계는 일반적으로 다음과 같은 몇 가지 중요한 특징을 가집니다.

주로 직교 배열표를 이용하여 설계되며, 제어 인자들의 한 실험 조건에서 2 개 이상의 특성치를 얻습니다. 이렇게 제어 인자들의 한 실험 조건에서 여러 개의 특성치를 얻는 이유는 제어하기 어려운 변량인자의 영향을 파악하기 위해서이고 이렇게 특성치를 반복해서 얻는 방법으로 2 가지가 있습니다.

잡음인자들을 그대로 둔 상태에서 특성치를 반복해서 측정하는 것

직교 배열을 외측배열으로 이용하여 잡음인자를 배치하는 것

분산 분석 시에 성능 특성치에 대해서 분석하지 않고 SN 비로 분석합니다.

제품 설계나 공정 설계의 대상이 되는 시스템에 대하여 특성치에 영향을 주리라고 예상되는 모든 제어 인자를 포함시키고, 비 제어 인자로서 잡음인자, 블록인자 등을 배치하되 너무 많이 배치하지 않도록 합니다.

망소 / 망대 특성에 대한 파라미터 설계방법

제어 인자들로 이루어진 실험을 구성합니다. (교 배열표 사용)

각 실험 조건의 반복측정치로부터 SN 비를 계산합니다.

SN 비에 대한 분산분석을 통하여 SN 비에 영향을 미치는 제어인자를 찾습니다.

SN 비를 최대로 하는 수준 조합이 최적 수준 조합이 됩니다. SN 비에 유의한 영향을 미치지 못하는 제어인자는 경제성, 작업성 등을 고려하여 적절한 수준을 선택합니다.

최적 수준 조합에서 특성치의 모평균을 추정하여 보고 확인 실험을 실시하여 재현성이 있는가를 조사합니다.

망목 특성에 대한 파라미터 설계방법

제어 인자들로 이루어진 실험을 구성합니다. (직교 배열표 사용)

각 실험 조건의 반복 측정치로부터 SN 비와 Sn 을 계산합니다. 여기서 Sn 은 민감도를 나타내는 양으로 y 의 평균에 유의한 인자를 찾기 위하여 정의된 통계량입니다.

SN 비에 대한 분산분석을 통해 SN 비에 유의한 영향을 주는 제어 인자를 찾아냅니다.

Sn 에 대한 분산분석 등을 통하여 y 의 평균에 영향을 주는 제어인자를 찾아냅니다. SN 비에 대한 분산분석과 Sn 에 대한 분산분석을 통해서 제어인자를 다음과 같이 세가지로 분류할 수 있습니다

산포제어인자(dispersion control factor) : SN 비에 유의한 영향을 주는 인자

평균조정인자(mean adjustment factor) : y 의 평균 에만 유의한 영향을 주는 인자

기타 제어인자 : SN 비나 y 의 평균에 동시에 영향을 주지 못하는 인자

만약 하나의 제어인자가 SN 비와 y 의 평균에 동시에 영향을 준다면 이는 산포제어인자로 분류합니다.

산포 제어인자는 SN 비를 최대로 하는 수준에 놓고, y 의 평균이 목표치에 접근하도록 평균 조정인자의 수준을 조정합니다. 기타 제어인자에 대해서는 경제성, 작업성 등을 고려하여 적절한 수준을 선택합니다.

위에서 구한 최적수준조합에서의 특성치의 모평균을 추정하여 보고, 확인실험을 실시하여 재현성이 충분한가를 조사합니다.

6.4.3 플롯

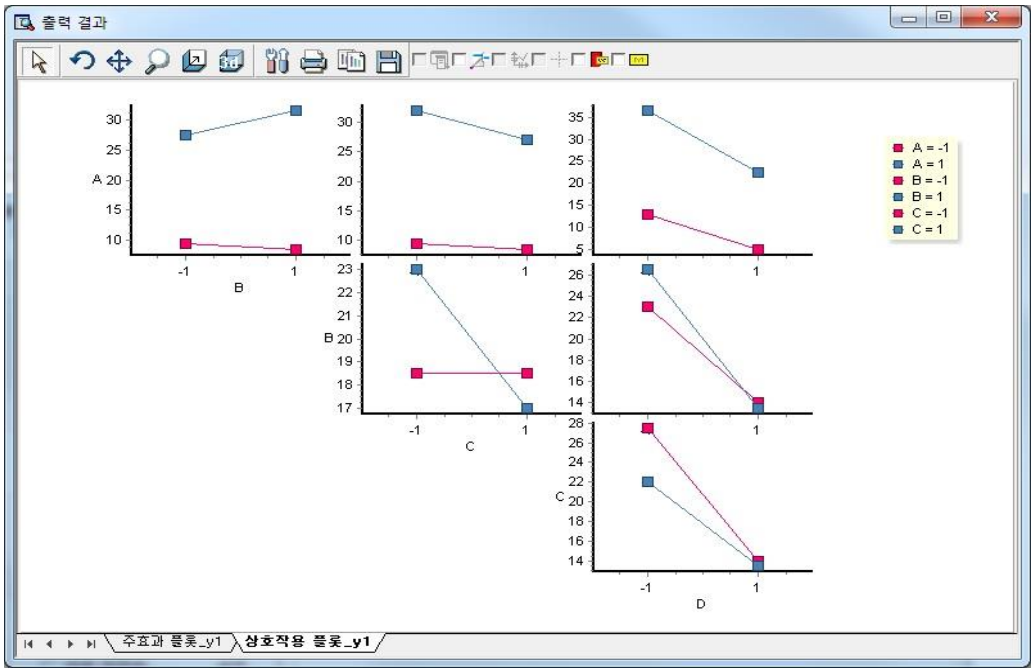
6.4.3.1. 주효과 플롯

주효과 플롯은 요인 설계(2 수준 요인 설계, Plackett Burman 설계, 일반 완전 요인 설계) 및 혼합물 설계에서 공정변수를 추가한 설계를 만들 때 나타나는 플롯입니다. 주효과 플롯을 이용하면 특정 수준에서, 반응 값 혹은 Regression Model 에서의 적합치가 어떠한 값을 갖는지를 직관적으로 파악할 수 있습니다. 요인 설계에서는 데이터 평균(반응 값을 직접 사용) 혹은 적합 평균(Regression Model 에서의 적합치 사용)을 선택할 수 있고, 혼합물 설계에서 공정변수를 추가한 설계에서는 데이터의 평균 만을 사용할 수 있습니다.

6.4.3.2. 상호 작용 플롯

이는 AB, BC 와 같이 두 가지 요인의 상호작용을 표시하기 위한 플롯입니다. 예를 들어 요인 A 가 -1,1 을 가질 수 있고 요인 B 가 -1,1 을 가질 수 있다면 요인 A 가 -1 이고 B 가 -1 일 때의

모든 실험의 반응 값(혹은 적합치)의 평균과 요인 A 가 -1 이고 요인 B 가 1 일 때의 모든 실험의 반응 값(혹은 적합치)의 평균을 이어서 하나의 선을 그립니다. 요인 A 가 1 이고 요인 B 가 -1 일 때의 모든 실험의 반응 값의 평균과 요인 A 가 1 이고 요인 B 가 1 일 때의 모든 실험의 반응 값(혹은 적합치)의 평균을 이어서 또 하나의 선을 그립니다. 이를 한 화면에 그리면 AB 의 상호작용을 볼 수 있습니다. 만약 요인이 3 개이면 AB, AC, BC 에 대해서 플롯을 그릴 수 있고 요인이 늘어나면 그에 따라서 플롯을 추가적으로 그려야 합니다.

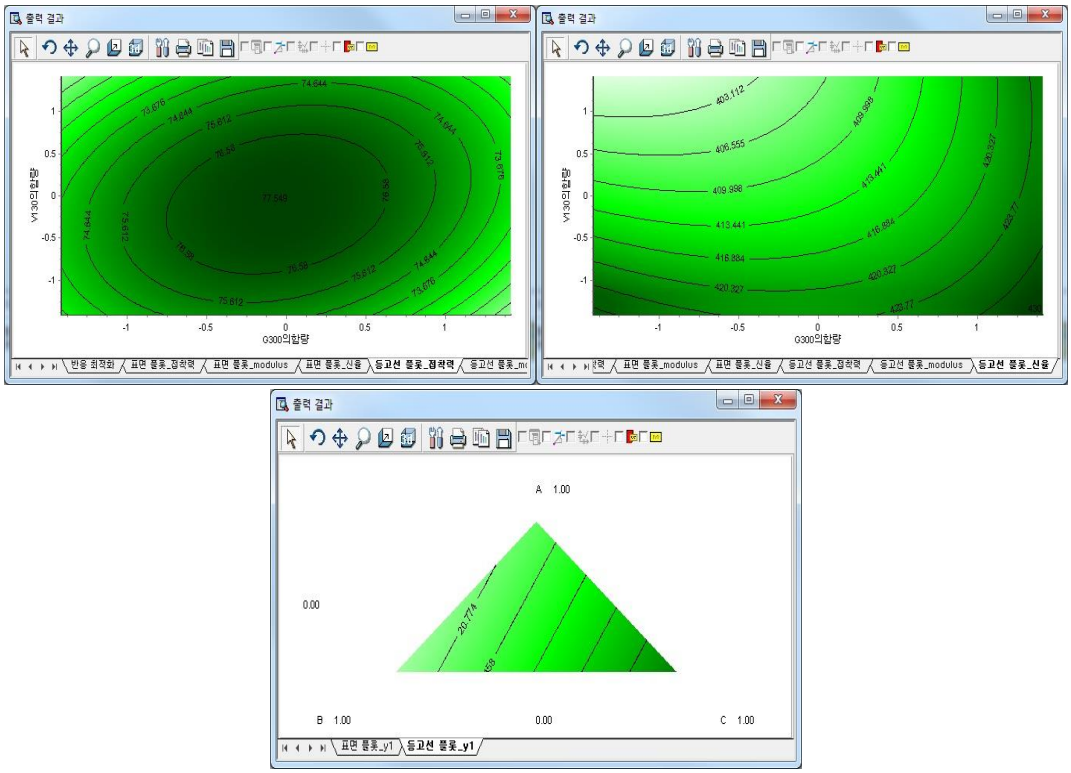


6.4.3.3. 등고선 플롯

Regression 을 통해서 다음과 같은 Regression Model 을 추정하였다고 합니다.

$$y = f(x_1, x_2, \ldots, x_n)$$

이러한 모델의 모양을 2 차원 평면에 표현하기 위해서는 n-2 개의 요인(혼합물 설계에서는 n-3 개의 성분)을 고정시킨 후 나머지 2 개의 요인(혼합물 설계에서는 3 개의 성분)이 가질 수 있는 값의 영역에서 y 값을 구하여 y 값이 크면 클수록 짙은 색, 작으면 작을수록 옅은 색으로 영역을 표현합니다. 이를 통해 반응 표면의 모양을 직관적으로 이해할 수 있습니다.

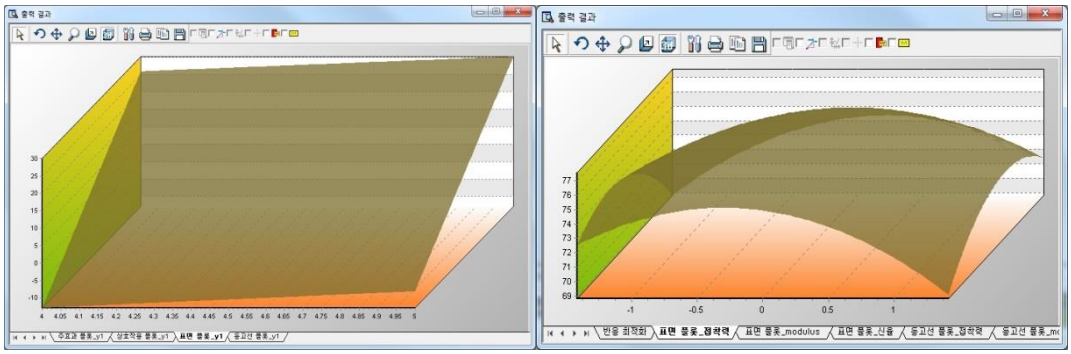


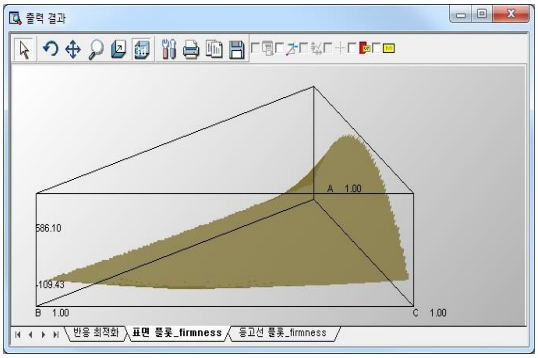
6.4.3.4. 표면 플롯

Regression 을 통해서 다음과 같은 Regression Model 을 추정하였다고 합니다.

$$y = f(x_1, x_2, \ldots, x_n)$$

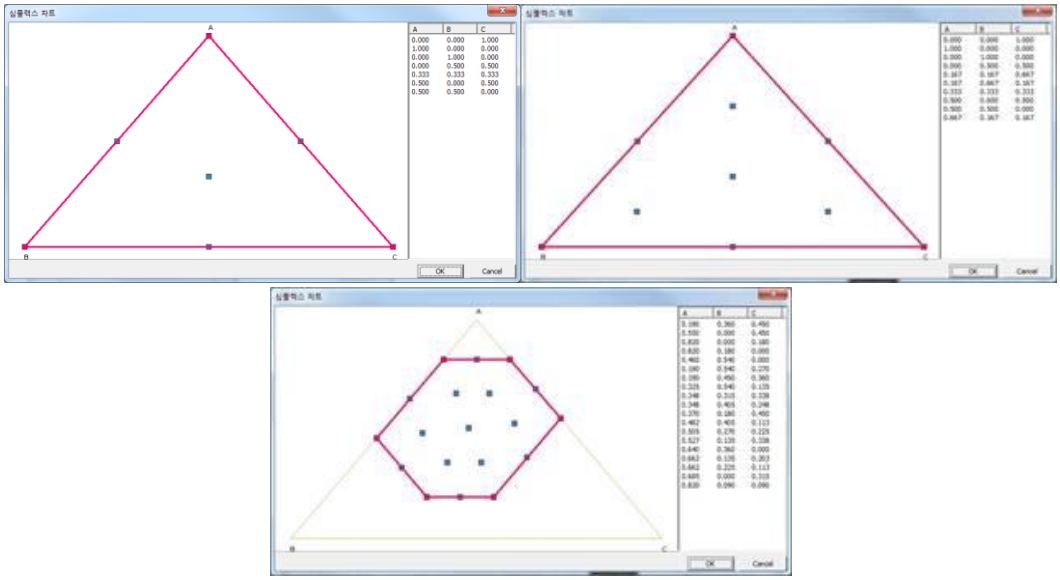
이러한 모델을 3 차원에 표현하기 위해서는 n-2 개의 요인(혼합물 설계에서는 n-3 개의 성분)을 고정시킨 후 나머지 2 개의 요인(혼합물 설계에서는 3 개의 성분)이 가질 수 있는 값의 영역에서 y 값을 구하여 이를 3 차원 그래프의 높이로 생각하여 하나의 평면을 만들어 그립니다. 이는 등고선 플롯과 함께 Regression 을 통해 얻은 반응 표면을 가장 쉽게 이해하는 도구라고 할 수 있습니다.





6.4.3.5. 심플렉스 설계 플롯

주효과 플롯, 상호 작용 플롯, 표면 플롯, 등고선 플롯이 모두 반응 값에 관심을 갖는 플롯인데 반하여 혼합물 설계에서만 제공하는 심플렉스 설계 플롯은 실험 점이 어떠한 형태로 배치되어 있는지를 보여주는 플롯입니다. DOE 의 혼합물 설계 방법론은 성분의 합이 일정한 성분들로 실험 점을 배치하는 방법을 제시합니다. 혼합물 설계에는 심플렉스 중심 설계, 심플렉스 격자 설계, 꼭지점 설계가 있는데 이러한 각 설계에서 실험점이 어떻게 정해져서 공간상에 어떻게 배치되어 있는지를 보여줌으로써 실험 점이 공간상에 적당하게 배치되어 있는지를 체크할 수 있습니다.



6.4.4 반응최적화

기본적으로 DOE 는 실험을 설계하고, 실험 점에서 실험한 결과를 가지고 Regression Model 을 만드는 것이 1 차적인 목표입니다. 왜냐하면 이 Regression Model 만 있어도 충분히 유의미한 결과를 얻어낼 수 있기 때문입니다. 조금 더 첨가를 하자면 Regression Model 의 유의미성을 판단하기 위해서 잔차 분석, 분산 분석 등을 시행합니다. 이것으로 DOE 의 과정은 어느 정도 일단락 지어집니다.

하지만 사용자의 편의 및 해석의 용이성을 위해서 플롯을 그려서 결과를 보다 시각적이고 직관적으로 표현할 수 있습니다. 이를 위해서 요인 설계의 경우는 주 효과 플롯, 상호 작용 플롯, 표면 플롯, 등고선 플롯을 제공하고 반응 표면 설계의 경우는 표면 플롯, 등고선 플롯을 제공합니다. 혼합물 설계의 경우에는 표면 플롯, 등고선 플롯을 제공하는데 공정변수가 추가될 경우에는 공정 변수에 대한 주 효과 플롯 및 상호 작용 플롯을 제공합니다.

하지만 DOE 의 최종적인 Step 은 반응 최적화라고 할 수 있습니다. 반응 최적화는 단순히 Regression Model 의 반응 값을 최대화 혹은 최대화시키는 것을 목적으로 하지 않습니다. 물론 이와 같은 부분적인 목적은 여러 입력 값의 조정을 통해서 충분히 달성할 수 있습니다. DOE 에서 말하는 반응 최적화는 보다 일반적인 목적을 충족시키는데 유용한 도구입니다. 예를 들어서 하나의 실험에서 두 개 이상의 반응 값이 나온다고 할 때 하나의 반응 값은 크게 하고 싶고 하나의 반응 값은 작게 하고 싶다고 하면 이는 어떠한 새로운 함수를 만들어서 이 함수를 증가시키는 문제를 풂으로써 해결될 수 있습니다. 물론 이 함수는 첫 번째 반응 값은 클 때 커지고, 두 번째 반응 값은 작을 때 커질 것입니다. 이 후에는 이를 어떻게 수학적으로 수식화하여 목표에 도달할 수 있는지에 대해서 설명합니다.

6.4.4.1. 바람직성 함수 (Desirability Function)

하나의 반응열(column)에 대해서 하나의 바람직성 함수가 존재하는데 반응열이 하나일 경우는 이 하나의 바람직성 함수를 최적화하는 요인들의 수준 조합을 찾으면 됩니다. 만약에 반응열이 여러 개이고 여러 개의 반응열을 모두 고려한 최적 요인 수준 조합을 찾고 싶으면 여러 개의 바람직성 함수를 모두 고려한 하나의 종합 바람직성 함수를 만들어서 그것을 최적화하면 됩니다. 반응열이 m 개이고 그것에 대한 바람직성 함수가 각각 d1,d2,..,dm 라고 하고, 각각의 반응열에 대한 중요도(importance)가 각각 r1,r2,..,rm 이라고 하면 그것을 모두 고려한 종합 바람직성 함수는

$$D = \left(d_1^{r_1} d_2^{r_2} \dots d_m^{r_m}\right)^{\frac{1}{\sum_{i=1}^{m} r_i}}$$

이 됩니다.

그런데 각각의 di 는 요인들 x1,x2,…의 함수이고 그렇기 때문에 D 또한 x1,x2,…의 함수라고 할 수 있습니다. 최종 목표는 D 를 최대화하는 x1,x2,…의 조합을 찾는 것이라고 할 수 있습니다.

최소화

위에서 설명하였듯이 하나의 반응열에는 하나의 바람직성 함수가 대응됩니다. 반응을 최소화 하고 싶든, 최대화하고 싶든, 혹은 목표 값에 적중시키고 싶든 이 바람직성 함수를 최대화하는 것이 목적임은 물론입니다. 따라서 바람직성 함수는 사용자가 원하는 최적화의 종류에 따라 달라야 할 것입니다. 최소화를 선택하였을 때는 상한과 목표 값을 설정할 수 있습니다. 상한을 U, 목표 값을 T 라고 했을 때 최소화에서의 바람직성 함수는

$$d = 1 \quad \text{if} \quad y < T$$

$$d = \left(\frac{U - y}{U - T}\right)^{\lambda} \quad \text{if } T \le y \le U$$

$$d = 0 \quad y > U$$

입니다. 여기서 λ 는 가중치입니다.

최대화

최대화를 선택하였을 때는 하한과 목표 값을 선택할 수 있습니다. 하한을 L, 목표 값을 T 라고 했을 때 최대화에서의 바람직성 함수는

$$d = 0 \text{ if } y < L$$

$$d = \left(\frac{y - L}{T - L}\right)^{\lambda} \quad L \le y \le T$$

$$d = 1 \quad y > T$$

이다. 여기서 λ 는 가중치입니다.

목표 값 적중

목표 값 적중을 선택하였을 때는 상한과 하한 및 목표 값을 선택할 수 있습니다. 하한을 L, 상한을 U, 목표 값을 T 라고 하였을 때

$$d = 0 \quad y < L$$

$$d = \left(\frac{y - L}{T - L}\right)^{\lambda} \quad L \le y \le T$$

$$d = \left(\frac{U - y}{U - T}\right)^{\lambda} \quad T \le y \le U$$

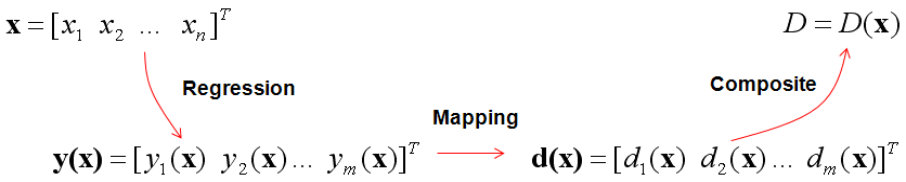
$$d = 0 \quad y > U$$

여기서 λ 는 가중치입니다.

위에서 공통적으로 나타나는 y 는 Regression Model 을 통해서 만든 회귀식입니다. 따라서

$$y = f(x_1, x_2, \ldots, x_k)$$

이고, 그러므로 d 또한 x1,x2,...의 함수가 됩니다. 이를 통해 각각의 d 를 구할 수 있고 여기에 가중치를 고려하여 계산하면 종합 바람직성 함수를 구할 수 있습니다. 이 과정을 종합하면 다음과 같습니다.

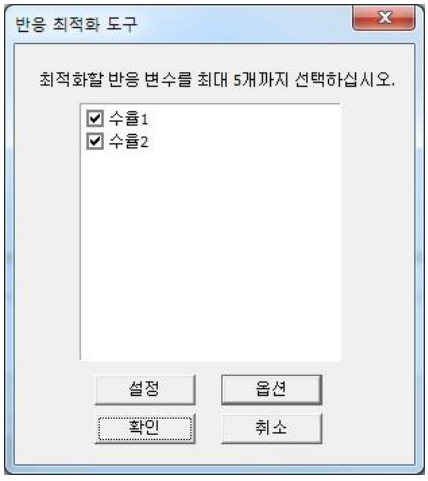


이제 최적화 알고리즘을 도입하여 이 종합 바람직성 함수를 최대화해야 합니다. 이 종합 바람직성 함수는 사용자가 흔히 접하는 함수의 모양과 다릅니다. 많은 점에서 미분이 불가능하고 따라서 Derivative Based Optimization Algorithm 을 쓸 수가 없습니다. ECMiner™ DOE 에서는 Box. M.J 의 Constrained Simplex Algorithm 을 사용합니다. 이 알고리즘을 통해서 종합 바람직성 함수를 최대화 할 수 있습니다.

6.4.4.2. ECMiner™ DOE 반응 최적화

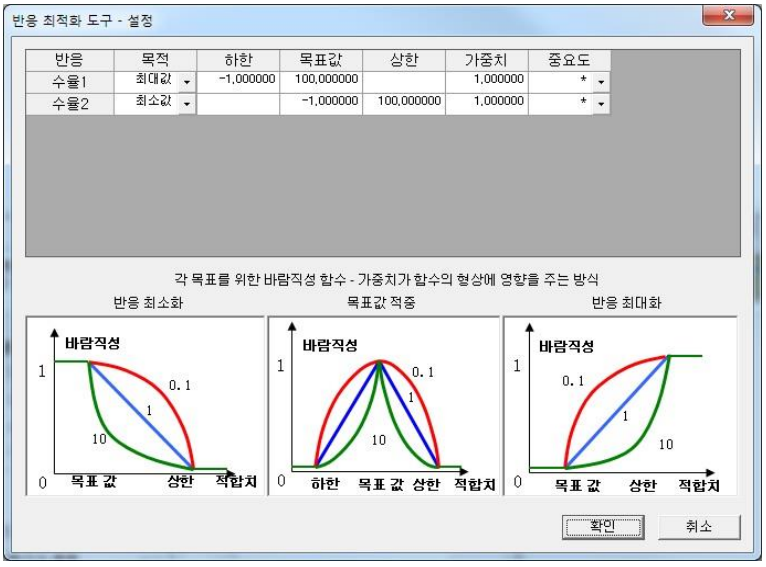
반응 변수 선택

ECMiner™ DOE 반응 최적화는 최적화를 할 반응 변수를 선택하는 것부터 시작합니다. 다변량 반응 최적화를 위해서 반응 변수 2 개를 선택하도록 합니다.

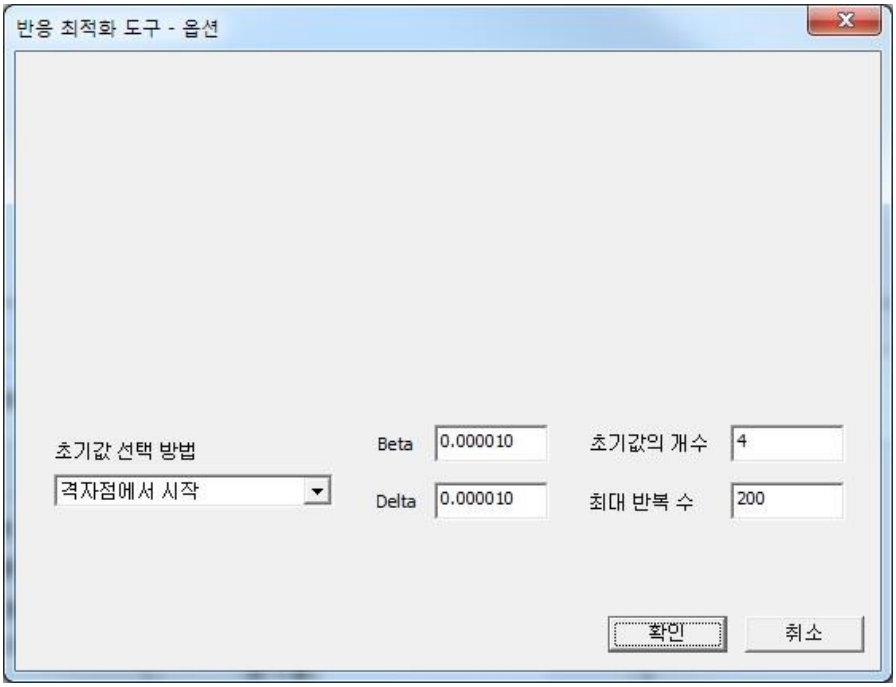


반응 최적화 설정

설정 단추를 누르면 다음과 같은 화면이 나타나고, 수율 1 은 최대화하고, 수율 2 는 최소화하고 싶을 때 다음과 같이 입력합니다. 이 때 가중치는 바람직성 함수 그림에서 보여주듯이 곡선의 곡률을 결정하고 중요도는 각 반응 변수의 중요성의 정도를 나타내는 지표입니다.



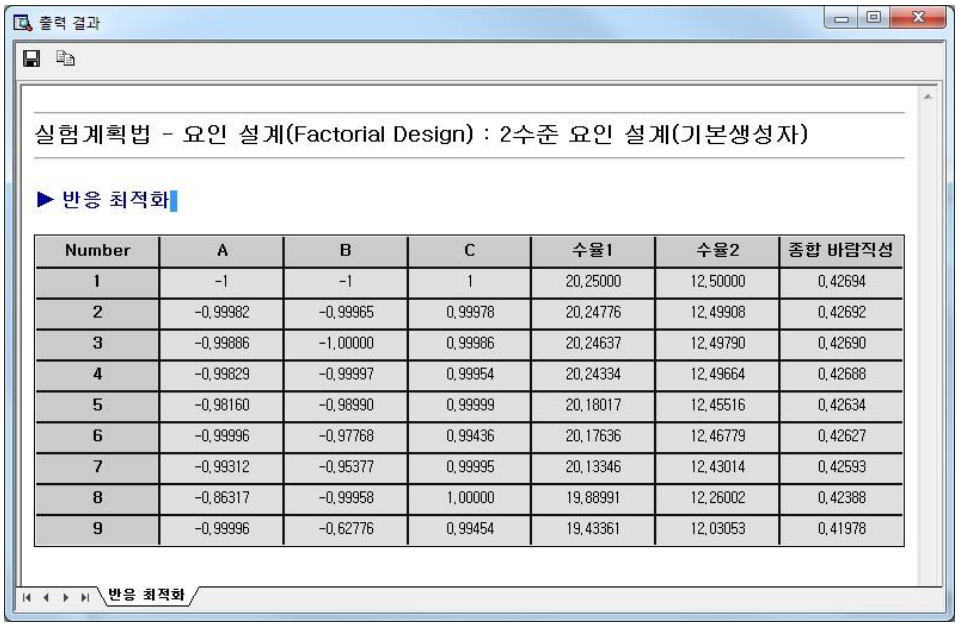
반응 최적화 옵션



옵션에서는 최적화 알고리즘에서 필요한 옵션을 선택하도록 합니다.

- 초기값 선택 방법: 모든 최적화 알고리즘에서 초기값은 매우 중요한 역할을 합니다. 초기값에 따라 알고리즘이 local optimum 을 찾을지, global optimum 을 찾을 지가 결정될 수 있기 때문입니다. ECMiner™ DOE 에서는 요인 설계, 반응 표면 설계에서는 격자 점에서 시작, 랜덤 점에서 시작, 사용자 정의 점에서 시작이라는 선택방법을 제공하고 혼합물 설계에서는 랜덤 점에서 시작 선택 방법을 제공합니다.
- Beta: 알고리즘의 특성상 여러 점에서 구한 Composite Desirability 를 저장해 놓는데 이 때 수렴 조건은 여러 점에서 구한 Composite Desirability 사이의 최대 차이가 Beta 보다 작아야 한다는 것입니다. 따라서 보다 정확한 값을 원하는 사용자의 경우, 이 Beta 값을 아주 작게 하면 원하는 목적을 이룰 수 있습니다.
- Delta: Delta 는 어떤 점이 제한 영역 밖으로 나갔을 때 이를 다시 제한 영역 안으로 들어오도록 해야 하는데 이 때 가까운 제한 영역 평면의 어느 정도 영역 안까지 점을 들어오게 할지를 결정하는 measure 입니다. 하지만 경험적으로 볼 때 measure 가 알고리즘의 성능에 결정적인 영향을 미치지는 않습니다.
- 초기값의 개수: 서로 다른 몇 개의 초기 값에 대해서 알고리즘을 수행할지를 결정하는 것입니다.
- 최대 반복 수: 알고리즘을 최대 몇 번까지 반복할지를 결정하는 것입니다. 반복 수가 많을수록 더 좋은 해를 찾을 가능성이 높아집니다.

최적화 결과


출력 결과

실험계획법 - 요인 설계(Factorial Design) : 2수준 요인 설계(기본생성자)

▶ 반응 최적화

Number	A	B	C	수율1	수율2	종합 바람직성
1	-1	-1	1	20,25000	12,50000	0,42694
2	-0,99982	-0,99965	0,99978	20,24776	12,49908	0,42692
3	-0,99886	-1,00000	0,99986	20,24637	12,49790	0,42690
4	-0,99829	-0,99997	0,99954	20,24334	12,49664	0,42688
5	-0,98160	-0,98990	0,99999	20,18017	12,45516	0,42634
6	-0,99996	-0,97768	0,99436	20,17636	12,46779	0,42627
7	-0,99312	-0,95377	0,99995	20,13346	12,43014	0,42593
8	-0,86317	-0,99958	1,00000	19,88991	12,26002	0,42388
9	-0,99996	-0,62776	0,99454	19,43361	12,03053	0,41978

반응 최적화

ECMiner™ DOE 에서는 위와 같이 최적화 결과를 보여줍니다. 각 요인(혹은 성분)의 어떤 값에서 Composite Desirability(종합 바람직성)이 최대화가 되는지를 보여줌으로써 사용자는 목적을 달성할 수 있습니다.

제 7 장. 확률 분포

7.1. 개요

여러 상황에서 접하는 Data 를 단순히 히스토그램을 그려 살펴 본다고 할 때 Data 가 발생되는 상황에 따라 매우 다른 형태를 갖는다는 것을 경험적으로 알고 있습니다. 확률 분포는 바로 이렇게 서로 다른 형태를 갖는 데이터를 설명하기 위해 도입되었습니다. 각 데이터는 형태에 따라 잘 맞는 분포가 있습니다. 분석자는 데이터 형태와 가장 잘 맞는 분포를 선택함으로써 보다 일반적으로 데이터를 설명할 수 있습니다.

7.2. 확률 분포의 종류

7.2.1. 베타 분포 (Beta distribution)

베타 분포는 연속 확률 분포의 일종입니다. 이는 두 모수를 갖는 분포로 Dirichlet 분포의 특정한 경우입니다. 베타 분포는 여러 분포를 표현할 수 있는 능력이 있습니다. 예를 들어 Beta(1,1)의 경우는 [0,1]의 연속 균일 분포와 동일한 것입니다. 즉 이처럼 여러 분포와의 연관성이 높고, Bayesian 통계, 정보 이론(Information Theory) 등과 같은 폭 넓은 적용성 때문에 많이 사용되고 있습니다.

베타분포의 pdf

$$f(x \mid a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

베타분포의 cdf

$$F(x \mid a,b) = \frac{1}{B(a,b)} \int_0^x t^{a-1}(1-t)^{b-1} dt$$

베타분포의 평균, 분산

$$E(X) = \frac{a}{a+b}$$

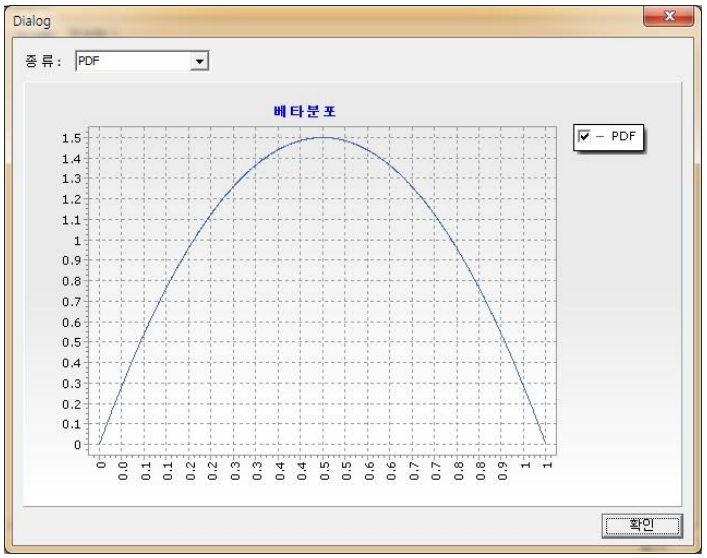
$$Var(X) = \frac{ab}{\{(a+b+1)(a+b)^2\}}$$

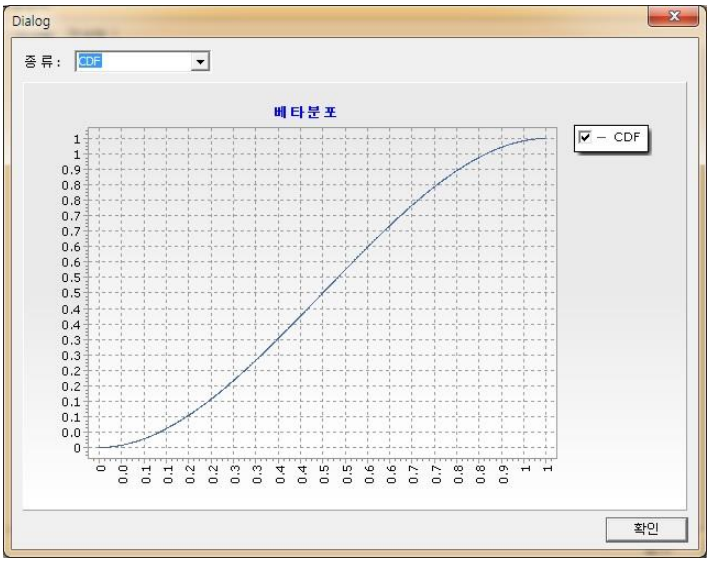
베타 분포의 누적 분포 함수의 역함수

$$x = F^{-1}(p \mid a,b) = \{x : F(x \mid a,b) = p\}$$

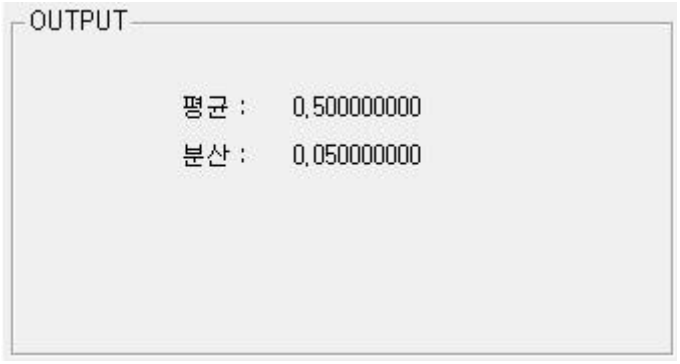
예시

$a=2,b=2$ 일 때 pdf, cdf 화면





$a=2,b=2$ 일 때 평균, 분산



$a = 2, b = 2$, p=0.648 일 때 누적 분포 함수의 역함수

7.2.2. 이항분포 (Binomial distribution)

이항분포는 베르누이 분포에서 유도됩니다. 베르누이 확률 변수는 다음과 같은 성질을 갖습니다.

$$X_k = \begin{cases} 1 & \text{with probabilit y p} \\ 0 & \text{with probabilit y 1 - p} \end{cases}$$

이 때 이항분포는 서로 독립인 베르누이 확률 변수로부터 다음과 같이 만들어 집니다.

$$Y = X_1 + X_2 + ... + X_n$$

이 때 $P(Y = k)$ 를 해석하면 각 시행에서 성공(1)이 일어날 확률이 p 이고 실패(0)할 확률이 1-p 인 실험을 동시에 n 번할 때 k 번이 성공하고 n-k 번이 실패할 확률이라고 할 수 있습니다. 이항분포에 대한 자세한 설명은 다음과 같습니다.

이항분포의 pmf

$$f(x \mid n, p) = {}_nC_x p^x (1-p)^{n-x}$$

이항분포의 cdf

$$F(x \mid a, b) = \sum_{i=1}^{x} {}_nC_i p^i (1-p)^{n-i}$$

이항분포의 평균,분산

$$E(X) = np$$

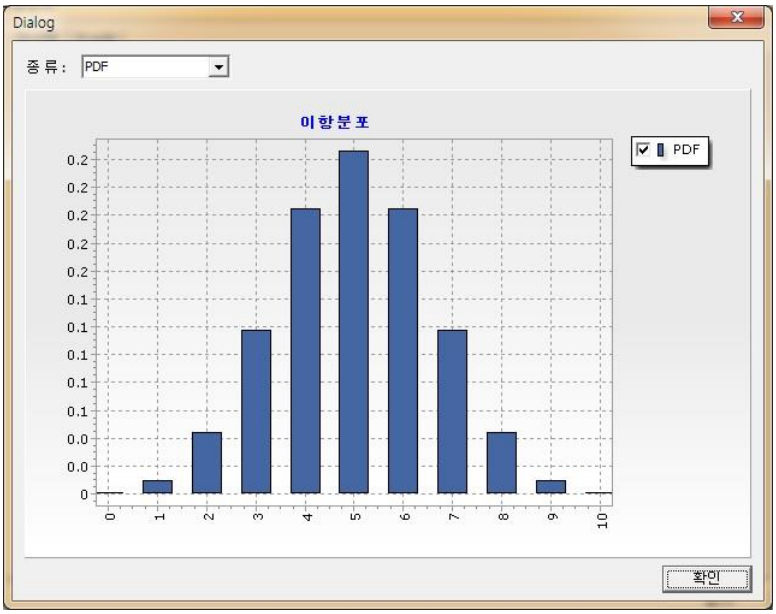
$$Var(X) = np(1-p)$$

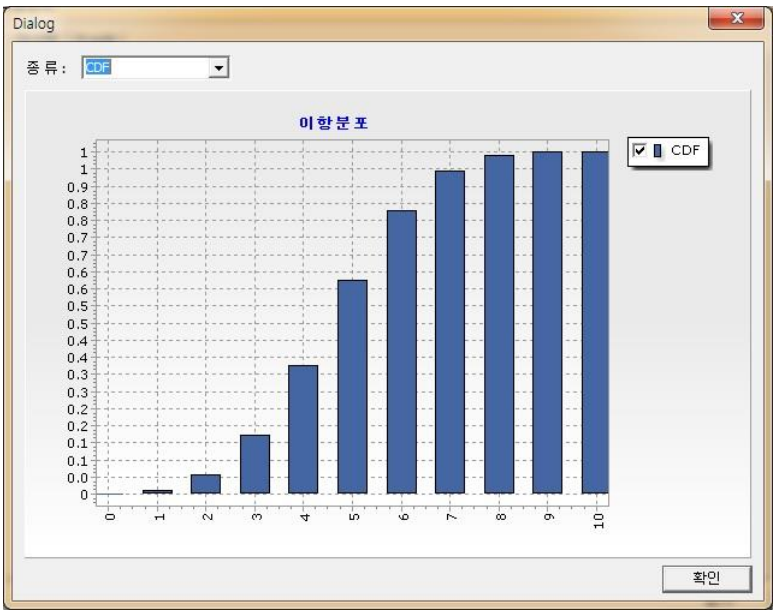
이항 분포의 누적 분포 함수의 역함수

$$x = F^{-1}(p \mid a, b)$$

예시

- $n = 10, p = 0.5$ 일 때 pmf, cdf 그래프





▪ $n=10, p=0.5$ 일 때 평균, 분산. **X=5 일 때 pdf, cdf**

▪ $n = 10, p = 0.5$ 일 때 **y= 0.5** 일 때 누적 분포 함수의 역함수

7.2.3. 카이 제곱 분포 (Chi-squared distribution)

카이 제곱 분포는 $Z_1, Z_2, ..., Z_n$가 iid 인 표준 정규 분포일 때

$$\chi^2 = \sum_{i=1}^{n} Z_i^{\,2}$$

가 따르는 분포입니다. 위와 같은 경우 카이제곱분포의 자유도는 n 이 되는데 일반적으로는 자유도 ν (실수)에 대해서 정의됩니다. 카이 제곱 분포는 Goodness of Fit Test, Likelihood Ratio Test 등 많은 통계 분야에 사용됩니다. 정규 분포와 더불어 가장 많이 사용되는 분포 중 하나입니다.

카이 제곱 분포의 pdf

$$f(x \mid \nu) = \frac{x^{\frac{\nu-2}{2}} e^{-\frac{x}{2}}}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})}$$

카이 제곱 분포의 cdf

$$\mathrm{F(x|v)} = \int_0^{\mathrm{x}} \frac{\mathrm{t}^{\frac{\mathrm{v}-2}{2}} \mathrm{e}^{-\frac{\mathrm{t}}{2}}}{2^{\frac{\mathrm{v}}{2}} \Gamma(\frac{\mathrm{v}}{2})} \mathrm{dt}$$

카이 제곱 분포의 평균, 분산

$$\mathrm{E(X)} = \int_0^{\infty} \mathrm{t} \frac{\mathrm{t}^{\frac{\mathrm{v}-2}{2}} \mathrm{e}^{-\frac{\mathrm{t}}{2}}}{2^{\frac{\mathrm{v}}{2}} \Gamma(\frac{\mathrm{v}}{2})} \mathrm{dt} = \mathrm{v}$$

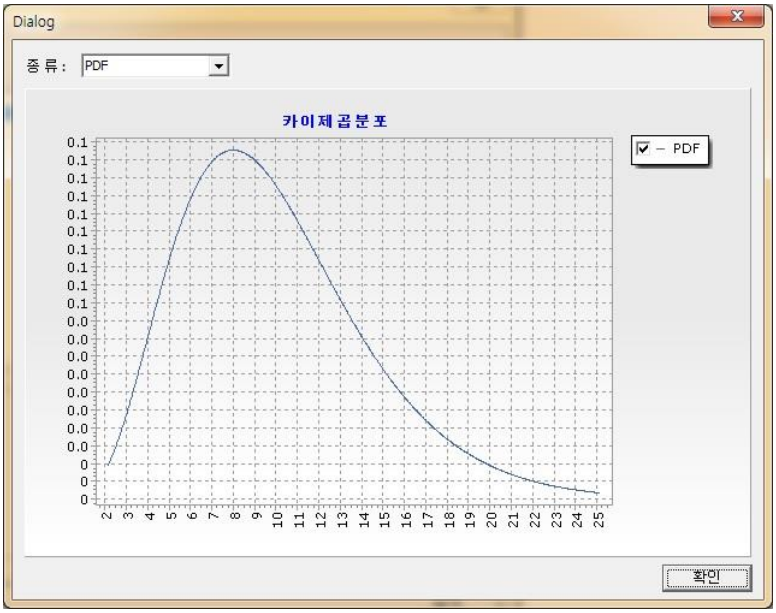
$$\mathrm{Var(X)} = \mathrm{E(X^2)} - (\mathrm{E(X)})^2 = \int_0^{\infty} \mathrm{t}^2 \frac{\mathrm{t}^{\frac{\mathrm{v}-2}{2}} \mathrm{e}^{-\frac{\mathrm{t}}{2}}}{2^{\frac{\mathrm{v}}{2}} \Gamma(\frac{\mathrm{v}}{2})} \mathrm{dt} - \mathrm{v}^2 = 2\mathrm{v}$$

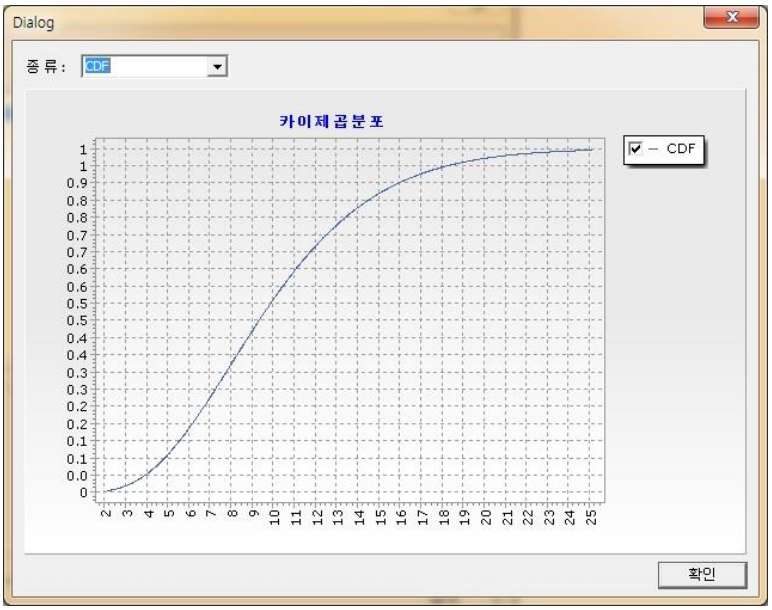
카이 제곱 분포의 누적 분포함수의 역함수

$$x = F^{-1}(p \mid \nu)$$

예시

- $\nu = 10$ 일 때 **pdf, cdf** 의 그래프





- $\nu = 10$ 일 때 평균, 분산. **X=5** 일 때의 **pdf, cdf**

- $v = 10, p = 0.5$ 일 때의 누적 확률 분포의 역함수

7.2.4. 극단값 분포 (Extreme value distribution)

극단값 분포는 연속 확률 분포의 일종입니다. 극단값 분포는 안 좋은 영향을 끼칠 수 있는 매우 희귀하게 일어나는 event 를 모델링 하는데 사용됩니다. 이러한 예로는 기온이 매우 높은 경우, 환율이 매우 심하게 변동하는 경우, 주식 시장이 붕괴하는 경우 등이 있습니다. 이러한 상황들을 모델링 하기 위해서 경제학자, 기업 분석가, 공학자들은 극단값 분포를 활용하고 있습니다.

극단값 분포의 pdf

$$f(x \mid \mu,\sigma) = \frac{1}{\sigma}\exp(\frac{x-\mu}{\sigma})\exp(-\exp(\frac{x-\mu}{\sigma}))$$
$$-\infty < x < \infty, \sigma > 0$$

극단값 분포의 cdf

$$F(x \mid \mu,\sigma) = 1-\exp(-\exp(\frac{x-\mu}{\sigma}))$$
$$-\infty < x < \infty, \sigma > 0$$

극단값 분포의 평균, 분산

$$E(X) = \int_{-\infty}^{\infty} x f(x \mid \mu,\sigma)dx = \mu\psi(1)\sigma$$

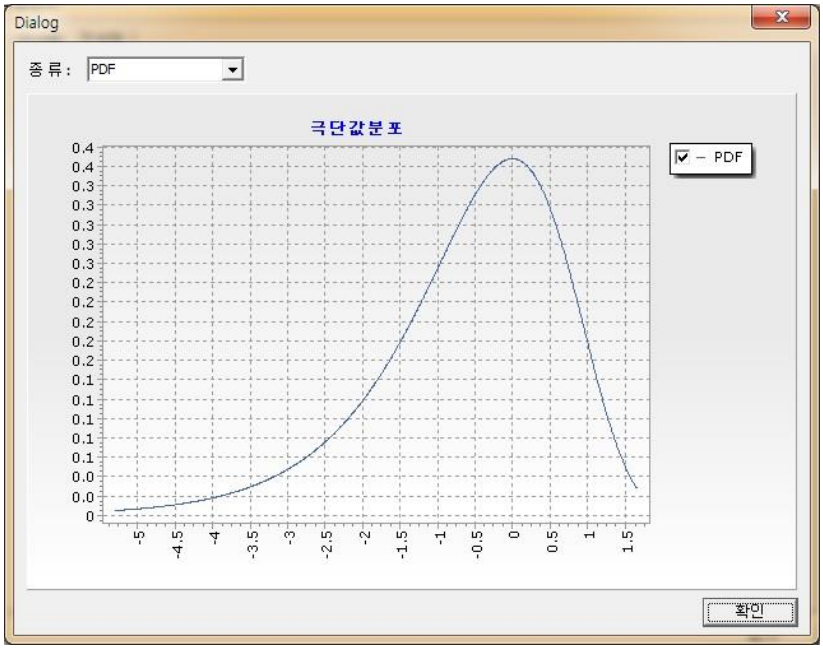
$$Var(X) = E(X^2) - (E(X))^2 = \frac{(\pi\sigma)^2}{6}$$

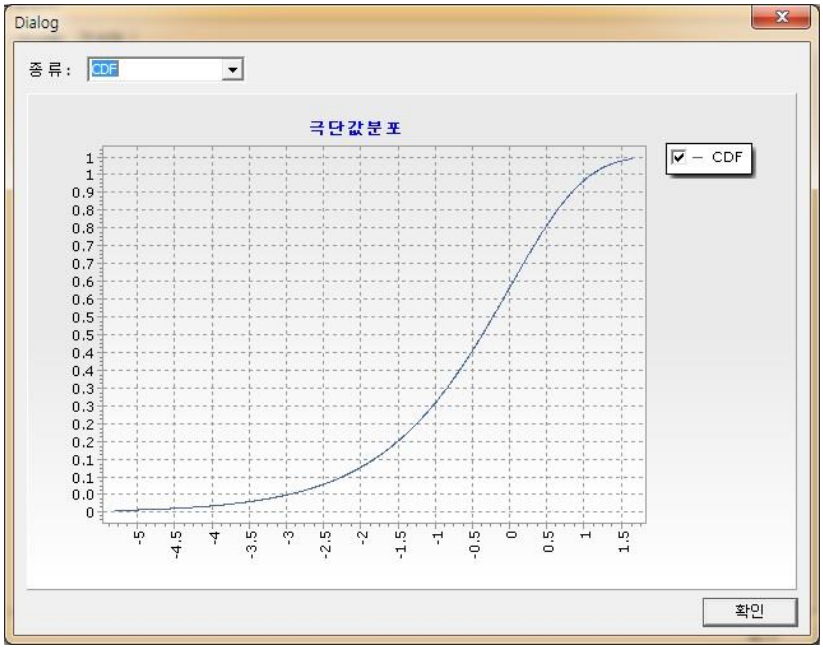
극단값 분포의 누적 분포함수의 역함수

$$x = F^{-1}(p \mid \mu,\sigma) = \mu + \sigma\ln(-\ln(1-p))$$

예시

- $\mu = 0, \sigma = 1$ **일 때의 pdf, cdf 그래프**





- $\mu = 0, \sigma = 1$ 일 때의 평균, 분산. $x = 2$ 일 때의 **pdf, cdf**

- $\mu = 0, \sigma = 1, \mathrm{p} = 0.5$ 일 때 누적 분포 함수의 역함수

7.2.5. 지수 분포 (Exponential distribution)

지수 분포는 연속 확률 분포의 일종으로 포아송 process 에 사용되는 분포입니다. 포아송 process 를 $N(t)$ 라고 할 때 이는 t 시점까지 event 가 일어난 회수로 해석할 수 있습니다. 이 때 하나의 event 가 발생한 시점부터 다음 event 가 발생할 시점까지의 시간이 지수 분포를 따른다고 가정하는 것이 포아송 Process 입니다. 지수 분포는 이뿐 아니라 여러 대기 이론(Queuing Theory)에 많이 활용되는데 이는 Memoryless Property 와 같이 현실 문제에 적합한 특징들을 갖고 있기 때문입니다.

지수 분포의 pdf

$$f(x \mid \mu) = \frac{1}{\mu}\exp(-\frac{x}{\mu})$$

$$x > 0, \mu > 0$$

지수 분포의 cdf

$$F(x \mid \mu) = \int_0^x \frac{1}{\mu}\exp(-\frac{t}{\mu})dt = 1 - \exp(-\frac{x}{\mu})$$

$$x > 0, \mu > 0$$

지수 분포의 평균, 분산

$$E(X) = \int_0^\infty x f(x \mid \mu)dx = \mu$$

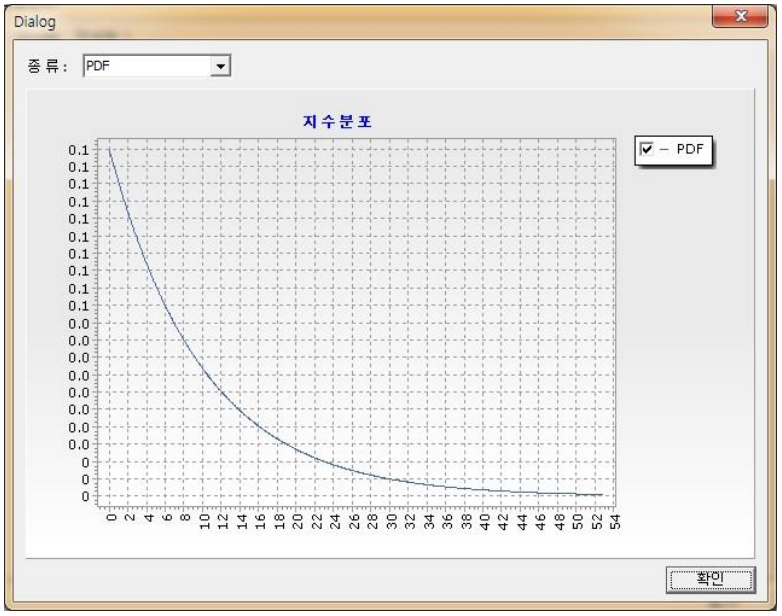
$$Var(X) = E(X^2) - (E(X))^2 = \mu^2$$

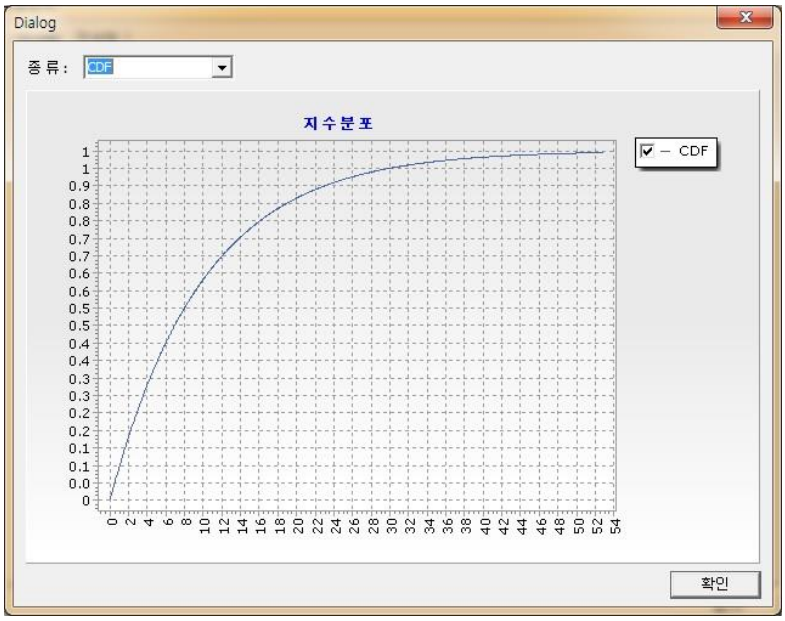
지수 분포 누적 분포 함수의 역함수

$$x = F^{-1}(p \mid \mu) = -\mu \ln(1-p)$$

예시

- $\mu = 10$ **일 때 pdf, cdf 그래프**





- $\mu = 10$ 일 때 평균, 분산. $x = 5$ 일 때 **pdf, cdf**

- $\mu = 10, p = 0.5$ 일 때 누적 분포 함수의 역함수

7.2.6. F 분포 (F – distribution)

F 분포는 연속 확률 분포의 일종으로 통계적 추정 및 검증에서 가장 많이 사용되는 분포 중의 하나입니다. F 분포는 ANOVA(Analysis of Variance : 분산 분석)에서 유의성을 검증하는데 주로 사용되고 있습니다. 이와 함께 Likelihood Ratio Test 에서도 카이 제곱 분포, 정규 분포와 함께 많이 사용됩니다.

F 분포의 확률 변수는 다음과 같이 정의됩니다.

$$F = \frac{U_1 / d_1}{U_2 / d_2}$$

이 때 U_1, U_2 는 서로 독립이고 자유도가 각각 d_1, d_2 인 카이제곱 분포입니다. 이러한 형태가 ANOVA 와 Likelihood Ratio Test 에서 나타나기 때문에 이러한 Test 에서 F 분포가 사용되는 것입니다.

F 분포의 pdf

$$f(x \mid \nu_1, \nu_2) = \frac{\Gamma[(\nu_1 + \nu_2)/2]}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} (\nu_1/\nu_2)^{\nu_1/2} \frac{x^{(\nu_1-2)/2}}{[1+(\nu_1/\nu_2)x]^{(\nu_1+\nu_2)/2}}$$

$$\nu_1, \nu_2 > 0, x > 0$$

F 분포의 cdf

$$F(x \mid \nu_1, \nu_2) = \int_0^x \frac{\Gamma[(\nu_1 + \nu_2)/2]}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} (\nu_1/\nu_2)^{\nu_1/2} \frac{t^{(\nu_1-2)/2}}{[1+(\nu_1/\nu_2)t]^{(\nu_1+\nu_2)/2}} dt$$

$$\nu_1, \nu_2 > 0, x > 0$$

F 분포의 평균, 분산

$$E(X) = \frac{\nu_2}{\nu_2 - 2} \quad \nu_2 > 2$$

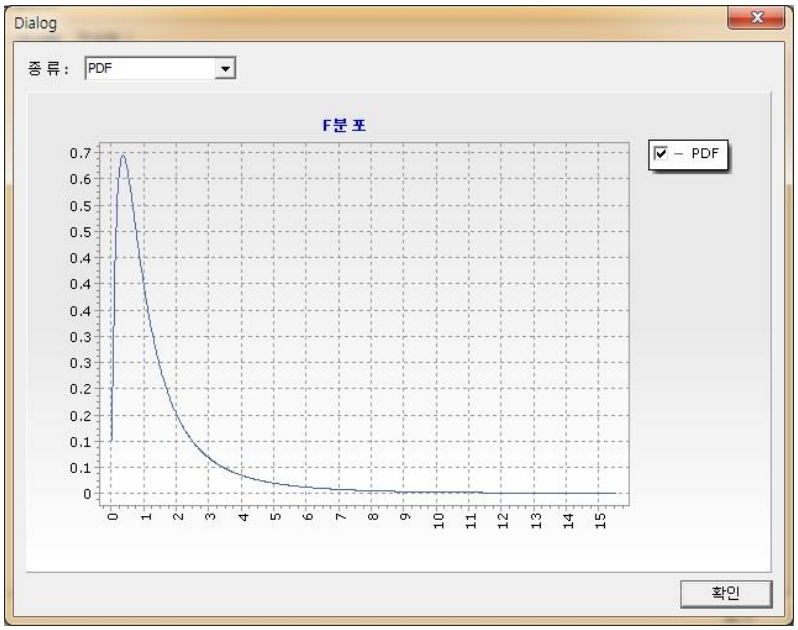
$$Var(X) = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)} \quad \nu_2 > 4$$

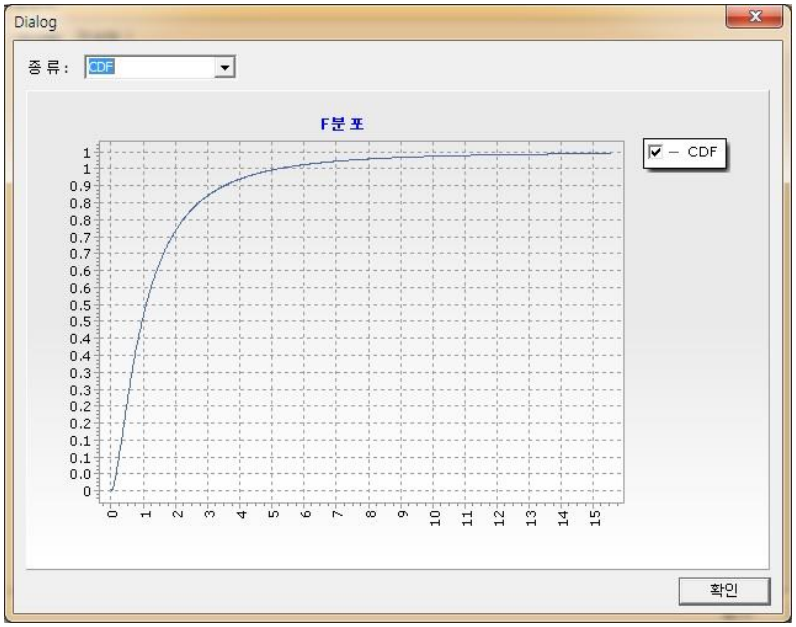
F 분포의 누적 분포 함수의 역함수

$$x = F^{-1}(p \mid \nu_1, \nu_2)$$

예시

- $\nu_1 = 4, \nu_2 = 5$ **일 때 pdf, cdf**





▪ $\nu_1 = 4, \nu_2 = 5$ 일 때 평균, 분산. $x = 3$ 일 때 **pdf, cdf.**

▪ $\nu_1 = 4, \nu_2 = 5, p = 0.5$ 일 때 누적 분포 함수의 역함수

7.2.7 감마 분포 (Gamma distribution)

감마 분포는 두 개의 모수를 가지는 연속 확률 분포 함수의 일종입니다. 감마 분포를 이해하는 가장 좋은 방법은 지수 분포를 활용하는 것입니다. 모수가 λ 인 서로 독립인 지수 분포 확률 변수를 $X_1, X_2, ..., X_n$ 을 더한 다음과 같은 확률 변수 Y 는 감마 분포를 따릅니다.

$$Y = X_1 + X_2 + ... + X_n \rightarrow Y \sim Gamma(n, \lambda)$$

이러한 감마 분포는 한 사람이 사망하기에 이르는 시간과 같은 waiting time 을 모델링 하는데 많이 활용됩니다.

감마 분포의 pdf

$$f(x \mid a,b) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-\frac{x}{b}}$$

$$a,b > 0, x > 0$$

감마 분포의 cdf

$$F(x \mid a,b) = \frac{1}{b^a \Gamma(a)} \int_0^x t^{a-1} e^{-\frac{t}{b}} dt$$

감마 분포의 평균, 분산

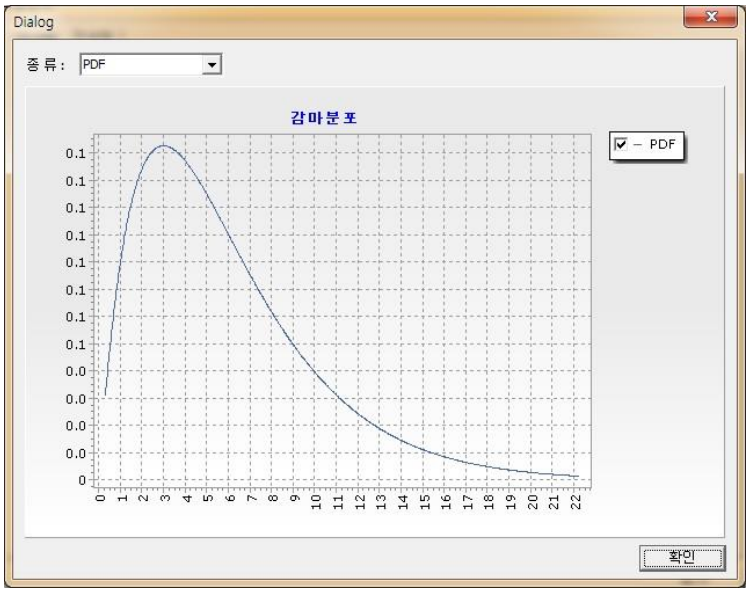
$$E(X) = ab$$
$$Var(X) = ab^2$$

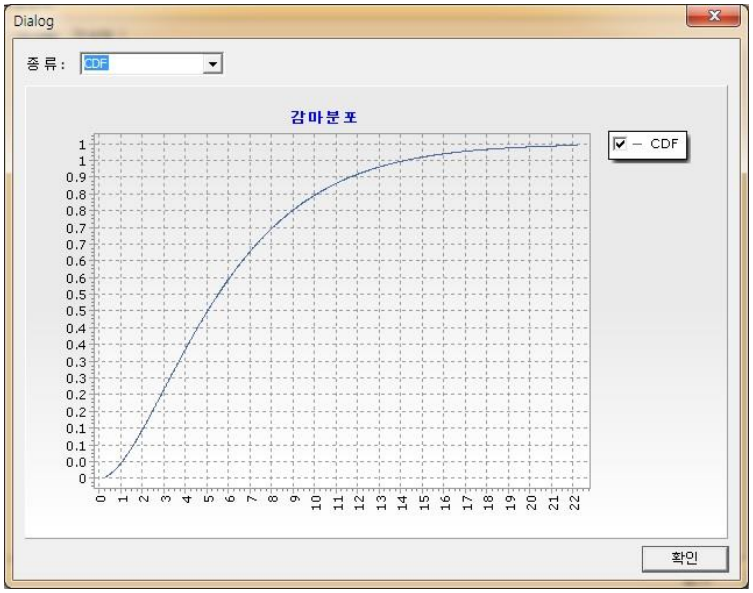
감마 분포의 누적 분포 함수의 역함수

$$x = F^{-1}(p \mid a,b)$$

예시

$a = 2, b = 3$ **일 때 pdf, cdf 의 그래프**





- $a=2, b=3$ 일 때 평균, 분산. $x=4$ 일 때의 **pdf, cdf**

- $a = 2, b = 3, p = 0.5$ 일 때 누적 분포 함수의 역함수

7.2.8 기하 분포 (Geometric distribution)

기하 분포는 이산형 확률 분포의 일종입니다. 기하 분포 확률 변수는 한번의 성공이 일어날 때까지 발생하는 실패 수로 해석할 수 있습니다. 한번의 event 가 성공할 확률을 p, 실패할 확률을 1-p 라고 할 때 기하 분포에 대한 자세한 설명은 아래와 같습니다.

기하 분포의 pmf

$$f(x \mid p) = p(1-p)^x$$
$$x = 0,1,2,...$$

기하 분포의 cdf

$$F(x \mid p) = \sum_{i=0}^{floor(x)} p(1-p)^i$$
$$x = 0,1,2,...$$

기하 분포의 평균, 분산

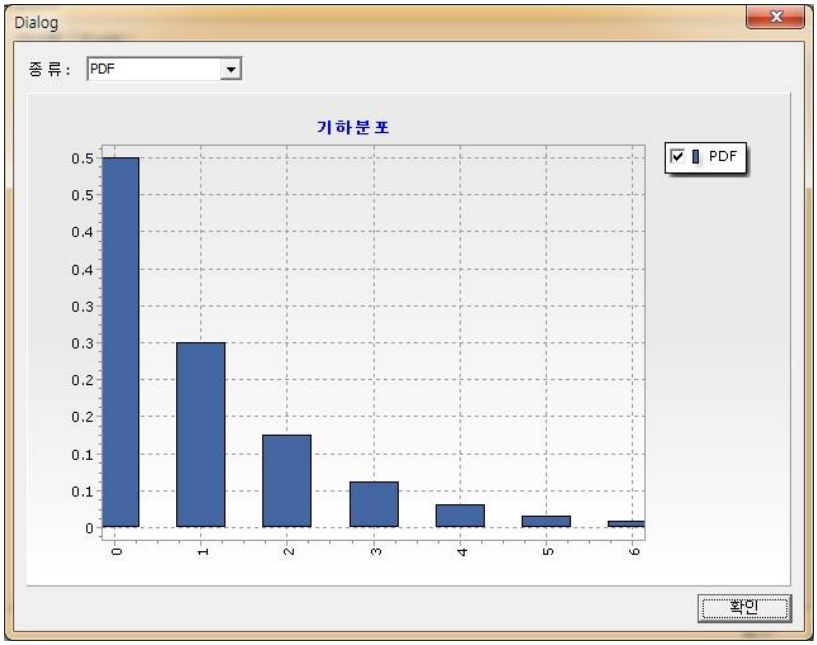
$$E(X) = (1-p)/p$$
$$Var(X) = (1-p)/p^2$$

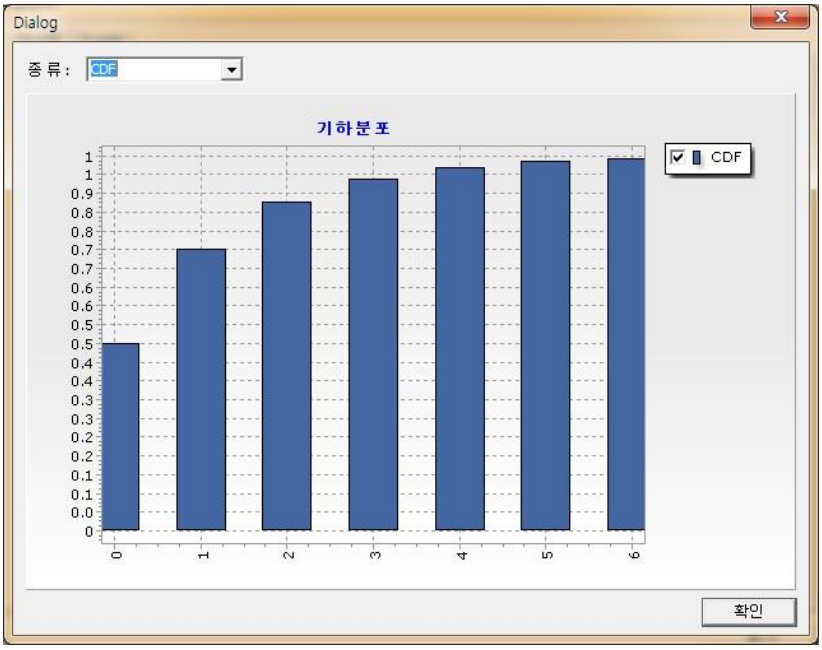
기하 분포의 누적 분포 함수의 역함수

$$x = F^{-1}(y \mid p)$$
$$y : 확률(0 \le y \le 1)$$

예시

- $p = 0.5$ **일 때 pdf, cdf 의 그래프**





- $p = 0.5$ 일 때 평균, 분산. $x = 1$일 때 **pdf, cdf**

- $p = 0.5, y = 0.75$ 일 때 누적 분포 함수의 역함수

7.2.9 일반화 극단값 분포 (Generalized extreme value distribution)

일반화 극단값 분포는 연속 확률 분포의 일종입니다. 이 분포는 Gumbel, Weibull, Frechet 분포와 관련이 있습니다. 분포에 쓰이는 즉 k 의 값에 따라 k=0 일 때 Gumbel, k>0 일 때 Frechet, k<일 때 Weibull 분포라고 합니다. 이를 각각 type 1, type 2, type 3 극단값 분포라고도 부릅니다.

일반화 극단값 분포의 pdf

$$f(x\mid\mu,\sigma,k)=\begin{cases}\frac{1}{\sigma}[1+k(\frac{x-\mu}{\sigma})]^{(-1/k)-1}\exp\{-[1+k(\frac{x-\mu}{\sigma})]^{(-1/k)}\}\text{ if } k\neq 0\\ \frac{1}{\sigma}(\exp(-\frac{x-\mu}{\sigma}))^{k+1}\exp(-\exp(-\frac{x-\mu}{\sigma}))\text{if } k=0\end{cases}$$

$$1+k(x-\mu)/\sigma>0$$

$$x\in[\mu-\sigma/k,+\infty) \text{ if } k>0$$

$$x\in(-\infty,+\infty) \text{ if } k=0$$

$$x\in(-\infty,\mu-\sigma/k] \text{ if } k<0$$

일반화 극단값 분포의 cdf

$$F(x\mid\mu,\sigma,k)=\begin{cases}\exp\{-[1+k(\frac{x-\mu}{\sigma})]^{(-1/k)}\}\text{if } k\neq 0\\ \exp(-\exp(-\frac{x-\mu}{\sigma}))\text{if } k=0\end{cases}$$

$$1+k(x-\mu)/\sigma>0$$

$$x\in[\mu-\sigma/k,+\infty) \text{ if } k>0$$

$$x\in(-\infty,+\infty) \text{ if } k=0$$

$$x\in(-\infty,\mu-\sigma/k] \text{ if } k<0$$

일반화 극단값 분포의 평균, 분산

$$E(X)=\begin{cases}\mu-\dfrac{\sigma}{k}+\dfrac{\sigma}{k}g_1 \text{ if } k\neq 0, k<1\\ \mu+\sigma\gamma \text{ if } k=0\\ \text{not exists if } k\geq 1\end{cases}$$

$$Var(X)=\begin{cases}\dfrac{\sigma^2}{k^2}(g_2-g_1{}^2) \text{ if } k\neq 0, k<\dfrac{1}{2}\\ \sigma^2\dfrac{\pi^2}{6} \text{ if } k=0\\ \text{not exists if } k\geq\dfrac{1}{2}\end{cases}$$

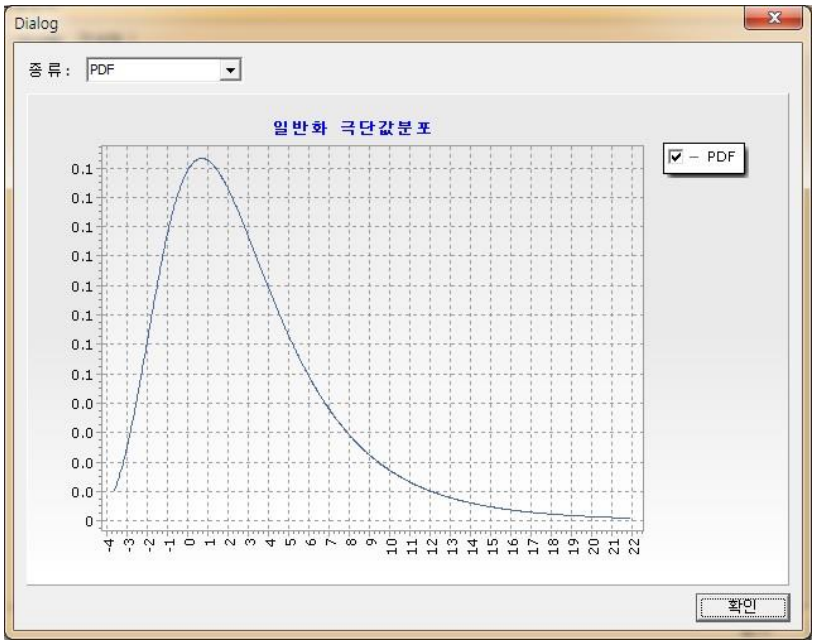
$$g_i=\Gamma(1-ik), k=1,2,3,4$$

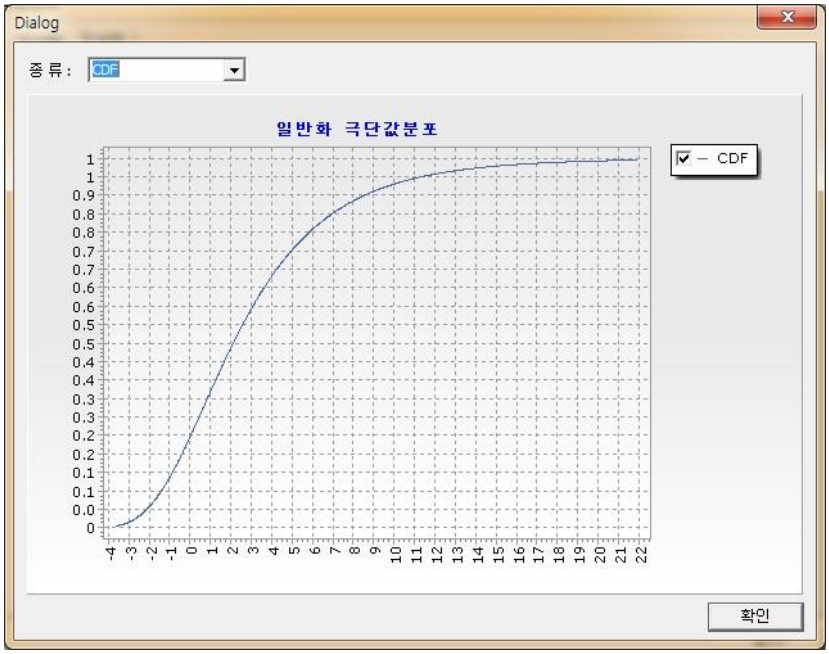
일반화 극단값 분포의 누적 분포 함수의 역함수

$$x=F^{-1}(p\mid\mu,\sigma,k)=\begin{cases}\mu-\sigma\log(\log(\dfrac{1}{p})) \text{ if } k=0\\ \mu+\dfrac{\sigma}{k}((-\log(p))^{-k}-1) \text{ if } k\neq 0\end{cases}$$

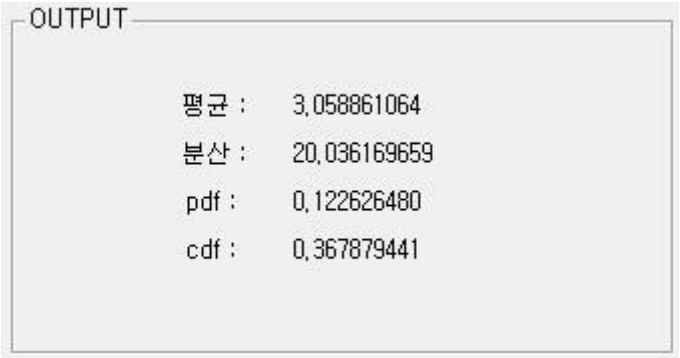
예시

- $\mu=1,\sigma=3,k=0.1$ 일 때 **pdf, cdf**





- $\mu = 1, \sigma = 3, k = 0.1$ 일 때 평균, 분산. $x = 1$ 일 때 **pdf, cdf.**



- $\mu = 1, \sigma = 3, k = 0.1, p = 0.4$ 일 때 누적 분포 함수의 역함수

7.2.10 일반화 파레토 분포 (Generalized pareto distribution)

일반화 파레토 분포는 연속 확률 분포의 일종입니다. 파레토 분포는 파레토 법칙을 설명하는데 사용되는 분포로 파레토 법칙이란 전체 부의 80%는 20%의 사람들에게 집중되어 있다고 설명합니다. 이러한 파레토 법칙을 설명한 파레토 분포를 더욱 일반화 한 것이 일반화 파레토 분포입니다.

일반화 파레토 분포 pdf

$$f(x\,|\,k,\sigma,\theta)=\begin{cases}\frac{1}{\sigma}(1+k\frac{(x-\theta)}{\sigma})^{-1-1/k} & \text{if } k>0, x>\theta. \text{if } k<0, \theta<x<\theta-\frac{\sigma}{k}\\ \frac{1}{\sigma}\exp(-\frac{(x-\theta)}{\sigma}) & \text{if } k=0. x>\theta\end{cases}$$

일반화 파레토 분포 cdf

$$F(x\,|\,k,\sigma,\theta)=\begin{cases}1-\left(1+k\left(\frac{x-\theta}{\sigma}\right)\right)^{-1/k} & \text{if } k>0, x>\theta. \text{if } k<0, \theta<x<\theta-\frac{\sigma}{k}\\ 1-\exp\left(-\frac{x-\theta}{\sigma}\right) & \text{if } k=0. x>\theta\end{cases}$$

일반화 파레토 분포 평균, 분산

$$E(X)=\begin{cases}\theta+\frac{\sigma}{1-k} & \text{if } k<1 \text{ and } k\neq 0\\ \text{not defined} & \text{if } k\geq 1\\ \theta+\sigma & \text{if } k=0\end{cases}$$

$$V(X)=\begin{cases}\frac{\sigma^2}{(1-k)^2(1-2k)} & \text{if } k<1/2 \text{ and } k\neq 0\\ \text{not defined} & \text{if } k\geq 1/2\\ \sigma^2 & \text{if } k=0\end{cases}$$

일반화 파레토 분포의 누적 분포 함수의 역함수

$k=0$ 일 때

$$x = F^{-1}(p \mid k,\sigma,\theta) = \begin{cases} \theta & \text{if } p = 0 \\ \infty & \text{if } p = 1 \\ \theta - \sigma \ln(1-p) & \text{otherwise} \end{cases}$$

$k > 0$ 일 때

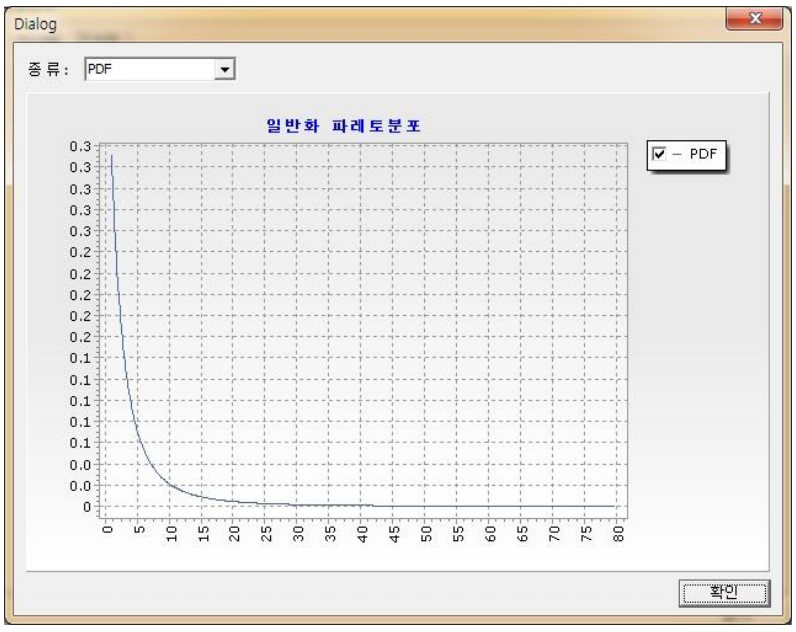
$$x = F^{-1}(p \mid k,\sigma,\theta) = \begin{cases} \theta & \text{if } p = 0 \\ \infty & \text{if } p = 1 \\ \theta + \frac{\sigma}{k}((1-p)^{-k} - 1) & \text{otherwise} \end{cases}$$

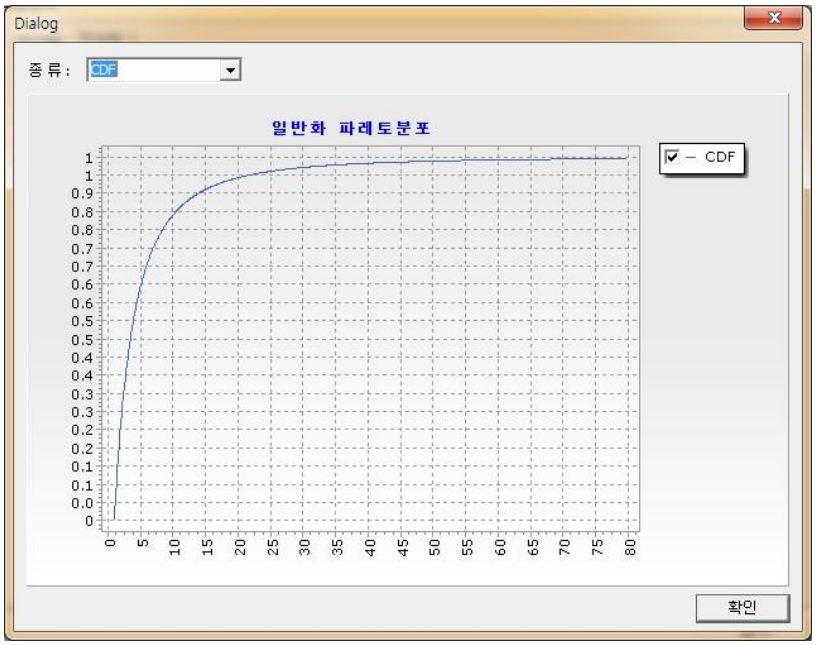
$k < 0$ 일 때

$$x = F^{-1}(p \mid k,\sigma,\theta) = \begin{cases} \theta & \text{if } p = 0 \\ \theta - \frac{\sigma}{k} & \text{if } p = 1 \\ \theta + \frac{\sigma}{k}((1-p)^{-k} - 1) & \text{otherwise} \end{cases}$$

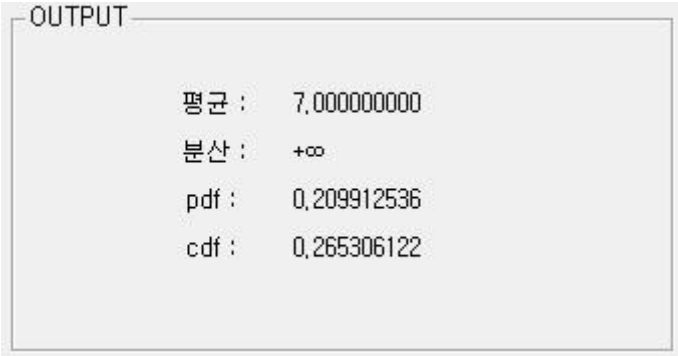
예시

- $k = 0.5, \sigma = 3, \theta = 1$일 때 **pdf, cdf.**

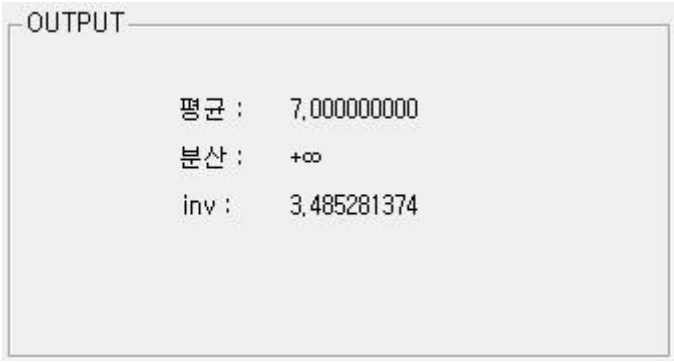




- $k=0.5, \sigma=3, \theta=1$ 평균, 분산. $x=2$ 일 때 **pdf, cdf** 값.

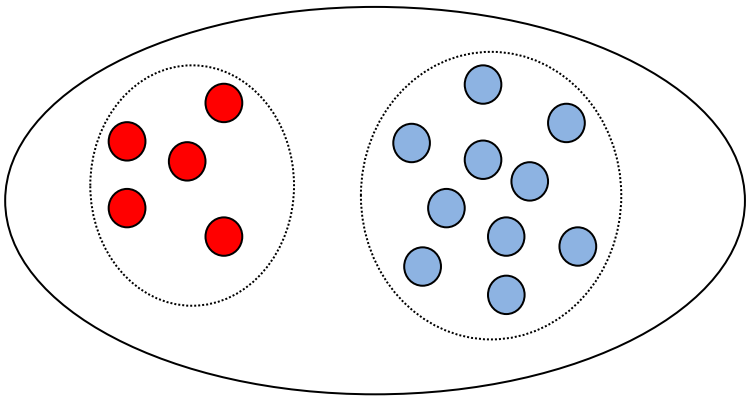


- $k=0.5, \sigma=3, \theta=1, p=0.5$ 일 때 누적 분포 함수의 역함수



7.2.11 초기하 분포 (Hypergeometric distribution)

초기하 분포는 이산 확률 분포의 일종입니다. 이에 대해 설명하기 위해서 다음과 같은 그림을 생각하는 것이 유용합니다.



위와 같이 큰 주머니에 M 개의 공이 들어있다고 가정합니다. 빨간색 공(성공) K 개와 파란색 공(실패) M-K 개가 섞여 있을 때, N 개의 공을 뽑을 때 그 중 빨간색 공의 개수가 바로 초기하 분포의 확률 변수라고 할 수 있습니다. 이항 분포를 복원 추출로 공을 뽑는 것으로 생각하고, 초기하 분포를 비복원 추출로 공을 뽑는 것으로 생각할 때 이 두 분포는 유사한 분포라고 할 수 있습니다.

초기하 분포의 pmf

$$f(x|M,K,N)=\frac{\binom{K}{x}\binom{M-K}{N-x}}{\binom{M}{N}}$$

초기하 분포의 cdf

$$F(x|M,K,N)=\sum_{i=0}^{floor(x)}\frac{\binom{K}{i}\binom{M-K}{N-i}}{\binom{M}{N}}$$

초기하 분포의 평균, 분산

$$E(X)=\frac{NK}{M}$$

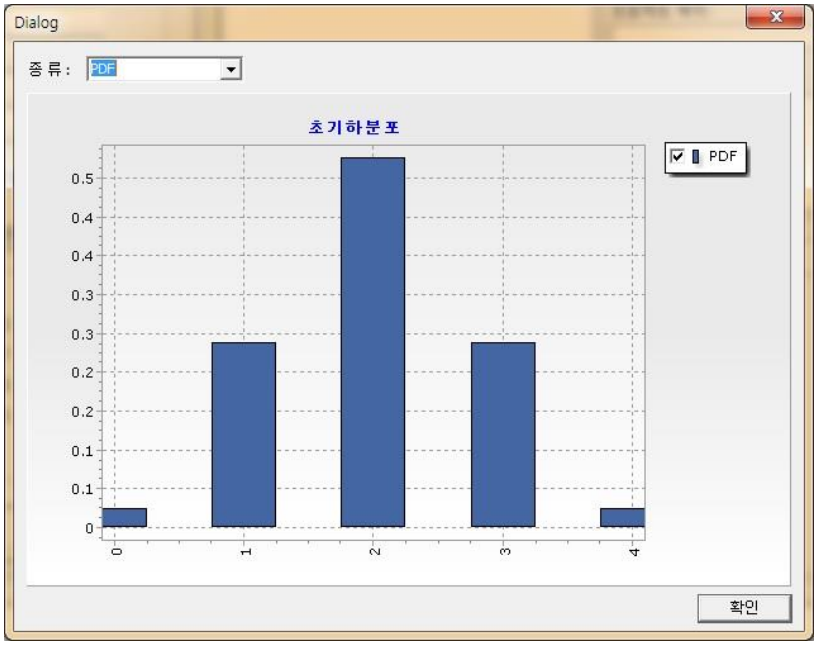
$$Var(X)=\frac{NK(M-K)(M-N)}{M^2(M-1)}$$

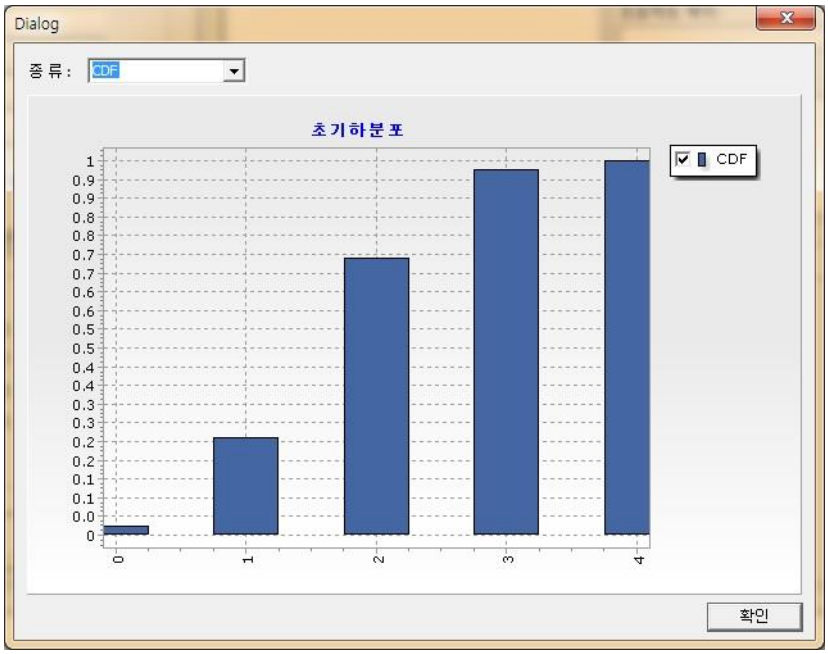
초기하 분포의 누적 확률 분포의 역함수

$$x = F^{-1}(p \mid M, K, N)$$

예시

$m = 10, k = 5, n = 4$ 일 때 pmf, cdf 의 그래프





$m = 10, k = 5, n = 4$ 일 때의 평균, 분산. $x = 3$ 일 때의 pmf, cdf

$m=10, k=5, n=4, p=0.5$ 일 때 누적 분포 함수의 역함수

7.2.12 로그 정규 분포 (Lognormal distribution)

로그 정규 분포는 연속 확률 분포의 일종으로 정규 분포에서 파생된 분포라고 할 수 있습니다. 정규 분포 확률 변수를 Y 라고 할 때,

$$X = \exp(Y)$$

는 로그 정규 분포를 갖습니다. 이러한 로그 정규 분포는 특히 Mathematical Finance 에서 많이 사용합니다. Brownian Motion 을 따르는 확률 과정(Stochastic Process)은 특정 시점에서 로그 정규 분포를 따릅니다. Mathematical Finance 에서는 기초 자산이 Brownian Motion 을 따른다는 가정을 많이 하기 때문에 이를 기초로 하여 만든 금융 파생 상품의 가격을 책정하는데 널리 사용되고 있습니다.

로그 정규 분포의 pdf

$$f(x \mid \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right)$$

로그 정규 분포의 cdf

$$F(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^x \frac{\exp\left(-\frac{(\ln(t) - \mu)^2}{2\sigma^2}\right)}{t} dt$$

로그 정규 분포의 평균, 분산

$$E(X) = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

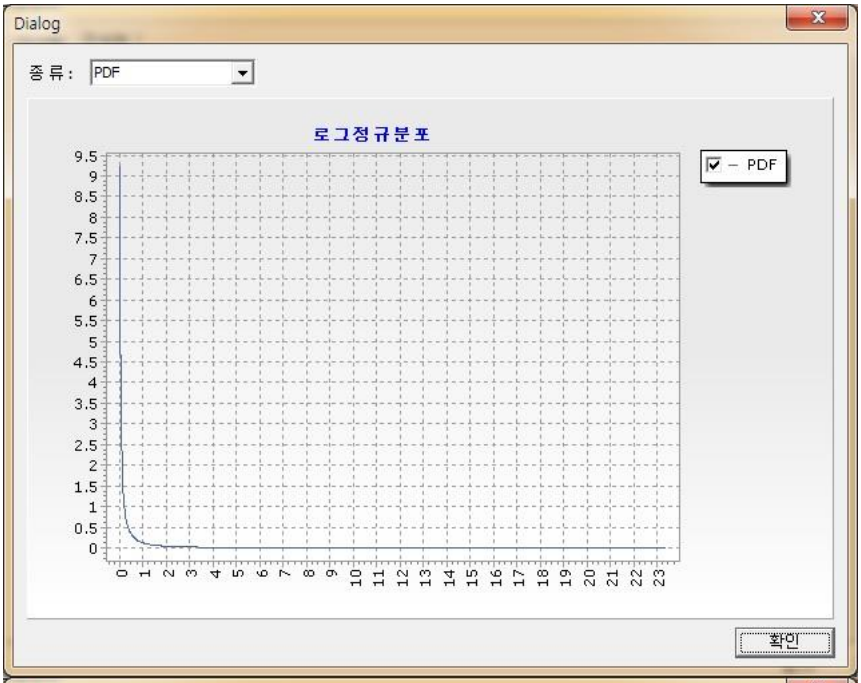
$$Var(X) = \exp(2\mu + \frac{\sigma^2}{2})(\exp(\sigma^2) - 1)$$

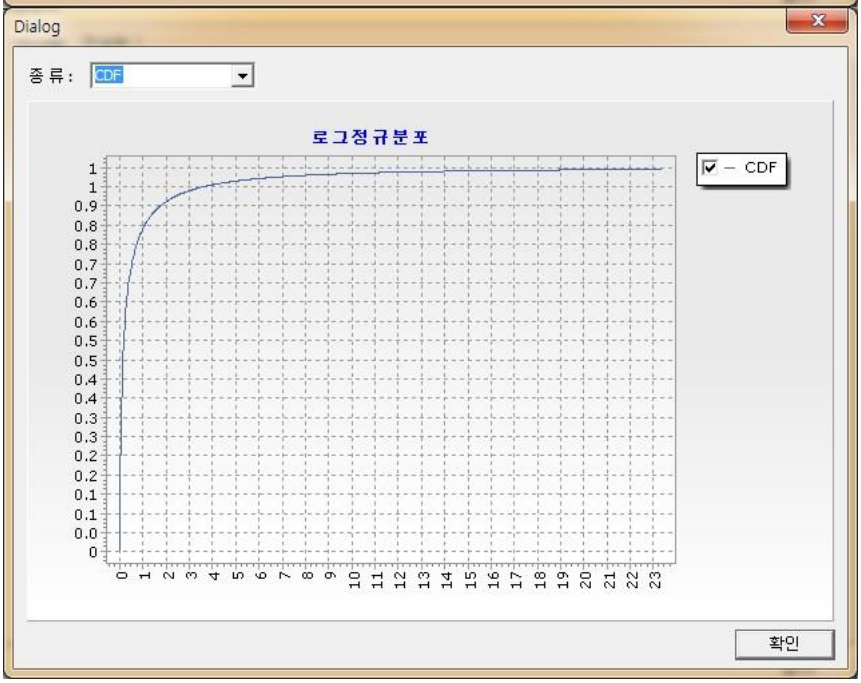
로그 정규 분포의 누적 분포 함수의 역함수

$$x = F^{-1}(p \mid \mu, \sigma)$$

예시

$\mu = -2, \sigma = 2$ 일 때, pdf, cdf 의 그래프





$\mu = -2, \sigma = 2$ 일 때, 평균, 분산. $x = 1$ 일 때 pdf, cdf.

$\mu = -2, \sigma = 2, p = 0.5$ 일 때 누적 분포 함수의 역함수

7.2.13 음이항 분포 (Negative binomial distribution)

음이항 분포는 이산 확률 분포의 일종입니다. 각 시행이 베르누이 분포를 따를 때 r 번의 성공이 일어나기 위해 수반되어야 하는 실패 횟수를 음이항 분포 확률 변수라고 할 수 있습니다. 이러한 음이항 분포는 r=1 일 때(즉 성공이 1 번일 때)는 기하 분포가 되는 것으로 보아 기하 분포의 일반적인 형태라고 할 수 있고, 음이항 분포의 확률 변수는 서로 독립인 기하분포의 확률 변수를 r 개 더함으로써 만들어 질 수 있습니다.

음이항 분포의 pmf

$$f(x \mid r, p) = \binom{r+x-1}{x} p^r (1-p)^x$$

음이항 분포의 cdf

$$F(x \mid r, p) = \sum_{i=0}^{floor(x)} \binom{r+i-1}{i} p^r (1-p)^i$$

음이항 분포의 평균, 분산

$$E(X) = \frac{r(1-p)}{p}$$

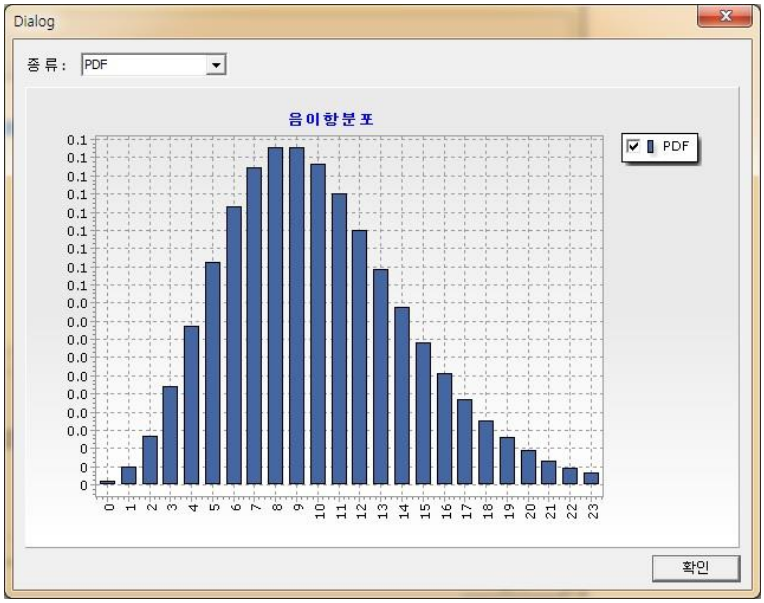
$$Var(X) = \frac{r(1-p)}{p^2}$$

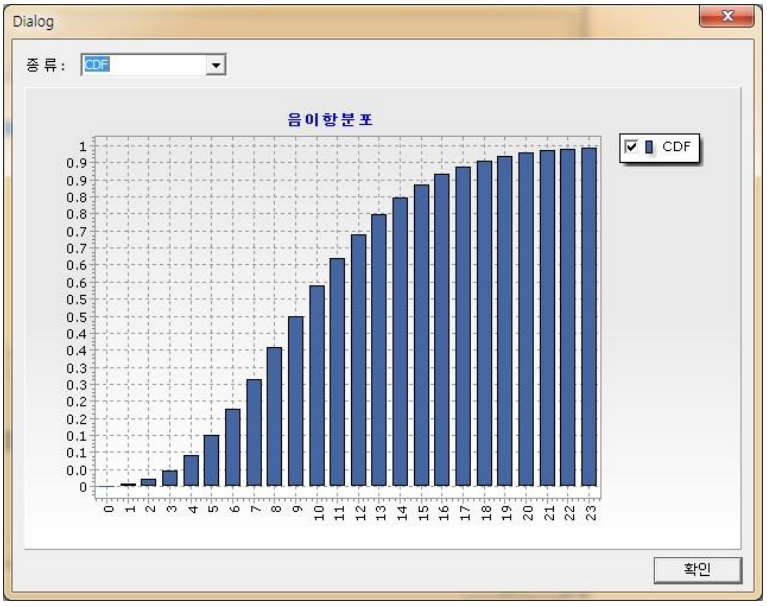
음이항 분포의 누적 분포 함수의 역함수

$$x = F^{-1}(y \mid r, p)(0 \le y \le 1)$$

예시

$r = 10, p = 0.5$ 일 때 pdf, cdf 그래프





$r=10, p=0.5$ 일 때 평균, 분산. $x=10$ 일 때 pdf, cdf.

OUTPUT

평균 : 10,000000000
분산 : 20,000000000
pdf : 0,088098526
cdf : 0,588098526

$r=10, p=0.5, y=0.5$ 일 때 누적 분포 함수의 역함수

OUTPUT

평균 : 10,000000000
분산 : 20,000000000
inv : 9,000000000

7.2.14 비중심 f 분포 (Non-central F-distribution)

만약 X 가 noncentral parameter λ 와 자유도 ν_1 을 갖는 비중심 카이 제곱 확률 변수이고, Y 가 ν_2 의 자유도를 가질 때

$$F = \frac{X/\nu_1}{Y/\nu_2}$$

를 비중심 f 분포를 갖는 확률 변수라고 합니다.

비중심 f 분포의 pdf

$$f(x \mid \lambda, \nu_1, \nu_2) = \sum_{k=0}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^k}{B(\nu_2/2, \nu_1/2+k)k!} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}+k} \left(\frac{\nu_2}{\nu_2+\nu_1 x}\right)^{\frac{\nu_1+\nu_2}{2}+k} x^{\frac{\nu_1}{2}-1+k}$$

$$x \geq 0, \nu_1, \nu_2 > 0$$

비중심 f 분포의 cdf

$$F(x \mid \lambda, \nu_1, \nu_2) = \int_0^x f(t \mid \lambda, \nu_1, \nu_2) dt$$

비중심 f 분포의 평균, 분산

$$E(F) = \begin{cases} \dfrac{\nu_2(\nu_1+\lambda)}{\nu_1(\nu_2-2)} & \nu_2 > 2 \\ \text{not exist} & \nu_2 \leq 2 \end{cases}$$

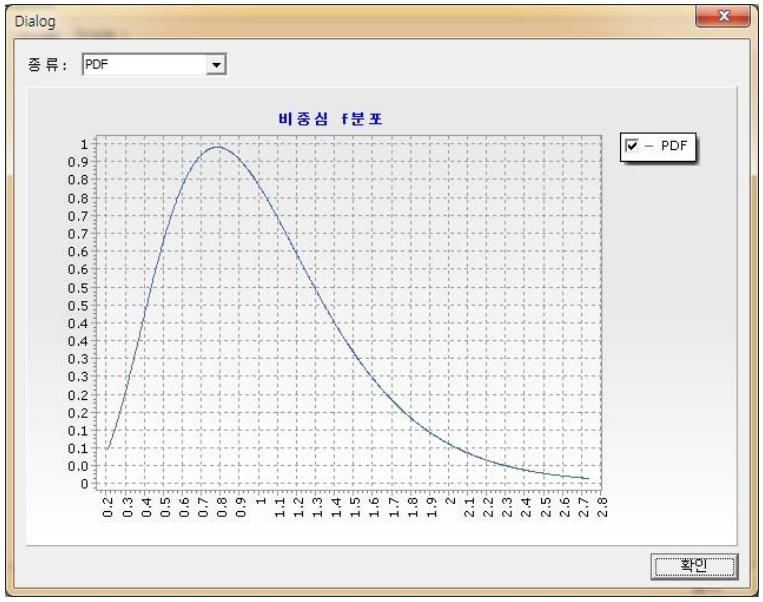
$$Var(F) = \begin{cases} 2\dfrac{(\nu_1+\lambda)^2 + (\nu_1+2\lambda)(\nu_2-2)}{(\nu_2-2)^2(\nu_2-4)} \left(\dfrac{\nu_2}{\nu_1}\right)^2 & \nu_2 > 4 \\ \text{not exist} & \nu_2 \leq 4 \end{cases}$$

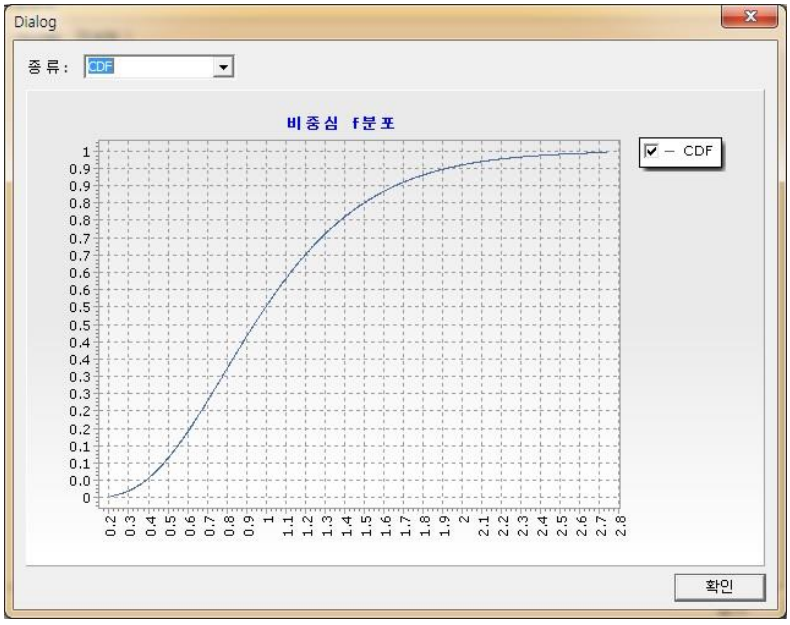
비중심 f 분포의 누적 분포 함수의 역함수

$$x = F^{-1}(p \mid \lambda, \nu_1, \nu_2)$$

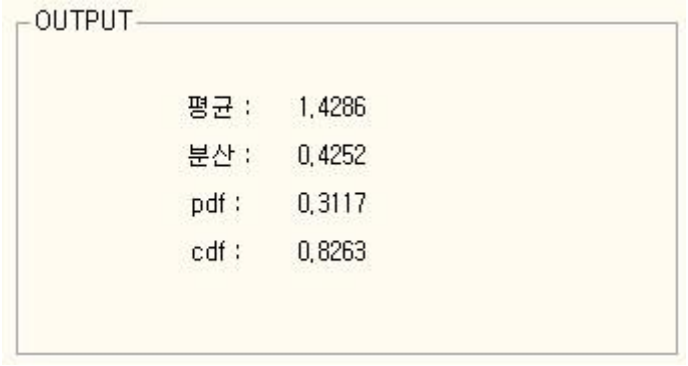
예시

$\nu_1 = 10, \nu_2 = 100$ 일 때 pdf, cdf 그래프





$\nu_1 = 10, \nu_2 = 100, \delta = 4$ 일 때 평균, 분산. $x = 2$ 일 때의 pdf, cdf



$\nu_1 = 10, \nu_2 = 100, \delta = 4, \mathrm{p} = 0.5$ 일 때 누적 분포 함수의 역함수

7.2.15 비중심 t 분포 (Non-central T-distribution)

Z 라는 확률 변수가 표준 정규 확률 변수라고 하고, V 가 자유도 ν 를 갖는 카이 제곱 분포일 때 새롭게 만드는 다음과 같은 T 라는 확률 변수는

$$T = \frac{Z + \mu}{\sqrt{V/\nu}}$$

비중심 t 분포를 갖습니다.

비중심 t 분포의 pdf

$$f(x|\nu,\mu) = \frac{\nu^{\nu/2}\exp\left\{-\frac{\nu\mu^2}{2(x^2+\nu)}\right\}}{\sqrt{\pi}\Gamma(\nu/2)2^{(\nu-1)/2}(x^2+\nu)^{(\nu+1)/2}}\int_0^\infty y^\nu \exp\left\{-\frac{1}{2}\left(y-\frac{\mu x}{\sqrt{x^2+\nu}}\right)^2\right\}$$

비중심 t 분포 cdf

$$F_{\nu,\mu}(x) = \begin{cases} \tilde{F}_{\nu,\mu}(x) & \text{if } x \geq 0. \\ 1 - \tilde{F}_{\nu,-\mu}(-x) & \text{if } x < 0. \end{cases}$$

$$\tilde{F}_{\nu,\mu}(x) = \Phi(-\mu) + \frac{1}{2}\sum_{j=0}^{\infty}\left[p_j I_j\left(j+\frac{1}{2}, \frac{\nu}{2} \right) + q_j I_j\left(j+1, \frac{\nu}{2} \right) \right]$$

$I_y(a,b)$ 은 regularized incomplete beta function

$$y = \frac{x^2}{x^2+\nu}$$

$$p_j = \frac{1}{j!}\exp\left\{ -\frac{\mu^2}{2} \right\}\left(\frac{\mu^2}{2} \right)^j$$

$$q_j = \frac{\mu}{\sqrt{2}\Gamma(j+3/2)}\exp\left\{ -\frac{\mu^2}{2} \right\}\left(\frac{\mu^2}{2} \right)^j$$

비중심 t 분포의 평균, 분산

$$E[T] = \begin{cases} \mu\sqrt{\frac{\nu}{2}}\,\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}, & \text{if } \nu > 1 \\ \text{not exist} \quad , & \text{if } \nu \leq 1 \end{cases}$$

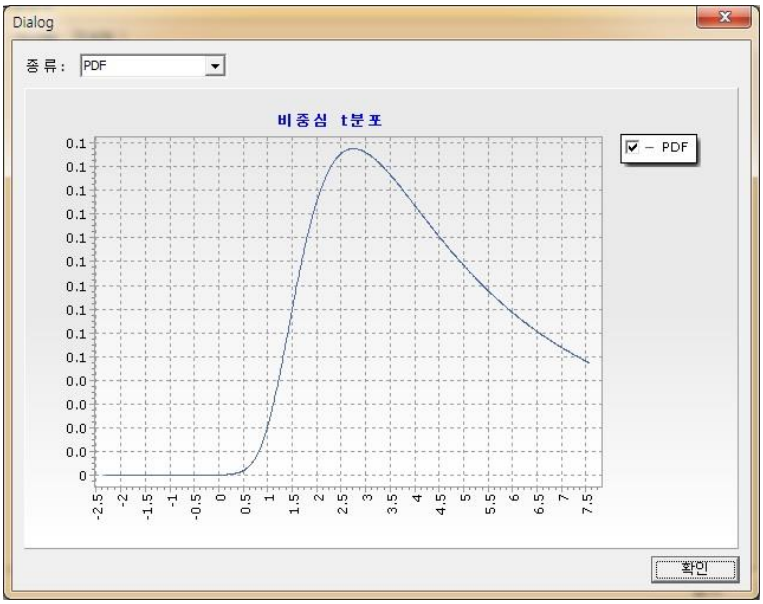
$$Var[T] = \begin{cases} \frac{\nu(1+\mu^2)}{\nu-2} - \frac{\mu^2\nu}{2}\left(\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)} \right)^2 & \text{if } \nu > 2 \\ \text{not exist} & \text{if } \nu \leq 2 \end{cases}$$

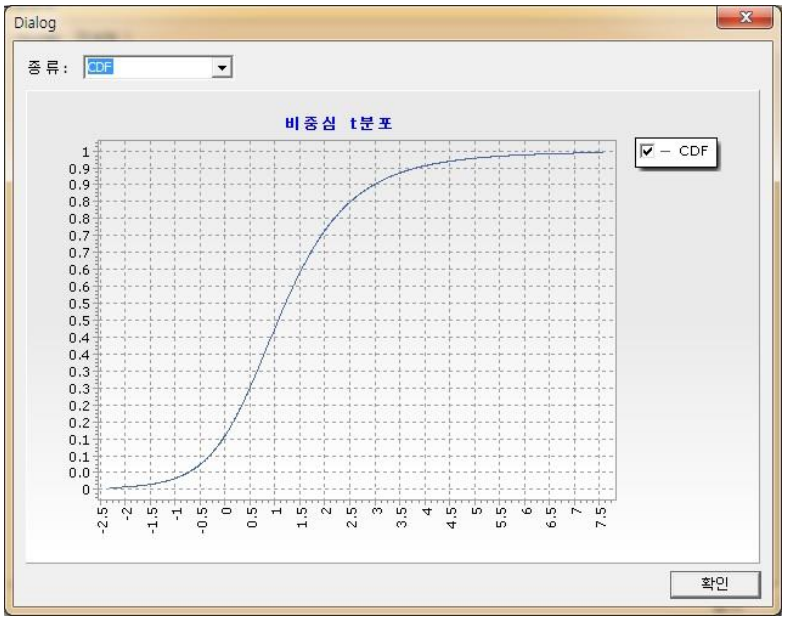
비중심 t 분포의 누적 분포 함수의 역함수

$$x = F^{-1}(p \mid \nu, \mu)$$

예시

$\nu = 4, \delta = 1$ 일 때 pdf, cdf 의 그래프





$\nu = 4, \delta = 1$ 일 때 평균, 분산. $x = 2$ 일 때의 pdf, cdf.

$\nu = 4, \delta = 1, p = 0.5$ 일 때 누적 분포 함수의 역함수

7.2.16 비중심 카이제곱 분포 (Non-central Chi-squared distribution)

$X_i \sim N(\mu_i, {\sigma_i}^2)$ iid 일 때, $\sum_{i=1}^{k}\left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$ 는 자유도가 k 인 카이제곱 분포를 따르는 확률변수입니다. 하지만 다음과 같이 변형을 하면 $\sum_{i=1}^{k}\left(\frac{X_i}{\sigma_i}\right)^2$ 비중심 카이제곱 분포를 따른다고 합니다.

일반적으로 비중심 카이제곱 분포에서 사용되는 parameter 는 λ, k 입니다.

비중심 카이 제곱 분포의 pdf

$$f(x \mid k, \lambda) = \frac{1}{2} e^{-\frac{x+\lambda}{2}} \left(\frac{x}{\lambda}\right)^{k/4-1/2} I_{k/2-1}(\sqrt{\lambda x})$$

$$k > 0, \lambda > 0, x \geq 0$$

$I_{,}$ 는 Bessel function

비중심 카이 제곱 분포의 cdf

$$F(x \mid k, \lambda) = \sum_{j=0}^{\infty} e^{-\lambda/2} \frac{(\lambda/2)^j}{j!} \frac{\gamma(j + k/2, x/2)}{\Gamma(j + k/2)}$$

$$k > 0, \lambda > 0, x \geq 0$$

$\gamma(,)$ 는 incomplete gamma function

비중심 카이 제곱 분포의 mgf

Mgf(moment generating function)은 Moment 를 구하는데 매우 유용한 함수입니다. 비중심 카이 제곱 분포의 mgf 는 다음과 같습니다.

$$E(e^{tX}) = \frac{\exp\left(\frac{\lambda t}{1-2t}\right)}{(1-2t)^{k/2}}$$

비중심 카이 제곱 분포의 평균, 분산

$$E(X) = k + \lambda$$

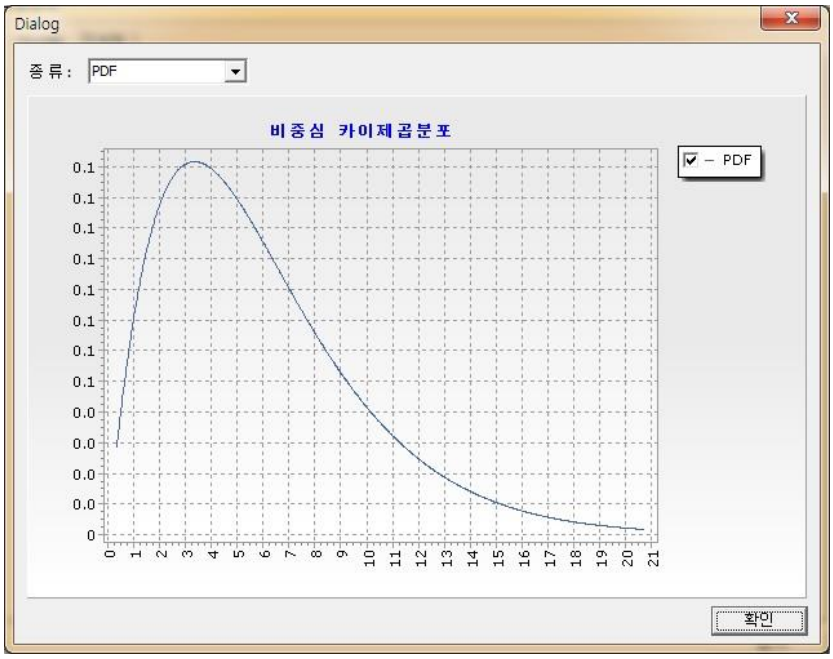
$$Var(X) = 2(k + 2\lambda)$$

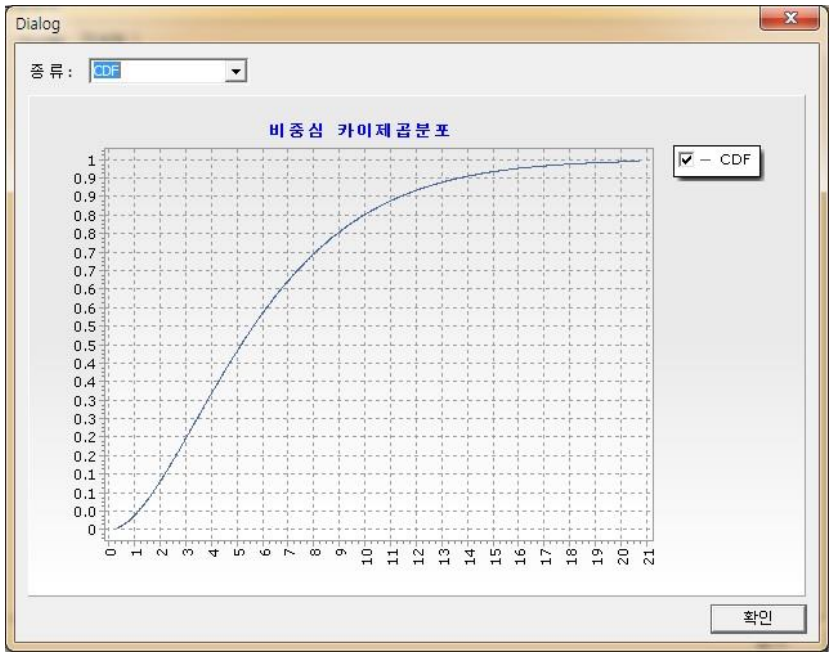
비중심 카이 제곱 분포의 누적 분포 함수의 역함수

$$x = F^{-1}(p \mid k, \lambda)$$

예시

$\nu = 4, \delta = 2$ 일 때 pdf, cdf 의 그래프





$\nu = 4, \delta = 2$ 일 때 평균, 분산. $x = 3$ 일 때의 pdf, cdf.

$\nu = 4, \delta = 2, p = 0.5$ 일 때 누적 분포 함수의 역함수

7.2.17 정규 분포 (Normal distribution)

정규 분포는 모든 분포 중에서 가장 많이 사용하는 분포입니다. 특히 Central Limit Theorem 에 의해서 $X_1, X_2, ..., X_n$ 가 iid 이고 평균이 μ, 분산이 σ^2 일 때

$$\sqrt{n}(\overline{X} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

가 성립합니다. 이 뿐 아니라 Maximum Likelihood Estimator 의 Asymptotic Normality 등 여러 부분에서 정규 분포는 유용하게 사용되고 있습니다.

정규 분포의 pdf

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{\sigma^2} \right)$$

$$-\infty < x < \infty$$

$$\sigma > 0$$

정규 분포의 cdf

$$F(x \mid \mu,\sigma) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t-\mu)^2}{\sigma^2}\right) dt = \frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right]$$

$$-\infty < x < \infty$$

$$\sigma > 0$$

$erf()$ 는 error function

정규 분포의 평균, 분산

$$E(X) = \mu$$

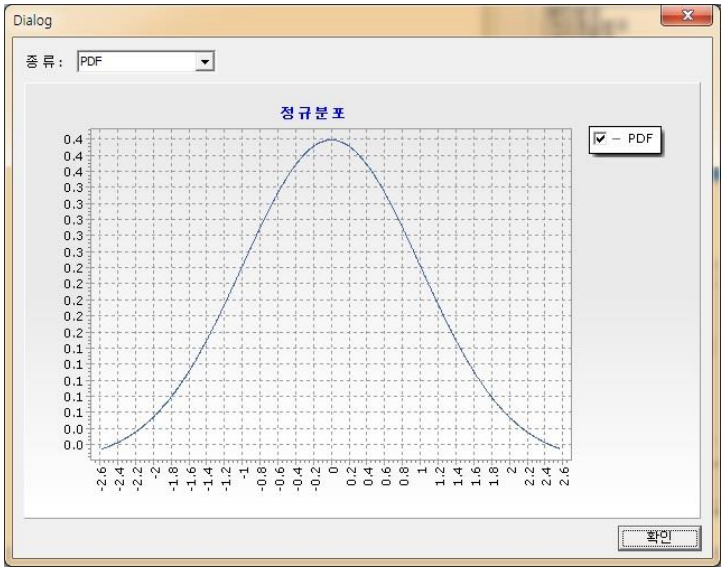
$$Var(X) = \sigma^2$$

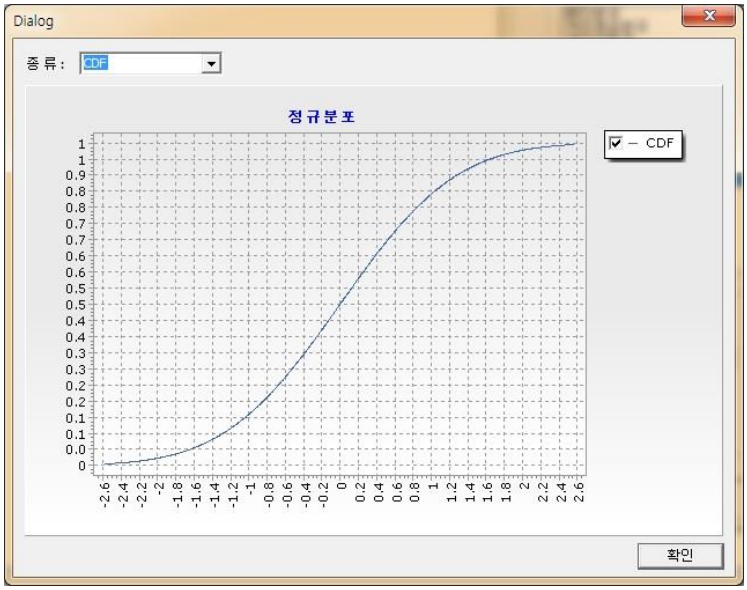
정규 분포의 누적 분포 함수의 역함수

$$x = F^{-1}(p \mid \mu,\sigma)$$

예시

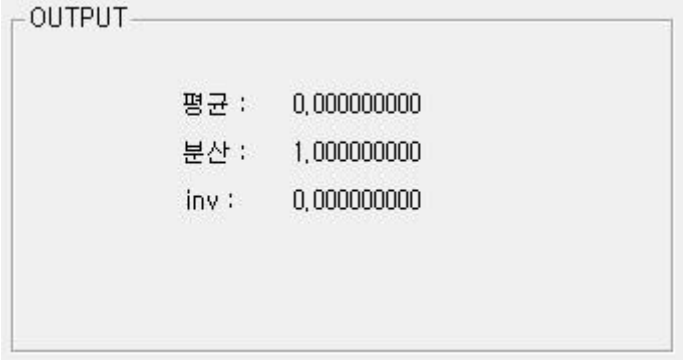
- $\mu = 0, \sigma = 1$일 때 pdf, cdf 의 그래프





- $\mu = 0, \sigma = 1$일 때 평균, 분산. $x = 0.5$일 때 pdf, cdf.

- $\mu = 0, \sigma = 1, p = 0.5$ 일 때 누적 분포 함수의 역함수



7.2.18 포아송 분포 (Poisson distribution)

포아송 분포는 이산 확률 분포의 일종으로 대기 이론(Queing Theory)에 많이 등장하는 분포입니다. event 가 발생하는 시간 간격이 서로 독립인 모수 λ 를 갖는 지수 분포 확률 변수라고 할 때 포아송 process $N(t)$ 는 다음과 같은 특징을 갖습니다.

$$P(N(t+\tau)-N(t)=k)=\frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}$$

이는 모수가 $\lambda\tau$ 인 포아송 분포입니다. 이와 같이 포아송 분포는 여러 확률 모델을 만드는데 유용하게 사용되고 있습니다.

포아송 분포의 pmf

$$f(x\mid\lambda)=\frac{\lambda^x}{x!}\exp(-\lambda)$$

$$x=0,1,2,\ldots \quad \lambda>0$$

포아송 분포의 cdf

$$F(x\mid\lambda)=\exp(-\lambda)\sum_{i=0}^{floor(x)}\frac{\lambda^i}{i!}$$

포아송 분포의 mgf

$$E(e^{tX})=\exp(\lambda(e^t-1))$$

포아송 분포의 평균, 분산

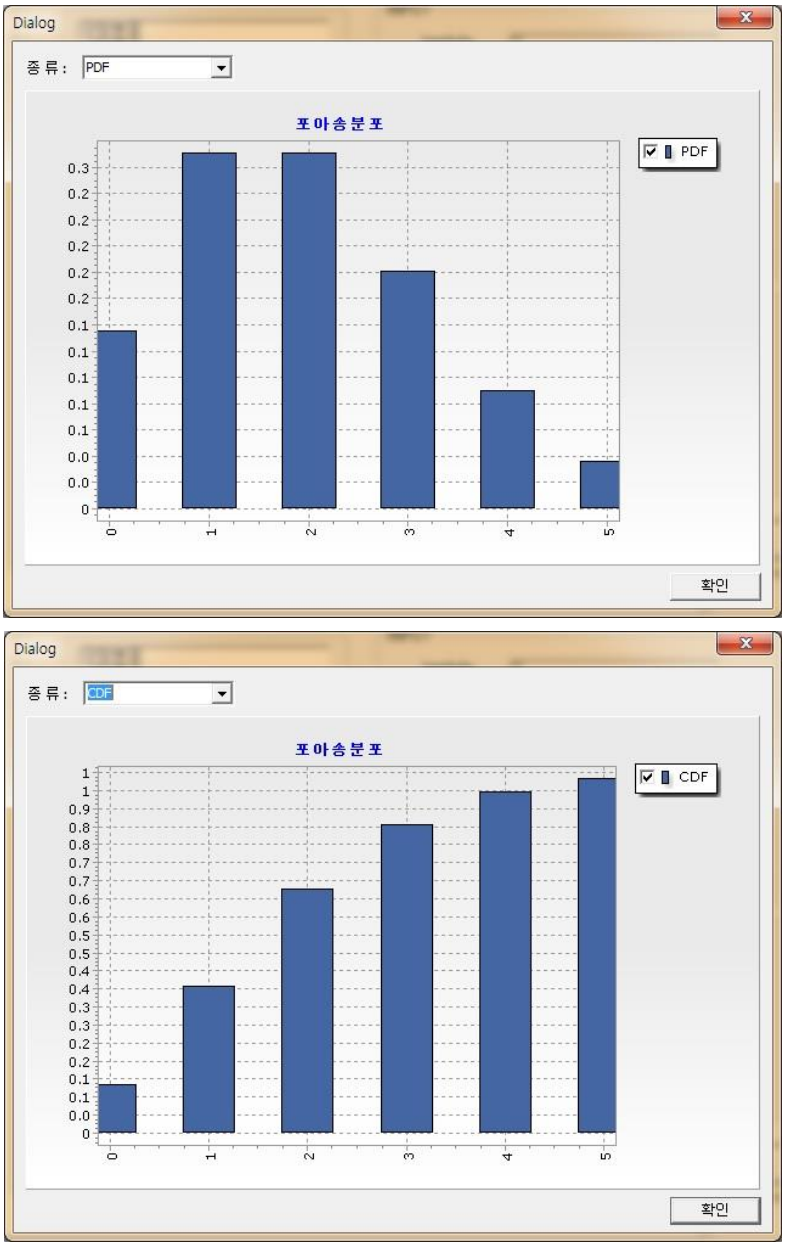
$$E(X)=\lambda$$
$$Var(X)=\lambda$$

포아송 분포의 누적 분포 함수의 역함수

$$x=F^{-1}(p\mid\lambda)$$

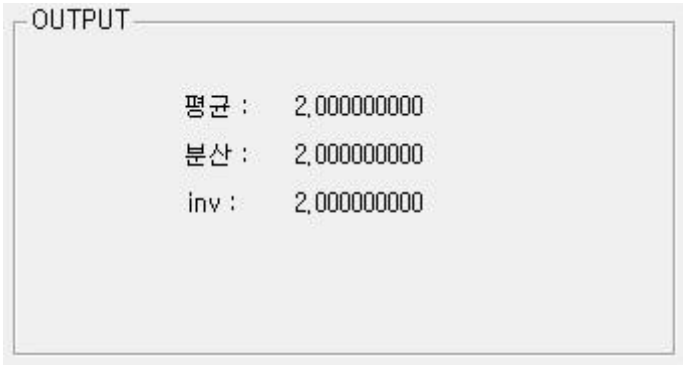
예시

- $\lambda=2$ 일 때 pdf, cdf



- $\lambda = 2$ 일 때 평균, 분산. $x = 1$ 일 때 pdf, cdf.

▪ $\lambda = 2, p = 0.5$ 일 때 누적 분포 함수의 역함수



7.2.19 레일리 분포 (Rayleigh distribution)

레일리 분포는 연속 확률 분포의 일종입니다. 레일리 분포는 오른쪽으로 꼬리가 긴 분포로 특히 정규 분포와 밀접한 관계가 있습니다.
어떠한 물체가 (0,0)을 기준으로 (X, Y) 의 위치에 있다고 합니다. 그리고 이 때 그 물체의 위치는 다음과 같은 distribution 에 의해서 결정된다고 합니다.

$$X \sim N(0,\sigma^2), Y \sim N(0,\sigma^2)$$

만약 이 두 확률 변수 X, Y 가 서로 독립이라면 (0,0)으로부터의 거리 또한 확률 변수가 되고 이 확률 변수는 레일리 분포가 됩니다.

$$R = \sqrt{X^2 + Y^2}$$

레일리 분포의 pdf

$$f(x|\sigma) = \frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right)$$

레일리 분포의 cdf

$$F(x|\sigma) = 1-\exp\left(-\frac{x^2}{2\sigma^2}\right)$$

$$x \geq 0\ \sigma > 0$$

레일리 분포의 평균, 분산

$$E(X) = \sigma\sqrt{\frac{\pi}{2}}$$

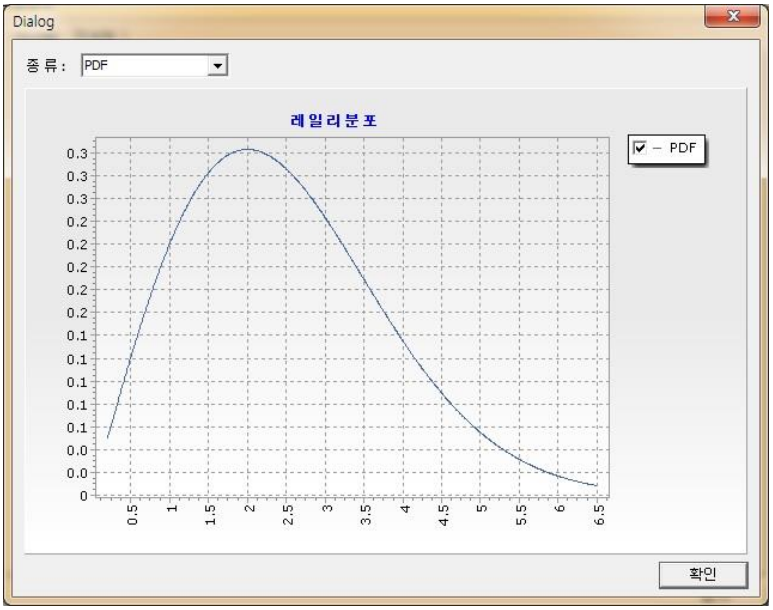
$$Var(X) = \frac{4-\pi}{2}\sigma^2$$

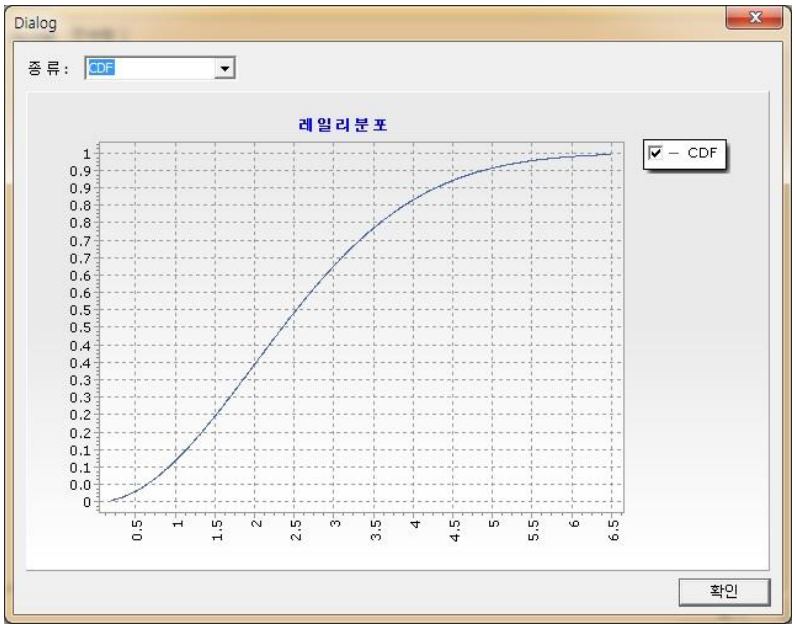
레일리 분포의 누적 분포 함수의 역함수

$$x = F^{-1}(p|\sigma) = \sqrt{-2\sigma^2\ln(1-p)}$$

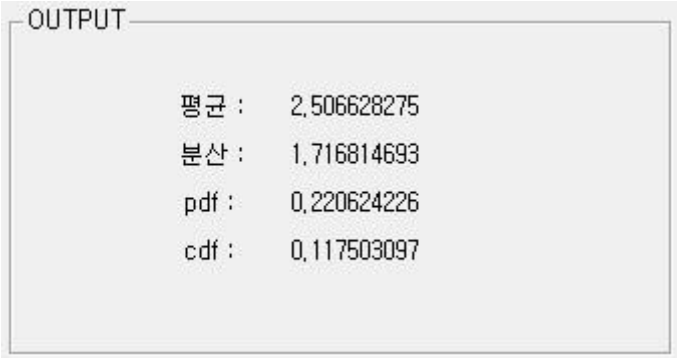
예시

- $b = 2$일 때 pdf, cdf 의 그래프





- $b=2$일 때 평균, 분산. $x=1$일 때 pdf, cdf.



- $b=2, p=0.5$일 때 누적 분포 함수의 역함수

7.2.20 t 분포 (T-distribution)

Z 가 표준 정규 확률 변수이고, V 가 자유도 ν 를 갖는 카이 제곱 분포일 때 이를 조합하여 만든 새로운 확률 변수 T 는

$$T = \frac{Z}{\sqrt{V/\nu}}$$

t 분포를 따른다고 합니다.(이 때 Z, V 는 서로 독립인 확률 변수입니다.)

t 분포의 pdf

$$f(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

t 분포의 cdf

$$F(x|\nu) = \int_{-\infty}^{x} f(t|\nu)dt = I_x\left(\frac{\nu}{2},\frac{\nu}{2}\right) = \frac{B\left(x;\frac{\nu}{2},\frac{\nu}{2}\right)}{B\left(\frac{\nu}{2},\frac{\nu}{2}\right)}$$

$B(x;\frac{\nu}{2},\frac{\nu}{2})$ 은 incomplete beta function

t 분포의 평균, 분산

$$E(X) = \begin{cases} 0 & \text{if } \nu > 1 \\ undefined & otherwise \end{cases}$$

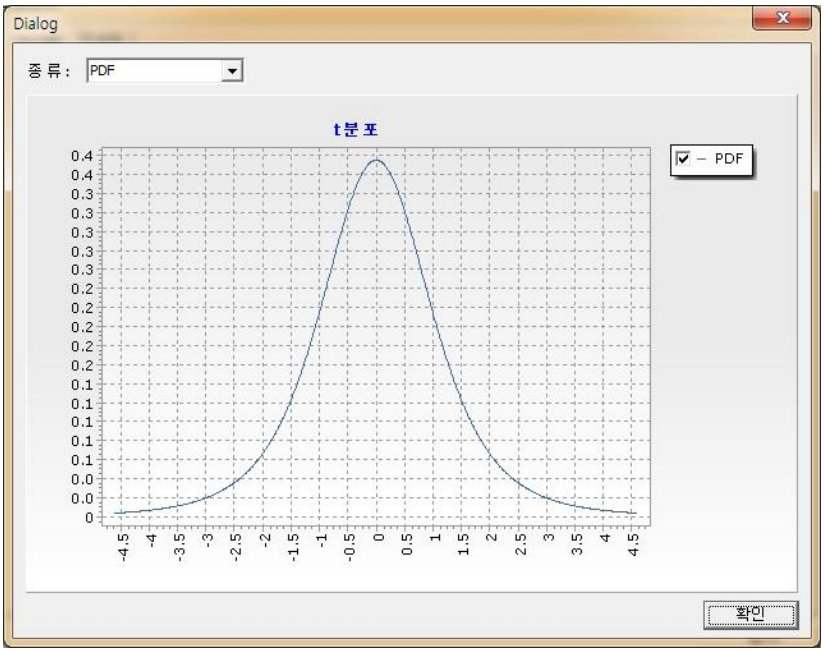
$$Var(X) = \begin{cases} \frac{\nu}{\nu-2} & \text{if } \nu > 2 \\ \infty & \text{if } 1 < \nu \le 2 \\ undefined & \text{otherwise} \end{cases}$$

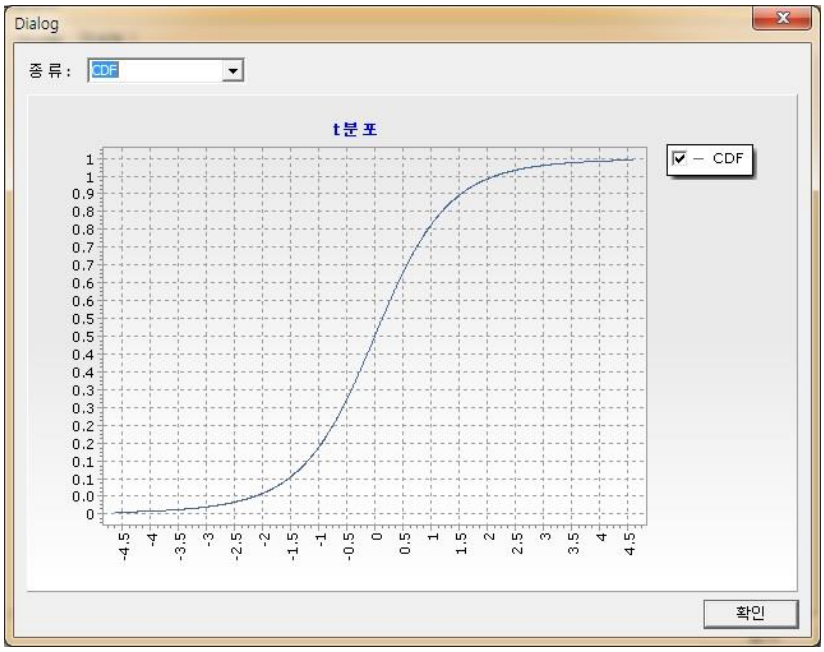
t 분포의 누적 분포 함수의 역함수

$$x = F^{-1}(p|\nu)$$

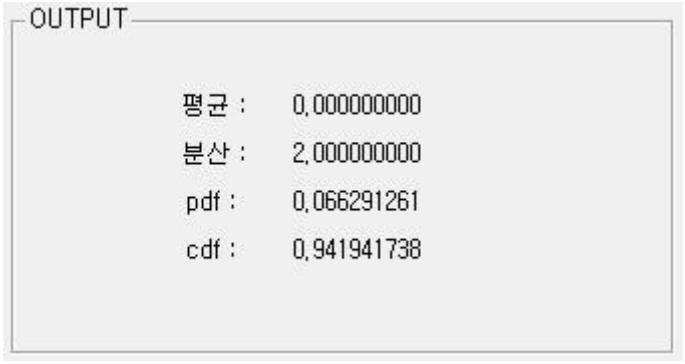
예시

- $\nu = 4$ 일 때 pdf, cdf 의 그래프





- $\nu = 4$ 일 때 평균, 분산. $x = 2$ 일 때 pdf, cdf.



▪ $\nu = 4, p = 0.5$일 때 누적 분포 함수의 역함수

7.2.21 이산 균일 분포 (Discrete uniform distribution)

이산 균일 분포는 이산 확률 분포의 일종으로 확률 변수가 가질 수 있는 값에서의 확률 값이 모두 동일한 분포입니다.

이산 균일 분포의 pmf

$$f(x \mid N) = \frac{1}{N}$$

이산 균일 분포의 cdf

$$F(x \mid N) = \frac{floor(x)}{N}$$

이산 균일 분포의 평균, 분산

$$E(X) = \frac{N+1}{2}$$

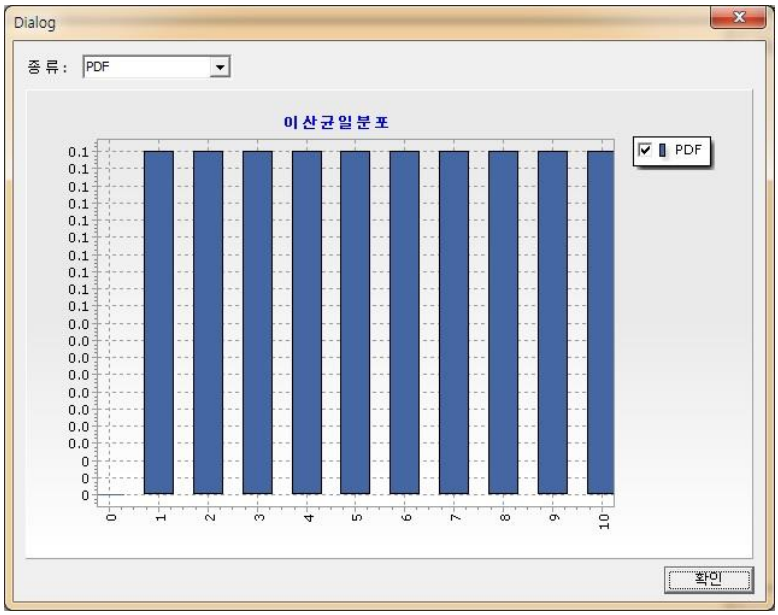
$$Var(X) = \frac{N^2 - 1}{12}$$

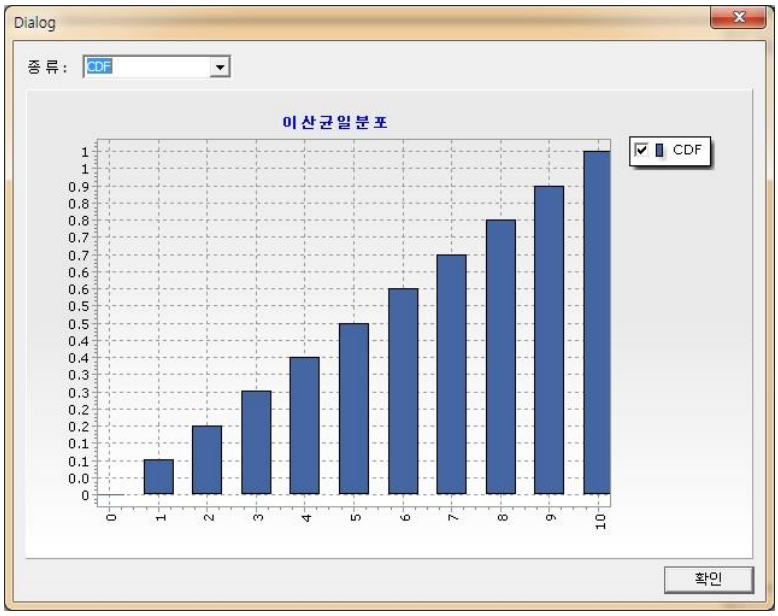
이산 균일 분포의 누적 분포 함수의 역함수

$$x = F^{-1}(p \mid N)$$

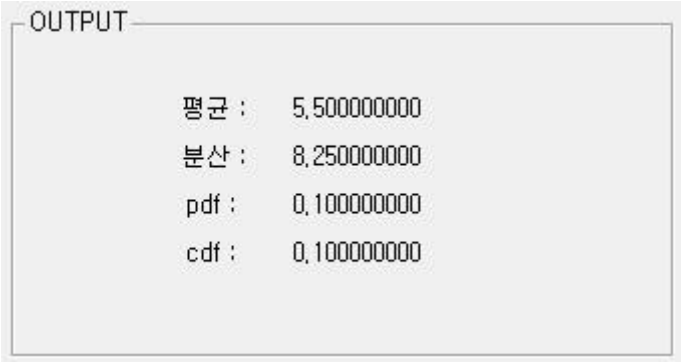
예시

$n = 10$ 일 때 pmf, cdf 의 그래프





▪ $n=10$ 일 때 평균, 분산. $x=1$ 일 때 pmf, cdf.



▪ $n=10, p=0.5$ 일 때 누적 분포 함수의 역함수

7.2.22 연속 균일 분포 (Continuous uniform distribution)

연속 균일 분포는 연속 확률 분포의 일종입니다. 매우 단순할 수 있는 이 분포는 매우 중요한 가치를 갖는데 모든 연속 확률 분포의 cdf 의 값이 바로 이 연속 균일 분포를 따르기 때문입니다.

$$U = F(X), X : \text{any random variable whose cdf is F}$$

이는 특정한 분포를 갖는 확률 변수의 random number 를 만드는데 중요한 역할을 합니다. 어떤 특정한 분포의 random number 는

$$F^{-1}(U)$$

의 식을 통해서 구해집니다. 이 때 U ~ Uniform[0,1] 입니다.

연속 균일 분포의 pdf

$$f(x \mid a,b) = \frac{1}{b-a} I_{(a,b)}(x)$$

$$a < b$$

연속 균일 분포의 cdf

$$F(x \mid a,b) = \begin{cases} 0 & \text{if } x \le a \\ \dfrac{x-a}{b-a} & \text{if } a < x < b \\ 1 & \text{if } x \ge b \end{cases}$$

연속 균일 분포의 평균, 분산

$$E(X) = \frac{a+b}{2}$$

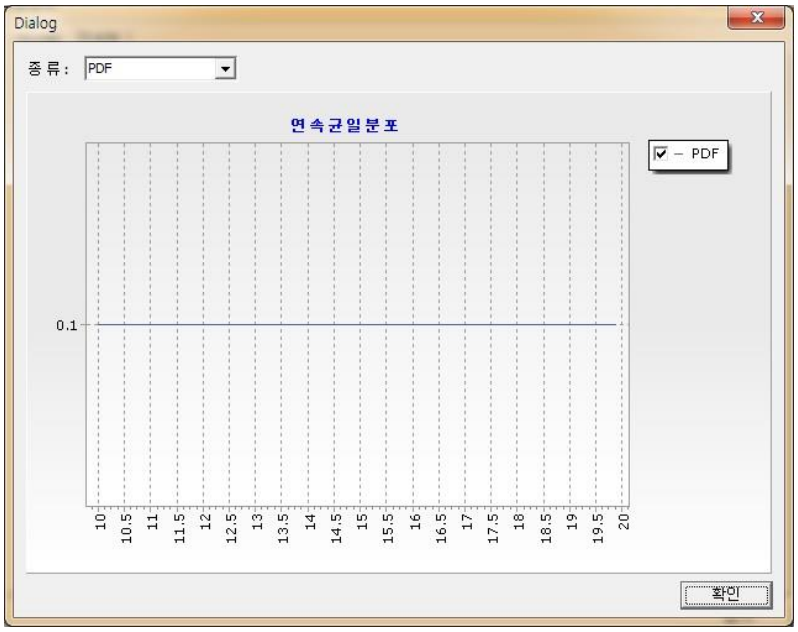
$$Var(X) = \frac{(b-a)^2}{12}$$

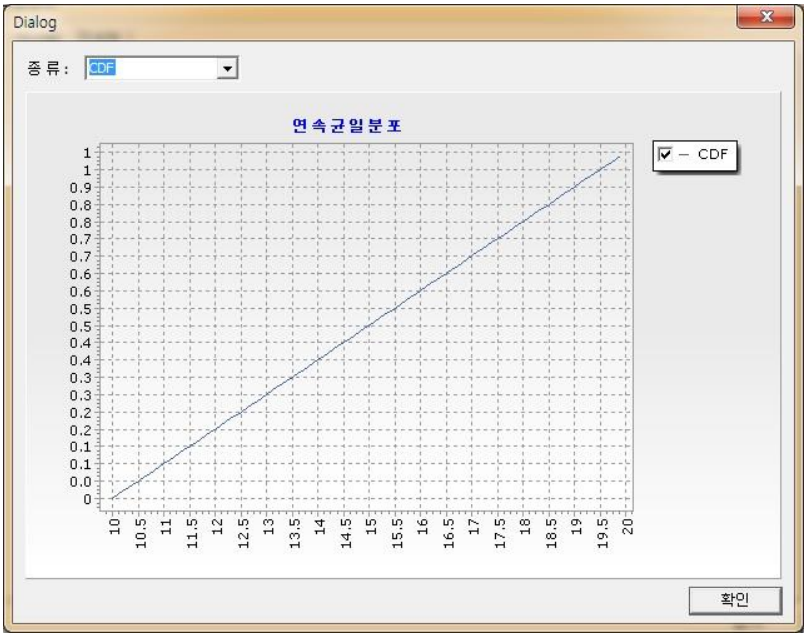
연속 균일 분포의 누적 분포 함수의 역함수

$$x = F^{-1}(p \mid a,b) = a + p(b-a)$$

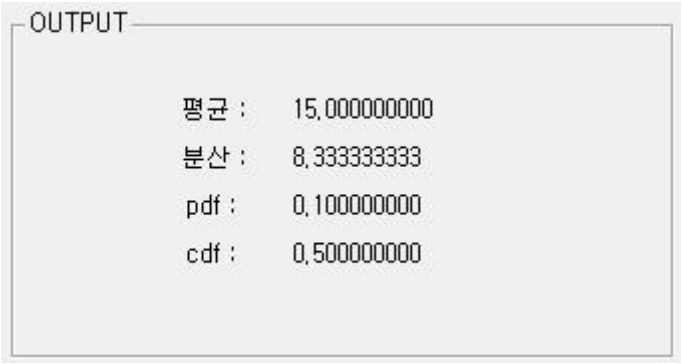
예시

- $a = 10, b = 20$ 일 때 pdf, cdf 의 그래프





- $a=10, b=20$ 일 때 평균, 분산. $x=15$ 일 때 pdf, cdf.



- $a=10, b=20, p=0.5$ 일 때 누적 분포의 역함수

7.2.23 와이블 분포 (Weibull distribution)

와이블 분포는 연속 확률 분포의 일종입니다. 이는 극단값 분포, 지수 분포 등과 밀접한 관련이 있는 분포로 생존 분석(Survival Analysis), 신뢰성 공학(Reliability Engineering), 기상 예측(Weather Forecasting)등에 많이 사용됩니다.

와이블 분포의 pdf

$$f(x \mid a,b) = \left(\frac{b}{a}\right)\left(\frac{x}{a}\right)^{a-1} \exp\left(-\left(\frac{x}{a}\right)^{b}\right)$$

$$x \geq 0, a > 0, b > 0$$

와이블 분포의 cdf

$$F(x \mid a,b) = 1 - \exp\left(-\left(\frac{x}{a}\right)^{b}\right)$$

$$x \geq 0, a > 0, b > 0$$

와이블 분포의 평균, 분산

$$E(X) = a\Gamma\left(1+\frac{1}{a}\right)$$

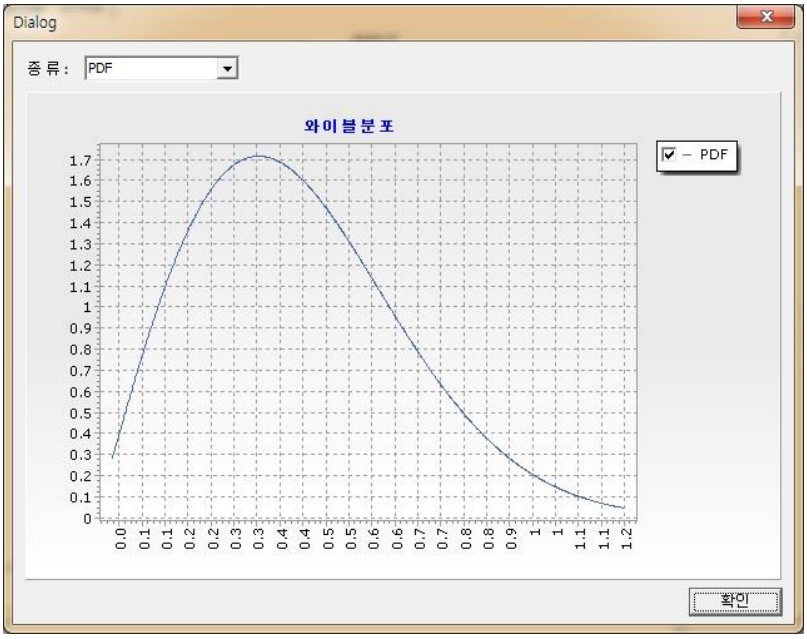
$$Var(X) = a^2\Gamma\left(1+\frac{2}{b}\right) - (E(X))^2$$

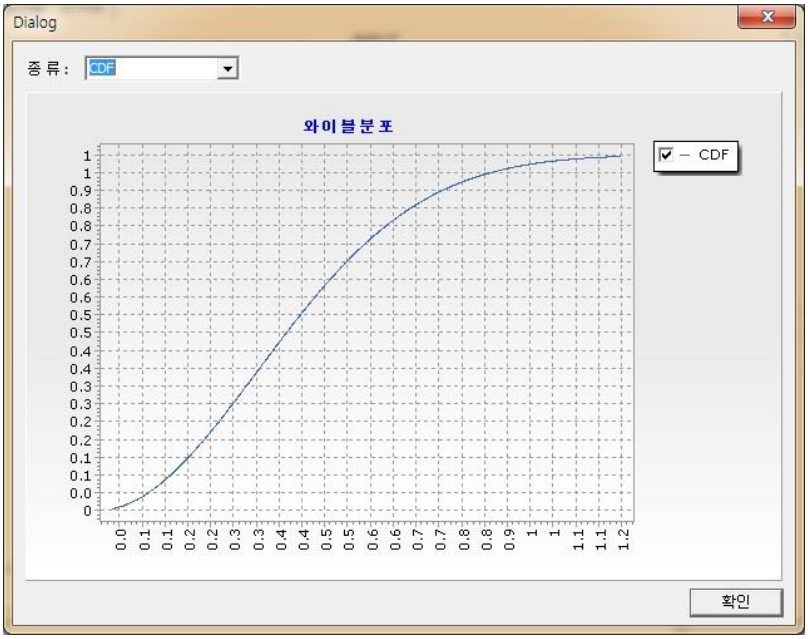
와이블 분포의 누적 분포 함수의 역함수

$$x = F^{-1}(p \mid a,b) = \begin{cases} \infty & \text{if } p = 1 \\ a(-\ln(1-p))^{1/b} & \text{otherwise} \end{cases}$$

예시

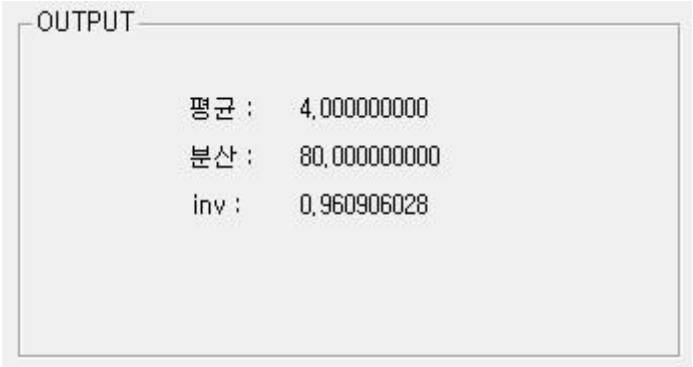
- $a = 2, b = 0.5$일 때 pdf, cdf 의 그래프





- $a=2, b=0.5$일 때 평균, 분산. $x=1$일 때 pdf, cdf.

- $a = 2, b = 0.5, p = 0.5$ 일 때 누적 분포 함수의 역함수



Appendix 1. 수식 편집기

수식 편집기는 파생변수, 필터, 선택, 채우기 등 다양한 노드에서 사용되며, 새로운 값을 계산하거나 조건을 위한 수식을 편집할 때 사용합니다. 마우스 조작만으로 수식을 편집할 수도 있으며 익숙해진 경우 수식 편집창에 직접 입력함으로써 보다 빠르게 수식을 편집할 수 있습니다.

특징

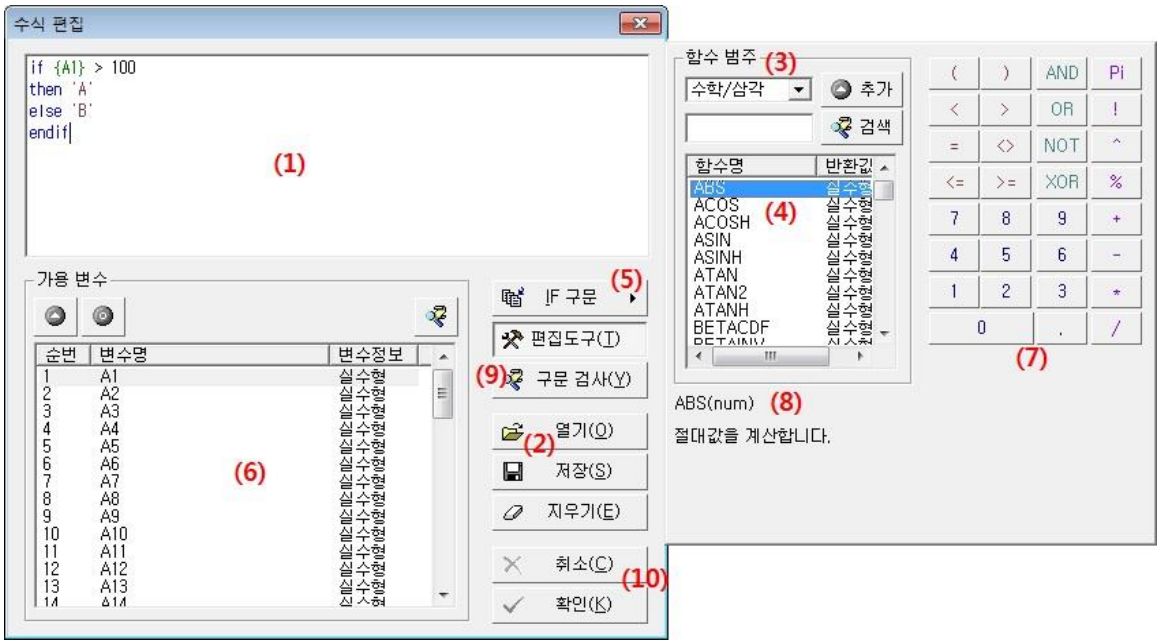
마우스 조작만으로 쉽게 ECMiner™에서 사용할 수식을 편집할 수 있습니다.

수식의 각 요소에 따라 색상을 달리 표시함으로써 가독성을 높였습니다.

다양한 수식 편집이 가능합니다.

저장 / 불러오기 기능을 이용하여 기 작성된 수식을 재 사용할 수 있습니다.

사용자 인터페이스



번호	설명	비고
(1)	수식 편집창. 수식을 편집할 수 있는 곳입니다. 임의로 입력할 수도 있으며 (3) ~ (7)을 이용하여 편집할 수도 있습니다. (3) ~ (7)을 이용하면 마우스 조작만으로 쉽게 ECMiner™에서 사용할 수식을 편집할 수 있으며 and 등의 논리 연산도 사용	

	가능합니다.	
(2)	현재 입력되어 있는 수식을 저장하거나 예전에 저장해 두었던 수식을 불러 와 다시 사용할 수 있습니다. 지우기 버튼을 누르면 현재 입력되어 있는 수식이 모두 지워집니다.	수식파일 확장자 : *.eaf
(3)	ECMiner™에서 제공하는 함수의 범주를 선택할 수 있습니다. ECMiner™에서는 변환함수, 수학/삼각, 텍스트, 날짜/시간, 변수통계, 정보 관련 함수들을 제공합니다. 이를 변경할 때마다 지정된 범주에 해당하는 함수들이 (4)에 목록 됩니다.	함수 목록 참조
(4)	지정된 함수범주에 해당하는 함수들이 나타납니다. 사용하고자 하는 함수를 선택하여 더블 클릭하거나 **추가 버튼**을 누르면 수식편집창에 입력됩니다. (8)에 각 함수의 사용법 및 간단한 설명이 나타납니다.	
(5)	IF 구문을 클릭하여 IF ~ THEN ~ ELSE ~ ENDIF 구문을 입력합니다. 편집도구를 클릭하면 오른쪽의 함수범주 창이 나타납니다.	
(6)	현재 파생변수 노드에서 사용할 수 있는 기존 변수 목록입니다. 입력하고자 하는 변수를 선택한 뒤 더블 클릭하여 편집창에 추가할 수 있습니다. 추가하면 {DATE} 같이 "{}"이 붙어 나타나며 이는 변수임을 의미하게 됩니다.	
(7)	마우스 조작만으로 수식을 편집할 수 있도록 하기 위해 많이 사용되는 숫자와 기호를 버튼으로 구성한 부분입니다. 버튼을 누르면 해당하는 숫자 혹은 기호가 편집창에 입력됩니다.	
(8)	현재 선택되어 있는 함수에 대한 간단한 설명 및 사용법이 나타납니다.	
(9)	현재 입력되어 있는 구문이 올바르게 되었는지 검사합니다. 구문이 틀렸을 경우 오류 메시지를 반환하며 이 오류 메시지를 보고 잘 못된 부분을 수정할 수 있습니다.	
(10)	현재 입력되어 있는 구문을 입력하거나 취소합니다. 편집 버튼을 누르면 입력되며, 입력하기 전에 구문검사를 수행합니다. 만약 잘못된 구문이 입력되었을 경우 프로그램이 제대로 동작하지 않을 수도 있습니다.	

스타일

문자열 → "123" or '123'- 진한 빨강

날짜형 → #2004-01-01# - 진한 빨강

숫자형 → 123 – 검정

함수, IF 구문 → ABS() - 파랑

변수 → {DATE} - 녹색

연산자 → +, -, *, / – 검정

매크로 → @ROWNUM – 진한 녹색

수식 편집창에서 수식이나 함수에 오류가 있으면 오류가 있는 부분이 빨간색으로 표시됩니다.

관련 항목

함수 목록

정의된 매크로

값 입력

A1.1 함수

ECMiner™에서 제공하는 함수입니다.

함수 범주

함수 범주	설명
변환함수	값의 형태를 변환합니다. 즉, 날짜형을 문자형으로 문자형을 숫자형으로 변환하는 함수입니다.
수학 / 삼각	sin, cos 등 삼각함수와 수학과 관련된 함수입니다.
텍스트	문자형 데이터를 처리하기 위한 함수입니다.
날짜 / 시간	날짜형 데이터에서 년도 수, 날짜, 시간 등의 값을 추출하는 함수입니다.
변수통계	현재 사용중인 변수에 통계학적 종류(연속형, 이산형)에 따른 기본 통계량이 필요할 때 사용합니다.
정보	입력된 값의 상태 및 형태 같은 정보를 산출하기 위한 함수입니다.
레코드	레코드 단위의 통계 처리를 위한 함수입니다.
기타	추가 필요사항에 의해 구현된 함수입니다.

- 아래 예시에 표현되는 결과값 형태는 함수 적용 결과 데이터 형태가 문자형일 경우 "결과값"의 형태로, 날짜형일 경우 #결과값#의 형태로 표현됩니다.(실제로 화면상에 표현되는 값은 ",#을 제외한 숫자/문자만 표현됩니다.)

A1.1.1 변환함수

ASC(text)

text: 문자열

입력된 문자열(text) 중 첫 문자의 ASCII Code 값을 반환합니다.

ASC("ABC") ➔ 65

CHR(num)

num: ASCII Code

ASCII Code 에 해당하는 문자를 반환합니다.

CHR(65) ➔ "A"

DATETOSTR(date)

date: 날짜형

날짜형을 문자형으로 변경합니다.

DATETOSTR(#2004-1-1#) ➔ "20040101"

IIF(cond,expr1,expr2)

cond: 조건문, expr1 : 수식 1, expr2 : 수식 2

조건이 참이라면 expr1, 거짓이라면 expr2 를 계산합니다.

IIF(1>0, 1+2, 1+3) ➔ 3

INT(num)

num: 실수

실수를 정수로 변환합니다. 소수점 이하의 숫자가 모두 무시됩니다.

INT(2.371) ➔ 2

STR(num[,num1[,num2]])

num: 숫자, num1: 최대 길이, num2 :소수점 자리수+

숫자를 문자로 변환합니다. num1 에 최대 길이, num2 에 소수점 자리수를 입력합니다.

STR(123.345,6,2) ➔ "123.35"

STRTODATE(text, fmt)

text: 문자열, fmt: 포맷

fmt 에 지정한 포맷에 따라 text 를 날짜형으로 변환합니다. 년도는 y, 월은 m, 일은 d, 시간은 H, 분은 M, 초는 S 로 대변됩니다. 예를 들어 문자열 '2004-10-23 23:28:11'의 경우 'yyyy-mm-dd HH:MM:SS'와 같이 fmt 를 지정합니다.

STRTODATE("2004-01-01 11:50",'yyyy-mm-dd HH:MM') ➔ #2004-01-01 11:50:00#

VAL(text)

text: 문자열

문자열을 숫자로 변환합니다.

VAL("123") ➔ 123

A1.1.2 수학/삼각함수

- 인수의 제약조건에서 벗어난 값이 입력될 경우, 결과값은 null(빈값)이 생성됩니다.

ABS(num)

num : 숫자

절대값을 계산합니다.

ABS(-123) ➔ 123

ACOS(num)

num: -1 이상 1 이하인 값

acos 을 계산합니다. 어떤 숫자의 acos 값은 cos 값을 구하면 그 숫자가 되는 각도입니다. 결과 각도는 라디안으로 표시되며 0 에서 pi 까지 입니다.

ACOS(-0.5) ➔ 2.09440(2*pi/3 라디안)

ACOSH(num)

num: 1 이상의 값

역하이퍼볼릭 코사인의 값을 계산합니다.

ACOSH(2) ➔ 1.31696

ASIN(num)

num: -1 이상 1 이하인 값

asin 을 계산합니다. 어떤 숫자의 asin 값은 sin 값을 구하면 그 숫자가 되는 각도입니다. 결과 각도는 라디안으로 표시되며 –pi/2 에서 pi/2 까지 입니다.

ASIN(-0.5) ➔ -0.52360(pi/6 라디안)

ASINH(num)

num: 숫자

역하이퍼볼릭 사인의 값을 계산합니다.

ASINH(2) ➔1.44364

ATAN(num)

num: 숫자

atan 을 계산합니다. 어떤 숫자의 atan 값은 tan 값이 그 숫자인 각도입니다. 표시 되는 각도는 -pi/2 에서 pi/2 까지의 라디안 값입니다.

ATAN(1) ➔ 0.78540(pi/4 라디안)

ATAN2(num1, num2)

num1: 숫자, num2: 숫자

atan2 을 계산합니다. 지정된 x, y 좌표의 아크탄젠트 값을 구합니다. 아크탄젠트는 원점에서 좌표(x_num, y_num)까지의 선과 x 축이 이루는 각도입니다. 이 각도는 -pi 에서 pi 사이의 라디안 값(-pi 제외)입니다.

ATAN2(1, 1)➔ 0.78540(pi/4 라디안)

ATANH(num1)

num1 : 숫자(-1 < num1 < 1)

역 하이퍼볼릭 탄젠트 값을 반환하는 함수

ATANH(0.2) ➔ 0.20273

BETACDF(num1, num2, num3)

num 1, num2, num3: 숫자

모수 num1, num2 에 해당하는 Beta 분포의 num3 에서의 CDF 값을 구합니다

BETACDF(2, 2, 1) ➔ 1.0

BETAINV(num1, num2, num3)

num 1, num2, num3: 숫자

모수 num1, num2 에 해당하는 Beta 분포 CDF 의 num3 에서의 역함수 값을 구합니다.

BETAINV(2, 2, 1) ➔ 1.0

BETAMEAN(num1, num2)

num 1, num2: 숫자

모수 num1, num2 에 해당하는 Beta 분포의 평균을 구합니다.

BETAMEAN(2,2) ➔ 0.5

BETAPDF(num1, num2, num3)

num 1, num2, num3: 숫자

모수 num1, num2 에 해당하는 Beta 분포의 num3 에서의 PDF 값을 구합니다

BETAPDF(2, 2, 1) ➔ 0

BETAVARIANCE(num1, num2)

num 1, num2: 숫자

모수 num1, num2 에 해당하는 Beta 분포의 분산을 구합니다.

BETAVARIANCE(2,2) ➔ 0.05

BINOCDF(num1, num2, num3)

num1, num2, num3 : 숫자

모수 num1(n), num2(p)에 해당하는 이항분포의 num3 에서의 CDF 를 구합니다.

BINOCDF(4, 0.3, 3) ➔ 0.9919

BINOINV(num1, num2, num3)

num1, num2, num3 : 숫자

모수 num1(n), num2(p)에 해당하는 이항분포 CDF 의 num3 에서의 역함수 값을 구합니다.

BINOINV(4, 0.3, 0.9919) ➔ 3

BINOMEAN(num1, num2)

num 1, num2: 숫자

모수 num1(n), num2(p)에 해당하는 이항분포의 평균을 구합니다.

BINOMEAN(10, 0.5) ➔ 5

BINOPDF(num1, num2, num3)

num1, num2, num3 : 숫자

모수 num1(n), num2(p)에 해당하는 이항분포의 num3 에서의 PDF 를 구합니다.

BINOPDF(4, 0.3, 3) ➔ 0.07560

BINOVARIANCE(num1, num2)

num 1, num2: 숫자

모수 num1, num2 에 해당하는 이항분포의 분산을 구합니다.

BINOVARIANCE(10, 0.5) ➔ 2.5

CEILING(num)

num: 숫자

num 보다 크거나 같은 가장 작은 정수를 계산합니다.

CEILING(2.1) ➔ 3

CEILING2(num1, num2)

num 1, num2: 숫자

num1 을 num2 의 배수가 되도록 올림한 값을 반환하는 함수입니다.

CEILING2(6.4) ➔ 8

CHICDF(num1, num2)

num1, num2 : 숫자

모수 num1(v) 에 해당하는 카이제곱분포의 num2 에서의 CDF 를 구합니다.

CHICDF(1, 0.3) ➔ 0.41612

CHIINV(num1, num2)

num1, num2 : 숫자

모수 num1(v) 에 해당하는 카이제곱분포 CDF 의 num2 에서의 역함수 값을 구합니다.

CHIINV(1, 0.41612) ➔ 0.3

CHIMEAN(num1)

num 1: 숫자

모수 num1 에 해당하는 카이제곱분포의 평균을 구합니다.

CHIMEAN(10) ➔ 10

CHIPDF(num1, num2)

num1, num2 : 숫자

모수 num1(v) 에 해당하는 카이제곱분포의 num2 에서의 PDF 를 구합니다.

CHIPDF(1, 0.3) ➔ 0.62691

CHIVARIANCE(num1)

num 1: 숫자

모수 num1 에 해당하는 카이제곱분포의 분산을 구합니다.

CHIVARIANCE(10) ➔ 20

COMB(num1, num2)

num 1, num2: 숫자

Combination 값을 계산합니다. Num1 개 중에서 num2 개를 선택하는 조합의 수를 계산. Num1>num2, 두 인수는 0 또는 양의 정수

COMB(4,2) ➔ 6

COS(num)

num: 라디안 단위의 각도

cos 을 계산합니다.

COS(1.047) ➔ 0.50017

COSH(num)

num: 숫자

cosh 을 계산합니다.

COSH(4) ➔ 27.30823

DEGREE(num)

num: 숫자

Radians 을 Degree 로 변환합니다.

DEGREE(3.14159265) ➔ 180

DIGAMMA(num)

num: 숫자 (num>0)

digamma 함수의 값을 반환하는 함수, lgamma 함수의 미분값.

DIGAMMA(1) ➔ -0. 57722

ERF(num)

num: 숫자

오차함수 값을 반환하는 함수

ERF(1) ➔ 0.8427

ERFC(num)

num: 숫자

오차함수 여함수값을 반환하는 함수

ERFC(1) ➔ 0.15776

EVEN(num)

num: 숫자

num 에 가장 가까운 짝수로 올림합니다.

EVEN(1.5) ➔ 2

EVCDF(num1, num2, num3)

num 1, num2, num3: 숫자

모수 num1, num2 에 해당하는 극단값 분포의 num3 에서의 CDF 를 구합니다.(num1 은 mu, num2 는 sigma)

EVCDF(0, 1, 0.4) ➔ 0.77504

EVINV(num1, num2, num3)

num 1, num2, num3: 숫자

모수 num1, num2 에 해당하는 극단값 분포 CDF 의 num3 에서의 역함수 값을 구합니다. (num1 은 mu, num2 는 sigma)

EVINV(0, 1, 0.77504) ➔ 0.40001

EVMEAN(num1, num2)

num 1, num2: 숫자

모수 num1, num2 에 해당하는 극단값 분포의 평균을 구합니다.(num1 은 mu, num2 는 sigma)

EVMEAN(0,1) ➔ 0.57722

EVPDF(num1, num2, num3)

num 1, num2, num3: 숫자

모수 num1, num2 에 해당하는 극단값 분포의 num3 에서의 PDF 를 구합니다.(num1 은 mu, num2 는 sigma)

EVPDF(0, 1, 0.4) ➔ 0.3356

EVVARIANCE(num1, num2)

num 1, num2: 숫자

모수 num1, num2 에 해당하는 극단값 분포의 분산을 구합니다.(num1 은 mu, num2 는 sigma)

EVVARIANCE(0,1) ➔ 1.64493

EXP(num)

num: 지수

exp 를 계산합니다. e = 2.71828182845904 이고 자연 로그의 밑입니다.

EXP(1) ➔ 2.71828

EXPCDF(num1, num2)

num 1, num2: 숫자

모수 num1 에 해당하는 지수 분포의 num2 에서의 CDF 를 구합니다.

EXPCDF(1, 0.4) ➔ 0.32968

EXPINV(num1, num2)

num 1, num2: 숫자

모수 num1 에 해당하는 지수 분포 CDF 의 num2 에서의 역함수 값을 구합니다.

EXPINV(1, 0.32968) ➔ 0.4

EXPMEAN(num1)

num 1: 숫자

모수 num1 에 해당하는 지수 분포의 평균을 구합니다.

EXPMEAN(10)➔ 10

EXPPDF(num1, num2)

num 1, num2: 숫자

모수 num1 에 해당하는 지수 분포의 num2 에서의 PDF 값을 구합니다.

EXPPDF(1, 0.4) ➔ 0.67032

EXPVARIANCE(num1)

num 1: 숫자

모수 num1 에 해당하는 지수 분포의 분산을 구합니다.

EXPVARIANCE(10) ➔ 100

FACT(num)

num: 숫자

Factorial 을 계산합니다.

FACT(4) ➔ 24

FISHER(num1)

num 1: 숫자

FISHER 변환값을 계산합니다. -1<num1<1

FISHER(0.5) ➔ 0.54931

FISHERINV(num1)

num 1: 숫자

FISHER 변환의 역변환값을 계산합니다.

FISHERINV(0.54931) ➔ 0.5

FLOOR(num)

num: 숫자

num 보다 작거나 같은 가장 큰 정수를 계산합니다.

FLOOR(2.1) ➔ 2

FCDF(num1, num2, num3)

num 1, num2, num3: 숫자

모수 num1, num2 에 해당하는 F 분포의 num3 에서의 CDF 값을 구합니다.

FCDF(2, 4, 3) ➔ 0.84

FINV(num1, num2, num3)

num 1, num2, num3: 숫자

모수 num1, num2 에 해당하는 F 분포 CDF 의 num3 에서의 역함수 값을 구합니다.

FINV(2, 4, 0.84) ➔ 3

FMEAN(num1, num2)

num 1, num2: 숫자

모수 num1, num2 에 해당하는 F 분포의 평균을 구합니다.

FMEAN(5, 5) ➔ 1.66667

FPDF(num1, num2, num3)

num 1, num2, num3: 숫자

모수 num1, num2 에 해당하는 F 분포의 num3 에서의 PDF 값을 구합니다.

FPDF(2, 4, 3) ➔ 0.064

FVARIANCE(num1, num2)

num 1, num2: 숫자

모수 num1, num2 에 해당하는 F 분포의 분산을 구합니다.

FVARIANCE(5, 5) ➔ 8.88889

GAMCDF(num1, num2, num3)

num 1, num2, num3: 숫자

모수 num1, num2 에 해당하는 감마분포의 num3 에서의 CDF 를 구합니다.

GAMCDF(1, 2, 1) ➔ 0.39347

GAMINV(num1, num2, num3)

num 1, num2, num3: 숫자

모수 num1, num2 에 해당하는 감마분포 CDF 의 num3 에서의 역함수 값을 구합니다.

GAMINV(1, 2, 0.39347) ➔ 1

GAMMEAN(num1, num2)

num 1, num2: 숫자

모수 num1, num2 에 해당하는 감마분포의 평균을 구합니다.

GAMMEAN(3, 3) ➔ 9

GAMPDF(num1, num2, num3)

num 1, num2, num3: 숫자

모수 num1, num2 에 해당하는 감마분포의 num3 에서의 PDF 를 구합니다.

GAMPDF(1, 2, 1) ➔ 0.30327

GAMVARIANCE(num1, num2)

num 1, num2: 숫자

모수 num1, num2 에 해당하는 감마분포의 분산을 구합니다.

GAMVARIANCE(3, 3) ➔ 27

GEOCDF(num1, num2)

num 1, num2: 숫자 (단, num1 은 0 과 1 사이의 실수, num2 는 0 이상의 정수)

모수 num1 에 해당하는 기하분포의 num2 에서의 CDF 값을 구합니다.

GEOCDF(0.3, 2) ➔ 0.657

GEOINV(num1, num2)

num 1, num2: 숫자 (단, num1 은 0 과 1 사이의 실수)

모수 num1 에 해당하는 기하분포 CDF 의 num2 에서의 역함수 값을 구합니다.

GEOINV(0.3, 0.657) ➔ 3

GEOMEAN(num1)

num 1: 숫자 (단, num1 은 0 과 1 사이의 실수)

모수 num1 에 해당하는 기하분포의 평균을 구합니다.

GEOMEAN(0.25) ➔ 3

GEOPDF(num1, num2)

num 1, num2: 숫자 (단, num1 은 0 과 1 사이의 실수, num2 는 0 이상의 정수)

모수 num1 에 해당하는 기하분포 PDF 의 num2 에서의 PDF 값을 구합니다.

GEOPDF(0.3, 2) ➔ 0.147

GEOVARIANCE(num1)

num 1: 숫자 (단, num1 은 0 과 1 사이의 실수)

모수 num1 에 해당하는 기하분포의 분산을 구합니다.

GEOVARIANCE(0.25) ➔ 12

GEVCDF(num1, num2, num3, num4)

num 1, num2, num3, num4: 숫자

모수 num1, num2, num3 에 해당하는 일반화 극단값 분포의 num4 에서의 CDF 값을 구합니다. 단 num1 은 mu, num2 는 sigma, num3 는 k.

GEVCDF(0, 1, 0, 2) ➔ 0.87342

GEVINV(num1, num2, num3, num4)

num 1, num2, num3, num4: 숫자

모수 num1, num2, num3 에 해당하는 일반화 극단값 분포 CDF 의 num4 에서의 역함수 값을 구합니다. 단 num1 은 mu, num2 는 sigma, num3 는 k.

GEVINV(0, 1, 0, 0.87342) ➔ 1.99997

GEVMEAN(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2, num3 에 해당하는 일반화 극단값 분포의 평균을 구합니다. 단 num1 은 mu, num2 는 sigma, num3 는 k.

GEVMEAN(1,3,0.1) ➔ 3.05886

GEVPDF(num1, num2, num3, num4)

num 1, num2, num3, num4: 숫자

모수 num1, num2, num3 에 해당하는 일반화 극단값 분포의 num4 에서의 PDF 값을 구합니다. 단 num1 은 mu, num2 는 sigma, num3 는 k.

GEVPDF(0, 1, 0, 2) ➔ 0.1182

GEVVARIANCE(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2, num3 에 해당하는 일반화 극단값 분포의 분산을 구합니다. 단 num1 은 mu, num2 는 sigma, num3 는 k.

GEVVARIANCE(1,3,0.1) ➔ 20.03617

GPCDF(num1, num2, num3, num4)

num 1, num2, num3, num4 : 숫자

모수 num1, num2, num3 에 해당하는 일반화 파레토 분포의 num4 에서의 CDF 를 구합니다. 단 num1 은 K, num2 는 sigma, num3 는 theta.

GPCDF(0, 1, 0, 2) ➔ 0.86466

GPINV(num1, num2, num3, num4)

num 1, num2, num3, num4: 숫자

모수 num1, num2, num3 에 해당하는 일반화 파레토 분포 CDF 의 num4 에서의 역함수 값을 구합니다. 단 num1 은 K, num2 는 sigma, num3 는 theta..

GPINV(0, 1, 0, 0.86466) ➔ 1.99997

GPMEAN(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2, num3 에 해당하는 일반화 파레토 분포의 평균을 구합니다. 단 num1 은 K, num2 는 sigma, num3 는 theta.

GPMEAN(0.1, 3, 1) ➔ 4.33333

GPPDF(num1, num2, num3, num4)

num 1, num2, num3, num4 : 숫자

모수 num1, num2, num3 에 해당하는 일반화 파레토 분포의 num4 에서의 PDF 값을 구합니다. 단 num1 은 K, num2 는 sigma, num3 는 theta.

GPPDF(0, 1, 0, 2) ➔ 0.13534

GPVARIANCE(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2, num3 에 해당하는 일반화 파레토 분포의 분산을 구합니다. 단 num1 은 K, num2 는 sigma, num3 는 theta.

GPVARIANCE(0.1, 3, 1) ➔ 13.88889

HYGECDF(num1, num2, num3, num4)

num 1, num2, num3, num4 : 숫자

모수 num1, num2, num3 에 해당하는 초기하 분포의 num4 에서의 CDF 값을 구합니다. 단 num1 은 m, num2 는 k, num3 는 n.

HYGECDF(100, 20, 10, 2) ➔ 0.68122

HYGEINV(num1, num2, num3, num4)

num 1, num2, num3, num4 : 숫자

모수 num1, num2, num3 에 해당하는 초기하 분포 CDF 의 num4 에서의 역함수 값을 구합니다. 단 num1 은 m, num2 는 k, num3 는 n.

HYGEINV(100, 20, 10, 0.68122) ➔ 2

HYGEMEAN(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2, num3 에 해당하는 초기하 분포의 평균을 구합니다. 단 num1 은 m, num2 는 k, num3 는 n.

HYGEMEAN(10, 5, 4) ➔ 2

HYGEPDF(num1, num2, num3, num4)

num 1, num2, num3, num4 : 숫자

모수 num1, num2, num3 에 해당하는 초기하 분포의 num4 에서의 PDF 값을 구합니다. 단 num1 은 m, num2 는 k, num3 는 n.

HYGEPDF(100, 20, 10, 2) ➔ 0.31817

HYGEVARIANCE(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2, num3 에 해당하는 초기하 분포의 평균을 구합니다. 단 num1 은 m, num2 는 k, num3 는 n.

HYGEVARIANCE(10, 5, 4) ➔ 0.66667

LGAMMA(num)

num : 양의 실수

감마함수의 자연로그값을 반환하는 함수

LGAMMA(3) ➔ 0.69315

LN(num)

num : 양의 실수

자연 로그값을 계산합니다.

LN(86) ➔ 4.45435

LOG(num,base)

num : 양의 실수 base : 1 이 아닌 양의 실수

지정한 base 에 대한 로그값을 반환합니다.

LOG(7,2) ➔ 2.80735

LOG10(num)

num : 양의 실수

상용 로그값을 계산합니다.

LOG10(86) ➔ 1.93450

LOGBETA(num1, num2)

num1,2: 양의 실수

베타함수의 자연로그값을 반환하는 함수

LOGBETA (5,3) ➔ -4.65396

LOGNCDF(num1, num2, num3)

num 1, num2, num3: 숫자

모수 num1, num2 에 해당하는 로그 정규 분포의 num3 에서의 CDF 값을 구합니다. 단 num1 은 mu, num2 는 sigma.

LOGNCDF(0, 1, 2) ➔ 0.75589

LOGNINV(num1, num2, num3)

num 1, num2, num3: 숫자

모수 num1, num2 에 해당하는 로그 정규 분포 CDF 의 num3 에서의 역함수 값을 구합니다. 단 num1 은 mu, num2 는 sigma.

LOGNINV(0, 1, 0.75589) ➔ 1.99999

LOGNMEAN(num1, num2)

num 1, num2: 숫자

모수 num1, num2 에 해당하는 로그 정규 분포의 평균을 구합니다. 단 num1 은 mu, num2 는 sigma.

LOGNMEAN(-2, 2) ➔ 1

LOGNPDF(num1, num2, num3)

num 1, num2, num3: 숫자

모수 num1, num2 에 해당하는 로그 정규 분포의 num3 에서의 PDF 값을 구합니다. 단 num1 은 mu, num2 는 sigma.

LOGNPDF(0, 1, 2) ➔ 0.15687

LOGNVARIANCE(num1, num2)

num 1, num2: 숫자

모수 num1, num2 에 해당하는 로그 정규 분포의 분산을 구합니다. 단 num1 은 mu, num2 는 sigma.

LOGNVARIANCE(-2, 2) ➔ 53.59815

MOD(num1, num2)

num1 : 숫자, num2 : 숫자

num1/ num2 시 나머지를 계산합니다.

MOD(5, 2) ➔ 1

NBINCDF(num1, num2, num3)

num 1, num2 , num3: 숫자

모수 num1, num2 에 해당하는 음이항 분포의 num3 에서의 CDF 값을 구합니다. 단 num1 은 r, num2 는 p.

NBINCDF(3, 0.5, 2) ➔ 0.5

NBININV(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2 에 해당하는 음이항 분포 CDF 의 num3 에서의 역함수 값을 구합니다. 단 num1 은 r, num2 는 p.

NBININV(3, 0.5, 0.5) ➔ 2

NBINMEAN(num1, num2)

num 1, num2 : 숫자

모수 num1, num2 에 해당하는 음이항 분포의 평균을 구합니다. 단 num1 은 r, num2 는 p.

NBINMEAN(10, 0.5) ➔ 10

NBINPDF(num1, num2, num3)

num 1, num2 , num3: 숫자

모수 num1, num2 에 해당하는 음이항 분포의 num3 에서의 PDF 값을 구합니다. 단 num1 은 r, num2 는 p.

NBINPDF(3, 0.5, 2) ➔ 0.1875

NBINVARIANCE(num1, num2)

num 1, num2 : 숫자

모수 num1, num2 에 해당하는 음이항 분포의 분산을 구합니다. 단 num1 은 r, num2 는 p.

NBINVARIANCE(10, 0.5) ➔ 20

NCFCDF(num1, num2, num3, num4)

num 1, num2, num3, num4 : 숫자

모수 num1, num2, num3 에 해당하는 비중심 F 분포의 num4 에서의 CDF 를 구합니다. 단 num1 은 nu1, num2 는 nu2, num3 는 delta.

NCFCDF(5, 20, 10, 2) ➔ 0.2719

NCFINV(num1, num2, num3, num4)

num 1, num2, num3, num4 : 숫자

모수 num1, num2, num3 에 해당하는 비중심 F 분포 CDF 의 num4 에서의 역함수 값을 구합니다. 단 num1 은 nu1, num2 는 nu2, num3 는 delta.

NCFINV(5, 20, 10, 0.2719) ➔ 2

NCFMEAN(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2, num3 에 해당하는 비중심 F 분포의 평균을 구합니다. 단 num1 은 nu1, num2 는 nu2, num3 는 delta. 단 num1 > 2 인 경우 평균 계산 가능

NCFMEAN(10, 100, 4) ➔ 1.42857

NCFPDF(num1, num2, num3, num4)

num 1, num2, num3, num4 : 숫자

모수 num1, num2, num3 에 해당하는 비중심 F 분포의 num4 에서의 PDF 를 구합니다. 단 num1 은 nu1, num2 는 nu2, num3 는 delta.

NCFPDF(5, 20, 10, 2) ➔ 0.26075

NCFVARIANCE(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2, num3 에 해당하는 비중심 F 분포의 분산을 구합니다. 단 num1 은 nu1, num2 는 nu2, num3 는 delta. 단 num1!=0, num2>4 인 경우 분산 계산 가능

NCFVARIANCE(10, 100, 4) ➔ 0.42517

NCTCDF(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2 에 해당하는 비중심 T 분포의 num3 에서의 CDF 값을 구합니다. 단 num1 은 nu, num2 는 delta.

NCTCDF(10, 1, 2) ➔ 0.80761

NCTINV(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2 에 해당하는 비중심 T 분포 CDF 의 num3 에서의 역함수 값을 구합니다. 단 num1 은 nu, num2 는 delta.

NCTINV(10, 1, 0.80761) ➔ 1.99999

NCTMEAN(num1, num2)

num 1, num2 : 숫자

모수 num1, num2 에 해당하는 비중심 T 분포의 분산을 구합니다. 단 num1 은 nu, num2 는 delta. 단, num1>2 인 경우 평균 계산 가능

NCTMEAN(4, 1) ➔ 1.25331

NCTPDF(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2 에 해당하는 비중심 T 분포의 num3 에서의 PDF 값을 구합니다. 단 num1 은 nu, num2 는 delta.

NCTPDF(10, 1, 2) ➔ 0.00006

NCTVARIANCE(num1, num2)

num 1, num2 : 숫자

모수 num1, num2 에 해당하는 비중심 T 분포의 분산을 구합니다. 단 num1 은 nu, num2 는 delta. 단, num1>2 인 경우 분산 계산 가능

NCTVARIANCE(4, 1) ➔ 2.4292

NCX2CDF(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2 에 해당하는 비중심 카이 제곱 분포의 num3 에서의 CDF 값을 구합니다. 단 num1 은 nu, num2 는 delta.

NCX2CDF(4, 2, 2) ➔ 0.13048

NCX2INV(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2 에 해당하는 비중심 카이 제곱 분포 CDF 의 num3 에서의 역함수 값을 구합니다. 단 num1 은 nu, num2 는 delta.

NCX2INV(4, 2, 0.13048) ➔ 2.00003

NCX2MEAN(num1, num2)

num 1, num2 : 숫자

모수 num1, num2 에 해당하는 비중심 카이 제곱 분포의 평균을 구합니다. 단 num1 은 nu, num2 는 delta.

NCX2MEAN(4, 2)➔ 6

NCX2PDF(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2 에 해당하는 비중심 카이 제곱 분포의 num3 에서의 PDF 값을 구합니다. 단 num1 은 nu, num2 는 delta.

NCX2PDF(4, 2, 2) ➔ 0.10763

NCX2VARIANCE(num1, num2)

num 1, num2 : 숫자

모수 num1, num2 에 해당하는 비중심 카이 제곱 분포의 분산을 구합니다. 단 num1 은 nu, num2 는 delta.

NCX2VARIANCE(4, 2) ➔ 16

NORMCDF(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2 에 해당하는 정규 분포의 num3 에서의 CDF 를 구합니다. 단 num1 은 mu, num2 는 sigma.

NORMCDF(-1, 1, 2) ➔ 0.99865

NORMDIST(num1, num2, num3)

num1, num2, num3 : 숫자

모수 num2, num3 에 해당하는 정규 분포의 num1 에서의 CDF 값을 계산합니다. 만약 num1 < num2 인 경우 1-CDF 값을 반환합니다.

NORMDIST(6, 5, 1) ➔ 0.84134

NORMDIST(4, 5, 1) ➔ 0.84134

NORMINV(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2 에 해당하는 정규 분포 CDF 의 num3 에서의 역함수 값을 구합니다. 단 num1 은 mu, num2 는 sigma.

NORMINV(-1, 1, 0.99865) ➔ 1.99998

NORMMEAN(num1, num2)

num 1, num2 : 숫자

모수 num1, num2 에 해당하는 정규 분포의 평균을 구합니다. 단 num1 은 mu, num2 는 sigma.

NORMMEAN(0, 2) ➔ 0

NORMPDF(num1, num2, num3)

num 1, num2, num3 : 숫자

모수 num1, num2 에 해당하는 정규 분포의 num3 에서의 PDF 를 구합니다. 단 num1 은 mu, num2 는 sigma.

NORMPDF(-1, 1, 2) ➔ 0.00443

NORMVARIANCE(num1, num2)

num 1, num2 : 숫자

모수 num1, num2 에 해당하는 정규 분포의 분산을 구합니다. 단 num1 은 mu, num2 는 sigma.

NORMVARIANCE(0, 2) ➔ 4

ODD(num)

num : 숫자

num 에 가장 가까운 홀수로 올림합니다.

ODD(1.1) ➔ 3

PERM(num1, num2)

num 1, num2: 숫자

Permuation 의 값을 계산합니다. Num1 개 중에서 num2 개를 선택하는 순열의 수를 계산합니다. (단 num1 > num2, 두 인수는 0 또는 양의정수)

PERM(4, 2) ➔ 12

POISCDF(num1, num2)

num 1, num2: 숫자

모수 num1 에 해당하는 포아송 분포의 num2 에서의 CDF 값을 구합니다. 단 num1 은 lambda.

POISCDF(4, 2) ➔ 0.2381

POISINV(num1, num2)

num 1, num2: 숫자

모수 num1 에 해당하는 포아송 분포 CDF 의 num2 에서의 역함수 값을 구합니다. 단 num1 은 lambda.

POISINV(4, 0.2381) ➔ 1

POISMEAN(num1)

num 1: 숫자

모수 num1 에 해당하는 포아송 분포의 평균을 구합니다. 단 num1 은 lambda.

POISMEAN(2) ➔ 2

POISPDF(num1, num2)

num 1, num2: 숫자

모수 num1 에 해당하는 포아송 분포의 num2 에서의 PDF 값을 구합니다. 단 num1 은 lambda.

POISPDF(4, 2) ➔ 0.14653

POISVARIANCE(num1)

num 1: 숫자

모수 num1 에 해당하는 포아송 분포의 분산을 구합니다. 단 num1 은 lambda.

POISVARIANCE(2) ➔ 2

POWER(num1, num2)

num1 : 숫자, num2 : 숫자

num1 을 num2 만큼 거듭제곱한 값을 계산합니다.

POWER(4, 0.5) ➔ 2

PRODUCT(num1, num2,…, numn)

num1, num2, … , numn : 숫자

인수들의 곱을 반환

PRODUCT(1, 2) ➔ 2

RADIANS(num)

num : 숫자

Degree 를 Radians 으로 변환합니다.

RADIANS(180) ➔ 3.14159(pi)

RAND()

0 에서 1000 사이의 난수를 만듭니다.

RAYLCDF(num1, num2)

num 1, num2: 숫자

모수 num1 에 해당하는 레일리 분포의 num2 에서의 CDF 값을 구합니다.

RAYLCDF(1, 2) ➔ 0.86466

RAYLINV(num1, num2)

num 1, num2: 숫자

모수 num1 에 해당하는 레일리 분포 CDF 의 num2 에서의 역함수 값을 구합니다.

RAYLINV(1, 0.86466) ➔ 1.99998

RAYLMEAN(num1)

num 1: 숫자

모수 num1 에 해당하는 레일리 분포의 평균을 구합니다.

RAYLMEAN(2) ➔ 2.50663

RAYLPDF(num1)

num 1, num2: 숫자

모수 num1 에 해당하는 레일리 분포의 num2 에서의 PDF 값을 구합니다.

RAYLPDF(1, 2) ➔ 0.27067

RAYLVARIANCE(num1)

num 1: 숫자

모수 num1 에 해당하는 레일리 분포의 분산을 구합니다.

RAYLVARIANCE(2) ➔ 1.71681

ROUND(num1, digit)

num1 : 숫자, digit : 자리수

num1 을 지정한 자리수(digit)로 반올림합니다.

ROUND(12.345, 2) ➔ 12.35

ROUNDDOWN(num1, digit)

num1 : 숫자, digit : 자리수

num1 을 지정한 자리수(digit)로 내림합니다.

ROUNDDOWN(2.50663, 2) ➔ 2.50

ROUNDUP(num1, digit)

num1 : 숫자, digit : 자리수

num1 을 지정한 자리수(digit)로 올림합니다.

ROUNDUP(2.50663, 2) ➔ 2.51

SIN(num)

num : 라디안 단위의 각도

sin 을 계산합니다.

SIN(3.141592) ➔ 0

SINH(num)

num : 숫자

sinh 를 계산합니다.

SINH(1) ➔ 1.17520

SQRT(num)

num : 숫자

양의 제곱근을 계산합니다.

SQRT(4) ➔ 2

TAN(num)

num : 라디안 단위의 각도

tan 를 계산합니다.

TAN(0.785) ➔ 0.99920

TANH(num)

num : 숫자

tanh 를 계산합니다.

TANH(-2) ➔ -0.96403

TCDF(num1, num2)

num 1, num2: 숫자

모수 num1 에 해당하는 t 분포의 num2 에서의 CDF 값을 구합니다.

TCDF(1, 0.5) ➔ 0.64758

TINV(num1, num2)

num 1, num2: 숫자

모수 num1 에 해당하는 t 분포 CDF 의 num2 에서의 역함수 값을 구합니다.

TINV(1, 0.64758) ➔ 0.49999

TMEAN(num1)

num 1: 숫자

모수 num1 에 해당하는 t 분포의 평균을 구합니다.

TMEAN(4) ➔ 0

TOMAX(var, limit)

var: 변수값

limit: 임의값

변수값(var)이 지정된 한계값(limit)보다 작거나 같을 때 변수값을 반환하며, 클경우 한계값을 반환합니다.

TOMAX(9, 10) ➔ 9

TOMAX(11, 10) ➔ 10

TPDF(num1, num2)

num 1, num2: 숫자

모수 num1 에 해당하는 t 분포의 num2 에서의 PDF 값을 구합니다.

TPDF(1, 0.5) ➔ 0.25465

TRIGAMMA(num1)

num 1: 숫자

Trigamma 함수값을 반환하는 함수. num1>0

TRIGAMMA(1) ➔ 1.64493

TRUNC(num1, num2)

num 1, num2: 숫자

지정한 자릿수만큼 소수점 아래에 남기고 나머지는 버린 값을 반환하는 함수

Num2 는 digit.(정수)

TRUNC(0.76,1) ➔ 0.7

TVARIANCE(num1)

num 1: 숫자

모수 num1 에 해당하는 t 분포의 분산을 구합니다.

TVARIANCE(4) ➔ 2

UNIDCDF(num1, num2)

num 1, num2: 숫자

모수 num1 에 해당하는 이산균일 분포의 num2 에서의 CDF 값을 구합니다.

UNIDCDF(50, 20) ➔ 0.4

UNIDINV(num1, num2)

num 1, num2: 숫자

모수 num1 에 해당하는 이산균일 분포 CDF 의 num2 에서의 역함수 값을 구합니다.

UNIDINV(50, 0.4) ➔ 20

UNIDMEAN(num1)

num 1: 숫자

모수 num1 에 해당하는 이산균일 분포의 평균을 구합니다.

UNIDMEAN(5) ➔ 3

UNIDPDF(num1, num2)

num 1, num2: 숫자

모수 num1 에 해당하는 이산균일 분포의 num2 에서의 PDF 값을 구합니다.

UNIDPDF(50, 20) ➔ 0.02

UNIDVARIANCE(num1)

num 1: 숫자

모수 num1 에 해당하는 이산균일 분포의 분산을 구합니다.

UNIDMEAN(5) ➔ 3

UNIFCDF (num1, num2, num3)

num1, num2, num3 : 숫자

모수 num1, num2 에 해당하는 연속 균일 분포의 num3 에서의 CDF 값을 구합니다.

UNIFCDF(-1, 1, 0.7) ➔ 0.85

UNIFINV (num1, num2, num3)

num1, num2, num3 : 숫자

모수 num1, num2 에 해당하는 연속 균일 분포 CDF 의 num3 에서의 역함수 값을 구합니다.

UNIFINV(-1, 1, 0.85) ➔ 0.7

UNIFMEAN (num1, num2)

num1, num2 : 숫자

모수 num1, num2 에 해당하는 연속 균일 분포의 평균을 구합니다.

UNIFMEAN(1,7) ➔ 4

UNIFPDF (num1, num2, num3)

num1, num2, num3 : 숫자

모수 num1, num2 에 해당하는 연속 균일 분포의 num3 에서의 PDF 값을 구합니다.

UNIFPDF(-1, 1, 0.7) ➔ 0.5

UNIFVARIANCE (num1, num2)

num1, num2 : 숫자

모수 num1, num2 에 해당하는 연속 균일 분포의 분산을 구합니다.

UNIFVARIANCE(1,7) ➔ 3

WBLCDF(num1, num2, num3)

num1, num2, num3 : 숫자

모수 num1, num2 에 해당하는 와이블 분포의 num3 에서의 CDF 값을 구합니다.

WBLCDF(0.15, 0.8, 0.5) ➔ 0.53846

WBLINV(num1, num2, num3)

num1, num2, num3 : 숫자

모수 num1, num2 에 해당하는 와이블 분포 CDF 의 num3 에서의 역함수 값을 구합니다.

WBLINV(0.15, 0.8, 0.53846) ➔ 0.5

WBLMEAN(num1, num2)

num1, num2 : 숫자

모수 num1, num2 에 해당하는 와이블 분포의 평균을 구합니다.

WBLMEAN(2, 0.5) ➔ 4

WBLPDF(num1, num2, num3)

num1, num2, num3 : 숫자

모수 num1, num2 에 해당하는 와이블 분포의 num3 에서의 PDF 값을 구합니다.

WBLPDF(0.15, 0.8, 0.5) ➔ 1.53846

WBLVARIANCE(num1, num2)

num1, num2 : 숫자

모수 num1, num2 에 해당하는 와이블 분포의 분산을 구합니다.

WBLVARIANCE(2, 80) ➔ 0.00100

A1.1.3 텍스트

ALLTRIM(text1[, text2])

text1: 문자열, text2: 제거할 문자열

text1 에 있는 text2 의 문자열을 모두 제거합니다. text2 가 입력되지 않으면 공백을 제거합니다.

ALLTRIM("AABACD","A") ➔ "BCD"

CHRTRAN(text1, text2, text3)

text1: 문자열, text2: 문자열, text3: 문자열

text1 에서 text2 을 찾아 text3 로 바꿉니다.

CHRTRAN("BBACD", "A", "C") ➔ "BBCCD"

FIND(text1, text2[, num])

text1 : 문자열, text2 : 문자열, num : 회수

text1 에서 text2 을 num 만큼 찾은 문자의 시작 위치를 반환합니다. num 이 입력되니 않으면 1 로 간주 됩니다. 문자열의 시작 위치는 1 입니다.

FIND("ABBACDAB", "A", 2) ➔ 4

KOR_GET1(text)

text : 문자열

text 에서 한글을 추출합니다.

KOR_GET1(“ABC ㄱㄴㄷ”) ➔ ”ㄱㄴㄷ”

KOR_GET2(text, num1[, num2])

Text: 문자열, num1,2: 임의값

Text 중 num1 번째 문자열부터 num2 번째 문자열을 잘라 한글문자만 반환합니다. num2 가 없으면 num1 부터 끝까지 잘라 반환합니다.

kor_get2("가나다 ab 라마",2,6) ➔ "나다라마"

LEFT(text1, num)

text1 : 문자열, num : 숫자

text1 의 왼쪽에서 num 만큼의 문자열을 잘라 반환합니다.

LEFT("ABBACDAB", 3) ➔ "ABB"

LEN(text1)

text1 : 문자열

text1 의 문자수를 반환합니다.

LEN("ABBACDAB") ➔ 8

LIKE(text1, text2)

text1: 문자열, text2: 문자열

text1 이 text2 와 비슷하면 TRUE 를 반환합니다. 와일드 문자 *, ?를 사용할 수 있습니다. ?는 한 문자, *는 여러 문자를 대변합니다.

LIKE("ABBACD", "*CD") ➔ TRUE

LOWER(text1)

text1 : 문자열

text1 을 소문자로 변환 합니다.

LOWER("AbCD") ➔ "abcd"

LTRIM(text1[, text2])

text1: 문자열, text2: 제거할 문자열

text1 의 왼쪽에서부터 text2 에 속한 모든 문자를 제거합니다. text2 가 입력되지 않으면 공백을 제거합니다.

LTRIM("AABACD", "A") ➔ "BACD"

OCCURS(text1, text2)

text1: 문자열, text2 : 문자열

text1 중에 text2 가 몇 회 있는지 계산합니다.

OCCURS("AABACD", "A") ➔ 3

REPT(text1, num)

text1 : 문자열, num : 숫자

text1 을 num 만큼 반복하여 새로운 문자열을 만듭니다.

REPT("A", 3) ➔ "AAA"

RIGHT(text1, num)

text1 : 문자열, num : 숫자

text1 중 오른쪽에서 num 만큼의 문자열을 잘라 반환합니다.

RIGHT("AABACD", 2) ➔ "CD"

RTRIM(text1 [,text2])

text1: 문자열, text2: 제거할 문자열

text1 의 오른쪽에서부터 text2 에 속한 모든 문자를 제거합니다. text2 가 입력되지 않으면 공백을 제거합니다.

RTRIM("DCABAA", A") ➔ "DCAB"

SPACE(num)

num: 숫자

num 만큼 공백을 반복한 문자열을 만듭니다. REPT(" ", num)와 같습니다.

SPACE(3) ➔ “ “

STRSORT(text)

text: 문자열

text 를 정렬합니다.

STRSORT(“adbc”) ➔”abcd”

SUBSTR(text1, num1 [,num2])

text1 : 문자열, num1 : 시작 위치, num2 : 길이

text1 중 num1 부터 num2 만큼의 문자열을 잘라 반환합니다. num2 가 없으면 num1 부터 끝까지 문자열을 잘라 반환합니다.

SUBSTR("AABACD", 2, 3) ➔ "ABA"

TRIM(text1 [,text2])

text1 : 문자열, text2 : 문자열

text1 의 양쪽 끝에서부터 text2 에 속한 모든 문자를 제거합니다. text2 가 입력되지 않으면 공백을 제거합니다.

TRIM("AABACDA", "A") ➔ "BACD"

UPPER(text1)

text1 : 문자열

text1 을 대문자로 변환합니다.

UPPER("aaBAcD") ➔ "AABACD"

UTF8(text1)

text1 : 문자열(유니코드형태)

text1 을 UTF8 로 Decode 합니다.(url 해석용)

UTF8("%EB%B9%B5") ➔ "빵"

A1.1.4 날짜/시간함수

CMONTH(date)

date : 날짜

날짜 중 월명을 반환합니다.

CMONTH(#2004-01-01#) ➔ "1 월 "

CWEEK(date)

date : 날짜

날짜 중 요일명을 반환합니다.

CWEEK(#2004-01-01#) ➔ "목요일"

DATE([nYear, nMonth, nDay])

nYear : 년도, nMonth : 1 ~ 12, nDay : 날수

날짜를 반환합니다. 인수가 없으면 현재 날짜를 반환합니다.

DATE(2004, 1, 1) ➔ #2004-01-01#

DATETIME([nYear, nMonth, nDay [, nHours [, nMinutes [, nSeconds]]]])

nYear: 년도, nMonth: 1 ~ 12, nDay: 날수, nHours: 시간, nMinutes: 분, nSeconds: 초

날짜/시간을 반환합니다. 인수가 없으면 현재 날짜/시간을 반환합니다.

DATETIME(2004, 1, 1, 12) ➔ #2004-01-01 12:00:00#

DAY(date)

date: 날짜

날짜를 반환합니다.

DAY(#2007-04-06#) ➔ 6

DAYOFYEAR(date)

date: 날짜

해당 해의 몇 번째 날인지 반환합니다.

DAYOFYEAR(#2004-02-01#) ➔ 32

DAYS(date1, date2)

date1: 날짜, date2: 날짜

date1(end date)과 date2(start date) 사이의 날수 차이를 반환합니다.

DAYS(#2004-01-01#, #2003-02-01#) ➔ 334

GODAY(date, day)

date: 날짜, day: 숫자

date 에 day 만큼 일수를 더하거나 뺀 날짜를 반환합니다.

GODAY(#2004-02-01#, 5) ➔ #2004-02-06#

GOMONTH(date, month)

date: 날짜, day: 숫자

date 에 month 만큼 달수를 더하거나 뺀 날짜를 반환합니다.

GOMONTH(#2004-02-01#, 5) ➔ #2004-07-01#

GOYEAR(date, year)

date: 날짜, day: 숫자

date 에 year 만큼 년수를 더하거나 뺀 날짜를 반환합니다.

GOYEAR(#2004-02-01#, 5) ➔ #2009-02-01#

HOUR(date)

date: 날짜

시간을 반환합니다.

HOUR(#2004-01-01 오후 1:30:20#) ➔ 13

HOURS(date1, date2)

date1: 날짜, date2: 날짜

date1(end date)과 date2(start date) 사이의 시간 차이를 시간 단위로 반환합니다.

HOURS(#2004-01-01 오후 7:00:00#, #2004-01-01 오후 1:00:00#) ➔ 6

MINUTE(date)

date: 날짜

분을 반환합니다.

MINUTE(#2004-01-01 오후 1:30:20#) ➔ 30

MINUTES(date1, date2)

date1: 날짜, date2:날짜

date1(end date)과 date2(start date) 사이의 시간 차이를 분 단위로 반환합니다.

MINUTES(#2004-01-01 오후 1:30:00#, #2004-01-01 오후 1:00:00#) ➔ 30

MONTH(date)

date: 날짜

월을 반환합니다.

MONTH(#2004-01-01 오후 1:30:20#) ➔ 1

MONTHS(date1, date2)

date1: 날짜, date2: 날짜

date1(end date)과 date2(start date) 사이의 차이를 월 단위로 반환합니다.

MONTHS(#2004-01-01#, #2003-01-01#) ➔ 12

SEC(date)

date: 날짜

초를 반환합니다.

SEC(#2004-01-01 오후 1:30:20#) ➔ 20

SECS(date1, date2)

date1: 날짜, date2: 날짜

date1(end date)과 date2(start date) 의 차이를 초 단위로 반환합니다.

SECS(#2004-01-01 오후 1:30:40#, #2004-01-01 오후 1:30:20#) ➔ 20

TIME()

스트림 실행 시점을 시:분:초 형태로 반환합니다.

TIME() ➔ #17:30:20#

TRIMDATE(date)

입력된 날짜에서 날짜를 제외한 시간만 반환합니다.

TRIMDATE(#2004-01-01 오후 1:30:20#) -> #1:30:20#

TRIMTIME (date)

입력된 날짜에서 시간을 제외한 날짜만 반환합니다.

TRIMTIME(#2004-01-01 오후 1:30:20#) -> #2004-01-01#

WEEK(date)

date: 날짜

입력된 날짜의 요일수를 반환합니다. 일요일은 1, 월요일은 2, ..., 토요일은 7 을 반환합니다.

WEEK(#2004-01-01 오후 1:30:20#) ➔ 5

YEAR(date)

date : 날짜

연도를 반환합니다.

YEAR(#2004-01-01 오후 1:30:20#) ➔ 2004

YEARS(date1, date2)

date1 : 날짜, date2 : 날짜

date1(end date)과 date2(start date) 사이의 년수를 반환합니다.

YEARS(#2004-01-01#, #2003-01-01#) ➔ 1

A1.1.5 변수통계

연속형 변수

_AVEDEV(num)

num : field_index

지정된 인덱스에 해당하는 변수에 대한 절대편차의 평균을 반환합니다.

_CORR(num1, num2)

num1, num2 : 변수 인덱스

지정된 인덱스에 해당하는 변수들의 선형 또는 등급 상관계수를 반환합니다.

_COUNT(num)

num: 변수 인덱스

지정된 인덱스에 해당하는 수치형 변수의 데이터 수를 반환합니다.(결측치 제외)

_CSS(num)

num: 변수 인덱스

지정된 인덱스에 해당하는 변수의 수정 제곱합을 반환합니다.

_CV(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 변동계수를 반환합니다.

_DEVSQ(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 표본평균으로부터의 편차의 제곱합을 반환합니다.

_GEOMEAN(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 기하평균을 반환합니다.

_HARMMEAN(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 조화평균을 반환합니다.

_IQR(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 IQR 를 반환합니다.

_KURTOSIS(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 첨도를 반환합니다.

_MAD(num,flag)

num : 변수 인덱스, flag : 0 or 1

지정된 인덱스에 해당하는 변수의 중위 절대편차를 반환하는 함수. flag = 0 이면 mean(abs(X-mean(X))), flag != 0 이면 median(abs(X-median(X)))

_MAX(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 최대값을 반환합니다.

_MAXINDEX(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 최대값에 해당하는 행의 인덱스를 반환합니다.

_MAXINGRP(idx1, val, idx2)

idx1 에 해당하는 변수의 값이 val 인 레코드에 대하여 idx2 에 해당하는 변수의 최대값을 구합니다.

_MEAN(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 평균을 반환합니다.

_MEDIAN(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 중앙값을 반환합니다.

_MIDRANGE(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 Midrange 를 반환합니다.

_MIN(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 최소값을 반환합니다.

_MININDEX(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 최소값에 해당하는 행의 인데스를 반환합니다.

_MININGRP(idx1, val, idx2)

idx1 에 해당하는 변수의 값이 val 인 레코드에 대하여 idx2 에 해당하는 변수의 최소값을 구합니다.

_MODE(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 최빈값을 반환하는 함수

_MOMENT(num1, num2)

num1 : 변수 인덱스 , num2 : 숫자(양의 정수)

지정된 인덱스에 해당하는 변수의 표본의 n 차 적률을 반환합니다.

_PERCENTRANK(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 백분율순위를 반환합니다.

_Q0(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 사분위수 중 첫번째 값을 반환합니다.

_Q1(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 사분위수 중 두번째 값을 반환합니다.

_Q2(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 사분위수 중 세번째 값을 반환합니다. (중앙값과 같음)

_Q3(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 사분위수 중 네번째 값을 반환합니다.

_Q4(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 사분위수 중 다섯번째 값을 반환합니다.

_RANGE(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 범위를 반환합니다.

_RMS(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 제곱평균제곱근을 반환하는 함수

_RSQ(num1, num2)

num1, num2: 변수 인덱스

지정된 인덱스에 해당하는 변수의 결정계수(피어슨 상관계수의 제곱)를 반환합니다.

_SKEWNESS(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 왜도를 반환합니다.

_SQRSUM(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 제곱의 합을 반환합니다.

_STD(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 표준편차를 반환합니다.

_STDP(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 모집단의 표준편차를 반환합니다.

_SUM(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 총합을 반환합니다.

_TRIMMEAN(num1, num2)

num1 : 변수 인덱스 , num2 : k

지정된 인덱스에 해당하는 변수의 절사평균(이상치를 제외한 평균)을 반환합니다.

_USS(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 제곱합을 반환합니다.

_VARIANCE(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 분산을 반환합니다.

_VARP(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 모집단의 분산을 반환합니다.

이산형 변수

_CLASSCNT(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수의 총 Class 개수를 반환합니다. 값이 몇가지 있는 지 알수 있습니다.

_CLASSTOVAL(num1, num2)

num1 : 변수 인덱스, num2: class index

지정된 인덱스에 해당하는 변수 중 num2 에 해당하는 Class 의 원래 값을 반환합니다.

_MAXCNTINDEX(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수 중 횟수가 가장 많은 Class 의 인덱스를 반환합니다.

_MINCNTINDEX(num)

num : 변수 인덱스

지정된 인덱스에 해당하는 변수 중 횟수가 가장 적은 Class 의 인덱스를 반환합니다.

A1.1.6 정보

BETWEEN(val, lval, hval)

val : 임의값, lval : 임의값, hval : 임의값

val 이 lval 과 hval 사이에 있다면 TRUE 를 반환합니다.

BETWEEN(1, 0, 2) ➔ TRUE

DATASIZE()

현재 사용하고 있는 데이터의 크기 (행 개수)를 반환합니다.

DATASIZE() ➔ 100

EQSTR(val1, val2[,…, valN])

val1~valN : 임의의 문자값

문자형 val1 과 valN 사이의 값이 모두 같다면 TRUE 를 반환합니다.

EQSTR(“A”, “A”) ➔ TRUE

ISALPHA(text)

text : 문자

text 가 알파벳이면 TRUE 를 반환합니다.

ISALPHA("A") ➔ TRUE, ISALPHA(“123”)->FALSE

ISBOOL(val)

val : 임의값

val 이 논리판단을 할 수 있는 Boolean 형태면 TRUE 를 반환합니다.

ISBOOL(1=2) ➔ TRUE

ISBOOL(123) ➔FALSE

ISDATE(val)

val : 임의값

val 이 날짜형이면 TRUE 를 반환합니다.

ISDATE(#2004-01-01#)➔ TRUE

ISDIGIT(text)

text : 문자

text 가 숫자면 TRUE 를 반환합니다.

ISDIGIT(“123”) ➔ TRUE

ISIN(val, val1[, …, valN])

val, val1 ~ valN : 임의값

val 의 값이 val1~valN 값 중에 일치하는 값이 존재한다면 TRUE 를 반환합니다.

ISIN(“BAG”, “BAG”, “Nice”,”Book”) ➔ TRUE

ISLOWER(text)

text : 문자

text 가 소문자면 TRUE 를 반환합니다.

ISLOWER("a") ➔ TRUE

ISNULL(val)

val : 임의값

val 이 결측치 혹은 NULL 값이면 TRUE 를 반환합니다.

ISNULL(@NULL) -> TRUE

ISNUMERIC(val)

val : 임의값

val 이 정수형 혹은 실수형이면 TRUE 를 반환합니다.

ISNUMERIC(1.2) ➔ TRUE

ISSTRING(val)

val : 임의값

val 이 문자형이면 TRUE 를 반환합니다.

ISSTRING("123")➔ TRUE

ISUPPER(text)

text : 문자

text 가 대문자면 TRUE 를 반환합니다.

ISUPPER("A") ➔ TRUE

MAX(val1, val2 [, ..., valN])

val1 ~ valN : 임의값

val1 과 valN 사이의 값 중 가장 큰 값을 반환합니다. val1, val2 는 필수 입력입니다.

MAX(1, 2, 3) ➔ 3

MIN(val1, val2 [, ..., valN])

val1 ~ valN : 임의값

val1 과 valN 사이의 값 중 가장 작은 값을 반환합니다. val1, val2 는 필수 입력입니다.

MIN(1, 2, 3) ➔ 1

SIGN(num)

num : 숫자

num 이 양수면 1, 음수면 -1, 0 이면 0 을 반환합니다.

SIGN(-123) ➔ -1

VARTYPE(val)

val : 임의값

val 의 형태에 따라 S(문자형), N(숫자형), B(Boolean), D(날짜형), X(NULL), U(알 수 없음)의 문자를 반환합니다.

VARTYPE("A") ➔ "S"

A1.1.7 레코드

RAVERAGE(f, [varIdx1, …, varIdxN])

f : 0 or 1

현재 열에서 지정된 인덱스에 해당하는 각 변수들의 평균을 계산합니다. 인덱스를 지정하지 않으면 전체 평균을 계산합니다. f=0 이면 결측치 미포함 f=1 이면 결측치 포함

RCNTBTN(val1, val2[varIdx1, …, varIdxN])

val1,2 : 임의값, varIdx: field_index

현재 열에서 전체 혹은 지명된 인덱스에 해당하는 변수의 값이 val1 보다 크거나 같고 val2 보다 작거나 같은 변수의 개수를 반환합니다.

RCNTEQ(val[,varIdx1, …, varIdxN])

val : 임의값, varIdx: field_index

현재 열에서 전체 혹은 지명된 인덱스에 해당하는 변수의 값이 입력한 값과 같은 변수의 개수를 반환합니다.

RCNTEQGT(val[,varIdx1, …, varIdxN])

val : 임의값, varIdx: field_index

현재 열에서 전체 혹은 지명된 인덱스에 해당하는 변수의 값이 입력한 값보다 크거나 같은 변수의 개수를 반환합니다.

RCNTEQLT(val[,varIdx1, …, varIdxN])

val : 임의값, varIdx: field_index

현재 열에서 전체 혹은 지명된 인덱스에 해당하는 변수의 값이 입력한 값보다 작거나 같은 변수의 개수를 반환합니다.

RCNTGT(val[,varIdx1, …, varIdxN])

val : 임의값, varIdx: field_index

현재 열에서 전체 혹은 지명된 인덱스에 해당하는 변수의 값이 입력한 값보다 큰 변수의 개수를 반환합니다.

RCNTLT(val[,varIdx1, …, varIdxN])

val : 임의값, varIdx: field_index

현재 열에서 전체 혹은 지명된 인덱스에 해당하는 변수의 값이 입력한 값보다 작은 변수의 개수를 반환합니다.

RCOUNT(f[,varIdx1, …, varIdxN])

f: 0 or 1 , varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 각 변수들의 개수를 반환합니다. 인덱스를 지정하지 않으면 전체 필드를 대상으로 계산합니다. f=0 이면 수치형 데이터, f=1 이면 전체 데이터(결측치 제외)

RCSS([,varIdx1, …, varIdxN])

val : 임의값, varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수들의 수정제곱합을 계산합니다. 인덱스를 지정하지 않으면 전체 변수 중 연속형인 변수의 수정제곱합을 계산합니다.

RDEVSQ([,varIdx1, …, varIdxN])

val : 임의값

현재 열에서 지정된 인덱스에 해당하는 각 변수들을 표본 평균으로부터 편차의 제곱합으로 계산하여 결과값을 반환합니다. 인덱스를 지정하지 않으면 전체에서 연속형인 변수들을 대상으로 계산합니다.

RMAXA([,varIdx1, …, varIdxN])

val : 임의값

현재 열에서 지정된 인덱스에 해당하는 변수 중 가장 큰 값을 반환합니다. 인덱스를 지정하지 않으면 전체 변수를 대상으로 합니다. 텍스트 논리값 포함(텍스트는 0 으로 인식, false =0, true = 1)

RMAXVAR([varIdx1, …, varIdxN])

varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수 중 가장 큰 값을 가지는 변수명을 반환합니다. 인덱스를 지정하지 않으면 전체 연속형 변수를 대상으로 합니다.

RMAXVARIDX([varIdx1, …, varIdxN])

varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수 중 가장 큰 값을 가지는 변수의 인덱스를 반환합니다. 인덱스를 지정하지 않으면 전체 연속형 변수를 대상으로 합니다.

RMEAN([varIdx1, …, varIdxN])

varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수들의 평균을 계산합니다. 인덱스를 지정하지 않으면 전체 변수 중 연속형인 변수의 평균을 계산합니다.

RMINA([varIdx1, …, varIdxN])

varIdx: field_index varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수 중 가장 작은 값을 반환합니다. 인덱스를 지정하지 않으면 전체 변수들을 대상으로 계산합니다. 텍스트 논리값 포함(텍스트 =0 으로 인식, false =0 , true =1)

RMINVAR([varIdx1, …, varIdxN])

varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수 중 가장 작은 값을 가지는 변수명을 반환합니다. 인덱스를 지정하지 않으면 전체 연속형 변수를 대상으로 합니다.

RMINVARIDX([varIdx1, …, varIdxN])

varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수 중 가장 작은 값을 가지는 변수의 인덱스를 반환합니다. 인덱스를 지정하지 않으면 전체 연속형 변수를 대상으로 합니다.

RRANK(ranking, [varIdx1, …, varIdxN])

Ranking: 순위, varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수들 중 ranking 번째 값의 변수값을 반환합니다. 인덱스를 지정하지 않으면 전체 변수 중 연속형인 변수만 고려합니다.

RRANKIDX(ranking, [varIdx1, …, varIdxN])

Ranking: 순위, varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수들 중 ranking 번째 값의 변수 인덱스를 반환합니다. 인덱스를 지정하지 않으면 전체 변수 중 연속형인 변수만 고려합니다.

RRMS([varIdx1, …, varIdxN])

varIdx: field_index

지정된 인덱스에 해당하는 변수들의 제곱평균제곱근을 계산합니다. 인덱스를 지정하지 않으면 전체 변수 중 연속형인 변수의 제곱평균제곱근을 계산합니다.

RSTD([varIdx1, …, varIdxN])

varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수들의 표준편차를 계산합니다. 인덱스를 지정하지 않으면 전체 변수 중 연속형인 변수의 표준편차를 계산합니다.

RSTDA([varIdx1, …, varIdxN])

varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수들을 표본집단의 표준편차로 계산하여 반환합니다. 인덱스를 지정하지 않으면 전체 변수들을 대상으로 계산합니다. 결측치 제외 텍스트 포함(텍스트 = 0)

RSTDP([varIdx1, …, varIdxN])

varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수들을 모집단의 표준편차로 계산하여 반환합니다. 인덱스를 지정하지 않으면 전체 변수들을 대상으로 계산합니다. 결측치 텍스트 제외.

RSTDPA([varIdx1, …, varIdxN])

varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수들을 모집단의 표준편차로 계산하여 반환합니다. 인덱스를 지정하지 않으면 전체 변수들을 대상으로 계산합니다. 결측치 제외, 텍스트 포함(텍스트 = 0)

RSUM(([varIdx1, …, varIdxN])

varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수들의 총합을 계산합니다. 인덱스를 지정하지 않으면 전체 변수 중 연속형인 변수의 총합을 계산합니다.

RUSS(([varIdx1, …, varIdxN])

varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 각 변수들의 제곱합을 계산합니다. 인덱스를 지정하지 않으면 전체 변수 중 연속형인 변수의 제곱평균제곱근을 계산합니다.

RVAR(([varIdx1, …, varIdxN])

varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수들의 분산을 계산합니다. 인덱스를 지정하지 않으면 전체 변수 중 연속형인 변수의 분산을 계산합니다.

RVARA(([varIdx1, …, varIdxN])

varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수들을 표본집단의 분산으로 계산하여 반환합니다. 인덱스를 지정하지 않으면 전체 변수들을 대상으로 계산합니다. 결측치 제외 텍스트 포함(텍스트 = 0)

RVARP(([varIdx1, …, varIdxN])

varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수들을 모집단의 분산을 계산하여 반환합니다. 인데스를 지정하지 않으면 전체 변수들을 대상으로 계산합니다. 텍스트 제외.

RVARPA(([varIdx1, …, varIdxN])

varIdx: field_index

현재 열에서 지정된 인덱스에 해당하는 변수들을 모집단의 분산을 계산하여 반환합니다. 인덱스를 지정하지 않으면 전체 변수들을 대상으로 계산합니다. 결측치 제외, 텍스트 포함(텍스트 = 0)

A1.2 매크로 및 기타 함수

매크로를 사용하여 기 정의된 값을 사용할 수 있습니다. 매크로를 사용하려면 @을 입력한 뒤 사용하고자 하는 값의 문자열을 입력합니다.

매크로

@NULL

NULL 값을 입력합니다.

@ROWNUM

현재 처리중인 레코드의 인덱스를 반환합니다.

@VAR

다중 파생필드 생성시 계산중인 변수를 가르킵니다. 다중 파생필드에서만 사용됩니다.

@VARIDX

다중 파생필드 생성시 계산중인 변수의 인덱스를 반환합니다. 다중 파생필드에서만 사용됩니다.

기타 함수

GETDEFDATE(key)

지정한 문자열 key 에 해당하는 기-정의된 날짜/시간 값을 반환합니다.

GETDEFNUM(key)

지정한 문자열 key 에 해당하는 기-정의된 숫자 값을 반환합니다.

GETDEFSTR(key)

지정한 문자열 key 에 해당하는 기-정의된 문자열 값을 반환합니다.

GETROWNUM(field_index, val)

val: 임의값

지정된 변수 인덱스(field_index)에 해당하는 변수에서 val 에 지정된 값을 찾아 해당 행의 번호를 반환합니다.

GETVALUE(field_index, row)

Row: 임의값

지정된 변수 인덱스(field_index)에 해당하는 row 번째 값을 반환합니다.

LARGEST(field_index, k)

K: 임의값

현재 열에서 지정된 인덱스(field_index)에 해당하는 변수들중 결측치를 제외한 데이터 중 k 번째로 큰 값을 반환합니다.

LOOKUP(field_index1, val, field_index2)

Val: 임의값

지정된 변수 인덱스 var1 에 해당하는 변수에서 val 에 지정된 값을 찾아 같은 위치(행)에 있는 var2 의 값을 반환합니다.

MOVINGAVERAGE(field_index, k)

K: 임의값

현재 열에서 지정된 인덱스에 해당하는 변수의 이동평균을 구해줍니다. k 는 이동평균의 길이입니다.

MOVINGMEDIAN(field_index, k)

K: 임의값

현재 열에서 지정된 인덱스에 해당하는 변수의 이동 중위값을 구해줍니다. k 는 이동 중위값의 길이입니다.

MOVINGSTD(field_index, k)

K: 임의값

지정된 인덱스에 해당하는 변수의 이동 표준편차를 구해줍니다. K 는 이동 표준편차의 길이입니다.

REGSLOPE(field_index, k)

K: 임의값

시간을 X 축으로 할 때, 지정된 인덱스에 해당하는 변수의 회귀 계수의 기울기를 반환합니다. k 는 시간의 길이입니다.

SMALLEST(field_index, k)

K: 임의값

현재 열에서 지정된 인덱스에 해당하는 변수들을 오름차순으로 정렬하여 k 번째 값의 변수값을 반환합니다.

* GETDEFDATE,GETDEFNUM,GETDEFSTR 함수의 경우 ECMiner2014 가 설치된 위치의 'ECMvalue.txt'파일에 기-정의된 내용이 있어야합니다. ECMvalue.txt 는 정해진 형태로 구성되어야 합니다. 첫번째 열에는 key 문자열이 구성되며, 두번째 열은 반환타입(D:date/N:number/S:string), 세번째 열은 반환값으로 구성되어 있어야합니다. 각 열의 구분은 tab 으로 구분하여야 합니다.

A1.3 값 입력

상수값

ECMiner™에서 상수값을 입력하기 위하여 다음과 같은 규칙을 따라야 합니다.

- 문자형 데이터

  문자형 데이터를 직접 입력하려면 따옴표를 사용합니다. 즉, "입력하려는 문자열" 혹은 '입력하려는 문자열'과 같이 입력합니다.

- 날짜/시간 데이터

  날짜/시간형 데이터를 입력하려면 #을 사용합니다. 즉, #2004-01-01 13:10:30#라고 입력하면 '2004 년 1 월 1 일 오후 1 시 10 분 30 초'로 인식합니다. 시간이 필요 없을 경우 #2004-01-01# 식으로 입력하면 날짜만 인식합니다.

- 숫자형 데이터

  아무런 기호 없이 숫자를 입력합니다. 예를 들어 1234.5, 1234, 1.234e3 등을 사용할 수 있습니다.

변수값

변수값은 {변수명}와 같이 입력합니다. 변수명이 ECMiner™에서 사용되는 예약어와 같은 경우를 방지하고 모든 형태의 변수명을 지원하기 위하여 {}을 사용하여 변수를 지정하도록 하였습니다.