



버전 18

# 다변량 방법

“진정 무엇인가를 발견하는 여행은 새로운 풍경을 바라보는  
것이 아니라 새로운 눈을 가지는 데 있다.”

Marcel Proust

JMP Statistical Discovery LLC  
920 SAS Campus Drive  
Cary, North Carolina 27513-2414

The correct bibliographic citation for this manual is as follows: JMP Statistical Discovery LLC 2024. *JMP® 18 Multivariate Methods*. Cary, NC: JMP Statistical Discovery LLC

## **JMP® 18 Multivariate Methods**

Copyright © 2024, JMP Statistical Discovery LLC, Cary, NC, USA

All rights reserved. Produced in the United States of America.

JMP Statistical Discovery LLC, 920 SAS Campus Drive, Cary, North Carolina 27513-2414.

March 2024

JMP® and all other JMP Statistical Discovery LLC product or service names are registered trademarks or trademarks of SAS Institute Inc. or JMP Statistical Discovery LLC in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

JMP software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For more information about third-party software distributed with JMP software, refer to <https://www.jmp.com/thirdpartysoftware>.

## **JMP 활용하기**

JMP를 처음 사용하든 오랫동안 사용해왔든, JMP와 관련해서 배워야 하는 것은 언제나 있습니다.

JMP.com을 방문하면 다음과 같은 유용한 자료를 볼 수 있습니다.

- JMP 시작 방법에 대한 라이브 및 녹화 웹 캐스트
- 새로운 기능과 고급 기법에 대한 비디오 데모 및 웹 캐스트
- JMP 교육 과정 등록에 대한 상세 정보
- 현지에서 개최되는 세미나 일정
- 다른 사용자들이 JMP를 사용하는 방법을 보여주는 성공 사례
- JMP 사용자 커뮤니티, 추가기능 및 스크립트 예를 포함한 사용자용 리소스, 포럼, 블로그, 컨퍼런스 정보 등

<https://www.jmp.com/getstarted>

# 목차

## 다변량 방법

---

<b>1</b>	<b>JMP 알아보기</b> .....	11
	<b>설명서 및 추가 리소스</b>	
	JMP Pro .....	13
	JMP 온라인 도움말 .....	13
	설명서 PDF 추가기능 .....	13
	JMP 도움말 메뉴 .....	21
	JMP 학습을 위한 추가 리소스 .....	22
	JMP 시작하기 .....	23
	JMP 검색 .....	23
	샘플 데이터 테이블 .....	23
	JSL 알아보기 .....	24
	JMP 툴팁 .....	24
	JMP 사용자 커뮤니티 .....	24
	무료 온라인 Statistical Thinking 교육 과정 .....	24
	JMP 새로운 사용자를 위한 입문 키트 .....	25
	Statistics Knowledge 포털 .....	25
	JMP 교육 과정 .....	25
	사용자가 작성한 JMP 설명서 .....	25
	JMP 시작하기 창 .....	25
	JMP 기술 지원 .....	26
<b>2</b>	<b>다변량 분석 소개</b> .....	27
	<b>다변량 기법 개요</b>	
<b>3</b>	<b>상관 및 다변량 기법</b> .....	29
	<b>변수의 다차원 동작 탐색</b>	
	다변량 플랫폼의 예 .....	31
	다변량 플랫폼 시작 .....	33
	다변량 보고서 .....	34
	다변량 플랫폼 옵션 .....	35
	산점도 행렬 .....	39
	이상치 분석 .....	40

항목 신뢰도 .....	42
부분 상관 다이어그램 .....	43
다변량 플랫폼의 추가 예 .....	44
항목 신뢰도의 예 .....	44
부분 상관의 예 .....	45
다변량 플랫폼에 대한 통계 상세 정보 .....	48
분산 추정 방법에 대한 통계 상세 정보 .....	49
Pearson 곱적률 상관계수에 대한 통계 상세 정보 .....	49
비모수 연관성 측도에 대한 통계 상세 정보 .....	50
역상관 행렬에 대한 통계 상세 정보 .....	51
거리 측도에 대한 통계 상세 정보 .....	52
Cronbach $\alpha$ 에 대한 통계 상세 정보 .....	54
<b>4 주성분</b> .....	<b>55</b>
<b>데이터 차원 축소</b>	
주성분 플랫폼 개요 .....	57
주성분 분석의 예 .....	57
주성분 플랫폼 시작 .....	58
결측 데이터 .....	62
주성분 보고서 .....	62
주성분 보고서 옵션 .....	63
이상치 분석 .....	73
주성분 플랫폼에 대한 통계 상세 정보 .....	75
분산 추정 방법에 대한 통계 상세 정보 .....	76
고급 방법에 대한 통계 상세 정보 .....	76
이상치 분석 계산에 대한 통계 상세 정보 .....	77
<b>5 판별 분석</b> .....	<b>79</b>
<b>연속형 변수를 기반으로 분류 예측</b>	
판별 플랫폼 개요 .....	81
판별 분석의 예 .....	81
판별 플랫폼 시작 .....	83
단계별 변수 선택 .....	84
판별 방법 .....	86
공분산 축소 .....	89
판별 분석 보고서 .....	89
주성분 .....	90
정준 그림 및 정준 구조 .....	91
판별 스코어 .....	94
스코어 요약 .....	95

판별 분석 옵션 .....	98
정준 상세 정보 표시 .....	102
정준 구조 표시 .....	104
새 수준 고려 .....	104
판별 행렬 저장 .....	104
JMP 및 JMP Pro 에서 검증 .....	105
판별 분석의 추가 예 .....	106
3D 정준 그림의 예 .....	106
단계별 변수 선택의 예 .....	107
판별 플랫폼에 대한 통계 상세 정보 .....	108
와이드 선형 알고리즘에 대한 통계 상세 정보 .....	109
저장된 계산식에 대한 통계 상세 정보 .....	109
다변량 검정에 대한 통계 상세 정보 .....	116
근사 F- 검정에 대한 통계 상세 정보 .....	117
그룹 간 공분산 행렬에 대한 통계 상세 정보 .....	118
<b>6 부분 최소 제곱 모형</b> .....	<b>119</b>
<b>Y와 X 간의 상관을 사용하여 모형 개발</b>	
부분 최소 제곱 플랫폼 개요 .....	121
부분 최소 제곱의 예 .....	122
부분 최소 제곱 플랫폼 시작 .....	125
중심화 및 척도화 .....	128
X 표준화 .....	128
모형 시작 제어판 .....	129
부분 최소 제곱 옵션 .....	130
부분 최소 제곱 보고서 .....	131
모형 비교 요약 .....	131
교차 검증 보고서 .....	132
모형 적합 보고서 .....	136
모형 적합 옵션 .....	137
변수 중요도 그림 .....	141
VIP 대 계수 그림 .....	142
부분 최소 제곱의 추가 예 .....	143
부분 최소 제곱 플랫폼에 대한 통계 상세 정보 .....	145
부분 최소 제곱에 대한 통계 상세 정보 .....	145
van der Voet $T^2$ 검정에 대한 통계 상세 정보 .....	146
$T^2$ 그림에 대한 통계 상세 정보 .....	147
X 스코어 산점도 행렬의 신뢰도 타원에 대한 통계 상세 정보 .....	147
예측 및 신뢰 한계에 대한 통계 상세 정보 .....	148
표준화된 스코어 및 적재량에 대한 통계 상세 정보 .....	149

	PLS 관별 분석에 대한 통계 상세 정보 .....	150
<b>7</b>	<b>다중 대응 분석</b> .....	151
	<b>범주형 변수 수준 간의 연관성 식별</b>	
	다중 대응 분석의 예 .....	153
	다중 대응 분석 플랫폼 시작 .....	156
	다중 대응 분석 보고서 .....	157
	다중 대응 분석 플랫폼 옵션 .....	158
	그림 표시 .....	161
	상세 정보 표시 .....	161
	조정 관성 표시 .....	161
	좌표 표시 .....	162
	요약 통계량 표시 .....	162
	관성에 대한 부분 기여도 표시 .....	163
	제공코사인 표시 .....	163
	Cochran Q 검정 .....	164
	교차표 .....	164
	다중 대응 분석의 추가 예 .....	164
	보조 변수를 사용한 예 .....	165
	보조 ID 를 사용한 예 .....	166
	Cochran Q 검정의 예 .....	167
	다중 대응 분석 플랫폼에 대한 통계 상세 정보 .....	168
	상세 정보 보고서에 대한 통계 상세 정보 .....	169
	조정 관성에 대한 통계 상세 정보 .....	170
	요약 통계량에 대한 통계 상세 정보 .....	170
	Cochran Q 통계량에 대한 통계 상세 정보 .....	171
<b>8</b>	<b>구조 방정식 모형</b> .....	173
	<b>구조 방정식 모형 적합</b>	
	구조 방정식 모형 개요 .....	175
	구조 방정식 모형의 유형 .....	175
	구조 방정식 모형의 예 .....	179
	구조 방정식 모형 플랫폼 시작 .....	182
	구조 방정식 모형 보고서 .....	183
	모형 규격 보고서 .....	184
	모형 비교 보고서 .....	192
	카이제곱 차이 검정 보고서 .....	193
	구조 방정식 모형 적합 보고서 .....	194
	구조 방정식 모형 플랫폼 옵션 .....	195
	모형 옵션 .....	197

경로 다이어그램 사용자 정의 .....	202
경로 다이어그램 팝업 메뉴 옵션 .....	202
경로 다이어그램 사용자 정의 옵션 .....	204
구조 방정식 모형의 추가 예 .....	205
잠재 경로 변수 모형의 예 .....	205
잠재 성장 곡선 모형의 예 .....	211
측정 모형 평가 보고서의 예 .....	213
다중 그룹 분석의 예 .....	216
중재 분석의 예 .....	218
구조 방정식 모형 플랫폼에 대한 통계 상세 정보 .....	222
추정 방법에 대한 통계 상세 정보 .....	223
적합 요약 측도에 대한 통계 상세 정보 .....	223
중재 분석에 대한 통계 상세 정보 .....	228
<b>9 요인 분석 .....</b>	<b>231</b>
<b>데이터의 잠재 변수 식별</b>	
요인 분석 플랫폼 개요 .....	233
요인 분석 플랫폼의 예 .....	233
요인 분석 플랫폼 시작 .....	236
요인 분석 보고서 .....	236
모형 시작 .....	238
회전 방법 .....	239
요인 분석 플랫폼 옵션 .....	240
요인 분석 모형 적합 옵션 .....	241
<b>10 다차원 척도법 .....</b>	<b>245</b>
<b>개체 집합 간의 근접성 시각화</b>	
다차원 척도법 플랫폼 개요 .....	247
다차원 척도법의 예 .....	247
다차원 척도법 플랫폼 시작 .....	250
다차원 척도 보고서 .....	251
다차원 척도 그림 .....	251
Shepard 다이어그램 .....	252
적합 상세 정보 .....	252
다차원 척도법 플랫폼 옵션 .....	252
Waern 링크 .....	253
다차원 척도법의 추가 예 .....	254
다차원 척도법 플랫폼에 대한 통계 상세 정보 .....	256
스트레스 함수에 대한 통계 상세 정보 .....	256
변환에 대한 통계 상세 정보 .....	256

속성 목록 형식에 대한 통계 상세 정보 .....	257
<b>11 다변량 임베딩</b> .....	259
<b>고차원 데이터를 저차원 공간에 매핑</b>	
다변량 임베딩 플랫폼 개요 .....	261
다변량 임베딩의 예 .....	262
다변량 임베딩 플랫폼 시작 .....	263
다변량 임베딩 보고서 .....	266
다변량 임베딩 플랫폼 옵션 .....	266
다변량 임베딩의 추가 예 .....	267
다변량 임베딩 플랫폼에 대한 통계 상세 정보 .....	269
t-SNE 방법에 대한 통계 상세 정보 .....	269
경사하강법 알고리즘에 대한 통계 상세 정보 .....	271
<b>12 항목 분석</b> .....	273
<b>항목 및 대상별 검정 결과 분석</b>	
항목 분석의 예 .....	275
항목 분석 플랫폼 시작 .....	277
로지스틱 3PL 모형 상세 정보 .....	278
데이터 형식 .....	278
항목 분석 보고서 .....	278
특성 곡선 .....	278
정보 그림 .....	279
쌍대 그림 .....	280
모수 추정값 .....	280
항목 분석 플랫폼 옵션 .....	281
항목 분석 플랫폼에 대한 통계 상세 정보 .....	281
항목 반응 곡선에 대한 통계 상세 정보 .....	282
항목 반응 곡선 모형에 대한 통계 상세 정보 .....	282
IRT 모형 가정에 대한 통계 상세 정보 .....	284
IRT 모형 적합에 대한 통계 상세 정보 .....	285
능력 계산식에 대한 통계 상세 정보 .....	286
<b>13 계층적 군집화</b> .....	287
<b>군집 트리를 사용하여 관측값 그룹화</b>	
계층적 군집화 플랫폼 개요 .....	289
관측값 군집화 플랫폼 개요 .....	289
계층적 군집화의 예 .....	291
계층적 군집화 플랫폼 시작 .....	293
계층적 군집화 보고서 .....	298
덴드로그램 .....	298

군집화 기록 .....	299
계층적 군집화 플랫폼 옵션 .....	299
계층적 군집화의 추가 예 .....	303
거리 행렬의 예 .....	304
공간 측도를 사용한 웨이퍼 결함 분류의 예 .....	306
계층적 군집화 플랫폼에 대한 통계 상세 정보 .....	308
공간 측도에 대한 통계 상세 정보 .....	309
거리 방법에 대한 통계 상세 정보 .....	310
근접 이웃 결합 주기에 대한 통계 상세 정보 .....	311
<b>14 K 평균 군집화 .....</b>	<b>313</b>
<b>거리를 사용하여 관측값 그룹화</b>	
K 평균 군집화 플랫폼 개요 .....	315
관측값 군집화 플랫폼 개요 .....	315
K 평균 군집화의 예 .....	316
K 평균 군집화 플랫폼 시작 .....	320
반복적 군집화 보고서 .....	320
반복적 군집화 옵션 .....	322
K 평균 보고서 .....	322
군집 비교 보고서 .....	322
K 평균 군집 수 보고서 .....	323
자기 조직화 지도 .....	325
자기 조직화 지도 제어판 .....	325
자기 조직화 지도 보고서 .....	326
SOM 알고리즘 설명 .....	326
K 평균 군집화의 추가 예 .....	327
<b>15 정규 혼합 .....</b>	<b>331</b>
<b>확률을 사용하여 관측값 그룹화</b>	
정규 혼합 플랫폼 개요 .....	333
관측값 군집화 플랫폼 개요 .....	333
정규 혼합 군집화의 예 .....	334
정규 혼합 플랫폼 시작 .....	336
정규 혼합 보고서 .....	337
정규 혼합 옵션 .....	338
개별 정규 혼합 보고서 .....	338
군집 비교 보고서 .....	339
정규 혼합 군집 수 보고서 .....	339
정규 혼합 플랫폼에 대한 통계 상세 정보 .....	341

<b>16</b>	<b>잠재 계층 분석</b> .....	343
	<b>범주형 변수의 관측값 그룹화</b>	
	잠재 계층 분석 플랫폼 개요 .....	345
	관측값 군집화 플랫폼 개요 .....	345
	잠재 계층 분석의 예 .....	346
	잠재 계층 분석 플랫폼 시작 .....	349
	잠재 계층 분석 보고서 .....	350
	군집 비교 보고서 .....	350
	잠재 계층 모형 보고서 .....	350
	잠재 계층 분석 플랫폼 옵션 .....	352
	잠재 계층 모형 옵션 .....	353
	잠재 계층 분석 플랫폼의 추가 예 .....	354
	잠재 계층 분석 플랫폼에 대한 통계 상세 정보 .....	355
	잠재 계층 모형 적합에 대한 통계 상세 정보 .....	355
	최대 군집 수에 대한 통계 상세 정보 .....	356
<b>17</b>	<b>변수 군집화</b> .....	357
	<b>유사한 변수를 대표 그룹으로 그룹화</b>	
	변수 군집화 플랫폼 개요 .....	359
	변수 군집화 플랫폼의 예 .....	359
	변수 군집화 플랫폼 시작 .....	361
	변수 군집화 보고서 .....	361
	상관 색상 맵 .....	362
	군집 요약 .....	362
	군집 멤버 .....	362
	표준화 성분 .....	363
	변수 군집화 플랫폼 옵션 .....	363
	변수 군집화 플랫폼의 추가 예 .....	364
	상관 색상 맵의 예 .....	364
	차원 축소를 위한 변수 군집화 플랫폼의 예 .....	365
	변수 군집화 플랫폼에 대한 통계 상세 정보 .....	368
<b>A</b>	<b>통계 상세 정보</b> .....	371
	<b>다변량 방법</b>	
	와이드 선형 방법 및 특이값 분해 .....	373
	특이값 분해 .....	373
<b>B</b>	<b>참조 자료</b> .....	377

# 1 장

## JMP 알아보기 설명서 및 추가 리소스

---

JMP Pro 지정, JMP 설명서 추가기능, 각 JMP 문서에 대한 설명, 도움말 메뉴 옵션, 기타 지원 제공 위치 등 JMP 설명서에 대해 알아봅니다.

## 목차

JMP Pro.....	13
JMP 온라인 도움말 .....	13
설명서 PDF 추가기능 .....	13
JMP 도움말 메뉴 .....	21
JMP 학습을 위한 추가 리소스 .....	22
JMP 시작하기 .....	23
JMP 검색.....	23
샘플 데이터 테이블 .....	23
JSL 알아보기.....	24
JMP 톨팁.....	24
JMP 사용자 커뮤니티 .....	24
무료 온라인 Statistical Thinking 교육 과정 .....	24
JMP 새로운 사용자를 위한 입문 키트.....	25
Statistics Knowledge 포털 .....	25
JMP 교육 과정 .....	25
사용자가 작성한 JMP 설명서 .....	25
JMP 시작하기 창 .....	25
JMP 기술 지원 .....	26

---

## JMP Pro

JMP Pro 전용 기능은 JMP Pro 아이콘 **JMP PRO** 으로 표시됩니다. JMP Pro 기능의 개요는 <https://www.jmp.com/software/pro>에서 확인할 수 있습니다.

---

## JMP 온라인 도움말

JMP 온라인 도움말에서는 JMP 기능, 통계적 방법 및 JSL(JMP 스크립트 언어)에 대한 정보를 검색할 수 있습니다. 다음과 같은 몇 가지 방법으로 JMP 온라인 도움말을 열 수 있습니다.

- Windows의 경우 **도움말 > JMP 온라인 도움말**을 선택합니다.
- macOS의 경우 **도움말 > JMP 도움말**을 선택합니다.
- Windows에서 F1 키를 누릅니다.
- 데이터 테이블 또는 보고서 창의 특정 부분에 대한 도움말을 보려면 **도움말 > 도움말 도구**를 선택합니다. 그런 다음 데이터 테이블 또는 보고서 창의 아무 곳이나 클릭합니다. 도움말 도구를 닫으려면 Esc 키를 누릅니다.
- JMP 창 내에서 **도움말** 버튼을 클릭합니다.

---

**참고:** JMP 도움말을 사용하려면 인터넷에 연결되어 있어야 합니다. 인터넷에 연결되지 않은 사용자는 설명서 추가기능을 설치하여 도움말을 볼 수 있습니다. 자세한 내용은 "[설명서 PDF 추가기능](#)"에서 확인하십시오.

---

---

## 설명서 PDF 추가기능

JMP 설명서 추가기능을 다운로드하여 설치할 수 있습니다. 설명서 추가기능에는 JMP 라이브러리에 있는 각 문서의 개별 PDF와 JMP 설명서 라이브러리 파일이 포함되어 있습니다. JMP 설명서 라이브러리 파일은 개별 설명서 PDF 파일이 포함된 단일 PDF 파일입니다. 사용자는 JMP 온라인 도움말과 유사한 단일 PDF 파일에서 모든 설명서를 검색할 수 있습니다.

설명서 추가기능이 설치되면 "도움말" 메뉴에 "설명서 PDF" 옵션이 추가되고 컴퓨터에 PDF 파일이 설치됩니다. 그러면 **도움말 > 설명서 PDF**를 선택하여 설명서에 로컬로 액세스할 수 있습니다. <https://www.jmp.com/doc-addin>에서 사용 가능한 설명서 추가기능을 다운로드하십시오.

다음 표에서는 설명서 추가기능에 포함된 각 문서의 용도와 내용을 설명합니다.

문서 제목	문서 용도	문서 내용
JMP 설명서 라이브러리	개별 설명서 PDF 파일의 단일 PDF를 제공합니다.	모든 JMP 설명서가 하나의 PDF에 포함되어 있습니다.
JMP 살펴보기	JMP에 익숙하지 않다면 이 설명서부터 시작하십시오.	JMP를 소개하고 데이터 생성 및 분석, 결과 공유를 시작하기 위한 정보를 제공합니다.
JMP 사용	JMP 데이터 테이블과 기본적인 작업을 수행하는 방법에 대해 알아봅니다.	데이터 가져오기, 열 특성 수정, 데이터 정렬, 워크플로우 빌더 사용 등 JMP의 모든 영역에 걸친 일반적인 JMP 개념과 기능을 다룹니다.
기본 분석	이 문서를 사용하여 기본적인 분석을 수행합니다.	<p>다음과 같은 분석 메뉴 플랫폼에 대해 설명합니다.</p> <ul style="list-style-type: none"> <li>• 분포</li> <li>• X로 Y 적합</li> <li>• 테이블 생성</li> <li>• 텍스트 탐색기</li> </ul> <p>분석 &gt; X로 Y 적합을 통해 이변량 분석, 일원 ANOVA 및 분할 분석을 수행하는 방법을 다룹니다. 또한 붓스트랩을 사용하여 표집 분포에 근사한 값을 산출하는 방법과 시뮬레이션 플랫폼을 사용하여 모수 재표집을 수행하는 방법도 설명합니다.</p>

문서 제목	문서 용도	문서 내용
그래프 분석	데이터에 이상적인 그래프를 찾습니다.	<p>다음과 같은 그래프 메뉴 플랫폼에 대해 설명합니다.</p> <ul style="list-style-type: none"> <li>• 그래프 빌더</li> <li>• 3D 산점도</li> <li>• 등고선 그림</li> <li>• 버블 그림</li> <li>• 평행 그림</li> <li>• 셀 그림</li> <li>• 산점도 행렬</li> <li>• 삼원 그림</li> <li>• 트리맵</li> <li>• 차트</li> <li>• 중첩 그림</li> </ul> <p>이 설명서에서는 배경 맵 및 사용자 맵을 생성하는 방법도 다룹니다.</p>
프로파일러	반응 표면의 횡단면을 볼 수 있게 해주는 대화식 프로파일링 도구의 사용 방법을 알아봅니다.	"그래프" 메뉴에 나열된 모든 프로파일러를 다룹니다. 랜덤 입력을 사용한 시뮬레이션 실행과 함께 잡음 요인 분석이 포함됩니다.
실험 설계 가이드	실험 설계 방법을 알아보고 적절한 표본 크기를 결정합니다.	"DOE" 메뉴의 모든 항목을 다룹니다.

문서 제목	문서 용도	문서 내용
선형 모형 적합	모형 적합 플랫폼과 이 플랫폼의 다양한 분석법에 대해 알아봅니다.	<p>"분석" 메뉴의 모형 적합 플랫폼에서 사용할 수 있는 다음 분석법에 대해 설명합니다.</p> <ul style="list-style-type: none"> <li>• 표준 최소 제곱</li> <li>• 단계별</li> <li>• 일반화 회귀</li> <li>• 혼합 모형</li> <li>• 일반화 선형 혼합 모형</li> <li>• 다변량 분산분석</li> <li>• 로그 선형 분산</li> <li>• 명목형 로지스틱</li> <li>• 순서형 로지스틱</li> <li>• 일반화 선형 모형</li> </ul>

문서 제목	문서 용도	문서 내용
예측 및 전문 모델링	추가 모델링 기법에 대해 알아봅니다.	<p>다음과 같은 분석 &gt; 예측 모델링 메뉴 플랫폼 폼에 대해 설명합니다.</p> <ul style="list-style-type: none"> <li>• 신경망</li> <li>• 파티션</li> <li>• 붓스트랩 포레스트</li> <li>• 부스티드 트리</li> <li>• K 최근접 이웃</li> <li>• 나이브 베이즈</li> <li>• 서포트 벡터 머신</li> <li>• 모형 비교</li> <li>• 모형 선별</li> <li>• 검증 열 생성</li> <li>• 계산식 저장소</li> </ul> <p>다음과 같은 분석 &gt; 전문 모델링 메뉴 플랫폼 폼에 대해 설명합니다.</p> <ul style="list-style-type: none"> <li>• 곡선 적합</li> <li>• 비선형</li> <li>• 함수 데이터 탐색기</li> <li>• 가우시안 과정</li> <li>• 시계열</li> <li>• 시계열 예측</li> <li>• 매칭 쌍</li> </ul> <p>다음과 같은 분석 &gt; 선별 메뉴 플랫폼에 대해 설명합니다.</p> <ul style="list-style-type: none"> <li>• 이상치 탐색</li> <li>• 결측값 탐색</li> <li>• 패턴 탐색</li> <li>• 반응 변수 선별</li> <li>• 예측 변수 선별</li> <li>• 연관성 분석</li> <li>• 공정 기록 탐색기</li> </ul>

문서 제목	문서 용도	문서 내용
다변량 방법	여러 변수를 동시에 분석하는 방법을 알아봅니다.	<p>다음과 같은 분석 &gt; 다변량 방법 메뉴 플랫폼에 대해 설명합니다.</p> <ul style="list-style-type: none"> <li>• 다변량</li> <li>• 주성분</li> <li>• 관별</li> <li>• 부분 최소 제곱</li> <li>• 다중 대응 분석</li> <li>• 구조 방정식 모형</li> <li>• 요인 분석</li> <li>• 다차원 척도법</li> <li>• 다변량 임베딩</li> <li>• 항목 분석</li> </ul> <p>다음과 같은 분석 &gt; 군집화 메뉴 플랫폼에 대해 설명합니다.</p> <ul style="list-style-type: none"> <li>• 계층적 군집화</li> <li>• K 평균 군집화</li> <li>• 정규 혼합</li> <li>• 잠재 계층 분석</li> <li>• 변수 군집화</li> </ul>

문서 제목	문서 용도	문서 내용
품질 및 공정 방법	공정 평가 및 개선을 위한 도구에 대해 알아봅니다.	<p>다음과 같은 분석 &gt; 품질 및 공정 메뉴 플랫폼에 대해 설명합니다.</p> <ul style="list-style-type: none"> <li>• 관리도 빌더 및 개별 관리도</li> <li>• 측정 시스템 분석 (EMP 및 유형 1 게이지)</li> <li>• 계량형 / 계수형 게이지 차트</li> <li>• 공정 변수 선별</li> <li>• 공정 능력</li> <li>• 모형 기반 다변량 관리도</li> <li>• 레거시 관리도</li> <li>• 파레토도</li> <li>• 다이어그램</li> <li>• 한계 관리</li> <li>• OC 곡선</li> </ul>
신뢰성 및 생존 방법	제품 또는 시스템의 신뢰도 평가 및 향상 방법과 사람 및 제품의 생존 데이터 분석 방법을 알아봅니다.	<p>다음과 같은 분석 &gt; 신뢰성 및 생존 메뉴 플랫폼에 대해 설명합니다.</p> <ul style="list-style-type: none"> <li>• 수명 분포</li> <li>• 수명 분포 적합</li> <li>• 누적 손상</li> <li>• 피로 모형</li> <li>• 재발 분석</li> <li>• 반복 측정 열화</li> <li>• 파괴 열화</li> <li>• 신뢰도 예측</li> <li>• 신뢰도 성장</li> <li>• 신뢰도 블록 다이어그램</li> <li>• 수리 가능 시스템 시뮬레이션</li> <li>• 생존</li> <li>• 모수 생존 모형 적합</li> <li>• 열화</li> <li>• 비례 위험 모형 적합</li> </ul>

문서 제목	문서 용도	문서 내용
소비자 조사	소비자 선호도를 연구하고 더 나은 제품과 서비스를 개발하는 방법을 알아봅니다.	다음과 같은 분석 > 소비자 조사 메뉴 플랫폼에 대해 설명합니다. <ul style="list-style-type: none"> <li>• 범주형</li> <li>• 선택</li> <li>• 최대차이</li> <li>• 업리프트</li> <li>• 다중 요인 분석</li> </ul>
유전학	유전자 데이터를 분석한 후 최적의 유전자 교배를 예측하기 위한 육종 프로그램을 시뮬레이션하는 방법을 알아봅니다.	다음과 같은 분석 > 유전학 메뉴 플랫폼에 대해 설명합니다. <ul style="list-style-type: none"> <li>• 표지자 통계량</li> <li>• 표지자 시뮬레이션</li> </ul>
Scripting Guide	강력한 JSL(JMP 스크립트 언어)에 대해 알아봅니다.	스크립트 작성/디버깅, 데이터 테이블 조작, 표시 상자 생성 및 JMP 응용 프로그램 생성 등의 다양한 주제를 다룹니다.
JSL Syntax Reference	JSL 함수 인수 및 메시지에 대해 알아봅니다.	JSL 명령의 구문, 예제 및 참고 사항이 포함되어 있습니다.
바로 가기 키	키보드를 사용하여 JMP를 빠르게 탐색하고 작업을 완료하는 방법을 알아봅니다.	Windows 및 macOS용 명령과 해당 키 입력이 포함되어 있습니다.
메뉴 설명	JMP의 메뉴 항목에 대해 알아봅니다.	Windows 및 macOS용 메뉴 옵션에 대해 설명합니다.

## JMP 도움말 메뉴

JMP 18부터 JMP 도움말 메뉴가 업데이트되었습니다.

메뉴 항목	설명
JMP 검색	JMP에서 통계적 검정 및 기타 기능을 검색할 수 있습니다. 자세한 내용은 <a href="#">"JMP 검색"</a> 에서 확인하십시오.
JMP 온라인 도움말	웹 브라우저에서 최신 버전의 도움말을 열 수 있습니다.
도움말 도구	데이터 테이블 또는 보고서 창의 아무 곳이나 클릭하여 도움말을 볼 수 있습니다.
빠른 시작	빠른 시작(이전의 "오늘의 유익한 정보")은 JMP의 기본 사항을 빠르게 학습하는 데 유용한 팁을 제공합니다. 자세한 내용은 <a href="#">"JMP 시작하기"</a> 에서 확인하십시오.
설명서 PDF	설치되면 JMP 설명서 PDF 파일에 로컬로 액세스할 수 있습니다. 자세한 내용은 <a href="#">"설명서 PDF 추가기능"</a> 에서 확인하십시오.  <b>참고 :</b> 이 메뉴 옵션은 설명서 추가기능을 다운로드하여 설치한 경우에만 나타납니다.
JMP 기능	JMP에서 사용할 수 있는 도구와 기능을 나열하는 웹 브라우저를 엽니다. 또한 자세한 정보를 확인할 수 있는 온라인 도움말 링크도 제공합니다.
JMP 알아보기	JMP 학습 자료로 연결되는 웹 브라우저를 엽니다. 짧은 비디오와 기타 리소스를 통해 JMP에 대해 알아볼 수 있습니다.

메뉴 항목	설명
JMP 사용자 커뮤니티	학습, 문제 해결, JMP 개선 아이디어 공유를 위해 다른 JMP 사용자와 교류할 수 있는 웹 브라우저를 엽니다. 자세한 내용은 " <a href="#">JMP 사용자 커뮤니티</a> "에서 확인하십시오.
JMP의 새로운 기능	최신 JMP 릴리스의 새로운 기능에 대해 알아볼 수 있는 웹 브라우저를 엽니다.
샘플 데이터 폴더	JMP 분석을 자세히 알아보기 위해 샘플 데이터에 액세스할 수 있습니다. 샘플 데이터 파일을 열고 스크립트를 실행하여 샘플 분석을 확인할 수 있습니다. 자세한 내용은 " <a href="#">샘플 데이터 테이블</a> "에서 확인하십시오.
샘플 인덱스	분석 유형 또는 분야에 따른 샘플 데이터 테이블, 교육 자료 및 추가 샘플 자료 링크를 찾을 수 있습니다. 자세한 내용은 " <a href="#">샘플 데이터 테이블</a> "에서 확인하십시오.
스크립트 인덱스	JMP 스크립트 명령을 검색하고 사용 방법을 배울 수 있습니다. 자세한 내용은 " <a href="#">JSL 알아보기</a> "에서 확인하십시오.
My JMP	웹에서 <a href="http://my.jmp.com">my.jmp.com</a> 을 엽니다.
JMP 정보	JMP 버전이 표시되고 소프트웨어 업데이트를 확인할 수 있습니다. 이 옵션은 Windows에서만 사용할 수 있습니다.

## JMP 학습을 위한 추가 리소스

JMP 도움말 외에도 다음 리소스를 사용하여 JMP에 대해 배울 수 있습니다.

- "[JMP 시작하기](#)"
- "[JMP 검색](#)"
- "[샘플 데이터 테이블](#)"
- "[JSL 알아보기](#)"

- "JMP 톨팁 "
- "JMP 사용자 커뮤니티 "
- " 무료 온라인 Statistical Thinking 교육 과정 "
- "JMP 새로운 사용자를 위한 입문 키트 "
- "Statistics Knowledge 포털 "
- "JMP 교육 과정 "
- " 사용자가 작성한 JMP 설명서 "
- "JMP 시작하기 창 "

## JMP 시작하기

JMP를 처음 시작하면 새로운 사용자가 JMP 사용을 시작하는 데 도움이 되는 "빠른 시작" 창이 나타납니다. "빠른 시작"을 해제하려면 **시작할 때 '빠른 시작' 표시** 체크박스를 선택 해제하십시오. 다시 표시하려면 **도움말 > 빠른 시작**을 선택합니다. 또는 "환경 설정" 창을 사용하여 이 기능을 해제할 수도 있습니다.

## JMP 검색

통계 절차의 위치를 잘 모르면 JMP에서 검색을 수행하십시오. 결과는 데이터 테이블 또는 보고서와 같이 검색 시작 창에 맞게 조정됩니다.

1. **도움말 > JMP 검색**을 클릭합니다. 또는 Ctrl+ 쉼표를 누릅니다.
2. 검색 텍스트를 입력합니다.
3. 원하는 절차가 포함된 결과를 클릭합니다.  
오른쪽에는 절차에 대한 설명과 위치가 표시됩니다.
4. 해당 버튼을 클릭하여 결과를 열거나 이동합니다.

## 샘플 데이터 테이블

JMP 설명서 모음에 포함된 모든 예에서는 샘플 데이터를 사용합니다. 샘플 데이터 디렉토리를 열려면 **도움말 > 샘플 데이터 폴더**를 선택하십시오.

샘플 데이터 테이블의 사전순 목록을 보거나 범주별로 샘플 데이터를 보려면 **도움말 > 샘플 인덱스**를 선택하십시오.

샘플 데이터 테이블은 다음 디렉토리에 설치되어 있습니다.

Windows: C:\Program Files\JMP\JMP\18\Samples\Data

macOS: /Library/Application Support/JMP/18/Samples/Data

JMP Pro의 경우에는 JMP 디렉터리가 아니라 JMPPRO 디렉터리에 샘플 데이터가 설치되어 있습니다.

샘플 데이터를 사용한 예를 보려면 **도움말 > 샘플 인덱스**를 선택하고 "교육용 예제"로 이동하십시오.

## JSL 알아보기

JSL 스크립트 및 예제에 대한 도움말을 보려면 **도움말 > 스크립트 인덱스**를 선택합니다. 스크립트 인덱스를 사용하여 JSL 함수, 개체 및 표시 상자에 대한 정보를 검색할 수 있습니다. 예제 스크립트를 편집 및 실행하고 명령에 대한 도움말을 볼 수 있습니다.

## JMP 툴팁

JMP에서 다음과 같은 항목을 커서로 가리키면 설명 툴팁 (또는 가리키기 라벨)이 제공됩니다.

- 메뉴 또는 도구 모음 옵션
- 그래프의 라벨
- 보고서 창의 텍스트 결과 (커서를 원 모양으로 움직이면 표시됨)
- 홈 창의 파일 또는 창
- 스크립트 편집기의 코드

---

**팁:** Windows의 경우 JMP 환경 설정에서 툴팁을 숨길 수 있습니다. **파일 > 환경 설정 > 일반**을 선택한 후 **메뉴 팁 표시**를 선택 취소합니다. 이 옵션은 macOS에서는 사용할 수 없습니다.

---

## JMP 사용자 커뮤니티

JMP 사용자 커뮤니티에서는 JMP에 대해 알아보고 다른 JMP 사용자와 교류하는 데 도움이 되는 다양한 옵션을 제공합니다. 한 페이지 분량의 가이드, 자습서 및 데모로 구성된 학습 라이브러리부터 시작하는 것이 좋습니다. 다양한 JMP 교육 과정에 등록하여 학습을 계속할 수도 있습니다.

그 밖에도 토론 포럼, 샘플 데이터 및 스크립트 파일 교환, 웹 캐스트 및 소셜 네트워킹 그룹을 비롯한 리소스가 있습니다.

웹 사이트의 JMP 리소스에 액세스하려면 **도움말 > JMP 사용자 커뮤니티**를 선택하거나 <https://community.jmp.com>을 방문하십시오.

## 무료 온라인 Statistical Thinking 교육 과정

이 무료 온라인 교육 과정에서는 탐색적 데이터 분석, 품질 관리 방법, 상관 및 회귀 등의 항목에 대한 실용적인 통계적 기술을 배울 수 있습니다. 이 교육 과정은 짧은 비디오와 데모, 연습 등으로 구성되어 있습니다. 자세한 내용은 <https://www.jmp.com/statisticalthinking>에서 확인하십시오.

## JMP 새로운 사용자를 위한 입문 키트

JMP 새로운 사용자를 위한 입문 키트는 JMP의 기본 사항을 빨리 익힐 수 있도록 돕기 위한 것입니다. 30개의 짧은 데모 비디오 및 작업을 마치면 좀 더 편하게 소프트웨어 사용 방법을 익히고 세계 최대 규모의 JMP 사용자 온라인 커뮤니티와 연결할 수 있습니다. 자세한 내용은 <https://www.jmp.com/welcome>에서 확인하십시오.

## Statistics Knowledge 포털

Statistics Knowledge 포털에서는 방문자가 확실한 기초를 토대로 통계적 기술을 쌓을 수 있도록 간략한 통계 설명과 함께 명확한 예시 및 그래픽을 제공합니다. 자세한 내용은 <https://www.jmp.com/skp>에서 확인하십시오.

## JMP 교육 과정

JMP에서는 숙련된 JMP 전문가 팀의 주도로 다양한 주제에 대한 교육 과정을 제공합니다. 공개 교육, 라이브 웹 교육, 현장 교육 등이 제공되며, 온라인 e-learning 구독을 선택하여 편리한 시간에 학습할 수도 있습니다. 자세한 내용은 <https://www.jmp.com/training>에서 확인하십시오.

## 사용자가 작성한 JMP 설명서

JMP 웹 사이트에서는 JMP 사용자가 작성한 추가 JMP 사용 설명서가 제공됩니다. 자세한 내용은 <https://www.jmp.com/books>에서 확인하십시오.

## JMP 시작하기 창

JMP 또는 데이터 분석에 익숙하지 않다면 먼저 "JMP 시작하기" 창을 살펴보십시오. 이 창에는 옵션이 범주별로 설명되어 있으며 버튼을 클릭하여 옵션을 시작할 수 있습니다. "JMP 시작하기" 창에는 "분석", "그래프", "테이블" 및 "파일" 메뉴에 있는 다양한 옵션이 포함됩니다. 또한 이 창에는 JMP Pro의 기능 및 플랫폼도 나열됩니다.

- "JMP 시작하기" 창을 열려면 **보기** (macOS의 경우 **창**) > **JMP 시작하기**를 선택합니다.
- Windows에서 JMP를 열 때 자동으로 "JMP 시작하기"를 표시하려면 **파일** > **환경 설정** > **일반**을 선택한 후 "초기 JMP 창" 목록에서 **JMP 시작하기**를 선택합니다. macOS에서는 **JMP** > **환경 설정** > **일반** > **초기 JMP 시작하기 창**을 선택합니다.

---

## JMP 기술 지원

JMP 기술 지원은 JMP 교육을 받은 통계학자와 엔지니어가 제공하며, 이들 중 상당수는 통계 또는 기타 기술 분야의 석사 학위를 갖고 있습니다.

기술 지원 전화 번호를 포함한 많은 기술 지원 옵션이 <https://www.jmp.com/support>에서 제공됩니다.

## 다변량 분석 소개 다변량 기법 개요

다변량 방법에서는 여러 변수를 동시에 분석하는 다음 기법에 대해 설명합니다.

- 다변량 플랫폼에서는 여러 변수가 서로 어떤 관계가 있는지 검토합니다. 자세한 내용은 "[상관 및 다변량 기법](#)"에서 확인하십시오.
- 주성분 플랫폼에서는 측정된 변수 집합 중 원래 변수의 변동성을 가능한 한 많이 포착하는 소수의 독립적 선형 결합 (주성분) 을 도출합니다. 이 방법은 유용한 탐색 기법이며 예측 모형을 생성하는 데 도움이 될 수 있습니다. 자세한 내용은 "[주성분](#)"에서 확인하십시오.
- 관별 플랫폼에서는 알려진 연속형 반응 (Y) 을 기반으로 명목형 또는 순서형 분류 (X) 변수를 예측하는 방법을 찾습니다. 이 방법은 MANOVA(다변량 분산분석)의 역추정 예측으로 간주할 수 있습니다. 자세한 내용은 "[관별 분석](#)"에서 확인하십시오.
- 부분 최소 제곱 플랫폼에서는 요인, 즉 설명 변수 (X)의 선형 결합을 기반으로 선형 모형을 적합시킵니다. PLS는 X와 Y 사이의 상관을 이용하여 기본 잠재 구조를 나타냅니다. 자세한 내용은 "[부분 최소 제곱 모형](#)"에서 확인하십시오.
- MCA(다중 대응 분석) 플랫폼에서는 여러 범주형 변수를 사용하여 해당 변수 수준 간의 연관성을 식별합니다. MCA는 특히 프랑스와 일본에서 사회 과학 분야에 주로 사용됩니다. 설문 조사 분석에서 질문 합치도를 식별하는 데 사용할 수 있습니다. 자세한 내용은 "[다중 대응 분석](#)"에서 확인하십시오.
- **JMP PRO** 구조 방정식 모형 플랫폼에서는 확증적 요인 분석, 잠재 변수를 사용하거나 사용하지 않는 경로 모형, 측정 오차 모형 등의 다양한 모형을 적합시킬 수 있습니다. 자세한 내용은 "[구조 방정식 모형](#)"에서 확인하십시오.
- 요인 분석 플랫폼을 사용하면 더 큰 관측 변수 집합에서 요인을 구성할 수 있습니다. 이러한 요인은 관측 변수 부분집합의 선형 결합으로 표현됩니다. 요인 분석을 사용하면 측정 및 관측된 변수 집합으로 설명되는 요인의 수, 요인과 변수 간 관계의 강도 등을 탐색할 수 있습니다. 자세한 내용은 "[요인 분석](#)"에서 확인하십시오.
- MDS(다차원 척도법) 플랫폼을 사용하면 개체 집합 간의 근접성 패턴(유사성, 비유사성 또는 거리)을 시각적으로 표현할 수 있습니다. 자세한 내용은 "[다차원 척도법](#)"에서 확인하십시오.
- **JMP PRO** 다변량 임베딩 플랫폼을 사용하면 고차원 공간의 데이터를 저차원 공간에 매핑할 수 있습니다. 자세한 내용은 "[다변량 임베딩](#)"에서 확인하십시오.
- 항목 분석 플랫폼을 사용하면 항목 반응 이론 모형을 적합시킬 수 있습니다. IRT(항목 반응 이론) 방법은 검사 및 설문지와 같이 측정기의 분석 및 스코어링에 사용됩니다. IRT는 개인이 어떤 항목을 지지하거나 올바르게 응답할 확률을 특성 또는 능력과 연관시키기 위해 모형

시스템을 사용합니다. 표준화 검사, 인지 발달 및 소비자 선호도를 연구할 때 IRT 를 사용할 수 있습니다. 자세한 내용은 "항목 분석"에서 확인하십시오.

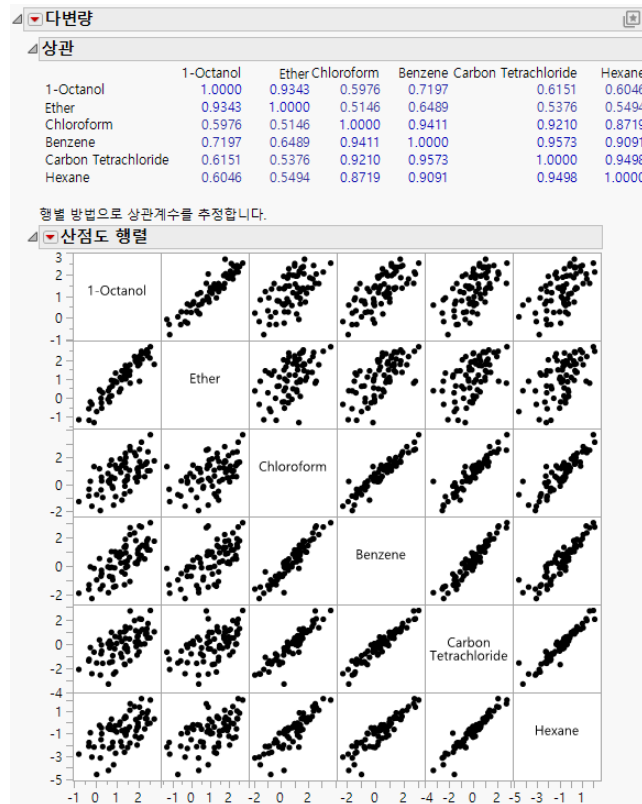
- 계층적 군집화 플랫폼에서는 여러 변수에 대해 유사한 값을 공유하는 행을 함께 그룹화합니다. 이 방법은 데이터의 군집 구조를 이해하는 데 도움이 되는 유용한 탐색 기법입니다. 자세한 내용은 "계층적 군집화"에서 확인하십시오.
- K 평균 군집화 플랫폼에서는 여러 변수에 대해 유사한 값을 공유하는 관측값을 그룹화합니다. 자세한 내용은 "K 평균 군집화"에서 확인하십시오.
- 정규 혼합 플랫폼을 사용하면 중첩 정규 분포의 데이터를 가져올 때 관측값을 군집화할 수 있습니다. 자세한 내용은 "정규 혼합"에서 확인하십시오.
- 잠재 계층 분석 플랫폼에서는 범주형 반응 변수에 대한 관측값 군집을 찾습니다. 모형은 다항 혼합 모형의 형태를 취합니다. 자세한 내용은 "잠재 계층 분석"에서 확인하십시오.
- 변수 군집화 플랫폼에서는 유사한 변수를 대표 그룹으로 그룹화합니다. 변수 군집화를 차원 축소 방법으로 사용할 수 있습니다. 큰 변수 집합을 사용하여 모델링하는 대신 군집에서 가장 대표적인 변수의 군집 성분을 사용하여 데이터 변동의 대부분을 설명할 수 있습니다. 자세한 내용은 "변수 군집화"에서 확인하십시오.

## 상관 및 다변량 기법 변수의 다차원 동작 탐색

다변량 데이터에는 하나(단변량) 또는 두 개(이변량)가 아니라 많은 변수가 포함됩니다. 다변량 플랫폼을 사용하여 여러 변수가 서로 어떤 관련이 있는지 탐색할 수 있습니다. 다변량 플랫폼에서는 각 반응 변수 쌍 간의 선형 관계 강도를 요약하고 검정하는 다양한 기법을 제공합니다. 모수 상관 검정과 비모수 상관 검정 모두 플랫폼에서 사용할 수 있습니다. 또한 산점도 행렬과 색상 맵 같은 그래픽 기능을 사용하여 변수 사이의 종속성, 이상치 및 군집을 식별할 수도 있습니다.

주성분 분석, 이상치 분석 및 항목 신뢰도를 포함하여 변수 간의 관계를 더 자세히 조사할 수 있는 추가 다변량 분석 기법이 있습니다. 이러한 기법은 "다변량" 보고서를 통해 사용할 수 있습니다. JMP의 주성분 분석 플랫폼과 이상치 분석 플랫폼을 사용하여 이러한 기법을 더 심층적으로 구현할 수 있습니다.

**그림 3.1** 다변량 보고서의 예



## 목차

다변량 플랫폼의 예 .....	31
다변량 플랫폼 시작 .....	33
다변량 보고서 .....	34
다변량 플랫폼 옵션 .....	35
산점도 행렬 .....	39
이상치 분석 .....	40
항목 신뢰도 .....	42
부분 상관 다이어그램 .....	43
다변량 플랫폼의 추가 예 .....	44
항목 신뢰도의 예 .....	44
부분 상관의 예 .....	45
다변량 플랫폼에 대한 통계 상세 정보 .....	48
분산 추정 방법에 대한 통계 상세 정보 .....	49
Pearson 곱적률 상관계수에 대한 통계 상세 정보 .....	49
비모수 연관성 측도에 대한 통계 상세 정보 .....	50
역상관 행렬에 대한 통계 상세 정보 .....	51
거리 측도에 대한 통계 상세 정보 .....	52
Cronbach $\alpha$ 에 대한 통계 상세 정보 .....	54

---

## 다변량 플랫폼의 예

이 예에서는 다변량 플랫폼에서 상관 행렬과 색상 맵을 생성하여 다양한 신체 측정값 간의 관계를 검토합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Body Measurements.jmp** 를 엽니다.

2. **분석 > 다변량 방법 > 다변량**을 선택합니다.

3. **Mass** 를 제외한 모든 열을 선택하고 **Y, 열**을 클릭합니다.

4. **확인**을 클릭합니다.

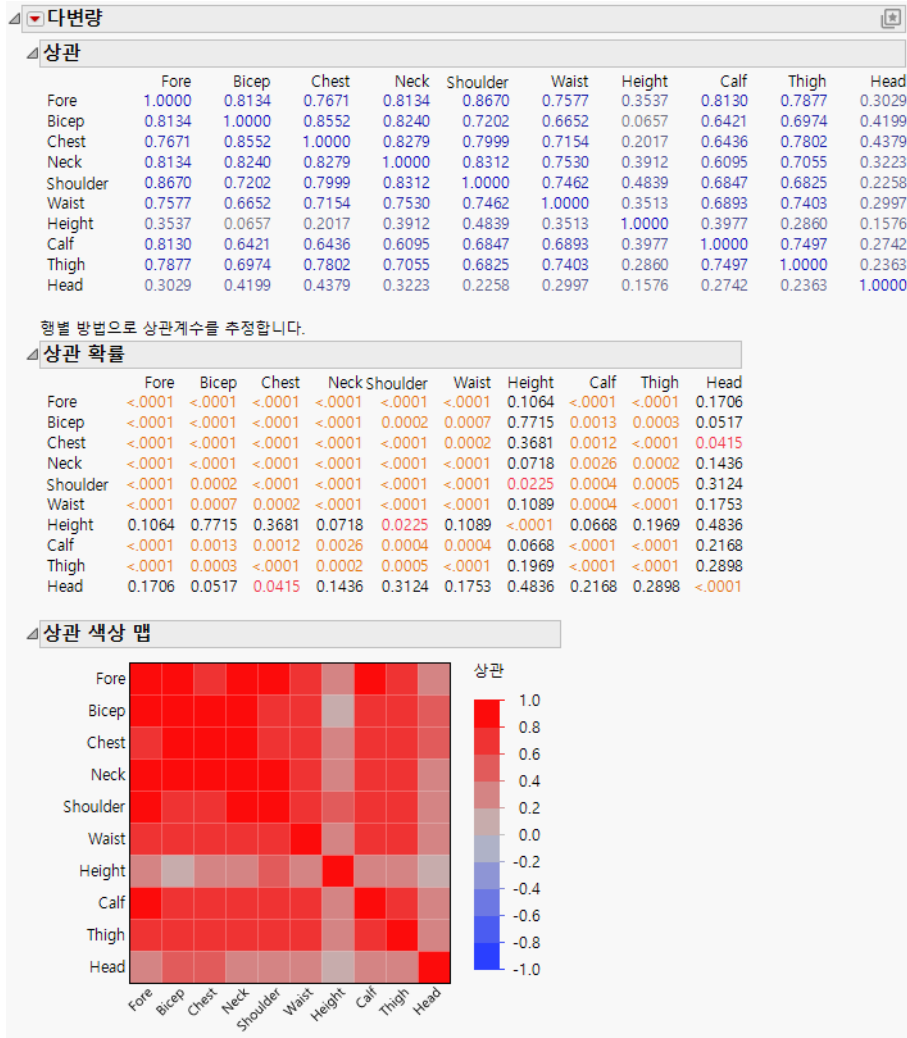
초기 다변량 보고서에는 상관 행렬과 산점도 행렬이 포함됩니다. 행별 분산 추정 방법이 사용되었음을 알려 주는 노트도 있습니다. 모든 변수는 양의 상관관계가 있지만 강도가 다양합니다.

5. "다변량"의 빨간색 삼각형을 클릭하고 **산점도 행렬**을 선택 취소합니다.

6. "다변량"의 빨간색 삼각형을 클릭하고 **상관 확률**을 선택합니다.

7. "다변량"의 빨간색 삼각형을 클릭하고 **색상 맵 > 상관 색상 맵**을 선택합니다.

그림 3.2 Body Measurements 데이터에 대한 다변량 보고서

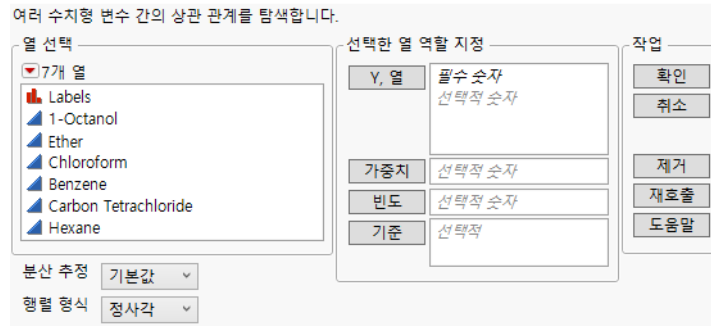


"상관 색상 맵" 보고서는 상관 행렬의 정보를 더 간결하게 나타냅니다. 그림의 대부분은 진한 빨간색으로 표시되어 대부분의 변수가 높은 상관관계가 있음을 나타냅니다. 더 연하게 표시된 두 행과 열은 Height(키)와 Head(머리 둘레) 측정값이 다른 변수와 큰 상관관계가 없음을 나타냅니다. 이 사실은 "상관 확률" 테이블에서 Height와 Head의  $p$  값이 대부분 유의하지 않음을 통해 뒷받침됩니다.

## 다변량 플랫폼 시작

분석 > 다변량 방법 > 다변량을 선택하여 다변량 플랫폼을 시작합니다.

그림 3.3 다변량 시작 창



"열 선택"의 빨간색 삼각형 메뉴에 포함된 옵션에 대한 자세한 내용은 **JMP 사용**에서 확인하십시오.

**Y, 열** 하나 이상의 반응 열을 식별합니다. 반응 열의 데이터 유형은 숫자여야 하지만 모델링 유형은 연속형 또는 순서형일 수 있습니다.

**참고 :** 순서형 반응 변수를 지정하는 경우 시작 창에서 " 확인 " 을 클릭하면 **JMP** 경고가 나타납니다. 이 경고는 어떤 변수가 순서형인지 나타내고 사용자가 분석에 순서형 변수를 포함하려고 했음을 확인합니다.

**가중치** 분석의 각 행에 가중치를 할당하는 숫자 값이 포함된 하나의 열을 식별합니다.

**빈도** 분석의 각 행에 빈도를 할당하는 숫자 값이 포함된 하나의 열을 식별합니다.

**기준** 기준 변수의 각 수준에 대해 개별 보고서를 생성합니다. 기준 변수가 둘 이상 할당되면 기준 변수의 가능한 각 수준 조합에 대해 개별 보고서가 생성됩니다.

**분산 추정** 상관 계산 방법을 지정합니다. "REML" 과 "쌍별" 방법이 가장 자주 사용됩니다. 이러한 방법 중 일부는 결측 데이터를 처리합니다. "행렬" 을 제외한 다른 방법을 사용하고 "결측 데이터 대치" 명령을 선택하여 결측값을 추정할 수도 있습니다. 자세한 내용은 "**결측 데이터 대치**" 에서 확인하십시오.

**기본값** "기본값" 옵션은 "행렬", "쌍별" 또는 "REML" 방법을 사용합니다.

- 행렬 추정은 결측값이 없는 데이터 테이블에 사용됩니다.
- 쌍별 추정은 결측값이 있고 열이 10 개, 행이 5,000 개를 각각 초과하거나 행보다 열이 더 많은 데이터 테이블에 사용됩니다.
- REML 추정은 그 외의 경우에 사용됩니다.

**참고 :** "기본값" 옵션이 "REML" 추정을 사용하지만 적합이 올바르게 수렴되지 않으면 플랫폼이 "쌍별" 방법으로 돌아갑니다. 이는 결측값이 있는 데이터 테이블에서 열이 10 개 미만이고, 행이 5,000 개 미만이며, 열이 행보다 적어야 한다는 모든 조건을 충족할 때 발생할 수 있습니다. 표시된 분산 추정 방법이 "쌍별" 이면 REML 적합이 수렴되지 않았음을 의미합니다.

**REML** REML(제한 최대 가능도) 추정은 결측값이 있는 경우에도 모든 데이터를 사용합니다. 데이터 집합이 크고 결측값이 많은 경우 이 방법을 사용하면 편향 수정 요인 때문에 처리 속도가 느립니다. 따라서 REML 은 작은 데이터 집합에 가장 유용합니다. 데이터에 결측 셀이 없으면 REML 및 ML 추정값이 동일하고 표본 공분산 행렬과 같습니다. 결측 셀이 있으면 REML 의 분산 및 공분산 추정값이 ML 추정의 추정값보다 덜 편향됩니다. 통계 상세 정보는 "REML" 에서 확인하십시오.

**ML** ML(최대 가능도) 추정은 결측값이 있는 경우에도 모든 데이터를 사용합니다. ML 추정값은 빠르게 생성되므로 이 방법은 결측 데이터가 있는 큰 데이터 테이블에 가장 유용합니다.

**로버스트** 로버스트 추정은 결측값이 있는 경우에도 모든 데이터를 사용합니다. 이 방법은 극단값의 비중을 낮추므로 이상치가 있을 수 있는 데이터 테이블에 유용합니다. 통계 상세 정보는 "로버스트" 에서 확인하십시오.

**행별** 행별 추정은 각 열 쌍에 대한 Pearson 상관을 계산합니다. 통계 상세 정보는 "Pearson 곱적률 상관계수에 대한 통계 상세 정보" 에서 확인하십시오. 행별 추정에서는 결측값이 있는 행을 사용하지 않습니다. 이 방법은 결측 데이터가 있는 관측값을 제외하는 데 유용합니다.

**쌍별** 쌍별 추정은 결측값이 있는 경우에도 모든 데이터를 사용합니다. 이 분산 추정 방법은 두 열에 대해 결측값이 없는 모든 관측값을 사용하여 각 열 쌍에 대한 Pearson 상관을 계산합니다. 통계 상세 정보는 "Pearson 곱적률 상관계수에 대한 통계 상세 정보" 에서 확인하십시오. 쌍별 추정은 결측값이 있는 데이터 테이블에서 행보다 열이 더 많거나, 열이 10 개를 초과하거나, 행이 5,000 개를 초과하는 경우에 가장 유용합니다.

**참고 :** "REML", "ML" 또는 "로버스트" 를 선택한 경우 데이터 테이블에 행보다 열이 더 많고 결측값이 있으면 JMP 에서 "분산 추정" 이 "쌍별" 로 전환됩니다.

**행렬 형식** 산점도 행렬의 형식 옵션을 선택합니다. "정사각" 옵션은 정렬된 모든 열 조합에 대한 그림을 표시합니다. "하삼각" 은 대각선 아래에 그림을 표시하며 처음  $n-1$  개 열이 가로 축에 나타납니다. "상삼각" 은 대각선 위에 그림을 표시하며 처음  $n-1$  개 열이 세로 축에 나타납니다.

## 다변량 보고서

기본 다변량 보고서에는 표준 상관 행렬, 산점도 행렬, 상관 추정에 사용된 방법을 나타내는 노트가 표시됩니다. 경우에 따라 지정된 방법이 사용된 이유를 설명하는 정보도 표시됩니다. 플랫

폼 메뉴에는 여러 변수를 검토하기 위한 추가 상관 옵션 및 기타 기법이 나열되어 있습니다. 자세한 내용은 "다변량 플랫폼 옵션"에서 확인하십시오.

---

## 다변량 플랫폼 옵션

"다변량"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

**다변량 상관** "상관" 테이블을 표시하거나 숨깁니다. 이 테이블은 각 반응 ( $Y$ ) 변수 쌍 간의 선형 관계 강도가 요약된 상관계수 행렬입니다. 이 옵션은 기본적으로 설정되어 있습니다. 자세한 내용은 "Pearson 곱적률 상관계수에 대한 통계 상세 정보"에서 확인하십시오.

---

**참고:** 이 상관 행렬은 시작 창에서 선택한 방법으로 계산됩니다.

**상관 확률** "상관 확률" 테이블을 표시하거나 숨깁니다. 이 테이블은  $p$  값 행렬입니다. 각  $p$  값은 변수 사이의 실제 상관이 0이라는 귀무가설 검정에 해당합니다. 즉, 두 반응 변수 사이에 선형 관계가 없는지에 대한 검정입니다.

**상관 CI** 상관의 양쪽 꼬리 신뢰 구간을 표시하거나 숨깁니다.

---

**팁:** 기본 신뢰 계수는 95%입니다. 신뢰 계수를 변경하려면 " $\alpha$  수준 설정" 옵션을 사용하십시오.

**역상관** 역상관 행렬을 나타내는 "역상관" 보고서를 표시하거나 숨깁니다. 행렬의 대각 요소는 변수가 다른 변수의 선형 함수에 얼마나 가까운지에 대한 함수입니다. 역상관 테이블에서 대각선은  $1/(1 - R^2)$ 의 값입니다.  $R^2$ 은 다른 모든 변수를 기반으로 해당 변수를 모델링하는 단순 선형 모형에서 계산됩니다. 다중상관이 0이면 대각선 역요소가 1입니다. 다중상관이 1이면 역요소가 무한대가 되고 결국으로 보고됩니다. 역상관에 대한 통계 상세 정보는 "역상관 행렬에 대한 통계 상세 정보"에서 확인하십시오.

**부분 상관** 부분 상관 행렬을 나타내는 "부분 상관" 보고서를 표시하거나 숨깁니다. 부분 상관 행렬에는 다른 모든 변수의 효과를 조정 한 후 변수 쌍 사이의 관계에 대한 측도가 표시됩니다. 이 테이블은 역상관 행렬에 음수를 취하고, 단위 대각값으로 척도화된 것입니다. 즉, 대각 요소가 1이 되도록 행렬이 척도화됩니다.

**부분 상관 확률**  $p$  값 행렬을 나타내는 "부분 상관 확률" 보고서를 표시하거나 숨깁니다. 각  $p$  값은 변수 사이의 실제 부분 상관이 0이라는 귀무가설 검정에 해당합니다. 즉, 다른 변수의 효과를 조정 한 후 두 반응 변수 사이에 선형 관계가 없는지에 대한 검정입니다.

---

**참고:** 자유도가 충분하지 않으면 "부분 상관 확률" 옵션을 사용할 수 없습니다. 이는 관측 값보다 변수가 더 많을 때 발생할 수 있습니다.

**공분산 행렬** 변수 쌍이 함께 변화하는 정도를 측정하는 공분산 행렬을 표시하거나 숨깁니다.

**쌍별 상관** 각  $Y$  변수 쌍에 대한 Pearson 곱적률 상관계수가 나열된 "쌍별 상관" 보고서를 표시하거나 숨깁니다. 상관은 쌍별 제거 방법으로 계산됩니다. 변수 쌍 중 어느 한 변수에 대해

결측값이 있으면 개수 값이 달라집니다. 또한 "쌍별 상관" 보고서는 유의 확률을 보여 주고 막대 차트로 상관관계를 비교합니다. 모든 결과는 쌍별 방법을 기반으로 합니다.

---

**참고 :** 이 옵션은 고려 중인 변수 중 하나에 대해 결측인 행을 제외합니다.

---

**Hotelling  $T^2$  검정** Y로 입력한 변수의 다변량 분포 평균에 대한 1 표본 검정을 수행할 수 있습니다. "Hotelling  $T^2$  검정" 옵션을 선택하면 귀무가설 하에서 평균 벡터를 지정할 수 있는 창이 나타납니다. 각 변수에 대해 가설 평균을 입력합니다. 이 검정에서는 Y 변수의 다변량 정규성을 가정합니다. "Hotelling  $T^2$  검정" 보고서에는 다음 정보가 제공됩니다.

**변수** Y로 입력된 변수입니다.

**평균** 각 변수의 표본 평균입니다.

**가설 평균** 지정한 귀무가설 평균입니다.

**검정 통계량** Hotelling  $T^2$  통계량 값입니다.

**F 비** 검정 통계량 값입니다.  $n$  개의 행과  $k$  개의 변수가 있는 경우 F 비는 다음과 같이 정의됩니다.

$$\frac{n-k}{k(n-1)} T^2$$

**Prob > F** 검정의  $p$  값입니다. 귀무가설 하에서 F 비는  $k$  및  $n-k$  자유도의 F 분포를 따릅니다.

---

**참고 :** 보고서를 제거하려면 "Hotelling  $T^2$  검정" 옆의 빨간색 삼각형을 클릭하고 **검정 제거**를 선택합니다.

---

**단순 통계량** 이 메뉴에는 각 열에 대한 단순 통계량 (평균, 표준편차, 합, 최소값 및 최대값)을 표시하거나 숨기는 두 가지 옵션이 포함되어 있습니다. 결측값이 있거나 "로버스트" 방법을 사용하는 경우 단변량 및 다변량 단순 통계량이 다를 수 있습니다.

**단변량 단순 통계량** 다른 열의 값이나 결측값에 관계없이 각 열에 대해 계산되는 통계량을 표시합니다. 이러한 값은 분포 플랫폼에서 생성되는 값과 일치합니다.

**다변량 단순 통계량** 시작 창에서 선택한 분산 추정 방법과 결측 데이터가 있는지 여부에 부합하는 통계량을 표시합니다. 결측 관측값이 없으면 "로버스트" 방법에만 이 옵션을 사용할 수 있습니다. 결측 관측값이 있으면 "쌍별"을 제외한 모든 분산 추정 방법에 이 옵션을 사용할 수 있습니다. "REML", "ML" 또는 "로버스트" 방법의 경우 선택한 방법으로 평균 벡터와 공분산 행렬이 추정됩니다. "행별" 방법의 경우 하나 이상의 결측값이 있는 모든 행이 평균 및 분산 계산에서 제외됩니다.

**비모수 상관** 쌍별 상관에 대한 비모수 측도가 포함된 하위 메뉴를 표시합니다. 각 옵션은 선택한 연관성 측도에 대한 유의 확률을 제공하고 연관성 값을 막대 차트에 표시하는 비모수 보고서를 표시하거나 숨깁니다. 세 가지 비모수 상관 측도가 있습니다.

**Spearman Rho** 데이터 값 대신 데이터 값의 순위에 대해 계산된 상관계수입니다.

**Kendall Tau** 관측값의 부합 / 비부합 쌍의 개수를 기반으로 합니다. 관측값의 X 값이 더 클 때 Y 값도 더 큰 쌍이 부합 쌍이고, 관측값의 X 값이 더 클 때 Y 값이 더 작은 쌍은 비부합 쌍입니다. 동일 쌍, 즉 X 값이 같거나 Y 값이 같은 관측값 쌍에 대해서는 수정이 적용됩니다.

**Hoeffding D** -0.5 ~ 1 범위의 통계 척도입니다. 큰 양수 값은 종속성을 나타냅니다. 이 통계량은 2x2 분류 테이블에 대한 카이제곱 통계량의 관측값에 대해 가중 합에 근사한 값을 산출합니다. 2x2 테이블은 각 데이터 값을 임계값으로 설정하여 만들어집니다. 이 통계량은 일반적인 독립성 이탈을 더 많이 감지합니다.

---

**참고:** 비모수 상관은 시각 창에서 다른 분산 추정 방법을 선택한 경우에도 쌍별 방법을 사용하여 계산됩니다.

---

**참고:** 가중치 변수가 지정된 경우 결측 및 0 값 가중치는 비모수 상관 계산에서 제외됩니다. 다른 모든 가중치 값은 1로 처리됩니다.

---

이 세 가지 방법에 대한 통계 상세 정보는 "[비모수 연관성 측도에 대한 통계 상세 정보](#)"에서 확인하십시오.

**$\alpha$  수준 설정** 상관 신뢰 구간에 대한 알파 값을 지정할 수 있습니다. 0.01, 0.05, 0.10 및 0.50의 네 가지 알파 값이 나열됩니다. 다른 값을 입력하려면 **기타**를 선택합니다.

**산점도 행렬** 각 반응 변수 쌍의 산점도 행렬을 표시하거나 숨깁니다. 이 옵션은 기본적으로 설정되어 있습니다. 자세한 내용은 "[산점도 행렬](#)"에서 확인하십시오.

**색상 맵** "색상 맵" 메뉴에는 7개 유형의 색상 맵이 포함되어 있습니다. 메뉴의 각 옵션은 해당 유형의 색상 맵을 표시하거나 숨깁니다. 다음 유형의 색상 맵을 사용할 수 있습니다.

**상관 색상 맵** 빨간색 (+1) ~ 파란색 (-1)의 척도로 변수 간 상관을 보여 주는 셀 그림을 생성합니다.

**p 값 색상 맵**  $p=0$ (빨간색) ~  $p=1$ (파란색)의 척도로 상관 유의성을 보여 주는 셀 그림을 생성합니다.

**상관 군집화** 유사한 변수를 함께 군집화하는 셀 그림을 생성합니다. 상관은 "상관 색상 맵"의 상관과 같지만 변수 위치가 다를 수 있습니다.

**쌍별 상관 색상 맵** 파란색 (+1) ~ 빨간색 (-1)의 척도로 변수 간 쌍별 상관을 보여 주는 셀 그림을 생성합니다.

**Spearman  $\rho$  색상 맵** 파란색 (+1) ~ 빨간색 (-1)의 척도로 변수 간 Spearman  $\rho$  비모수 상관을 보여 주는 셀 그림을 생성합니다.

**Kendall  $\tau$  색상 맵** 파란색 (+1) ~ 빨간색 (-1)의 척도로 변수 간 Kendall  $\tau$  비모수 상관을 보여 주는 셀 그림을 생성합니다.

**Hoeffding D 색상 맵** 파란색 (+1) ~ 빨간색 (-1)의 척도로 변수 간 Hoeffding D 비모수 상관을 보여 주는 셀 그림을 생성합니다.

**평행 좌표 그림** 변수의 평행 좌표 그림을 표시하거나 숨깁니다.

**타원 3D 그림** 95% 신뢰 타원체가 있는 3 차원 산점도를 표시하거나 숨깁니다. 이 옵션을 선택하면 세 개의 변수와 해당 축을 지정해야 합니다.

**부분 상관 다이어그램** 부분 상관 다이어그램 보고서를 표시하거나 숨깁니다. 이 옵션은 부분 상관 행렬에 대해 고유값 분해를 수행하고, 결과를 사용하여 부분 상관의 시각적 표현을 제공합니다. 다이어그램 표시를 변경할 수 있는 몇 가지 옵션이 있습니다. 자세한 내용은 "**부분 상관 다이어그램**"에서 확인하십시오.

**이상치 분석** 이 메뉴에는 Mahalanobis 거리, 잭나이프 거리 및  $T^2$  통계량 방법 중 하나를 사용하여 다변량 관점의 거리를 측정하는 그림을 각각 표시하거나 숨기기 위한 옵션이 포함되어 있습니다. 자세한 내용은 "**이상치 분석**"에서 확인하십시오.

**항목 신뢰도** 이 메뉴에는 각각 항목 신뢰도 보고서를 표시하거나 숨기기 위한 옵션이 포함되어 있습니다. 보고서는 Cronbach  $\alpha$  또는 표준화  $\alpha$  를 사용하여 일련의 측정기에서 전체 반응을 얼마나 일관되게 측정하는지 나타냅니다. 자세한 내용은 "**항목 신뢰도**"에서 확인하십시오.

**결측 데이터 대치** ( 데이터 테이블에 결측값이 포함된 경우에만 사용 가능 ) 데이터 테이블을 복제하고 모든 결측값을 추정값으로 대체하는 새 데이터 테이블을 생성합니다. 대치된 값은 각 행의 비결측값에 따른 조건부 기대값입니다. 시작 창에서 선택한 방법으로 추정된 평균 및 공분산 행렬이 결측값 대치 계산에 사용됩니다. 그런 다음 모든 다변량 검정 및 옵션을 대치된 데이터 집합에 사용할 수 있습니다.

**결측값 대치 계산식 저장** ( 데이터 테이블에 결측값이 포함된 경우에만 사용 가능 ) 결측값이 포함된 열에 대해 결측값 추정에 사용된 계산식을 포함하는 새 열을 데이터 테이블에 저장합니다. 새 열 이름은 결측값 대치됨 < 열 이름 > 입니다.

다음 옵션에 대한 자세한 내용은 **JMP 사용**에서 확인하십시오.

**로컬 데이터 필터** 특정 보고서에서 사용되는 데이터를 필터링할 수 있는 로컬 데이터 필터를 표시하거나 숨깁니다.

**다시 실행** 분석을 반복하거나 다시 시작할 수 있는 옵션이 포함되어 있습니다. 이 기능을 지원하는 플랫폼에서 "자동 재계산" 옵션은 해당하는 보고서 창에서 데이터 테이블에 대한 변경 사항을 즉시 반영합니다.

**플랫폼 환경 설정** 현재 플랫폼 환경 설정을 보거나, 현재 JMP 보고서의 설정과 일치하도록 플랫폼 환경 설정을 업데이트할 수 있는 옵션이 포함되어 있습니다.

**스크립트 저장** 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다.

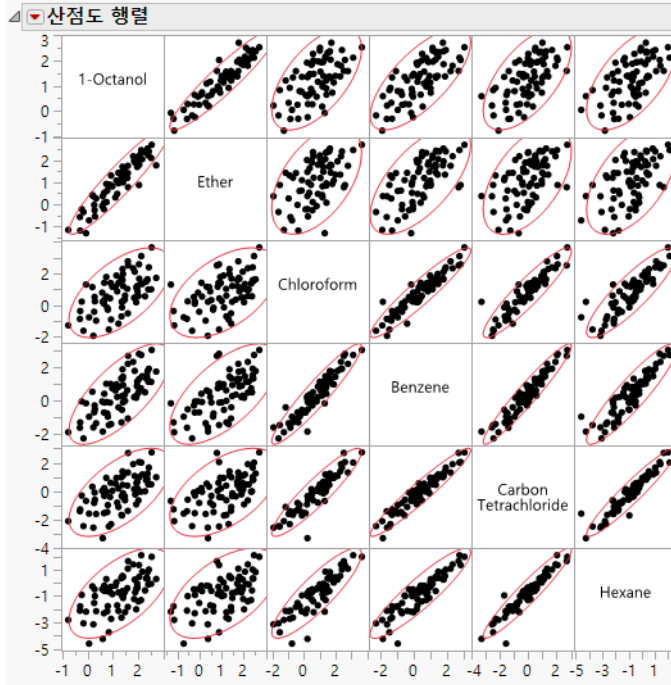
**그룹별 스크립트 저장** 기준 변수의 모든 수준에 대한 플랫폼 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다. 시작 창에서 기준 변수를 지정한 경우에만 사용할 수 있습니다.

**참고:** 이 플랫폼의 추가 옵션은 스크립트를 통해 사용할 수 있습니다. 자세한 내용은 "도움말" 메뉴의 "스크립트 인덱스"에서 여십시오. 또한 "스크립트 인덱스"에서 이 섹션에 설명된 옵션의 스크립트 예제도 찾을 수 있습니다.

## 산점도 행렬

다변량 플랫폼의 "산점도 행렬" 보고서는 각 반응 변수 쌍 사이의 상관을 시각화하는 데 도움이 됩니다.

그림 3.4 산점도 행렬



"밀도 타원" 옵션을 설정하면 각 산점도에 95% 이변량 정규 밀도 타원이 표시됩니다. 각 변수 쌍이 이변량 정규분포를 따른다고 가정하면 이 타원은 전체 점의 약 95%를 둘러쌉니다. 타원의 너비는 변수의 상관관계 정도를 반영합니다. 타원이 상당히 둥근 형태이고 대각선 방향이 아니면 해당 변수 간에 상관관계가 없는 것입니다. 타원이 좁고 대각선 방향이면 해당 변수 간에 상관관계가 높은 것입니다.

### 팁:

- 한 셀의 크기를 조정하면 모든 셀의 크기가 조정됩니다.
- 라벨 셀을 다른 라벨 셀로 드래그하면 행렬이 재정렬됩니다.
- 산점도 행렬에서 패턴을 찾을 때 변수가 상관관계에 따라 그룹으로 군집화되는 것을 확인할 수 있습니다. 그림 3.4에서는 두 개의 상관 군집을 보여 줍니다 (왼쪽 위의 두 변수와 오른쪽 아래의 네 변수).

## 산점도 행렬 옵션

"산점도 행렬"의 빨간색 삼각형 메뉴 옵션을 사용하면 색상과 밀도 타원을 사용하고  $\alpha$  수준을 설정하여 행렬을 사용자 정의할 수 있습니다.

**점 표시** 산점도에 점을 표시하거나 숨깁니다.

**선형 적합** 회귀선 및 적합 회귀선에 대한 95% 수준 신뢰 곡선을 표시하거나 숨깁니다.

**밀도 타원** 산점도에 95% 밀도 타원을 표시하거나 숨깁니다. "타원  $\alpha$ " 메뉴를 사용하여  $\alpha$  수준을 변경할 수 있습니다.

**음영 타원** 각 타원에 색상을 적용합니다. "타원 투명도" 및 "타원 색상" 메뉴를 사용하여 투명도와 색상을 변경할 수 있습니다.

**상관 표시** 각 산점도의 왼쪽 위에 각 변수 쌍의 상관을 표시하거나 숨깁니다.

**행렬 옵션** (시작 창에서 "정사각" 행렬 형식을 선택한 경우에만 사용 가능) 산점도 행렬의 오른쪽 상단 삼각형 영역 모양을 변경하는 옵션이 포함된 하위 메뉴를 표시합니다. 다음 옵션을 한 번에 하나만 선택할 수 있습니다.

**유의성 원** 산점도 행렬의 오른쪽 상단 삼각형 영역에 상관 원을 표시하거나 숨깁니다. 각 원의 색상은 각 변수 쌍 간의 상관관계를 빨간색 (+1) ~ 파란색 (-1) 의 척도로 나타냅니다. 각 원의 크기는 변수 간의 유의성 검정을 나타냅니다. 원이 클수록 더 유의한 관계를 나타냅니다.

**히트맵** 산점도 행렬의 오른쪽 상단 삼각형 영역에 상관 히트맵을 표시하거나 숨깁니다. 히트맵의 각 셀 색상은 각 변수 쌍 간의 상관관계를 빨간색 (+1) ~ 파란색 (-1) 의 척도로 나타냅니다.

**히스토그램 표시** 라벨 셀에 가로 또는 세로 히스토그램을 표시하거나 숨깁니다. 히스토그램을 추가한 후 **개수 표시**를 선택하여 히스토그램의 각 막대에 개수로 라벨을 지정할 수 있습니다. **수평** 또는 **수직**을 선택하여 히스토그램 방향을 변경하거나 히스토그램을 제거합니다.

**타원  $\alpha$**  타원에 사용되는  $\alpha$  수준을 설정합니다. 메뉴에서 표준  $\alpha$  수준 중 하나를 선택하거나, **기타**를 선택하여 다른 값을 입력합니다.

**타원 투명도** 색상이 적용되는 타원의 투명도를 설정합니다. 기본 수준 중 하나를 선택하거나, **기타**를 선택하여 다른 값을 입력합니다. 기본값은 0.2 입니다.

**타원 색상** 색상이 적용되는 타원의 색상을 설정합니다. 색상 팔레트에서 하나를 선택하거나, **기타**를 선택하여 다른 색상을 사용합니다. 기본값은 빨간색입니다.

**비모수 밀도** 데이터 점의 밀도를 묘사하는 평활 비모수 이변량 표면을 기반으로 음영 밀도 등고선을 표시하거나 숨깁니다. 비모수 표면의 10% 및 50% 분위수에 대한 등고선이 표시됩니다.

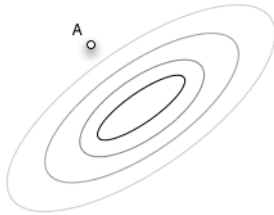
## 이상치 분석

"다변량" 빨간색 삼각형 메뉴의 "이상치 분석" 옵션에는 이상치 감지에 사용할 세 가지 거리 척도의 하위 메뉴가 포함되어 있습니다. 하위 메뉴의 각 옵션은 상관 구조와 관련하여 다변량 관점

에서 거리를 측정하는 그림을 표시하거나 숨깁니다. 그림 아래쪽에 표시된 유의 수준으로 검정이 수행됩니다.

그림 3.5에서 점 A는 좌표 방향에서는 이상치가 아니지만 상관 구조를 벗어나므로 이상치입니다.

그림 3.5 이상치의 예



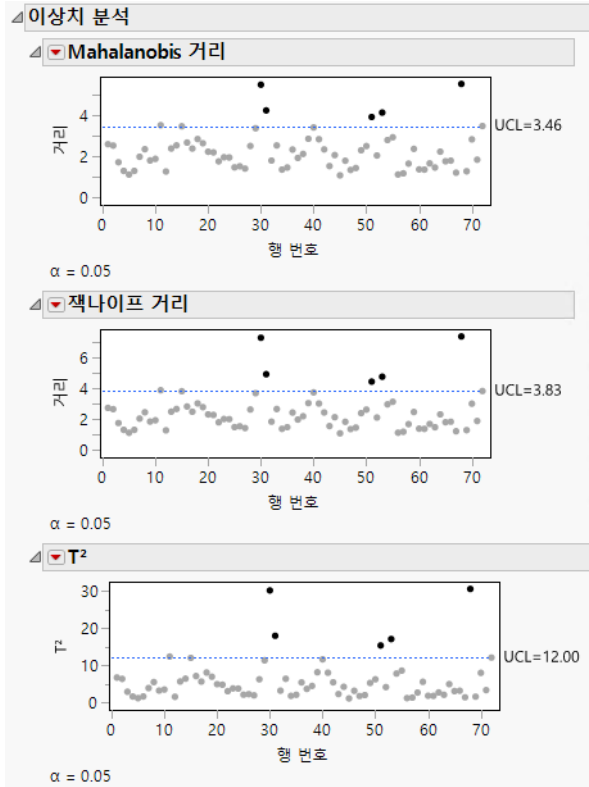
다음과 같은 거리 측도 옵션을 사용할 수 있습니다.

**Mahalanobis 거리** 다변량 평균 (중심) 과 각 점 사이의 Mahalanobis 거리를 표시하거나 숨깁니다. 표준 Mahalanobis 거리는 데이터의 평균, 표준편차 및 상관관계 추정값에 따라 달라집니다. 각 관측값 번호에 대해 거리가 표시됩니다. 거리 값이 가장 큰 점을 강조 표시하여 극단 다변량 이상치를 식별할 수 있습니다. 자세한 내용은 "[Mahalanobis 거리 측도](#)" 에서 확인하십시오.

**잭나이프 거리** 잭나이프 기법을 사용하여 계산된 거리를 표시하거나 숨깁니다. 각 관측값의 거리는 관측값 자체를 포함하지 않는 평균, 표준편차 및 상관 행렬의 추정값을 사용하여 계산됩니다. 잭나이프 거리는 이상치가 있을 때 유용합니다. 이 경우 Mahalanobis 거리는 왜곡되어 이상치를 감추거나 다른 점이 이상치보다 더 멀리 보이게 하는 경향이 있습니다. 자세한 내용은 "[잭나이프 거리 측도](#)" 에서 확인하십시오.

**T<sup>2</sup>** Mahalanobis 거리의 제곱인 거리를 표시하거나 숨깁니다. 이 그림은 다변량 관리도에서 선호됩니다. 계산된 T<sup>2</sup> 통계량 값과 관리 상한이 그림에 포함됩니다. 이 한계를 벗어나는 값은 이상치일 수 있습니다. 자세한 내용은 "[T<sup>2</sup> 거리 측도](#)" 에서 확인하십시오.

그림 3.6 이상치 분석 그림



## 거리 및 값 저장

그림의 빨간색 삼각형 메뉴에서 **저장** 옵션을 선택하여 거리를 데이터 테이블에 저장할 수 있습니다.

**참고:** 잭나이프 거리 열에는 계산식이 저장되지 않습니다. 즉, 데이터 테이블을 수정해도 거리가 다시 계산되지 않습니다. 열을 추가 또는 삭제하거나 데이터 테이블의 값을 변경하는 경우 **분석 > 다변량 방법 > 다변량**을 다시 선택하여 새 잭나이프 거리를 계산해야 합니다.

각 행의 거리 값을 저장하는 것 외에도 지정된 이상치 분석 유형에 대한 UCL(관리 상한) 값을 포함하는 열 특성이 생성됩니다.

## 항목 신뢰도

다변량 플랫폼에서 직접 항목 신뢰도 분석을 수행할 수 있습니다. 항목 신뢰도는 일련의 측정기에서 전체 반응을 얼마나 일관되게 측정하는지 나타냅니다. Cronbach  $\alpha$ (Cronbach 1951)는 신

뢰도 측도 중 하나입니다. Cronbach  $\alpha$  가 적용되는 두 가지 주요 분야는 공업용 측정기 신뢰도와 설문지 분석입니다.

Cronbach  $\alpha$  는 측정 척도에서 항목의 평균 상관을 기반으로 합니다. 이는 데이터 테이블에서 모든 반복 상관의 평균을 계산하는 것과 같습니다. 항목의 분산이 광범위한 경우 표준화  $\alpha$  를 요청할 수 있습니다.

**참고 :** Cronbach  $\alpha$  는 유의 수준  $\alpha$  와 관련이 없습니다. 또한 항목 신뢰도는 생존 시간 신뢰성 분석과 관련이 없습니다.

JMP에서 개별 항목의 영향을 검토하기 위해 해당 항목을 계산에서 제외하고 Cronbach  $\alpha$  값의 효과를 표시합니다. 변수(항목)를 제외할 때  $\alpha$  가 증가하면 해당 변수는 다른 변수와 높은 상관 관계가 없습니다.  $\alpha$  가 감소하면 해당 변수가 척도의 다른 항목과 상관되어 있다는 결론을 내릴 수 있습니다.

계산에 대한 자세한 내용은 "[Cronbach a에 대한 통계 상세 정보](#)"에서 확인하십시오.


## 부분 상관 다이어그램

다변량 플랫폼의 "부분 상관 다이어그램" 보고서에서는 부분 상관을 시각적으로 표현합니다. 다이어그램에는 각 변수에 대한 노드와 각 변수 쌍에 대한 엣지가 포함되어 있습니다. 다이어그램에서 노드는 "노드 방향" 옵션에 따라 배치됩니다. "순환"과 "고유 벡터"라는 두 가지 옵션을 사용할 수 있습니다. "순환" 방향은 노드 간격을 원형(다이어그램이 정사각형이 아닌 경우에는 타원형)으로 균등하게 유지합니다. "고유 벡터" 방향은 고유 벡터를 기반으로 노드의 간격을 둡니다. 각 고유 벡터에 해당 고유값의 제곱근을 곱하여 척도화된 고유 벡터 집합을 생성합니다. 다이어그램의 좌표는 다이어그램 아래에 지정된 차원의 척도화된 고유 벡터입니다.

각 변수 노드는 선으로 다른 노드에 연결됩니다. 선 색상은 변수 간에 양의 부분 상관(빨간색) 또는 음의 부분 상관(파란색)이 있음을 나타냅니다. 선 너비는 변수 간 부분 상관의 강도를 나타냅니다. 선이 두꺼울수록 부분 상관이 더 강합니다.

다음 옵션을 사용하여 다이어그램을 조정할 수 있습니다.

**노드 방향** 노드 방향을 "고유 벡터" 또는 "순환"으로 변경합니다.

**회전** 회전 버튼  을 사용하여 다이어그램을 회전합니다.

**절대 부분 상관** 절대 부분 상관 값이 지정된 값 미만인 선을 숨깁니다. "절대 부분 상관" 옆의 상자 값에 값을 입력하거나, 슬라이더 막대를 사용하여 경계를 지정할 수 있습니다.

**색상 분리** 선 색상을 조정합니다.

**너비** 선 너비를 조정합니다.

**투명도** 선 투명도를 조정합니다.

**재설정** 모든 표시 옵션을 기본값으로 재설정합니다.

**차원 선택** (고유값 방향에만 사용 가능) 노드 간격을 설정하는 데 사용할 고유 벡터 차원을 지정합니다.

## 부분 상관 다이어그램 옵션

"부분 상관 다이어그램"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

**고유값** 부분 상관 행렬의 분해를 통해 얻은 정렬된 고유값 테이블을 표시하거나 숨깁니다.

**고유 벡터** 부분 상관 행렬의 분해를 통해 얻은 고유 벡터 테이블을 표시하거나 숨깁니다.

**부분 상관 표시** 다이어그램의 해당 호 위에 부분 상관 값을 표시하거나 숨깁니다.

**색상 테마** 다이어그램의 색상 테마를 선택할 수 있습니다.

**좌표 저장** 척도화된 고유 벡터 (차원)를 새 데이터 테이블에 저장합니다. 척도화된 고유 벡터는 고유 벡터에 해당 고유값의 제곱근을 곱한 값입니다. 이러한 좌표는 부분 상관 다이어그램에 사용됩니다.

---

## 다변량 플랫폼의 추가 예

이 섹션에는 다변량 플랫폼을 사용하는 예가 포함되어 있습니다.

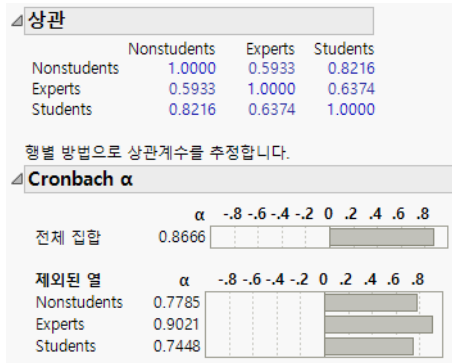
- "항목 신뢰도의 예"
- "부분 상관의 예"

### 항목 신뢰도의 예

이 예에서는 다변량 플랫폼의 항목 신뢰도 분석을 사용하여 세 그룹의 사람들이 인식한 위험 수준에 따라 30 개 항목의 순위를 매기는 방식으로 합치도를 평가합니다. 값이 각 그룹에 대해 동일한 순위 집합을 나타내는 이 유형의 예에서는 데이터를 표준화하는 것이 아무 효과가 없습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Danger.jmp**를 엽니다.
2. **분석 > 다변량 방법 > 다변량**을 선택합니다.
3. **Activity**를 제외한 모든 열을 선택하고 **Y, 열**을 클릭합니다.
4. **확인**을 클릭합니다.
5. "다변량"의 빨간색 삼각형을 클릭하고 **항목 신뢰도 > Cronbach  $\alpha$** 를 선택합니다.
6. (선택 사항)"다변량"의 빨간색 삼각형을 클릭하고 **산점도 행렬**을 선택하여 해당 그림을 숨깁니다.

그림 3.7 Cronbach  $\alpha$  보고서



Cronbach  $\alpha$  결과에는 전체  $\alpha$  0.8666이 표시되어 세 그룹 간의 순위 값이 높은 상관관계가 있음을 나타냅니다. 또한 분석에서 Experts(전문가 그룹)를 제거하면 Nonstudents와 Students는 Cronbach  $\alpha$  스코어가 각각 0.7785와 0.7448로, 위험 순위를 거의 동일하게 평가했습니다.

## 부분 상관의 예

다변량 플랫폼을 사용하여 금속 막대의 여러 구간에서 코팅 두께 사이의 관계를 검토합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Quality Control/Thickness.jmp 를 엽니다.
2. **분석 > 다변량 방법 > 다변량**을 선택합니다.
3. Thickness 열을 모두 선택하고 **Y, 열**을 클릭합니다.
4. **확인**을 클릭합니다.
5. (선택 사항) "다변량"의 빨간색 삼각형을 클릭하고 **산점도 행렬**을 선택하여 보고서에서 산점도 행렬을 제거합니다.  
산점도 행렬은 분석에 사용되지 않으므로 보고서를 정리하기 위해 이 작업을 수행합니다.
6. "다변량"의 빨간색 삼각형을 클릭하고 **부분 상관**을 선택합니다.

그림 3.8 Thickness 에 대한 상관 및 부분 상관

다변량												
상관												
	Thickness 01	Thickness 02	Thickness 03	Thickness 04	Thickness 05	Thickness 06	Thickness 07	Thickness 08	Thickness 09	Thickness 10	Thickness 11	Thickness 12
Thickness 01	1.0000	0.9815	0.9624	0.9590	0.9669	0.9793	0.9825	0.9803	0.9807	0.9521	0.9615	0.8250
Thickness 02	0.9815	1.0000	0.9720	0.9568	0.9598	0.9678	0.9633	0.9558	0.9513	0.9198	0.9304	0.8490
Thickness 03	0.9624	0.9720	1.0000	0.9942	0.9928	0.9867	0.9726	0.9671	0.9613	0.9127	0.9215	0.8872
Thickness 04	0.9590	0.9568	0.9942	1.0000	0.9949	0.9878	0.9776	0.9719	0.9665	0.9158	0.9241	0.8880
Thickness 05	0.9669	0.9598	0.9928	0.9949	1.0000	0.9961	0.9858	0.9824	0.9782	0.9237	0.9335	0.8663
Thickness 06	0.9793	0.9678	0.9867	0.9878	0.9961	1.0000	0.9926	0.9891	0.9854	0.9291	0.9401	0.8462
Thickness 07	0.9825	0.9633	0.9726	0.9776	0.9858	0.9926	1.0000	0.9977	0.9919	0.9392	0.9497	0.8383
Thickness 08	0.9803	0.9558	0.9671	0.9719	0.9824	0.9891	0.9977	1.0000	0.9960	0.9496	0.9588	0.8312
Thickness 09	0.9807	0.9513	0.9613	0.9665	0.9782	0.9854	0.9919	0.9960	1.0000	0.9673	0.9746	0.8237
Thickness 10	0.9521	0.9198	0.9127	0.9158	0.9237	0.9291	0.9392	0.9496	0.9673	1.0000	0.9986	0.8259
Thickness 11	0.9615	0.9304	0.9215	0.9241	0.9335	0.9401	0.9497	0.9588	0.9746	0.9986	1.0000	0.8247
Thickness 12	0.8250	0.8490	0.8872	0.8880	0.8663	0.8462	0.8383	0.8312	0.8237	0.8259	0.8247	1.0000

행별 방법으로 상관계수를 추정합니다.

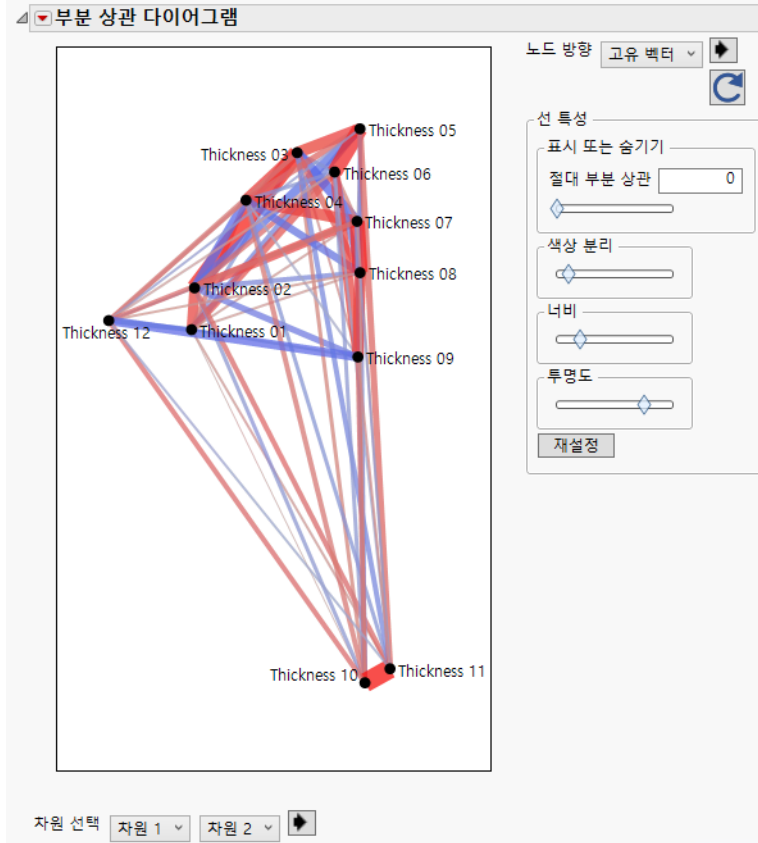
부분 상관												
	Thickness 01	Thickness 02	Thickness 03	Thickness 04	Thickness 05	Thickness 06	Thickness 07	Thickness 08	Thickness 09	Thickness 10	Thickness 11	Thickness 12
Thickness 01	.	0.4726	-0.1404	0.1830	-0.2349	0.3387	0.0529	0.0493	-0.0563	0.0331	0.0609	-0.1690
Thickness 02	0.4726	.	0.7829	-0.4282	-0.3929	0.2099	0.4214	-0.1933	-0.2868	-0.1195	0.2325	-0.0784
Thickness 03	-0.1404	0.7829	.	0.5479	0.5054	-0.1240	-0.4829	0.2775	0.0827	0.1277	-0.2094	0.1728
Thickness 04	0.1830	-0.4282	0.5479	.	0.2432	-0.1254	0.5071	-0.3655	-0.0370	0.1735	-0.1360	0.1292
Thickness 05	-0.2349	-0.3929	0.5054	0.2432	.	0.7058	0.0297	0.0909	-0.1508	-0.1659	0.2545	-0.0860
Thickness 06	0.3387	0.2099	-0.1240	-0.1254	0.7058	.	0.1242	-0.2577	0.4602	-0.1098	-0.0539	0.0399
Thickness 07	0.0529	0.4214	-0.4829	0.5071	0.0297	0.1242	.	0.7569	0.0580	-0.1731	0.0644	0.2092
Thickness 08	0.0493	-0.1933	0.2775	-0.3655	0.0909	-0.2577	0.7569	.	0.5104	0.0341	-0.0850	0.0711
Thickness 09	-0.0563	-0.2868	0.0827	-0.0370	-0.1508	0.4602	0.0580	0.5104	.	0.1882	0.0916	-0.4119
Thickness 10	0.0331	-0.1195	0.1277	0.1735	-0.1659	-0.1098	-0.1731	0.0341	0.1882	.	0.9449	0.2184
Thickness 11	0.0609	0.2325	-0.2094	-0.1360	0.2545	-0.0539	0.0644	-0.0850	0.0916	0.9449	.	-0.0516
Thickness 12	-0.1690	-0.0784	0.1728	0.1292	-0.0860	0.0399	0.2092	0.0711	-0.4119	0.2184	-0.0516	.

모든 다른 변수와 관련한 부분변수

부분 상관 행렬에서는 변수 간에 양의 상관관계와 음의 상관관계가 둘 다 있음을 보여 줍니다. 이는 양의 상관관계만 표시된 표준 상관 행렬과 다릅니다. 부분 상관은 다른 모든 변수의 효과를 조정한 후 변수 쌍 사이의 관계 강도를 측정하기 때문입니다. 이렇게 하면 변수 쌍 사이의 실제 관계를 더 명확하게 파악할 수 있습니다.

7. "다변량"의 빨간색 삼각형을 클릭하고 **부분 상관 다이어그램**을 선택합니다.

그림 3.9 부분 상관 다이어그램



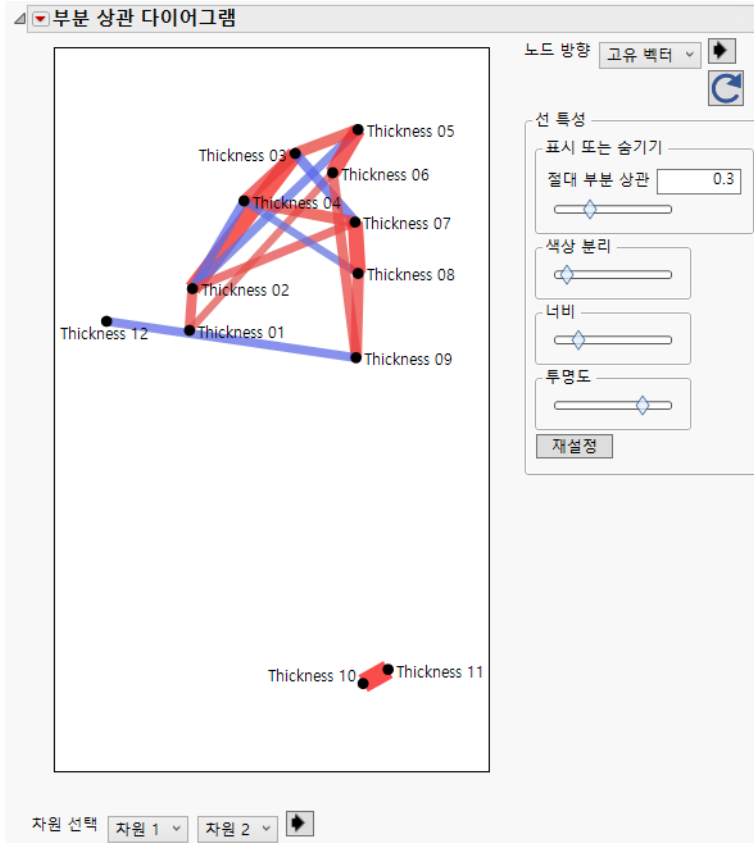
부분 상관 다이어그램은 부분 상관의 시각적 표현을 제공합니다. 그림 3.9의 다이어그램에는 각 두께 변수에 대한 노드가 있으며, 좌표는 척도화된 첫 번째 고유 벡터와 두 번째 고유 벡터에 해당합니다. Thickness 10과 Thickness 11은 처음 두 차원에서 서로 가깝고 나머지 두께 변수와 떨어져 있습니다. Thickness 10과 Thickness 11의 노드를 연결하는 두꺼운 빨간색 선은 두 변수 사이에 강한 양의 부분 상관도 있음을 나타냅니다.

나머지 노드는 함께 그룹화되어 해석하기 어려우므로 절대 부분 상관 값을 지정하여 표시되는 선을 조절할 수 있습니다. 이렇게 하면 지정된 절대 부분 상관 값보다 큰 부분 상관이 있는 선만 표시됩니다.

8. "절대 부분 상관" 옆의 상자에 "1"을 입력합니다.

그러면 다이어그램에서 모든 선이 제거됩니다. 부분 상관이 가장 강한 변수를 쉽게 확인하려면 "절대 부분 상관" 아래의 슬라이더를 사용하여 다이아몬드를 왼쪽으로 천천히 드래그합니다. 그림 3.10의 다이어그램에는 부분 상관 절대값이 0.3보다 큰 변수를 연결하는 선이 표시되어 있습니다.

그림 3.10 0.3 보다 큰 부분 상관



## 다변량 플랫폼에 대한 통계 상세 정보

이 섹션에는 다변량 플랫폼에 대한 통계 상세 정보가 포함되어 있습니다.

- " 분산 추정 방법에 대한 통계 상세 정보 "
- "Pearson 곱적률 상관계수에 대한 통계 상세 정보 "
- " 비모수 연관성 측도에 대한 통계 상세 정보 "
- " 역상관 행렬에 대한 통계 상세 정보 "
- " 거리 측도에 대한 통계 상세 정보 "
- "Cronbach a 에 대한 통계 상세 정보 "

## 분산 추정 방법에 대한 통계 상세 정보

이 섹션에는 다변량 플랫폼에서 사용되는 분산 추정 방법에 대한 통계 상세 정보가 포함되어 있습니다.

### REML

데이터에 결측값이 포함된 경우 REML(제한 최대 가능도) 추정값은 ML(최대 가능도) 추정 방법보다 덜 편향됩니다. REML 방법은 오차 대비를 기반으로 주변 가능도를 최대화합니다. REML 방법은 주로 분산 및 공분산을 추정하는 데 사용됩니다. 주성분 플랫폼의 REML 방법은 비정형 공분산 행렬을 사용하는 반복 측정 데이터에 대한 혼합 모형의 REML 추정과 동일합니다. 혼합 모형의 REML 추정에 대한 자세한 내용은 SAS Institute Inc. (2023e)의 "MIXED Procedure" 장에서 확인하십시오.

### 로버스트

기본적으로 이 방법은 이상치의 비중을 크게 줄여 해당 값을 무시합니다. 데이터의 반복 재가중 적합 시퀀스는 다음 가중치를 사용하여 수행됩니다.

$$Q < K \text{ 인 경우 } w_i = 1.0, \text{ 그렇지 않은 경우 } w_i = K/Q$$

여기서  $K$ 는 자유도가 데이터 테이블의 열 수와 같은 카이제곱 분포의 0.75 분위수와 동일한 상수입니다.  $Q$ 는 다음과 같이 정의됩니다.

$$Q = (y_i - \mu)^T (S)^{-1} (y_i - \mu)$$

이 방정식에서  $y_i = i$ 번째 관측값에 대한 반응,  $\mu$  = 평균 벡터의 현재 추정값,  $S$  = 공분산 행렬의 현재 추정값,  $T$  = 전치 행렬 연산입니다. 마지막 단계는 분산 행렬의 편향 축소입니다.

이 방법을 사용하면 데이터에 이상치가 많지 않을 때는 분산 추정값이 더 높지만 이상치가 있을 때 훨씬 더 정확한 분산 추정값을 얻을 수 있습니다.

## Pearson 곱적률 상관계수에 대한 통계 상세 정보

다변량 플랫폼에서 Pearson 곱적률 상관계수는 두 변수 간의 선형 관계 강도를 측정합니다. 반응 변수  $X$ 와  $Y$ 에 대해  $r$ 로 표시되고 다음과 같이 계산됩니다.

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2} \sqrt{\sum(y - \bar{y})^2}}$$

두 변수가 정확히 선형 관계에 있는 경우 상관계수는 변수가 양의 관계인지 아니면 음의 관계인지에 따라 1 또는 -1입니다. 선형 관계가 없으면 상관계수가 0에 가까워집니다.

## 비모수 연관성 측도에 대한 통계 상세 정보

다변량 플랫폼에서는 Spearman, Kendall 또는 Hoeffding 상관이라는 세 가지 비모수 연관성 측도를 제공합니다. 이러한 상관을 계산하기 위해 먼저 데이터에 순위를 매깁니다. 그런 다음 데이터 값 순위에 대해 계산이 수행됩니다. 동점일 경우 평균 순위가 사용됩니다.

**참고 :** 가중치 변수가 지정된 경우 결측 및 0 값 가중치는 비모수 상관 계산에서 제외됩니다. 다른 모든 가중치 값은 1로 처리됩니다.

### Spearman $\rho$ (Rho) 계수

Spearman  $\rho$  상관계수는 앞에서 설명한 Pearson 상관 계산식을 사용하여 데이터 순위에 대해 계산됩니다.

### Kendall $\tau_b$ 계수

Kendall  $\tau_b$  계수는 부합/비부합 쌍의 개수를 기반으로 합니다. 두 변수의 행 쌍은 더 큰 변수가 일치할 때 부합 쌍입니다. 그렇지 않은 경우 비부합 쌍 또는 동점입니다.

아래 계산식은

$$\tau_b = \frac{\sum_{i < j} \text{sgn}(x_i - x_j) \text{sgn}(y_i - y_j)}{\sqrt{(T_0 - T_1)(T_0 - T_2)}}$$

Kendall  $\tau_b$  를 계산합니다. 다음은 각 요소에 대한 설명입니다.

$$T_0 = (n(n-1))/2$$

$$T_1 = \sum((t_i)(t_i-1))/2$$

$$T_2 = \sum((u_i)(u_i-1))/2$$

다음 사항에 유의하십시오.

- $\text{sgn}(z)$  은  $z > 0$  인 경우 1,  $z = 0$  인 경우 0,  $z < 0$  인 경우 -1 입니다.
- $t_i(u_i)$  는 동점  $x$  (각  $y$ ) 값의  $i$  번째 그룹에서 동점  $x$  (각  $y$ ) 값의 수입니다.
- $n$  은 관측값 수입니다.
- Kendall  $\tau_b$  의 범위는 -1 ~ 1 입니다. 가중치 변수가 지정되면 무시됩니다.

계산은 다음과 같은 방식으로 진행됩니다.

- 첫 번째 변수의 값에 따라 순서대로 관측값 순위가 매겨집니다.
- 그런 다음 두 번째 변수의 값에 따라 관측값 순위가 다시 매겨집니다.
- 첫 번째 변수의 교환 횟수를 사용하여 Kendall  $\tau_b$  를 계산합니다.

### Hoeffding D 통계량

Hoeffding  $D(1948)$  계산식은 다음과 같습니다.

$$D = 30 \left( \frac{(n-2)(n-3)D_1 + D_2 - 2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)} \right)$$

다음은 각 요소에 대한 설명입니다.

$$D_1 = \sum_i (Q_i - 1)(Q_i - 2)$$

$$D_2 = \sum_i (R_i - 1)(R_i - 2)(S_i - 1)(S_i - 2)$$

$$D_3 = \sum_i (R_i - 2)(S_i - 2)(Q_i - 1)$$

다음 사항에 유의하십시오.

- $R_i$  와  $S_i$  는  $x$  및  $y$  값의 순위입니다.
- $Q_i$  (이변량 순위라고도 함) 는  $x$  값과  $y$  값이 모두  $i$  번째 점보다 작은 점의 수에 1 을 더한 값입니다.
- $x$  값 또는  $y$  값 중 하나만 같은 점은 다른 값이  $i$  번째 점의 해당 값보다 작으면  $Q_i$  에 1/2 을 적용합니다.  $x$  와  $y$  둘 다 같은 점은  $Q_i$  에 1/4 을 적용합니다.

관측값 중에서 동점이 없으면  $D$  통계량의 값은 -0.5에서 1 사이이며, 여기서 1은 완전 종속을 나타냅니다. 가중치 변수가 지정되면 무시됩니다.

### 역상관 행렬에 대한 통계 상세 정보

다변량 플랫폼에서 역상관 행렬은 유용한 다변량 정보를 제공합니다. 역상관 행렬의 대각 요소는 변수가 다른 변수의 선형 함수에 얼마나 가까운지에 대한 함수이며, VIF(분산 팽창 계수)라고도 합니다. 구체적으로 말하면 상관 행렬을  $\mathbf{R}$  로 나타내고 역상관 행렬을  $\mathbf{R}^{-1}$  로 나타낼 경우 대각 요소는  $r_{ii}$  로 나타내고 다음과 같이 계산됩니다.

$$r_{ii} = \text{VIF}_i = \frac{1}{1 - R_i^2}$$

여기서  $R_i^2$  은 다른 설명 변수를 기반으로  $i$  번째 설명 변수를 모델링하는 모형의 변동 계수입니다. 따라서  $r_{ii}$  가 크면  $i$  번째 변수가 다른 변수와 높은 상관관계가 있음을 나타냅니다.

## 거리 측도에 대한 통계 상세 정보

이 섹션에는 다변량 플랫폼의 이상치 분석 그림에 사용되는 거리 측도에 대한 통계 상세 정보가 포함되어 있습니다.

### Mahalanobis 거리 측도

Mahalanobis 거리는 데이터의 상관 구조와 개별 척도를 고려합니다. 각 값에 대해 Mahalanobis 거리는  $M_i$  로 나타내고 다음과 같이 계산됩니다.

$$M_i = \sqrt{(Y_i - \bar{Y})S^{-1}(Y_i - \bar{Y})}$$

다음은 각 요소에 대한 설명입니다.

$Y_i$  =  $i$  번째 행의 데이터

$\bar{Y}$  = 평균 행

$S$  = 데이터에 대한 추정된 공분산 행렬

"Mahalanobis 거리" 그림에 표시된 UCL 참조선 (Mason 과 Young 2002) 은 다음과 같이 계산됩니다.

$$UCL_{Mahalanobis} = \sqrt{\frac{(n-1)^2}{n} \beta \left[ 1 - \alpha; \frac{p}{2}; \frac{n-p-1}{2} \right]}$$

다음은 각 요소에 대한 설명입니다.

$n$  = 관측값 수

$p$  = 변수 (열) 수

$\beta \left[ 1 - \alpha; \frac{p}{2}; \frac{n-p-1}{2} \right]$  = 베타  $\left( \frac{p}{2}, \frac{n-p-1}{2} \right)$  분포의  $(1-\alpha)$  번째 분위수

변수가 다른 변수의 정확한 선형 결합인 경우 상관 행렬은 특이 행렬이고 해당 변수의 행과 열은 0으로 설정됩니다. 결과로 생성되는 일반화 역행렬은 거리를 구성하는 데 유효합니다.

### 잭나이프 거리 측도

잭나이프 거리는 관측값 자체를 포함하지 않는 평균, 표준편차 및 상관 행렬의 추정값을 사용하여 계산됩니다. 각 값에 대해 잭나이프 거리는 다음과 같이 계산됩니다.

$$J_i = \sqrt{\frac{(n-2)n^2}{(n-1)^3} \times \frac{M_i^2}{1 - \frac{nM_i^2}{(n-1)^2}}}$$

다음은 각 요소에 대한 설명입니다.

$n$  = 관측값 수

$p$  = 변수 (열) 수

$M_i$  =  $i$  번째 관측값의 Mahalanobis 거리

" 잭나이프 거리 " 그림에 표시된 UCL 참조선 (Penny 1996) 은 다음과 같이 계산됩니다.

$$UCL_{Jackknife} = \sqrt{\frac{(n-2)n^2}{(n-1)^3} \times \frac{UCL_{Mahalanobis}^2}{1 - \frac{n \cdot UCL_{Mahalanobis}}{(n-1)^2}}}$$

### T<sup>2</sup> 거리 측도

T<sup>2</sup> 거리는 Mahalanobis 거리의 제곱이므로  $T_i^2 = M_i^2$ 입니다.

T<sup>2</sup> 거리의 UCL 은 다음과 같습니다.

$$UCL_{T^2} = \frac{(n-1)^2}{n} \beta_{\left[1-\alpha; \frac{p}{2}; \frac{n-p-1}{2}\right]} = (UCL_{Mahalanobis})^2$$

다음은 각 요소에 대한 설명입니다.

$n$  = 관측값 수

$p$  = 변수 (열) 수

$\beta_{\left[1-\alpha; \frac{p}{2}; \frac{n-p-1}{2}\right]}$  = 베타  $\left(\frac{p}{2}, \frac{n-p-1}{2}\right)$  분포의  $(1-\alpha)$  번째 분위수

다변량 거리는 여러 차원에서 이상치를 찾는 데 유용합니다. 그러나 변수가 다변량 관점에서 높은 상관관계가 있는 경우에는 점을 차원의 부분집합에서 비정상적으로 보지 않고 다변량 공간의

이상치로 간주할 수 있습니다. 즉, 값이 상관되어 있는 경우 하나 또는 두 개의 축을 따라 보면 점이 특별하지 않을 수 있지만 상관관계를 위반하여 이상치가 될 수 있습니다.

## Cronbach $\alpha$ 에 대한 통계 상세 정보

"다변량" 빨간색 삼각형 메뉴의 "항목 신뢰도" 옵션에 사용되는 통계량은 Cronbach  $\alpha$  입니다. Cronbach  $\alpha$  는 다음과 같이 정의됩니다.

$$\alpha = \frac{kc}{v + (k-1)c}$$

다음은 각 요소에 대한 설명입니다.

$k$  = 척도의 항목 수

$c$  = 항목 간의 평균 공분산

$v$  = 항목 간의 평균 분산

항목이 일정한 분산을 가지도록 표준화되는 경우 계산식은 다음과 같습니다.

$$\alpha = \frac{k(r)}{1 + (k-1)r}$$

다음은 각 요소에 대한 설명입니다.

$r$  = 항목 간의 평균 상관

전체  $\alpha$  계수가 클수록 항목이 신뢰할 수 있는 척도 또는 검정에 기여한다는 것을 더 확신할 수 있습니다. 상관관계가 높은 항목이 많은 경우 계수가 1.0에 근접할 수 있습니다.

# 4 장

## 주성분 데이터 차원 축소

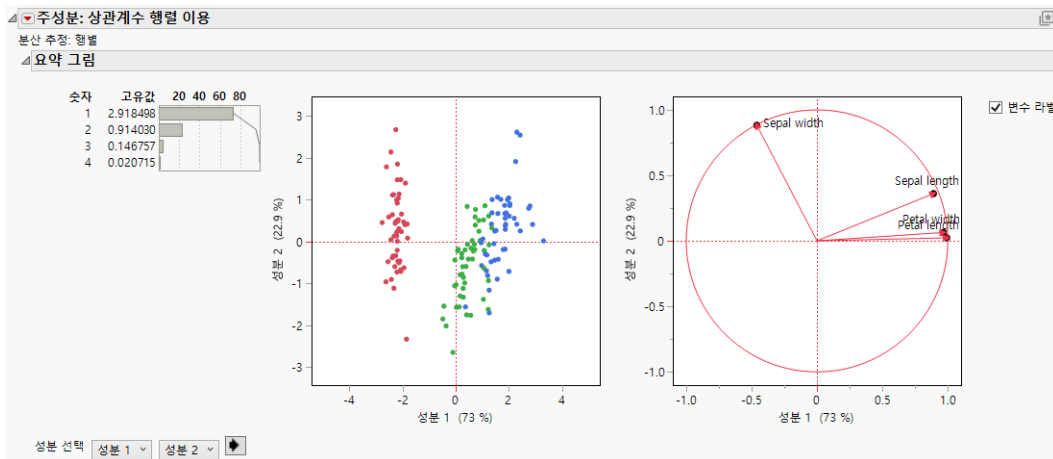
주성분 분석의 목적은 측정된 변수 집합 중 원래 변수의 변동성을 가능한 한 많이 포착하는 소수의 독립적 선형 결합(주성분)을 도출하는 것입니다. 주성분 분석은 탐색적 데이터 분석 도구가 차원 축소 기법입니다. 주성분 분석은 주성분 분석 회귀(PCA 회귀 또는 PCR이라고도 함)에서와 같이 예측 모형을 생성하는 데도 유용합니다.

주성분 플랫폼은 주성분을 얻기 위한 두 가지 범주의 방법을 제공합니다. "전통적" 방법은 공분산 행렬, 상관 행렬 또는 비척도화/비중심화 행렬을 사용하여 주성분을 얻습니다. "고급" 방법은 다양한 특이값 분해 기법을 사용하여 주성분을 얻습니다.

고급 방법은 변수의 수가 매우 많은 데이터에 유용하며, 효율적인 계산 시간으로 주성분을 계산할 수 있습니다. 대부분의 고급 방법을 사용할 때는 추정할 성분 수를 지정해야 합니다. 일반적으로 이 값은 데이터의 차원보다 훨씬 작습니다. 그런 다음 이러한 주성분을 PCA 회귀에 사용할 수 있습니다.

주성분 플랫폼에서는 요인 분석도 지원합니다. JMP에서는 추출된 성분을 해석하는 데 도움이 되는 여러 유형의 직교 및 사각 요인 분석 스타일 회전을 제공합니다. 요인 분석에 대한 자세한 내용은 "[요인 분석](#)"에서 확인하십시오.

그림 4.1 주성분의 예



**목차**

주성분 플랫폼 개요 .....	57
주성분 분석의 예 .....	57
주성분 플랫폼 시작 .....	58
결측 데이터.....	62
주성분 보고서.....	62
주성분 보고서 옵션 .....	63
이상치 분석.....	73
주성분 플랫폼에 대한 통계 상세 정보.....	75
분산 추정 방법에 대한 통계 상세 정보.....	76
고급 방법에 대한 통계 상세 정보.....	76
이상치 분석 계산에 대한 통계 상세 정보.....	77

---

## 주성분 플랫폼 개요

주성분 분석은 변수 집합의 변동을 원래 변수보다 더 적은 수의 주성분(원래 변수의 독립적 선형 결합)으로 모델링합니다.

상관된 많은 변수에 대한 점 배열을 보려면 주성분 분석을 사용하여 가장 눈에 띄는 고차원 데이터의 방향을 표시할 수 있습니다. 주성분 분석을 사용하면 데이터 집합의 차원이 축소됩니다. 주성분은 가능한 한 적은 수의 변수를 사용하여 데이터 구조를 최대한 완벽하게 나타낼 수 있는 방법입니다.

변수가  $p$  개인 경우 다음과 같은 방식으로  $p$  개의 주성분이 구성됩니다.

- 첫 번째 주성분은 표준화된 원래 변수 중에서 가능한 가장 큰 분산을 갖는 선형 결합입니다.
- 이후의 각 주성분은 가능한 분산이 가장 큰 변수 선형 결합이며 이전에 정의된 모든 성분과 상관관계가 없습니다.

각 주성분은 상관 행렬(또는 공분산 행렬이나 제곱합 및 교차곱 행렬)의 고유 벡터를 변수와 선형 결합하여 계산됩니다. 고유값은 각 성분의 분산을 나타냅니다.

주성분 플랫폼을 사용하면 상관 행렬, 공분산 행렬 또는 비척도화/비중심화 데이터에 대한 분석을 수행할 수 있습니다. 주성분 플랫폼 내에서 요인 분석을 수행할 수도 있습니다. 자세한 내용은 "[요인 분석](#)"에서 확인하십시오.

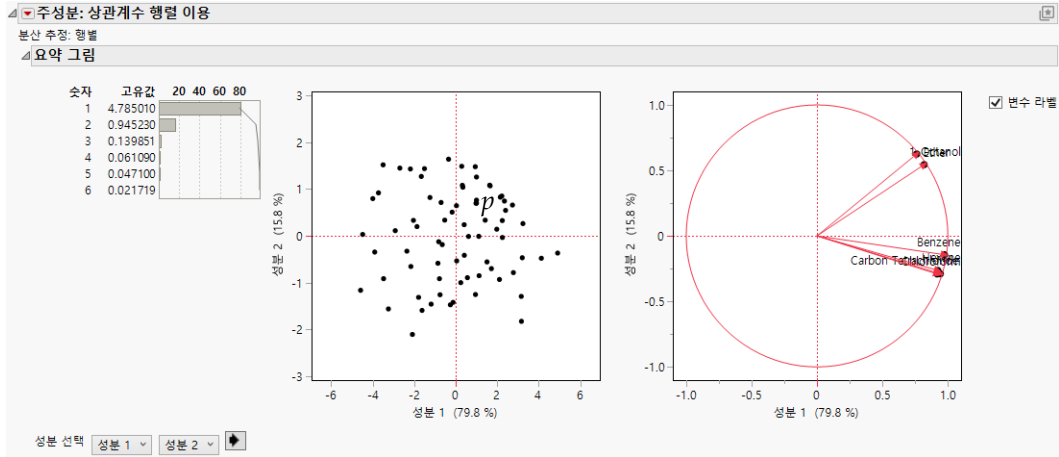
---

## 주성분 분석의 예

이 예에서는 6 개 요인에 대한 주성분 분석을 수행합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Solubility.jmp 를 엽니다.
2. **분석 > 다변량 방법 > 주성분**을 선택합니다.
3. 연속형 열을 모두 선택하고 **Y, 열**을 클릭합니다.
4. 6 개 요인만 분석하므로 "방법 유형" 옆에서 **전통적**을 선택합니다.
5. 분산 추정 기본값을 유지하고 **확인**을 클릭합니다.

그림 4.2 상관계수 행렬을 이용한 주성분 보고서



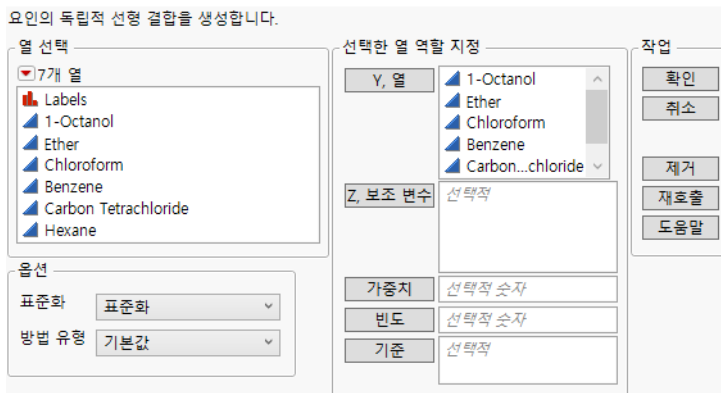
이 보고서에는 각 주성분이 설명하는 변동의 백분율을 나타내는 막대 차트와 고유값이 제공됩니다. 이 예에서는 첫 번째 주성분이 데이터 변동의 거의 80%를 설명합니다. 처음 두 주성분이 함께 데이터 변동을 거의 모두(95.5%) 설명합니다. 스코어 그림과 적재 그림도 있습니다. 자세한 내용은 "주성분 보고서"에서 확인하십시오.

## 주성분 플랫폼 시작

**분석 > 다변량 방법 > 주성분**을 선택하여 주성분 플랫폼을 시작합니다. 다변량 및 3D 산점도 플랫폼을 사용하여 주성분 분석을 사용할 수도 있습니다.

"주성분 분석의 예"에 설명된 예에서는 Solubility.jmp 샘플 데이터 테이블의 모든 연속형 변수를 사용합니다.

그림 4.3 주성분 시작 창



"열 선택"의 빨간색 삼각형 메뉴에 포함된 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

**Y, 열** 성분에 대해 분석할 변수입니다.

**Z, 보조 변수** 표시할 보조 변수입니다. 보조 변수는 주성분 계산에 포함되지 않으며 포함해도 결과에 영향을 주지 않습니다. 연속형 보조 변수는 적재 그림에 투영되어 해석을 향상시키는 데 사용할 수 있습니다.

**가중치** 분석의 각 행에 가중치를 할당하는 숫자 값이 포함된 하나의 열을 식별합니다.

---

**참고:** "와이드" 및 "최소" 분산 추정 방법에서는 "가중치" 역할이 무시됩니다.

---

**빈도** 분석의 각 행에 빈도를 할당하는 숫자 값이 포함된 하나의 열을 식별합니다.

---

**참고:** "와이드" 및 "최소" 분산 추정 방법에서는 "빈도" 역할이 무시됩니다.

---

**기준** 각 그룹에 대한 개별 분석을 수행할 수 있도록 "기준" 열에 지정된 각 값의 주성분 보고서를 생성합니다.

**표준화** 각 열의 중심화 및 표준화 여부를 지정합니다. 이 옵션은 주성분을 계산하는 데 사용되는 행렬을 결정합니다.

**표준화** 각 열을 개별적으로 중심화하고 표준화합니다. 상관 행렬을 기반으로 주성분이 계산됩니다.

**비척도화** 각 열을 개별적으로 중심화합니다. 공분산 행렬을 기반으로 주성분이 계산됩니다.

**비척도화 및 비중심화** 비척도화 / 비중심화 행렬을 기반으로 주성분이 계산됩니다.

**방법 유형** 데이터 유형을 지정합니다.

**기본값** 열 수가 500 개 미만이거나 행 수보다 적으면 "세로형 데이터"가 "방법 유형"으로 지정됩니다. 열 수가 500 개를 초과하거나 행 수보다 많으면 와이드 추정 방법을 권장하는 JMP 경고 창이 표시됩니다. 가로형 데이터 추정 방법을 사용하려면 **와이드 방법 (빠름)** 을 클릭하고, 세로형 데이터 추정 방법을 사용하려면 **기본 방법 (느림)** 을 클릭합니다.

**전통적** 공분산 행렬, 상관 행렬 또는 비척도화/비중심화 행렬을 사용하여 주성분을 얻습니다.

**고급** 특이값 분해를 사용하여 주성분을 얻습니다.

**분산 추정** ("전통적" 방법 유형을 지정한 경우에만 사용 가능) 상관 계산 방법을 지정합니다. 이러한 방법 중 일부는 결측 데이터를 처리합니다.

**기본값** "기본값" 옵션은 "행렬", "쌍별" 또는 "REML" 방법을 사용합니다. 해당하는 경우 "와이드" 방법으로 전환할 것을 권장하는 JMP 경고가 표시됩니다.

- 행렬 추정은 결측값이 없는 데이터 테이블에 사용됩니다.
- 쌍별 추정은 결측값이 있고 열이 10 개, 행이 5,000 개를 각각 초과하거나 행보다 열이 더 많은 데이터 테이블에 사용됩니다.
- REML 추정은 그 외의 경우에 사용됩니다.

**REML** REML(제한 최대 가능도) 추정은 결측값이 있는 경우에도 모든 데이터를 사용합니다. 데이터 집합이 크고 결측값이 많은 경우 이 방법을 사용하면 편향 수정 요인 때문에 처리 속도가 느립니다. 따라서 REML은 작은 데이터 집합에 가장 유용합니다. 데이터에 결측 셀이 없으면 REML 및 ML 추정값이 동일하고 표본 공분산 행렬과 같습니다. 결측 셀이 있으면 REML의 분산 및 공분산 추정값이 ML 추정의 추정값보다 덜 편향됩니다. 통계 상세 정보는 "**REML**"에서 확인하십시오.

**ML** ML(최대 가능도) 추정은 결측값이 있는 경우에도 모든 데이터를 사용합니다. ML 추정값은 빠르게 생성되므로 이 방법은 결측 데이터가 있는 큰 데이터 테이블에 가장 유용합니다.

**로버스트** 로버스트 추정은 결측값이 있는 경우에도 모든 데이터를 사용합니다. 이 방법은 극단값의 비중을 낮추므로 이상치가 있을 수 있는 데이터 테이블에 유용합니다. 통계 상세 정보는 "**로버스트**"에서 확인하십시오.

**행별** 행별 추정은 각 열 쌍에 대한 Pearson 상관을 계산합니다. 통계 상세 정보는 "**Pearson** 급적률 상관계수에 대한 통계 상세 정보"에서 확인하십시오. 행별 추정에서는 결측값이 있는 관측값을 사용하지 않습니다. 이 방법은 결측 데이터가 있는 관측값을 제외하는 데 유용합니다.

**쌍별** 쌍별 추정은 결측값이 있는 경우에도 모든 데이터를 사용합니다. 이 분산 추정 방법은 두 열에 대해 결측값이 없는 모든 관측값을 사용하여 각 열 쌍에 대한 Pearson 상관을 계산합니다. 통계 상세 정보는 "**Pearson** 급적률 상관계수에 대한 통계 상세 정보"에서 확인하십시오. 쌍별 추정은 결측값이 있는 데이터 테이블에서 행보다 열이 더 많거나, 열이 10 개를 초과하거나, 행이 5,000 개를 초과하는 경우에 가장 유용합니다.

- "REML", "ML" 또는 "로버스트"를 선택한 경우 데이터 테이블에 행보다 열이 더 많고 결측값이 있으면 JMP에서 "분산 추정"이 "쌍별"로 전환됩니다.
- "로버스트"를 선택한 경우 데이터 테이블에 행보다 열이 더 많고 결측값이 없으면 JMP에서 "분산 추정"이 "행별"로 전환됩니다.
- 데이터 테이블의 열이 500 개를 초과하고 행보다 열이 더 많은 경우 원래 선택한 방법에 관계없이 JMP에서 "분산 추정"이 "와이드"로 전환됩니다.

---

**참고:** 열이 500 개를 초과하고 행보다 열이 더 많은 데이터 테이블의 경우 JMP 경고 창이 표시되어 와이드 추정 방법을 사용하도록 권장합니다. 열 수가 많은 경우 다른 방법을 사용하면 계산 시간이 길어질 수 있기 때문입니다. 와이드 추정 방법으로 전환하려면 **와이드 방법 (빠름)** 을 클릭하고, 원래 선택한 방법을 사용하려면 **기본 방법 (느림)** 을 클릭합니다.

---

**성분 수** ("고급" 방법 유형을 지정한 경우에만 사용 가능) 추정할 성분 수를 지정합니다. 일반적으로 성분 수는 데이터의 차원보다 훨씬 적습니다.

**지정됨** 절단 SVD 추정 방법을 사용하여 지정된 수의 성분을 추정합니다. 절단 SVD 추정은 결측값이 있는 경우에도 모든 데이터를 사용합니다. 이 추정 방법은 부분 특이값 분해를 기반으로 하는 알고리즘을 사용하여 첫 번째 지정된 수의 특이값과 특이값 벡터만 계산합니다. 이 알고리즘은 공분산 행렬 및 불필요한 주성분을 계산하지 않으므로 계산이 효율적입니다. 데이터가 희소할 때, 즉 0 이 많이 포함되거나 데이터에 열이 많은 경우에 유용합니다. 통계 상세 정보는 "**절단 SVD**" 에서 확인하십시오.

---

**참고 :** JMP 17 이전에는 " 희소 " 추정 방법으로 알려졌습니다.

---

**모두** 완전 SVD 추정 방법을 사용하여 모든 성분을 추정합니다. 완전 SVD 추정은 결측값이 있는 관측값을 사용하지 않으므로 결측 셀이 포함된 행이 제외됩니다. 이 추정 방법은 전체 특이값 분해를 기반으로 하는 알고리즘을 사용합니다. 이 알고리즘은 공분산 행렬을 계산하지 않으므로 계산이 효율적입니다. 데이터에 열이 매우 많은 경우에 유용합니다. 통계 상세 정보는 "**완전 SVD**" 에서 확인하십시오.

---

**참고 :** JMP 17 이전에는 " 와이드 " 추정 방법으로 알려졌습니다.

---

**결측값 대치** (" 고급 " 방법 유형을 지정한 경우에만 사용 가능) 행렬 완성을 통해 결측값을 대치합니다.

**특수 방법** (" 고급 " 방법 유형을 지정하고, 추정할 성분 수가 지정된 경우에만 사용 가능) 지정된 수의 성분을 계산하기 위한 추가 방법을 제공합니다.

**빠른 근사** 랜덤화 특이값 분해를 사용하여 지정된 수의 성분을 추정합니다. 자세한 내용은 "**랜덤화 SVD**" 에서 확인하십시오.

**로버스트 PCA** 데이터 행렬 분해를 위해 특이값 분해 및 임계 지정 단계를 순차적으로 사용하여 지정된 수의 성분을 추정합니다. 이 방법은 이상치 탐색 플랫폼에서도 사용됩니다. 로버스트 PCA 방법에 대한 자세한 내용은 **예측 및 전문 모델링**에서 확인하십시오.

**JMP PRO 비음 주성분** 행렬 분해에 비음성 제약 조건을 적용하여 지정된 수의 성분을 추정합니다. 이로 인해 음수가 아닌 적재 및 스코어가 생성됩니다. 성분은 한 번에 하나씩 추출되는 것이 아니라 동시에 추출됩니다.

---

**참고 :** 비음 주성분 방법은 데이터가 음수가 아니라고 가정합니다. 데이터를 음수가 아닌 상태로 유지하기 위해 중심화도 허용되지 않습니다. 따라서 " 표준화 " 를 선택하면 열이 개별적으로 척도화되지만 중심화되지는 않습니다.

---

**JMP PRO 희소 성분 분석** 희소하거나 대부분 0 이 포함된 적재를 생성하는 행렬 분해를 사용하여 지정된 수의 성분을 추정합니다. 행렬 분해는 한 번에 하나씩 수행되는 것이 아니라 동시에 수행됩니다. 이 방법은 **Varimax 회전**과 비슷합니다. 자세한 내용은 **Chen & Rohe(2023)** 에서 확인하십시오.

---

**참고 :** 결측값이 있으면 희소 성분 분석 알고리즘을 반복할 때마다 데이터 결측값 대치가 수행됩니다. 분해의 이전 반복에서 다시 생성된 값으로 결측값이 대체됩니다.

---

## 결측 데이터

주성분 플랫폼에서 결측 데이터를 처리하는 방법은 분산 추정 방법에 따라 달라집니다. 다음과 같은 방법으로 플랫폼 외부에서 결측값을 추정할 수도 있습니다.

- **다변량 방법 > 다변량** 아래에 있는 "결측 데이터 대치" 옵션을 사용합니다. 자세한 내용은 "[결측 데이터 대치](#)"에서 확인하십시오.
- **분석 > 선별 > 결측값 탐색**에 있는 "다변량 정규 대치법" 또는 "다변량 SVD 대치법" 유틸리티를 사용합니다. 자세한 내용은 [예측 및 전문 모델링](#)에서 확인하십시오.

## 주성분 보고서

초기 보고서 유형은 시작 창에서 선택한 분산 추정 방법에 따라 달라집니다. 대부분의 방법에서 "주성분: 상관계수 행렬 이용" 보고서가 나타납니다. "와이드" 및 "희소" 방법의 경우 특정 보고서가 나타납니다.

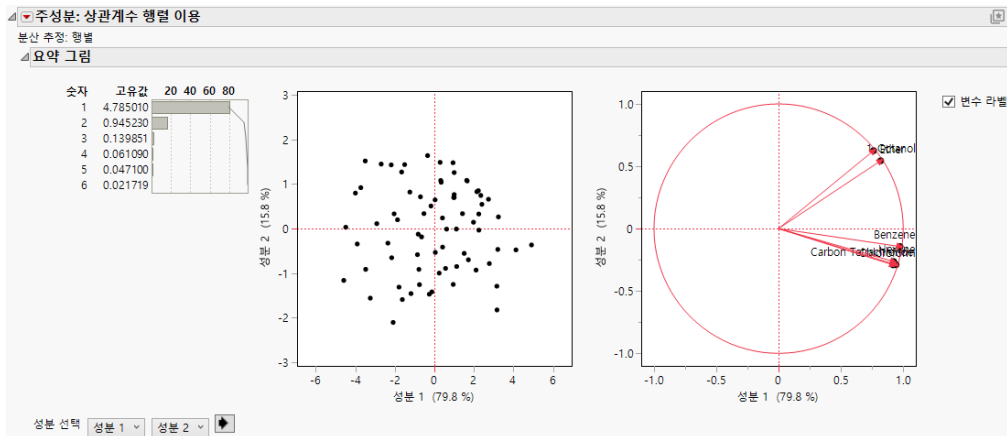
"주성분: 상관계수 행렬 이용" 보고서에는 지정된 Y 변수의 변동과 주성분이 요약되어 있습니다([그림 4.4](#) 참조). 빨간색 삼각형 메뉴에서 "주성분" 옵션을 선택하여 공분산 행렬 또는 비척도화/비중심화 데이터에 기반한 분석으로 전환할 수 있습니다.

선택한 옵션에 따라 다음 고유값 분해 중 하나에서 주성분이 도출됩니다.

- 상관 행렬
- 공분산 행렬
- 비척도화 / 비중심화 데이터에 대한 제곱합 및 교차곱 행렬

보고서에는 주성분이 데이터의 변동을 흡수하는 방식이 상세하게 표시됩니다. 변수의 고유 벡터 선형 결합에서 주성분 점이 파생됩니다.

**그림 4.4** 상관계수 행렬을 이용한 주성분 보고서



이 보고서에는 각 주성분이 설명하는 변동의 백분율을 나타내는 막대 차트와 고유값이 제공됩니다. 스코어 그림과 적재 그림도 있습니다. 고유값은 각 성분이 기여하는 분산의 양을 기반으로 추출된 총 성분 수를 나타냅니다.

스코어 그림에는 다른 성분과의 관계에 따라 평균 및 표준편차 값을 조정하여 계산된 각 성분의 값이 그래프로 표시됩니다.

적재 그림에는 변수와 성분 사이의 회전되지 않은 적재 행렬이 그래프로 표시됩니다. 값이 1에 가까울수록 성분이 변수에 미치는 효과가 커집니다.

기본적으로 이 보고서에는 처음 두 주성분에 대한 스코어 그림과 적재 그림이 표시됩니다. "성분 선택" 옆의 목록을 사용하여 스코어 그림 및 적재 그림에 그래프로 표시되는 주성분을 지정할 수 있습니다.

---

## 주성분 보고서 옵션

"주성분"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

---

**참고:** "와이드" 또는 "회소" 분산 추정 방법에서는 일부 옵션을 사용할 수 없습니다.

**주성분** ("와이드" 또는 "회소" 분산 추정 방법에는 사용 불가능) 상관계수, 공분산 또는 비척도화/비중심화 데이터를 기반으로 주성분을 생성할 수 있습니다.

**상관** ("와이드" 또는 "회소" 분산 추정 방법에는 사용 불가능) 변수 간의 상관 행렬입니다.

---

**참고:** 대각선 값은 1.0입니다.

**공분산 행렬** ("와이드" 또는 "회소" 분산 추정 방법에는 사용 불가능) 변수의 공분산을 표시하거나 숨깁니다.

**고유값** 각 주성분에 해당하는 고유값을 가장 큰 것부터 순서대로 나열합니다. 고유값은 다변량 표본에서 총 변동의 분할을 나타냅니다.

고유값의 척도는 주성분을 추출하기 위해 선택한 행렬에 따라 다릅니다.

- "상관계수 행렬 이용" 옵션을 선택하는 경우 고유값이 변수 수의 합계로 척도화됩니다.
- "공분산 행렬 이용" 옵션을 선택하는 경우 고유값이 척도화되지 않습니다.
- "비척도화 및 비중심화" 옵션을 선택하는 경우 고유값을 총 관측값 수로 나눕니다.

빨간색 삼각형 메뉴에서 **Bartlett 검정** 옵션을 선택하면 각 고유값에 대한 가설 검정 ([그림 4.6](#)) 이 제공됩니다 (Jackson 2003 참조).

그림 4.5 고유값

고유값							
숫자	고유값	백분율	20	40	60	80	누적 백분율
1	4.785010	79.750					79.750
2	0.945230	15.754					95.504
3	0.139851	2.331					97.835
4	0.061090	1.018					98.853
5	0.047100	0.785					99.638
6	0.021719	0.362					100.000

**고유 벡터** 각 주성분에 대한 고유 벡터 테이블을 왼쪽에서 오른쪽 순서로 표시하거나 숨깁니다. 이러한 계수를 사용하여 원래 변수의 선형 결합을 구성하면 주성분 변수가 생성됩니다. 표준 규칙에 따라 고유 벡터의 norm 은 1 입니다.

**참고 :** 표시된 고유 벡터 수는 상관 행렬의 계수와 같거나, " 회소 " 방법을 선택한 경우에는 시작 창에 지정된 성분 수와 같습니다.

**Bartlett 검정** (" 와이드 " 또는 " 회소 " 분산 추정 방법에는 사용 불가능) 동질성 검정 결과를 표시하거나 숨깁니다 (" 고유값 " 테이블에 추가됨). 이 검정은 검정에 대한 카이제곱, DF( 자유도) 및  $p$  값 (Prob>ChiSq) 을 계산하여 고유값의 분산이 동일한지 여부를 판별합니다. 자세한 내용은 Bartlett(1937, 1954) 연구 자료에서 확인하십시오.

그림 4.6 Bartlett 검정

고유값										
숫자	고유값	백분율	20	40	60	80	누적 백분율	카이제곱	DF	Prob>ChiSq
1	4.785010	79.750					79.750	701.245	11.243	<.0001*
2	0.945230	15.754					95.504	317.186	13.125	<.0001*
3	0.139851	2.331					97.835	58.444	9.270	<.0001*
4	0.061090	1.018					98.853	17.589	5.280	0.0044*
5	0.047100	0.785					99.638	9.723	1.899	0.0069*
6	0.021719	0.362					100.000	0.000	.	.

**적재 행렬** 각 성분에 대한 적재 테이블을 표시하거나 숨깁니다. 테이블 값의 투명도는 0 과 절대 적재 값 사이의 거리를 나타냅니다. 0 에 더 가까운 절대 적재 값은 0 에서 더 멀리 있는 절대 적재 값보다 투명합니다. 이 옵션은 각 성분에 대한 적재 그림도 표시합니다. 가로 막대와 세로 막대 간에 회전할 수 있는 버튼이 있습니다. " 적재 행렬 그림 " 의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

**그림 선택** 성분에 대한 적재를 그림에 표시하는 방법을 지정합니다. " 개요 " 옵션은 지정된 수의 성분에 대한 적재를 표시합니다. " 개별 " 옵션은 선택한 성분 하나의 적재를 표시합니다.

**막대 스타일** 그림의 막대 스타일을 나란히 또는 묶음으로 설정합니다.

보조 변수를 지정하면 각 연속형 보조 변수와 범주형 보조 변수의 각 수준에 대한 추가 좌표 테이블이 표시됩니다. 이러한 값은 적재 그림에서 연속형 보조 변수에 대한 그래프로 표시됩니다.

적재 및 좌표의 척도는 주성분을 추출하기 위해 선택한 행렬에 따라 다릅니다.

- "상관계수 행렬 이용" 옵션을 선택하는 경우 적재의  $i$  번째 열은  $i$  번째 고유 벡터에  $i$  번째 고유값의 제곱근을 곱한 값입니다.  $ij$  번째 적재는  $i$  번째 변수와  $j$  번째 주성분 간의 상관입니다.
- "공분산 행렬 이용" 옵션을 선택하는 경우 적재의  $i$  번째 열에서  $j$  번째 항목은  $i$  번째 고유 벡터에  $i$  번째 고유값의 제곱근을 곱하고  $j$  번째 변수의 표준편차로 나눈 값입니다.  $ij$  번째 적재는  $i$  번째 변수와  $j$  번째 주성분 간의 상관입니다.
- "비척도화 및 비중심화" 옵션을 선택하는 경우 적재의  $i$  번째 열에서  $j$  번째 항목은  $i$  번째 고유 벡터에  $i$  번째 고유값의 제곱근을 곱하고  $j$  번째 변수의 표준 오차로 나눈 값입니다.  $j$  번째 변수의 표준 오차는 제곱합 및 교차곱 행렬의  $j$  번째 대각 항목을 행 수로 나눈 값입니다 ( $X'X/n$ ).

**참고 :** 비척도화 및 비중심화 데이터를 분석하는 경우  $ij$  번째 적재는  $i$  번째 변수와  $j$  번째 주성분 간의 상관이 아닙니다.

**형식이 지정된 적재 행렬** 각 성분에 대한 적재 테이블을 표시하거나 숨깁니다. 이 테이블은 첫 번째 주성분에 대한 적재가 큰 것부터 내림차순으로 정렬됩니다. 따라서 변수는 첫 번째 성분에 대한 적재가 큰 것부터 내림차순으로 나열됩니다.

**그림 4.7** 형식이 지정된 적재 행렬

**형식이 지정된 적재 행렬**

	주성분1	주성분2	주성분3	주성분4	주성분5	주성분6
Benzene	0.974761	-0.143460	-0.081914	-0.023369	-0.108186	-0.101488
Carbon Tetrachloride	0.942849	-0.289099	0.069134	-0.059654	-0.099757	0.095746
Hexane	0.923483	-0.263641	0.256572	0.026770	0.099672	-0.034523
Chloroform	0.917422	-0.290351	-0.242516	0.075629	0.093454	0.027695
1-Octanol	0.819019	0.544318	-0.041397	-0.162739	0.068711	0.002762
Ether	0.761978	0.625283	0.044774	0.155130	-0.045337	0.016883

다음보다 작은 절대 적재 값 표시 안 함

텍스트 흐리게

**다음보다 작은 절대 적재 값 표시 안 함** "형식이 지정된 적재 행렬" 보고서에 사용할 수 없는 적재를 결정하는 값입니다. 텍스트 상자 또는 슬라이더를 사용하여 절대값이 선택한 값 미만인 적재를 흐리게 표시할 수 있습니다.

**텍스트 흐리게** "형식이 지정된 적재 행렬" 보고서에서 흐리게 표시된 값의 투명도입니다. 텍스트 상자 또는 슬라이더를 사용하여 흐리게 표시되는 적재의 투명도를 설정할 수 있습니다. 투명도 범위는 0~1 이며, 낮은 값이 높은 값보다 더 투명합니다. 예를 들어 투명도를 0 으로 설정하면 사용할 수 없는 적재가 행렬에서 완전히 제거되지만 투명도를 1 로 설정하면 적재를 계속 사용할 수 있습니다.

**변수의 제공코사인** 변수의 제공코사인을 포함하는 테이블을 표시하거나 숨깁니다. 보조 변수를 지정하면 보조 변수의 제공코사인 테이블이 추가로 표시됩니다. 전체 주성분에 대한 제공코사인 값의 합은 각 변수에 대해 1 과 같습니다. 제공코사인을 사용하면 주성분이 변수를 얼마나 잘 표현하는지 확인할 수 있습니다. 또한 특정 변수를 나타내는 데 필요한 주성분 수를 결정할 수 있습니다. 이 옵션은 성분에 대한 제공코사인 그림도 표시합니다. 가로 막대와 세로

막대 간에 회전할 수 있는 버튼이 있습니다. "변수의 제공코사인 그림"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

**그림 선택** 성분에 대한 제공코사인을 그림에 표시하는 방법을 지정합니다. "개요" 옵션은 지정된 수의 성분에 대한 제공코사인을 표시합니다. "개별" 옵션은 선택한 성분 하나의 제공코사인을 표시합니다.

**막대 스타일** 그림의 막대 스타일을 나란히 또는 묶음으로 설정합니다.

---

**참고 :** "최소" 분산 추정 방법을 사용하고 선택한 성분 수가 3 개 미만이면 지정된 수의 성분만 그림에 표시됩니다.

---

**변수의 부분 기여도** 변수의 부분 기여도를 포함하는 테이블을 표시하거나 숨깁니다. 부분 기여도를 사용하면 각 변수가 각 주성분에 기여하는 비율을 확인할 수 있습니다. 이 옵션은 성분에 대한 부분 기여도 그림도 표시합니다. 가로 막대와 세로 막대 간에 회전할 수 있는 버튼이 있습니다. "변수의 부분 기여도 그림"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

**그림 선택** 성분에 대한 부분 기여도를 그림에 표시하는 방법을 지정합니다. "개요" 옵션은 지정된 수의 성분에 대한 부분 기여도를 표시합니다. "개별" 옵션은 선택한 성분 하나의 부분 기여도를 표시합니다.

**막대 스타일** 그림의 막대 스타일을 나란히 또는 묶음으로 설정합니다.

---

**참고 :** "최소" 분산 추정 방법을 사용하고 선택한 성분 수가 3 개 미만이면 지정된 수의 성분만 그림에 표시됩니다.

---

**요약 그림** 생성된 요약 정보를 기본 보고서에 표시하거나 숨깁니다. 이 요약 정보에는 고유값 그림, 스코어 그림 및 적재 그림이 포함됩니다. 기본적으로 이 보고서에는 처음 두 주성분에 대한 스코어 그림과 적재 그림이 표시됩니다. 표시할 주성분을 지정하는 옵션이 보고서에 있습니다. 자세한 내용은 "[주성분 보고서](#)"에서 확인하십시오.

---

**참고:** 데이터에 결측값이 포함되어 있으면 결측값 대치된 스코어가 스코어 그림에 표시됩니다.

---

**팁 :** 적재 그림에서 화살표 끝을 선택하면 데이터 테이블에서 해당 열이 선택됩니다. Ctrl 키를 누른 채 화살표 끝을 클릭하면 열이 선택 취소됩니다.

---

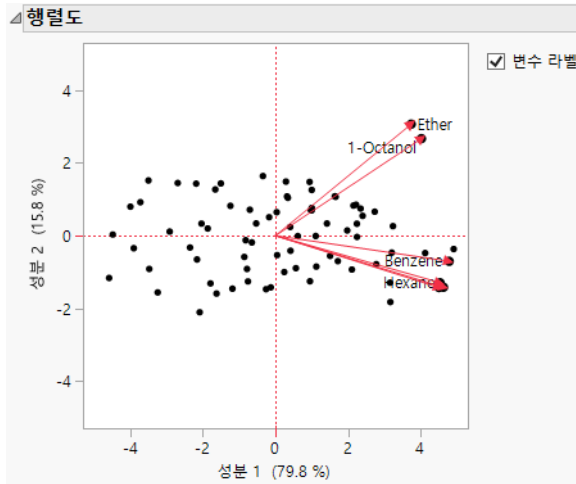
**행렬도** 지정된 수의 성분에 대한 스코어 그림과 적재 그림을 중첩하는 그림을 표시하거나 숨깁니다.

---

**참고 :** 데이터에 결측값이 포함되어 있으면 결측값 대치된 스코어가 행렬도에 표시됩니다.

---

그림 4.8 행렬도

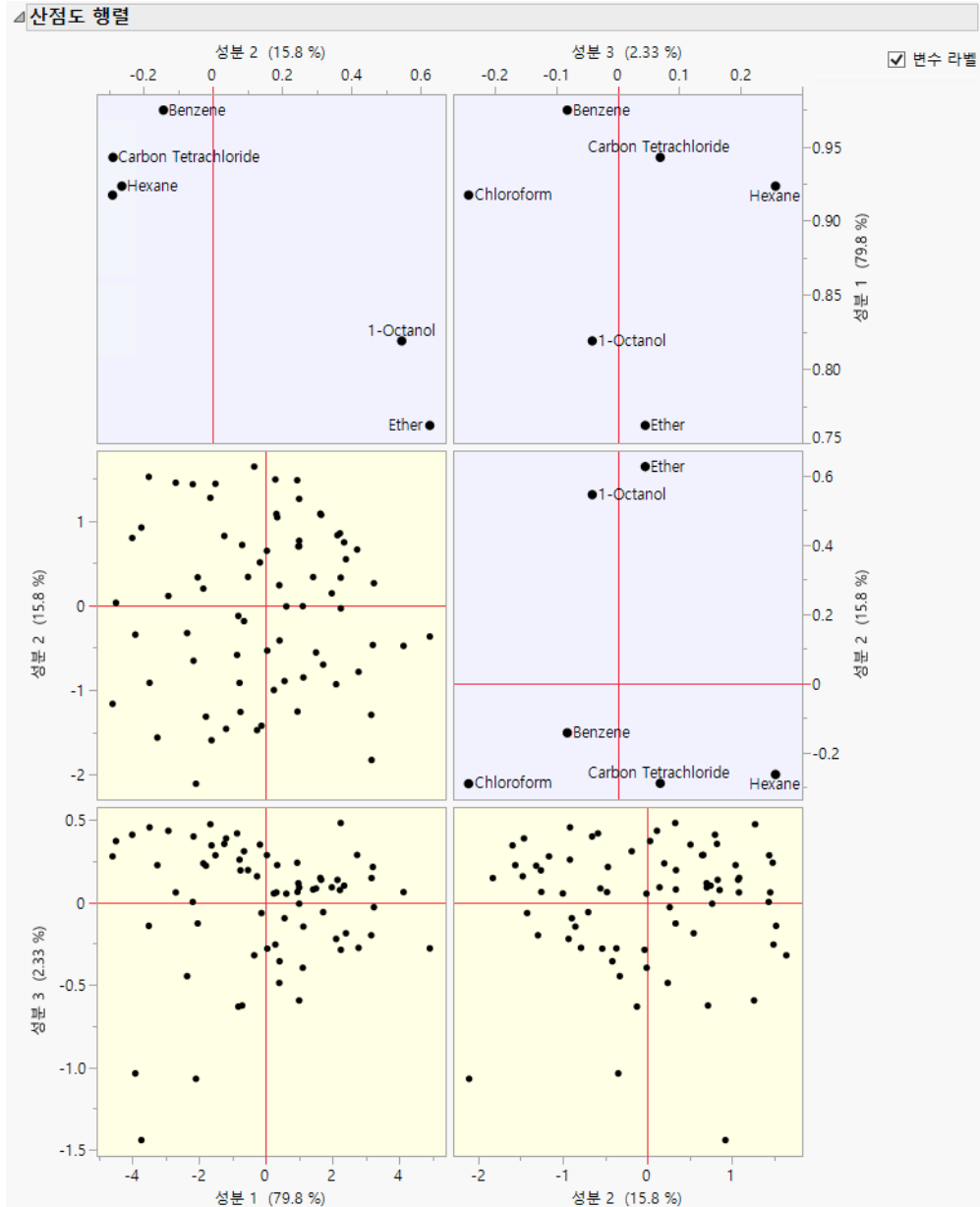


**참고:** 스코어 그림 표식은 점이고 적재 그림 표식은 다이아몬드입니다.

**산점도 행렬** 지정된 수의 주성분에 대한 스코어 그림과 적재 그림의 행렬을 표시하거나 숨깁니다. 산점도 행렬은 스코어 그림과 적재 그림을 모두 한 공간에 배열합니다. 스코어 그림에는 노란색 음영 배경이 있습니다. 적재 그림에는 파란색 음영 배경이 있습니다.

**참고:** 데이터에 결측값이 포함되어 있으면 결측값 대체된 스코어가 산점도 행렬에 표시됩니다.

그림 4.9 산점도 행렬



**참고 :** 산점도 행렬에 표시되는 적재 그림 행렬은 "적재 그림" 옵션을 선택하여 얻는 적재 그림 행렬의 전치입니다.

**스크리 그림** 각 성분에 대한 고유값 그래프를 표시하거나 숨깁니다. 이 스크리 그림은 데이터 공간의 차원을 시각화하는 데 도움이 됩니다.

**스코어 그림** 지정된 수의 성분에 대한 주성분 쌍의 스코어 산점도 행렬을 표시하거나 숨깁니다. 이 그림은 [그림 4.4](#)에 나와 있습니다 (맨 왼쪽 그림).

**적재 그림** 지정된 수의 성분에 대한 요인 적재의 2차원 표현 행렬을 표시하거나 숨깁니다. 변수가 30 개 이하이면 적재 그림의 변수에 라벨이 지정됩니다. 변수가 30 개를 초과하면 라벨이 기본적으로 해제되어 있습니다. 이 정보는 [그림 4.4](#)에 나와 있습니다 (맨 오른쪽 그림).

---

**팁:** 적재 그림에서 화살표 끝을 선택하면 데이터 테이블에서 해당 열이 선택됩니다. Ctrl 키를 누른 채 화살표 끝을 클릭하면 열이 선택 취소됩니다.

---

**결측값 대치를 사용한 스코어 그림** ("와이드" 또는 "최소" 분산 추정 방법에는 사용 불가능) 결측값을 대치하고 스코어 그림을 생성합니다. 이 옵션은 결측값이 있는 경우에만 사용할 수 있습니다.

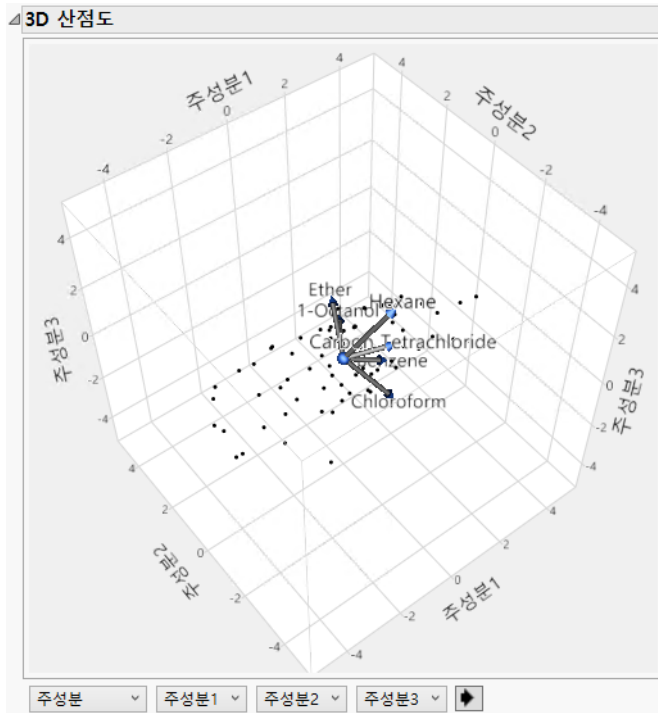
**3D 스코어 그림** ("와이드" 또는 "최소" 분산 추정 방법에는 사용 불가능) 세 가지 주성분 스코어의 3D 산점도를 표시하거나 숨깁니다. 명령을 처음 호출하면 처음 세 개의 주성분이 표시됩니다.

---

**참고:** 데이터에 결측값이 포함되어 있으면 결측값 대치된 스코어가 3D 스코어 그림에 표시됩니다.

---

그림 4.10 3D 산점도 스코어 그림



**그림 소스** 그림에 있는 데이터 점의 소스입니다. 사용 가능한 옵션은 "주성분", "회전된 주성분" 및 "데이터 열"입니다.

**축 컨트롤** 각 축의 콘텐츠입니다. "주성분" 옵션 또는 "회전된 성분" 옵션을 선택한 경우 "축 컨트롤" 옵션은 주성분입니다. "데이터 열" 옵션을 선택한 경우 옵션은 분석의 변수입니다.

**순환 버튼** 가능한 모든 축 콘텐츠 항목을 순환합니다.

변수는 그림에 선으로 표시됩니다. **행렬도 선**이라고 하는 이러한 선은 축에서 주성분의 함수로 변수를 근사화합니다. 변수가 두 개 또는 세 개만 있는 경우 선은 변수를 정확하게 나타냅니다. 이 선은 주성분 적재에 해당합니다.

**스코어 타원** 각 주성분 쌍에 대한 요약 스코어 그림에서 타원을 표시하거나 숨깁니다. 타원은 유의 수준에 기반한 신뢰도 타원 또는 관측값이 중심에서 얼마나 떨어져 있는지에 기반한 관리 한계 타원으로 구성됩니다. 기본적으로 타원은 95% 신뢰도 타원입니다.

**스코어 타원 범위** 스코어 타원의 구성 방법을 변경할 수 있는 하위 메뉴를 표시합니다. 신뢰 수준을 지정하거나, 중심과의 거리를  $k$ -시그마로 지정하여 스코어 타원을 지정합니다. 신뢰 수준  $p$  와  $k$ -시그마 간의 관계는  $p = 1 - \exp(-k^2/2)$  입니다.

**표시 옵션**

**화살표 선** 화살표를 표시할 수 있는 모든 그림에서 화살표를 표시하거나 숨깁니다. 변수가 1000 개 이하이면 화살표가 표시됩니다. 변수가 1000 개를 초과하면 화살표가 기본적으로 표시되지 않습니다.

**보조 변수 표시** (보조 변수를 지정한 경우에만 사용 가능) 행렬도, 스코어 그림 및 적재 그림에서 화살표 선 (연속형 보조 변수의 경우) 또는 라벨 표시 (범주형 보조 변수의 경우) 을 표시하거나 숨깁니다.

**이상치 분석**  $T^2$  및 기여도 통계량을 통해 데이터에서 이상치를 감지할 수 있는 "이상치 분석" 보고서를 표시하거나 숨깁니다. 자세한 내용은 "이상치 분석"에서 확인하십시오.

**요인 분석** ("와이드" 또는 "회소" 분산 추정 방법에는 사용 불가능) 주성분의 요인 분석 스타일 회전 또는 요인 분석을 수행합니다. 자세한 내용은 "요인 분석"에서 확인하십시오.

**JMP PRO 변수 군집화** ("와이드" 또는 "회소" 분산 추정 방법에는 사용 불가능) 변수를 서로 겹치지 않는 군집으로 나누어 변수에 대한 군집 분석을 수행합니다. 변수 군집화는 유사한 변수를 대표 그룹으로 그룹화하는 방법을 제공합니다. 그러면 각 군집을 단일 성분 또는 변수로 나타낼 수 있습니다. 성분은 군집에 있는 모든 변수의 선형 결합입니다. 아니면 군집을 해당 군집에서 가장 대표적인 멤버로 식별되는 변수로 나타낼 수도 있습니다. 자세한 내용은 "변수 군집화"에서 확인하십시오.

**참고:** 변수 군집화에서는 "공분산 행렬 이용" 또는 "비척도화 / 비중심화 행렬 이용" 옵션을 선택한 경우에도 모든 계산에 상관 행렬을 사용합니다.

**모형 기반 다변량 관리도** 지정된 수의 주성분에 대한 계산식을 저장하고 MDMCC (모형 기반 다변량 관리도) 시작 창을 실행합니다. 주성분 계산식은 MDMCC 시작 창에서 공정 열로 할당됩니다. 공정을 추가 또는 제거하고, 시간 ID 를 추가하거나, 과거 데이터 끝 위치를 설정한 후 "확인" 을 클릭할 수 있습니다. 자세한 내용은 품질 및 공정 방법에서 확인하십시오.

**예측값 프로파일러** 각 변수에 대한 예측 계산식이 포함된 "프로파일러" 시작 창을 실행합니다. 예측 계산식은 프로파일러 실행 전에 설정한 지정된 수의 주성분을 사용하는 각 변수에 대한 것입니다. 시작 창에서 "확인" 을 클릭하기 전에 잡음 요인 또는 다른 예측 계산식을 추가할 수 있습니다. 자세한 내용은 프로파일러에서 확인하십시오.

**열 저장** 저장 옵션 하위 메뉴를 제공합니다.

**주성분 저장** 지정한 수의 주성분을 각 성분에 대한 계산식과 함께 데이터 테이블에 저장합니다. 계산식은 결측값이 있는 행을 계산할 수 없습니다.

주성분 계산은 주성분을 추출하기 위해 선택한 행렬에 따라 다릅니다.

- "상관계수 행렬 이용" 옵션을 선택하는 경우  $i$  번째 주성분은  $i$  번째 고유 벡터의 항목을 계수로 사용하여 중심화 및 척도화된 관측값의 선형 결합입니다.
- "공분산 행렬 이용" 옵션을 선택하는 경우  $i$  번째 주성분은  $i$  번째 고유 벡터의 항목을 계수로 사용하여 중심화된 관측값의 선형 결합입니다.
- "비척도화 및 비중심화" 옵션을 선택하는 경우  $i$  번째 주성분은  $i$  번째 고유 벡터의 항목을 계수로 사용하는 원시 관측값의 선형 결합입니다.

**참고:** 지정된 성분 수가 상관 행렬의 계수를 초과하는 경우 저장되는 성분 수는 상관 행렬의 계수로 설정됩니다.

**주성분 값 저장** 지정된 수의 주성분을 데이터 테이블에 저장합니다. 이러한 열은 계산식 열이 아니며 성분 값만 포함합니다.

**팁:** 가로형 데이터의 경우 이 옵션을 사용합니다.

**낮은 계수 주성분 저장** (가로형 또는 세로형 데이터에 대해 시작 창에서 "로버스트 PCA" 추정 방법을 지정된 경우에만 사용 가능) 낮은 계수 행렬에서 이상치와 잡음이 제거된 주성분 스코어를 저장합니다.

**예측값 저장** 주성분 수가 지정되면 예측된 변수를 데이터 테이블의 새 열에 저장합니다.

**예측값을 성분 계산식으로 저장** 지정된 수의 성분을 선형 결합한 예측된 변수를 계산식으로 저장합니다. 계산식에 사용된 주성분도 데이터 테이블에 저장됩니다.

**정규화된 DModX 저장** 주성분 수가 지정되면 DModX(주성분 모형에 대한 관측값 거리)를 데이터 테이블의 새 열에 저장합니다. DModX 값이 클수록 데이터의 이상치가 사소하거나 중간 정도임을 나타냅니다. 자세한 내용은 **품질 및 공정 방법**에서 확인하십시오.

**개별 제공코사인 저장** 주성분 수가 지정되면 개별 제공코사인을 데이터 테이블의 새 열에 저장합니다.

**개별 부분 기여도 저장** 주성분 수가 지정되면 개별 부분 기여도를 데이터 테이블의 새 열에 저장합니다.

**회전된 성분 저장** ("와이드" 또는 "최소" 분산 추정 방법에는 사용 불가능) 회전된 성분을 성분에 대한 계산식과 함께 데이터 테이블에 저장합니다. 이 옵션은 "요인 분석" 옵션을 사용한 후에만 사용할 수 있습니다. 계산식은 결측값이 있는 행을 계산할 수 없습니다.

**결측값 대치를 사용한 주성분 저장** ("와이드" 또는 "최소" 분산 추정 방법에는 사용 불가능) 결측값을 대치하고 주성분을 데이터 테이블에 저장합니다. 결측값 대치 및 주성분 계산을 수행하기 위한 계산식이 열에 포함됩니다. 이 옵션은 결측값이 있는 경우에만 사용할 수 있습니다.

**결측값 대치를 사용한 회전된 성분 저장** ("와이드" 또는 "최소" 분산 추정 방법에는 사용 불가능) 결측값을 대치하고 회전된 성분을 데이터 테이블에 저장합니다. 결측값 대치 및 회전된 성분 계산을 수행하기 위한 계산식이 열에 포함됩니다. 이 옵션은 "요인 분석" 옵션을 사용한 후 결측값이 있는 경우에만 사용할 수 있습니다.

**결측값 대치 계산식 저장** (데이터 테이블에 결측값이 포함된 경우에만 사용 가능) 결측값이 포함된 열에 대해 결측값 추정에 사용된 계산식을 포함하는 새 열을 데이터 테이블에 저장합니다. 새 열 이름은 결측값 대치됨 < 열 이름 > 입니다.

**JMP PRO 성분 계산식 게시** 지정된 수의 주성분 계산식을 생성하여 계산식 저장소 플랫폼에 계산식 열 스크립트로 저장합니다. "계산식 저장소" 보고서가 열려 있지 않으면 "계산식 저장소" 보고서가 생성됩니다. 자세한 내용은 **예측 및 전문 모델링**에서 확인하십시오.

**JMP PRO** 정규화된 DModX 계산식 게시 지정된 수의 주성분에 기반한 정규화된 DModX 계산식을 계산식 저장소 플랫폼에 계산식 열 스크립트로 저장합니다. "계산식 저장소" 보고서가 열려 있지 않으면 "계산식 저장소" 보고서가 생성됩니다. 자세한 내용은 예측 및 전문 모델링에서 확인하십시오.

다음 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

**로컬 데이터 필터** 특정 보고서에서 사용되는 데이터를 필터링할 수 있는 로컬 데이터 필터를 표시하거나 숨깁니다.

**다시 실행** 분석을 반복하거나 다시 시작할 수 있는 옵션이 포함되어 있습니다. 이 기능을 지원하는 플랫폼에서 "자동 재계산" 옵션은 해당하는 보고서 창에서 데이터 테이블에 대한 변경 사항을 즉시 반영합니다.

**플랫폼 환경 설정** 현재 플랫폼 환경 설정을 보거나, 현재 JMP 보고서의 설정과 일치하도록 플랫폼 환경 설정을 업데이트할 수 있는 옵션이 포함되어 있습니다.

**스크립트 저장** 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다.

**그룹별 스크립트 저장** 기준 변수의 모든 수준에 대한 플랫폼 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다. 시작 창에서 기준 변수를 지정한 경우에만 사용할 수 있습니다.

---

**참고 :** 이 플랫폼의 추가 옵션은 스크립트를 통해 사용할 수 있습니다. 자세한 내용은 "도움말" 메뉴의 "스크립트 인덱스"에서 여십시오. 또한 "스크립트 인덱스"에서 이 섹션에 설명된 옵션의 스크립트 예제도 찾을 수 있습니다.

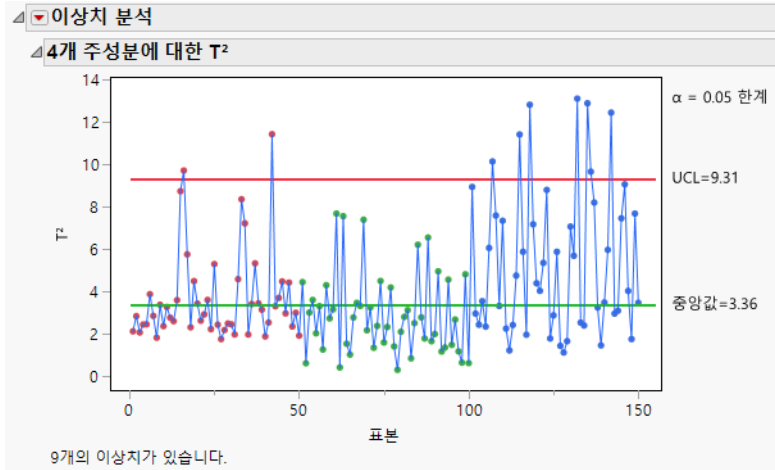
---

## 이상치 분석

주성분 플랫폼의 "이상치 분석" 보고서에는 "<A>개 주성분에 대한  $T^2$ " 그림이 기본적으로 표시됩니다. 이 그림에는 각 관측값에 대한  $T^2$  값이 표시되고 중앙값 및 UCL(관리 상한)을 나타내는 가로선이 있습니다.  $T^2$  값, 중앙값 및 UCL을 계산하는 방법에 대한 자세한 내용은 "이상치 분석 계산에 대한 통계 상세 정보"에서 확인하십시오.

UCL을 계산하는 데 사용된  $\alpha$  수준이 그림 옆에 표시됩니다. 감지된 이상치 수는 그림 아래에 표시됩니다. 이 수는  $T^2$  값이 UCL보다 큰 관측값 수입니다.

그림 4.11 이상치 분석 보고서



**팁 :**  $T^2$  그림의 점을 커서로 가리키면 해당 관측값에 대한  $T^2$  기여 비율 그림을 볼 수 있습니다.  $T^2$  기여 비율 그림을 클릭하면 보고서 창에 그림이 추가됩니다.

## 이상치 분석 보고서 옵션

"이상치 분석"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

**$T^2$  그림**  $T^2$  그림을 표시하거나 숨깁니다. 기본적으로 설정되어 있습니다.

**기여도 히트맵** 모든 관측값에 대한  $T^2$  기여도 값의 히트맵을 표시하거나 숨깁니다.

**기여 비율 히트맵** 모든 관측값에 대한  $T^2$  기여도 값을 개별 행의  $T^2$  비율로 표현한 히트맵을 표시하거나 숨깁니다. 개별 행의 비율은 기여도 제곱을 계산한 후 해당 개별 행에 대한 모든 기여도의 제곱합으로 나누어 구합니다.

**선택된 표본에 대한 기여도 그림** ( $T^2$  그림에서 점을 하나 이상 선택한 경우에만 사용 가능) 선택된 각 표본에 대한  $T^2$  기여도 그림이 포함된 보고서를 표시합니다.  $T^2$  기여도 그림은 표본의  $T^2$  통계량에 대한 각 변수의 기여도를 보여 줍니다. 기여도를 계산하는 방법에 대한 자세한 내용은 품질 및 공정 방법에서 확인하십시오. 기여도 그림을 사용하여 이상치를 조사할 수 있습니다. 양의 기여도 또는 음의 기여도가 가장 큰 변수가  $T^2$  값이 큰 표본에 가장 많이 기여하는 변수입니다. 빨간색 삼각형 메뉴 옵션에 대한 자세한 내용은 "**기여도 그림 보고서 옵션**"에서 확인하십시오.

**선택된 표본에 대한 기여 비율 그림** ( $T^2$  그림에서 점을 하나 이상 선택한 경우에만 사용 가능) 선택된 각 표본에 대한  $T^2$  기여 비율 그림이 포함된 보고서를 표시합니다.  $T^2$  기여 비율 그림에는 선택된 관측값에 대한 기여도 값이 개별 행의  $T^2$  비율로 표시됩니다. 이것은 기여 비율 히트맵의 정보를 또 다르게 표현한 것입니다. 빨간색 삼각형 메뉴 옵션에 대한 자세한 내용은 "**기여도 그림 보고서 옵션**"에서 확인하십시오.

**정규화된 DModX 그림** (성분 수가 변수 수보다 작은 경우에만 사용 가능) 정규화된 DModX 값의 그림을 표시하거나 숨깁니다. DModX 값은 데이터에서 중간 이상치를 감지하는 데 유용합니다.

**성분 수**  $T^2$  및  $T^2$  기여도 통계량에 사용되는 주성분 수를 지정할 수 있습니다. 성분 수를 변경하면  $T^2$  그림, 히트맵 및 정규화된 DModX 그림이 자동으로 업데이트됩니다.

**$\alpha$  수준 설정**  $\alpha$  수준을 지정할 수 있습니다.

**$T^2$  저장**  $T^2$  값을 데이터 테이블의 새 열에 저장합니다.

**기여도 저장**  $T^2$  기여도를 데이터 테이블의 새 열에 저장합니다. 각 Y 변수에 대해 열이 하나씩 있습니다.

**정규화된 DModX 저장** (성분 수가 변수 수보다 작은 경우에만 사용 가능) 정규화된 DModX 값을 데이터 테이블의 새 열에 저장합니다.

### 기여도 그림 보고서 옵션

"선택된 표본에 대한  $T^2$  기여도 그림", "선택된 표본에 대한  $T^2$  기여도 비율 그림" 및 "선택된 표본에 대한  $T^2$  평균 기여도 비율 그림"에는 다음과 같은 빨간색 삼각형 메뉴 옵션이 포함되어 있습니다.

**막대 라벨** 기여도 그림의 막대에 라벨을 지정하는 옵션이 포함된 하위 메뉴를 표시합니다. 라벨 옵션은 "라벨 없음", "값별 라벨" 및 "열별 라벨"입니다.

**그림 제거** 보고서에서 기여도 그림을 제거합니다.

**선택한 항목에 대한 관리도** (기여도 그림에서 막대 세그먼트를 하나 이상 선택한 경우에만 사용 가능) 선택한 각 공정 및 그룹에 대한 관리도 결과가 포함된 관리도 빌더 창을 엽니다.

---

**팁:** 기여도 그림에서 막대 세그먼트를 커서로 가리켜 관리도를 볼 수도 있습니다. 관리도를 클릭하면 관리도 빌더 창이 열립니다.

---

## 주성분 플랫폼에 대한 통계 상세 정보

이 섹션에는 주성분 플랫폼에 대한 통계 상세 정보가 포함되어 있습니다.

- "분산 추정 방법에 대한 통계 상세 정보"
- "고급 방법에 대한 통계 상세 정보"
- "이상치 분석 계산에 대한 통계 상세 정보"

## 분산 추정 방법에 대한 통계 상세 정보

이 섹션에는 주성분 시각 창에서 "전통적"을 "방법 유형"으로 지정한 경우 사용되는 분산 추정 방법에 대한 통계 상세 정보가 포함되어 있습니다.

### REML

데이터에 결측값이 포함된 경우 REML(제한 최대 가능도) 추정값은 ML(최대 가능도) 추정 방법보다 덜 편향됩니다. REML 방법은 오차 대비를 기반으로 주변 가능도를 최대화합니다. REML 방법은 주로 분산 및 공분산을 추정하는 데 사용됩니다. 주성분 플랫폼의 REML 방법은 비정형 공분산 행렬을 사용하는 반복 측정 데이터에 대한 혼합 모형의 REML 추정과 동일합니다. 혼합 모형의 REML 추정에 대한 자세한 내용은 SAS PROC MIXED 관련 문서에서 확인하십시오.

### 고급 방법에 대한 통계 상세 정보

이 섹션에는 주성분 시각 창에서 "고급"을 "방법 유형"으로 지정한 경우 사용되는 추정 방법에 대한 통계 상세 정보가 포함되어 있습니다. 이러한 방법은 SVD(특이값 분해)를 기반으로 합니다. 이 알고리즘은 공분산 행렬을 계산하지 않으므로 계산이 효율적입니다.

### 완전 SVD

완전 SVD 방법은 행렬의 완전 특이값 분해를 기반으로 하는 알고리즘을 사용합니다. 다음 표기를 고려하십시오.

- $n$  = 행 수
- $p$  = 변수 수
- $\mathbf{X}$  = 데이터 값의  $n \times p$  행렬

0이 아닌 고유값 수와 이에 따른 주성분 수는  $\mathbf{X}$ 의 상관 행렬 계수와 같습니다. 0이 아닌 고유값 수는  $n$ 과  $p$  중 더 작은 값을 초과할 수 없습니다.

데이터는 완전 SVD 방법을 구현하기 전에 표준화됩니다. 값을 표준화하려면 평균을 뺀 후 표준편차로 나눕니다. 표준화된 데이터 값의  $n \times p$  행렬을  $\mathbf{X}_s$ 로 나타냅니다. 그러면 표준화된 데이터의 공분산 행렬은  $\mathbf{X}$ 의 상관 행렬이며 다음과 같이 정의됩니다.

$$Cov = \mathbf{X}_s' \mathbf{X}_s / (n - 1)$$

특이값 분해를 사용하면  $\mathbf{X}_s$ 가  $\mathbf{U} \mathbf{D} \mathbf{V}'$ 로 표기됩니다. 이 표현은  $\mathbf{X}_s' \mathbf{X}_s$ 의 고유 벡터 및 고유값을 얻는 데 사용됩니다. 주성분 또는 스코어는  $\mathbf{X}_s \mathbf{V}$ 로 제공됩니다. 추가 배경 정보는 "[와이드 선형 방법 및 특이값 분해](#)"에서 확인하십시오.

### 절단 SVD

절단 SVD 방법은 특이값 분해를 기반으로 하지만 완전 분해를 구현하지 않는 알고리즘을 사용합니다. 대신 이 알고리즘은 특이값 분해에서 첫 번째 지정한 수의 특이값과 특이 벡터만 계산합니다. 따라서 첫 번째 지정한 수의 고유값과 주성분만 반환됩니다. 이 알고리즘에 대한 자세한 내용은 Baglama 와 Reichel(2005) 연구 자료에서 확인하십시오.

### 랜덤화 SVD

랜덤화 SVD 방법은 특이값 분해를 기반으로 하지만 완전 분해를 구현하지 않는 알고리즘을 사용합니다. 이 알고리즘은 낮은 계수 행렬 근사를 포함하는 2단계 프로세스입니다. 일반적으로 낮은 계수 행렬 근사의 목표는  $k$ 가  $p$ 보다 훨씬 작고  $\mathbf{A} \approx \mathbf{BC}$ 가 되도록  $m \times k$  행렬  $\mathbf{B}$ 와  $k \times p$  행렬  $\mathbf{C}$ 를 찾아  $m \times p$  행렬  $\mathbf{A}$ 에 근사한 값을 산출하는 것입니다.

"완전 SVD"에 설명된 것과 동일한 표기를 고려하십시오. 여기서 목표는  $\mathbf{X}_s$ 를 분해하는 것입니다.

랜덤화 SVD 알고리즘의 첫 번째 단계에서는 다음 조건을 충족하는 행렬  $\mathbf{Q}$ 를 찾습니다.

- $\mathbf{Q}$ 에  $l$ 개의 직교정규 열이 있음 ( $k \leq l \leq p$ )
- $\mathbf{X}_s \approx \mathbf{Q}\mathbf{Q}'\mathbf{X}_s$

가능한 한 적은 수의 열이 있는  $\mathbf{Q}$ 를 효율적으로 계산하기 위한 상세 정보는 Halko, Martinsson & Tropp(2011)에서 확인하십시오.

알고리즘의 두 번째 단계에서는  $\mathbf{Q}$ 를 사용하여  $\mathbf{X}_s$ 의 특이값 분해를 계산합니다.  $\mathbf{B} = \mathbf{Q}'\mathbf{X}_s$ 라고 하면 다음 조건을 충족합니다.

$$\mathbf{B} = \mathbf{Q}'\mathbf{X}_s \Rightarrow \mathbf{QB} = \mathbf{Q}\mathbf{Q}'\mathbf{X}_s \Rightarrow \mathbf{QB} = \mathbf{X}_s$$

그런 다음  $\mathbf{B}$ 의 SVD를 계산합니다( $\mathbf{X}_s$ 보다 훨씬 작음).

$$\mathbf{B} \approx \mathbf{U}'\text{Diag}(\Lambda)\mathbf{V}'$$

$$\mathbf{QB} \approx (\mathbf{Q}\mathbf{U}')\text{Diag}(\Lambda)\mathbf{V}'$$

$$\mathbf{A} \approx \mathbf{U}\text{Diag}(\Lambda)\mathbf{V}', \text{ 여기서 } \mathbf{U} = \mathbf{Q}\mathbf{U}'$$

완전 랜덤화 SVD 알고리즘에 대한 자세한 내용은 Halko, Martinsson & Tropp(2011)에서 확인하십시오.

### 이상치 분석 계산에 대한 통계 상세 정보

주성분 플랫폼에서 "이상치 분석" 보고서의 계산에 다음 표기가 사용됩니다.

$n$  = 관측값 수

$A$  = 주성분 수

$X_{ci}$  =  $i$  번째 관측값에 대한 표준화된 데이터

**T<sup>2</sup> 통계량**

$i$  번째 관측값에 대한  $T^2$  통계량은 다음과 같이 계산됩니다.

$$T_i^2 = X_{ci} P_A L^{-1} P_A^T X_{ci}^T$$

여기서  $P_A$ 는 처음  $A$ 개 고유 벡터를 포함하는 행렬이고,  $L$ 은 처음  $A$ 개 고유값을 포함하는 대각 행렬입니다.

$T^2$  그림의 중앙값과 UCL 은 다음과 같이 계산됩니다.

$$CL_{T^2, q} = \frac{(n-1)^2}{n} \beta \left[ q, \frac{A}{2}, \frac{n-A-1}{2} \right]$$

여기서,

$$\beta \left[ q, \frac{A}{2}, \frac{n-A-1}{2} \right] = \text{베타} \left( \frac{A}{2}, \frac{n-A-1}{2} \right) \text{ 분포의 } q \text{ 번째 분위수}$$

중앙값을 계산하려면  $q=0.5$ 를 사용합니다. UCL을 계산하려면  $q=(1-\alpha)$ 를 사용합니다.

기여도 통계량에 대한 자세한 내용은 품질 및 공정 방법에서 확인하십시오.

## 판별 분석

### 연속형 변수를 기반으로 분류 예측

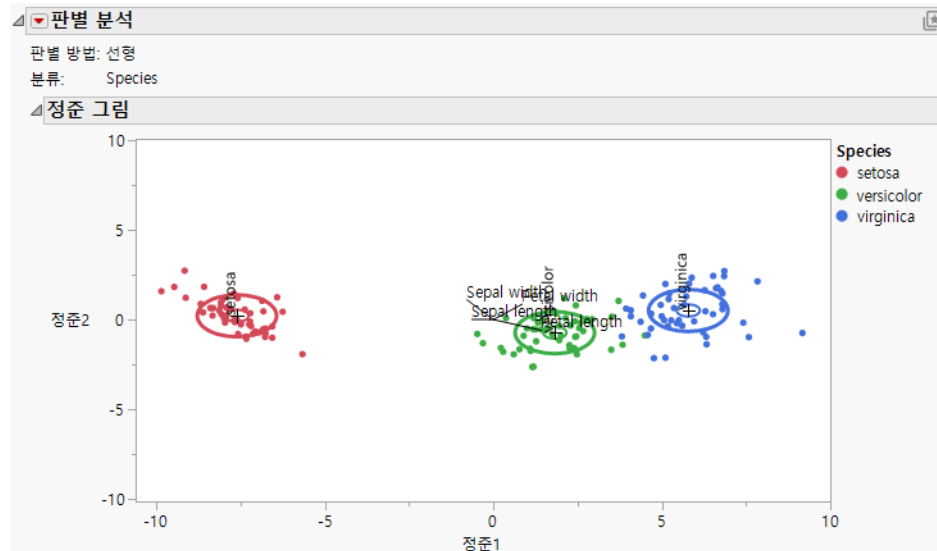
판별 분석은 여러 연속형 변수의 관측값을 기반으로 그룹 또는 범주의 소속을 예측합니다. 특히 판별 분석은 알려진 연속형 반응(Y)을 기반으로 범주형 분류(X) 변수를 예측합니다. 판별 분석을 위한 데이터는 소속 그룹이 알려진 관측값 표본과 연속형 변수에 대한 값으로 구성됩니다.

예를 들어 기대 수익성에 따라 대출 신청인을 저금리 대출, 장기 대출 또는 대출 없음이라는 세 가지 대출 범주(X)로 분류하려고 할 수도 있습니다. 현재 급여, 현재 직장에서의 근무 기간, 나이 및 부채 부담과 같은 연속형 변수(Y)를 사용하여 개인에게 가장 유리한 대출 범주를 예측할 수 있습니다. 판별 분석을 사용하여 개인을 대출 범주로 분류하는 예측 모형을 생성할 수 있습니다.

판별 플랫폼의 기능은 다음과 같습니다.

- 잘 구별하는 변수를 선택하는 데 도움이 되는 단계별 선택 옵션
- 적합 방법 선택: 선형, 2 차, 정규화 및 와이드 선형
- 정준 그림 및 오분류 요약
- 각 그룹에 대한 제곱 거리 및 판별 스코어
- 예측 거리와 확률을 데이터 테이블에 저장하는 옵션

그림 5.1 정준 그림



## 목차

판별 플랫폼 개요 .....	81
판별 분석의 예 .....	81
판별 플랫폼 시작 .....	83
단계별 변수 선택 .....	84
판별 방법 .....	86
공분산 축소 .....	89
판별 분석 보고서 .....	89
주성분 .....	90
정준 그림 및 정준 구조 .....	91
판별 스코어 .....	94
스코어 요약 .....	95
판별 분석 옵션 .....	98
정준 상세 정보 표시 .....	102
정준 구조 표시 .....	104
새 수준 고려 .....	104
판별 행렬 저장 .....	104
JMP 및 JMP Pro 에서 검증 .....	105
판별 분석의 추가 예 .....	106
3D 정준 그림의 예 .....	106
단계별 변수 선택의 예 .....	107
판별 플랫폼에 대한 통계 상세 정보 .....	108
와이드 선형 알고리즘에 대한 통계 상세 정보 .....	109
저장된 계산식에 대한 통계 상세 정보 .....	109
다변량 검정에 대한 통계 상세 정보 .....	116
근사 F-검정에 대한 통계 상세 정보 .....	117
그룹 간 공분산 행렬에 대한 통계 상세 정보 .....	118

---

## 판별 플랫폼 개요

판별 분석은 연속형 변수에 대한 값으로 설명되는 관측값을 그룹으로 분류하려고 시도합니다. 범주형 변수  $X$ 로 정의되는 소속 그룹은 연속형 변수에 의해 예측됩니다. 이러한 변수를 공변량이라고 하며  $Y$ 로 나타냅니다.

판별 분석은 로지스틱 회귀와 다릅니다. 로지스틱 회귀에서 분류 변수는 확률 변수이며 연속형 변수에 의해 예측됩니다. 판별 분석에서 분류는 고정되어 있으며 공변량( $Y$ )이 확률 변수의 구현입니다. 그러나 두 기법 모두 연속형 변수에 의해 범주형 값이 예측됩니다.

판별 플랫폼에서는 모형 적합을 위한 네 가지 방법을 제공합니다. 모든 방법은 Mahalanobis 거리를 사용하여 각 관측값에서 각 그룹의 다변량 평균(중심)까지의 거리를 추정합니다. 소속 그룹의 사전 확률을 지정할 수 있으며 이러한 확률은 거리 계산에서 고려됩니다. 관측값은 가장 근접한 그룹으로 분류됩니다.

다음과 같은 적합 방법을 사용할 수 있습니다.

- 선형 — 그룹 내 공분산 행렬이 같다고 가정합니다.  $X$ 에 의해 정의되는 그룹의 공변량 평균은 서로 다르다고 간주됩니다.
- 2차 — 그룹 내 공분산 행렬이 다르다고 가정합니다. 이 방법은 선형 방법보다 더 많은 모수를 추정해야 합니다. 그룹 표본 크기가 작으면 추정값이 불안정할 수 있습니다.
- 정규화 — 그룹 내 공분산 행렬이 다를 때 추정값에 안정성을 부여하는 두 가지 방법을 제공합니다. 이 옵션은 그룹 표본 크기가 작을 때 유용합니다.
- 와이드 선형 — 다른 방법을 사용하면 계산상 문제가 발생할 수 있는 다수의 공변량을 기반으로 모형을 적합시키는 데 유용합니다. 이 경우 모든 공분산 행렬이 같다고 가정합니다.

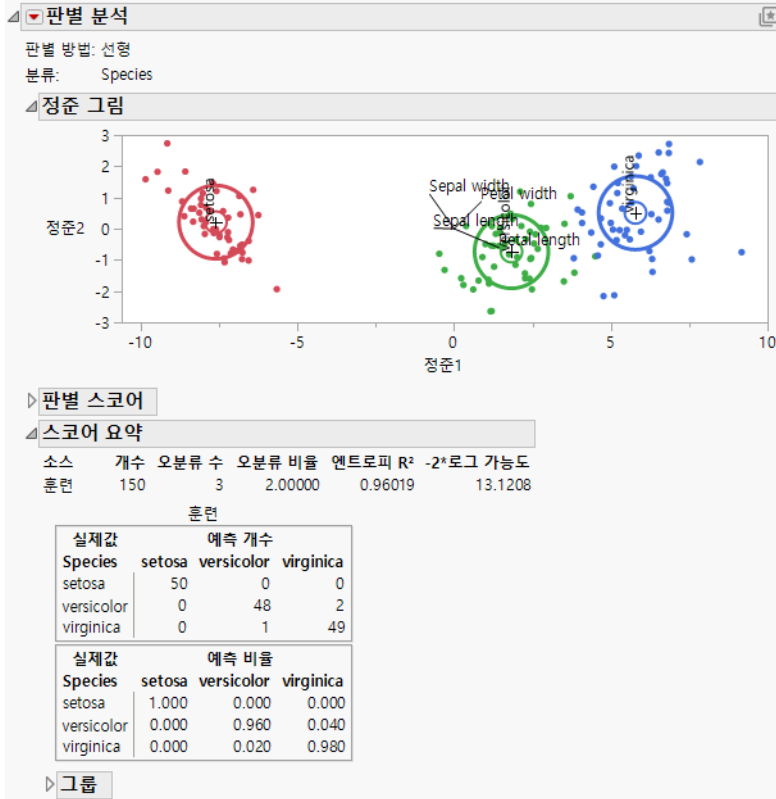
---

## 판별 분석의 예

이 예에서는 화초에서 추출한 네 개의 측정값을 사용하여 붓꽃의 종을 정확하게 식별하는 것이 목표입니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Iris.jmp 를 엽니다.
2. **분석 > 다변량 방법 > 판별**을 선택합니다.
3. Sepal length, Sepal width, Petal length 및 Petal width 를 선택하고 **Y, 공변량**을 클릭합니다.
4. Species 를 선택하고 **X, 범주**를 클릭합니다.
5. **확인**을 클릭합니다.

그림 5.2 판별 분석 보고서 창

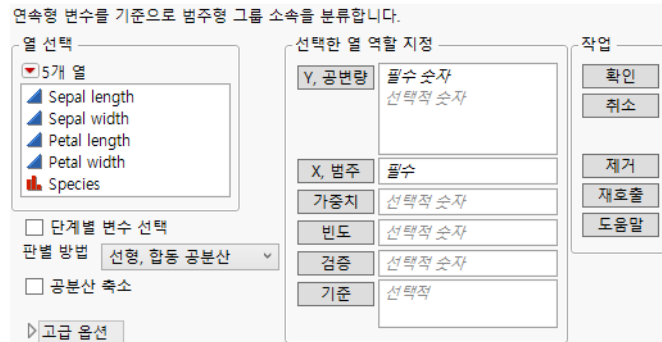


Species에 대한 클래스가 세 개이므로 두 개의 정준 변수가 있습니다. "정준 그림"에서 두 정준 좌표에 대해 각 관측값이 표시됩니다. 이 그림에서는 두 좌표에 세 가지 종이 구분되어 있음을 보여 줍니다. 검증 데이터 집합이 없으므로 "스코어 요약" 보고서에 훈련 데이터 집합에 대한 패널만 표시됩니다. 검증 데이터 집합이 없으면 전체 데이터 집합이 훈련 데이터 집합으로 간주됩니다. 150개 관측값 중 세 개만 잘못 분류되었습니다.

## 판별 플랫폼 시작

분석 > 다변량 방법 > 판별을 선택하여 판별 플랫폼을 시작합니다.

그림 5.3 Iris.jmp 에 대한 판별 시작 창



"열 선택"의 빨간색 삼각형 메뉴에 포함된 옵션에 대한 자세한 내용은 **JMP 사용**에서 확인하십시오.

**참고 :** "검증" 버튼은 JMP Pro에만 나타납니다. JMP에서는 제외된 행을 사용하여 검증 데이터 집합을 정의할 수 있습니다. 자세한 내용은 ["JMP 및 JMP Pro에서 검증"](#)에서 확인하십시오.

**Y, 공변량** 관측값을 범주로 분류하는 데 사용할 연속형 변수를 포함하는 열입니다.

**X, 범주** 관측값을 분류할 범주 또는 그룹을 포함하는 열입니다.

**가중치** 분석을 위해 각 행에 가중치를 할당하는 값이 들어 있는 열입니다.

**빈도** 분석을 위해 각 행에 빈도를 할당하는 값이 들어 있는 열입니다. 일반적으로 빈도 열의 효과는 데이터 테이블을 확장하여 정수 빈도가  $k$ 인 행을  $k$ 개의 행으로 확장하는 것입니다. 빈도 비율을 지정할 수 있습니다.

**JMP PRO 검증** 검증 데이터 집합을 정의하는 숫자 열입니다. 이 열에는 최대 세 개의 구분되는 값이 포함되어야 합니다.

- 두 개의 값이 있는 경우 더 작은 값이 훈련 데이터 집합을 정의하고 더 큰 값이 검증 데이터 집합을 정의합니다.
- 세 개의 값이 있는 경우 작은 값부터 순서대로 훈련 데이터 집합, 검증 데이터 집합, 테스트 데이터 집합을 각각 정의합니다.
- 검증 열에 네 개 이상의 수준이 있는 경우 가장 작은 세 개의 값을 포함하는 행이 검증 데이터 집합을 정의합니다. 다른 모든 행은 분석에서 제외됩니다.

판별 플랫폼에서는 "단계별 변수 선택"이 사용되는 경우 외에는 검증 열을 사용하여 모형을 훈련하고 평가합니다. 시작 창에서 "단계별 변수 선택" 옵션을 선택하는 경우 판별 플랫폼

에서는 검증 열을 사용하여 모형을 훈련 및 조정하거나 모형을 훈련, 조정 및 평가합니다. 검증에 대한 자세한 내용은 **예측 및 전문 모델링**에서 확인하십시오.

**팁:** "단계별 변수 선택" 을 사용하지 않는 경우 검증 열에는 구분되는 값이 두 개만 포함되어야 합니다.

"열 선택" 목록에서 아무 열도 선택하지 않은 상태로 "검증" 버튼을 클릭하면 데이터 테이블에 검증 열을 추가할 수 있습니다. "검증 열 생성" 유틸리티에 대한 자세한 내용은 **예측 및 전문 모델링**에서 확인하십시오.

**기준** 지정된 열의 각 수준에 대해 별도의 분석을 수행합니다.

**단계별 변수 선택** ("판별 방법" 에서 "와이드 선행" 을 선택한 경우에는 사용 불가능) 공분산 분석 및  $p$  값을 사용하여 단계별 변수 선택을 수행합니다. 자세한 내용은 "**단계별 변수 선택**" 에서 확인하십시오.

검증 데이터 집합을 지정한 경우 검증 데이터 집합에 대한 통계량도 나타냅니다. 검증 데이터 집합 통계량은 "시작" 버튼을 사용하는 경우 수행할 단계 수를 결정하는 데 사용됩니다.

**판별 방법** 판별 분석을 수행하기 위한 네 가지 방법을 제공합니다. 자세한 내용은 "**판별 방법**" 에서 확인하십시오.

**공분산 축소** 합동 그룹 내 공분산 행렬과 그룹 내 공분산 행렬의 비대각 요소를 축소합니다. 이로 인해 안정성이 향상되고 예측 분산을 줄일 수 있습니다. 자세한 내용은 "**공분산 축소**" 에서 확인하십시오.

**고급 옵션** 다음 옵션이 포함되어 있습니다.

**비중심화된 정준** 이전 버전 JMP 와의 호환성을 위해 정준 스코어 중심화를 제한합니다.

**유사역행렬 사용** 공분산 행렬이 특이 행렬일 때 Moore-Penrose 유사역행렬을 분석에 사용합니다. 결과 스코어에 모든 공변량이 포함됩니다. 이 옵션을 선택하지 않으면 **Y, 공변량** 목록에서 앞에 나오는 공변량의 선형 결합인 공변량이 분석에서 삭제됩니다.

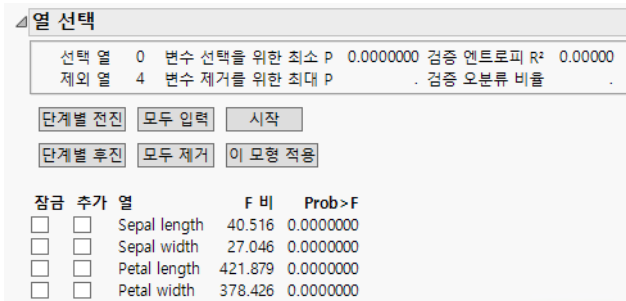
**제외된 행으로 교차 검증** 제외된 행이 적합 통계량이 계산되는 검증 데이터 집합을 구성하도록 지정합니다.

## 단계별 변수 선택

시작 창에서 "단계별 변수 선택" 옵션을 선택하면 "판별 분석" 보고서에 "열 선택" 패널이 포함됩니다. 버튼을 사용하여 변수를 선택하거나, "잠금" 및 "추가" 체크박스로 변수를 수동 선택하여 단계별 분석을 수행할 수 있습니다. 선택한 항목에 따라  $F$  비와  $p$  값이 업데이트됩니다. 업데이트 방식에 대한 자세한 내용은 "**F 비와 Prob>F 업데이트**"에서 확인하십시오.

검증 데이터 집합 유형을 지정하면 "시작" 버튼이 나타납니다. "시작"을 클릭하면 JMP에서 검증 데이터 집합 통계량을 사용하여 수행할 단계 수를 결정합니다.

그림 5.4 검증 데이터 집합을 사용한 Iris.jmp의 열 선택 패널



### F 비와 Prob>F 업데이트

모형에서 변수를 입력하거나 제거하면 다음과 같은 구조의 공분산 모형 분석을 기반으로 F 비와 Prob>F 값이 업데이트됩니다.

- 고려 중인 공변량이 반응입니다.
- 모형에 이미 입력된 공변량이 예측 변수입니다.
- 그룹 변수가 예측 변수입니다.

단계별 보고서에 제공된 "F 비"와 "Prob>F" 값은 그룹 변수에 대한 공분산 분석 검정의 F 비와  $p$  값입니다. 그룹 변수에 대한 공분산 분석 검정은 고려 중인 공변량에 상대적인 판별력을 나타내는 지표입니다.

### 통계량

**선택 열** 판별 모형에 입력하기 위해 현재 선택된 열의 수입입니다.

**제외 열** 판별 모형에 현재 입력할 수 있는 열의 수입입니다.

**변수 선택을 위한 최소 P** 모형에 입력할 수 있는 모든 공변량의  $p$  값 중 가장 작은  $p$  값입니다.

**변수 제거를 위한 최대 P** 모형에 입력하기 위해 현재 선택된 모든 공변량의  $p$  값 중 가장 큰  $p$  값입니다.

**검증 엔트로피 R<sup>2</sup>** 검증 데이터 집합에 대한 엔트로피 R<sup>2</sup>입니다. 값이 클수록 더 나은 적합을 나타냅니다. "엔트로피 R<sup>2</sup>" 값이 1 이면 분류가 완벽하게 예측되었음을 나타냅니다. 예측 확률의 불확도는 판별 모형에서 일반적으로 사용되므로 엔트로피 R<sup>2</sup> 값은 작은 경향이 있습니다. 자세한 내용은 "엔트로피 R<sup>2</sup>"에서 확인하십시오. 검증 데이터 집합이 사용되는 경우에만 사용할 수 있습니다.

**참고:** 검증 엔트로피 R<sup>2</sup>이 음수일 수 있습니다.

**검증 오분류 비율** 검증 데이터 집합에 대한 오분류 비율입니다. 값이 작을수록 더 나은 분류를 나타냅니다. 검증 데이터 집합이 사용되는 경우에만 사용할 수 있습니다.

**버튼**

**단계별 전진** 아직 입력되지 않은 공변량 중에서 가장 유의한 공변량을 입력합니다. 검증 데이터 집합을 사용하는 경우 "Prob>F" 값은 훈련 데이터 집합을 기반으로 합니다.

**단계별 후진** 입력되었지만 잠기지 않은 공변량 중에서 가장 유의하지 않은 공변량을 제거합니다. 검증 데이터 집합을 사용하는 경우 "Prob>F" 값은 훈련 데이터 집합을 기반으로 합니다.

**모두 입력** "추가" 열에 잠겨 있지 않은 모든 공변량을 선택하여 모든 공변량을 입력합니다.

**모두 제거** "추가" 열에서 잠겨 있지 않은 모든 공변량을 선택 취소하여 제거합니다.

**이 모형 적용** "추가" 열에 선택된 공변량을 기반으로 판별 분석 보고서를 생성합니다. "열 선택" 개요가 닫히고 "판별 분석" 창이 업데이트되어 선택한 판별 방법을 기반으로 분석 결과가 표시됩니다.

---

**팁 :** 이 모형 적용을 클릭하면 선택한 열이 "스코어 요약" 보고서 맨 위에 나타납니다.

---

**시작** 검증 엔트로피  $R^2$  이 감소하기 시작할 때까지 전진 선택 단계에서 공변량을 입력합니다. 검증 엔트로피  $R^2$  이 개선되지 않은 상태로 전진 선택 단계가 두 번 수행되면 입력이 종료됩니다. JMP 의 제외된 행 또는 JMP Pro 의 검증 열에 대해서만 사용할 수 있습니다.

**열**

**잠금** 버튼을 사용한 단계별 변수 선별에 관계없이 공변량을 강제로 현재 상태로 유지합니다. 다음 사항에 유의하십시오.

- 공변량을 입력한 후 해당 공변량에 대해 **잠금**을 선택하면 컨트롤 버튼을 사용하여 선택한 내용에 관계없이 공변량이 모형에 유지됩니다. 잠긴 공변량에 대한 **추가** 상자에는 체크 표시가 흐리게 표시되어 해당 공변량이 모형에 있음을 나타냅니다.
- 입력되지 않은 공변량에 대해 **잠금**을 선택하면 컨트롤 버튼을 사용하여 선택한 내용에 관계없이 공변량이 모형에 추가되지 않습니다.

**추가** 현재 모형에 있는 열을 나타냅니다. 모형에 있는 열 또는 없는 열을 수동으로 선택할 수 있습니다. 흐린 체크 표시는 모형에 입력된 잠긴 공변량을 나타냅니다.

**열** 관심 공변량입니다.

**F 비** 공분산 분석 모형을 사용하여 얻은 그룹 변수의 검정에 대한  $F$  비입니다. 자세한 내용은 "**F 비와 Prob>F 업데이트**"에서 확인하십시오.

**Prob > F** 공분산 분석 모형을 사용하여 얻은 그룹 변수의 검정에 대한  $p$  값입니다. 자세한 내용은 "**F 비와 Prob>F 업데이트**"에서 확인하십시오.

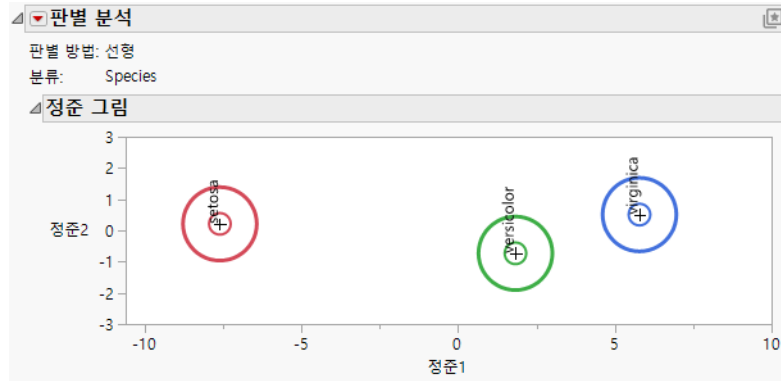
**판별 방법**

판별 플랫폼에서 판별 분석을 수행하는 데 사용할 수 있는 방법은 "선형", "2차", "정규화" 및 "와이드 선형"입니다. 처음 세 가지 방법은 기본 모형이 다릅니다. 와이드 선형 방법은 공변량 수가 많을 때 선형 모형을 적합시키는 효율적인 방법입니다.

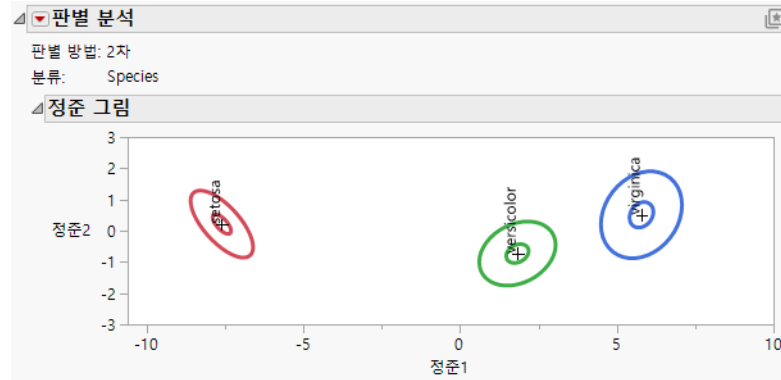
**참고 :** 공변량을 500 개 넘게 입력하면 와이드 선형 방법으로 전환할 것을 제안하는 JMP 경고가 나타납니다. 열 수가 많은 경우 다른 방법을 사용하면 계산 시간이 길어질 수 있기 때문입니다. 와이드 선형 방법으로 전환하려면 **와이드 선형**, **복수 열**을 클릭합니다. 원래 선택한 방법을 사용하려면 **계속**을 클릭합니다.

**그림 5.5** 선형, 2 차 및 정규화 판별 분석

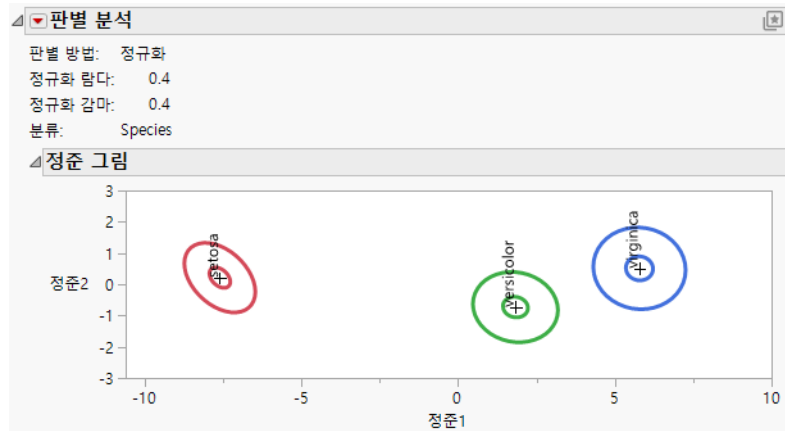
선형



2 차



정규화( $\lambda=0.4$ ,  
 $\gamma=0.4$ )



선형, 2차 및 정규화 방법이 **그림 5.5**에 나와 있습니다. 여기에서는 각 방법을 간략하게 설명합니다. 자세한 내용은 "**저장된 계산식에 대한 통계 상세 정보**"에서 확인하십시오.

**선형, 합동 공분산** 선형 판별 분석을 수행합니다. 이 방법에서는 그룹 내 공분산 행렬이 같다고 가정합니다. 자세한 내용은 "**선형 판별 방법**"에서 확인하십시오.

**2차, 각 공분산** 2차 판별 분석을 수행합니다. 이 방법에서는 그룹 내 공분산 행렬이 다르다고 가정합니다. 이 방법은 선형 방법보다 더 많은 모수를 추정해야 합니다. 그룹 표본 크기가 작으면 추정값이 불안정할 수 있습니다. 자세한 내용은 "**2차 판별 방법**"에서 확인하십시오.

공변량이 X 변수의 수준에 대해 일정하면 그룹 내 공분산 행렬의 관련 항목은 공분산이 0입니다. 역행렬을 계산하기 위해 공분산 0은 해당하는 합동 그룹 내 공분산으로 대체됩니다. 이 작업이 완료되면 문제가 있는 공변량과 X 수준을 식별하는 노트가 보고서 창에 나타납니다.

**팁:** 데이터 집합이 작은 경우 2차 방법의 단점이 있습니다. 가역적이고 안정적인 공분산 행렬을 생성하기 어려울 수 있습니다. 정규화 방법은 그룹 간 차이를 허용하면서 이러한 문제를 개선합니다.

**정규화, 절충 방법** 그룹 내 공분산 행렬이 다를 때 추정값에 안정성을 부여하는 두 가지 방법을 제공합니다. 이 옵션은 그룹 표본 크기가 작을 때 유용합니다. 자세한 내용은 "**정규화, 절충 방법**" 및 "**정규화 판별 방법**"에서 확인하십시오.

**와이드 선형, 복수 열** 다른 방법을 사용하면 계산상 문제가 발생할 수 있는 다수의 공변량을 기반으로 모형을 적합시키는 데 유용합니다. 이 방법에서는 그룹 내 공분산 행렬이 모두 같다고 가정합니다. 이 방법은 특이값 분해 방식을 사용하여 합동 그룹 내 공분산 행렬의 역행렬을 계산합니다. 자세한 내용은 "**와이드 선형 알고리즘에 대한 통계 상세 정보**"에서 확인하십시오.

**참고:** "와이드 선형" 옵션을 사용하면 다른 판별 방법에 일반적으로 나타나는 몇 가지 기능을 사용할 수 없습니다. 이 알고리즘은 매우 큰 합동 그룹 내 공분산 행렬을 명시적으로 계산하지 않기 때문입니다.

## 정규화, 절충 방법

정규화 판별 분석은 음수가 아닌 두 모수에 의해 관리됩니다.

- 첫 번째 모수 (**람다, 합동 공분산에 가깝게 축소**)는 개별 공분산 행렬과 그룹 공분산 행렬을 혼합하는 방법을 지정합니다. 이 모수의 경우 1은 선형 판별 분석에 해당하고 0은 2차 판별 분석에 해당합니다.
- 두 번째 모수 (**감마, 대각 행렬에 가깝게 축소**)는 비대각 요소 (변수 간 공분산)에 적용할 수 축소 크기를 지정하는 승수입니다. 1을 선택하면 공분산 행렬이 강제로 대각화됩니다.

이 두 모수에 각각 0을 할당하면 2차 판별 분석을 요청하는 것과 동일합니다. 마찬가지로 람다에 1을 할당하고 감마에 0을 할당하면 선형 판별 분석이 요청됩니다. **표 5.1**의 내용을 참조하면 정규화를 결정하는 데 도움이 됩니다. 선형, 2차 및 정규화 판별 분석의 예는 **그림 5.5**에서 확인하십시오.

**표 5.1** 정규화 판별 분석

더 작은 람다 사용	더 큰 람다 사용	더 작은 감마 사용	더 큰 감마 사용
공분산 행렬이 다름	공분산 행렬이 같음	변수에 상관관계가 있음	변수에 상관관계가 없음
행 수가 많음	행 수가 적음		
변수 수가 적음	변수 수가 많음		

## 공분산 축소

판별 시작 창에서 공분산 축소 옵션을 선택할 수 있습니다. 이 옵션은 일부 그룹의 관측값 수가 적을 때 권장됩니다. 판별 분석을 수행하려면 공분산 행렬의 역행렬을 계산해야 합니다. 비대각 항목을 축소하면 안정성이 향상되고 예측 분산이 줄어듭니다. 공분산 축소 옵션은 Schafer와 Strimmer(2005) 연구 자료에 설명된 방법을 사용하여 결정된 계수로 비대각 항목을 축소합니다.

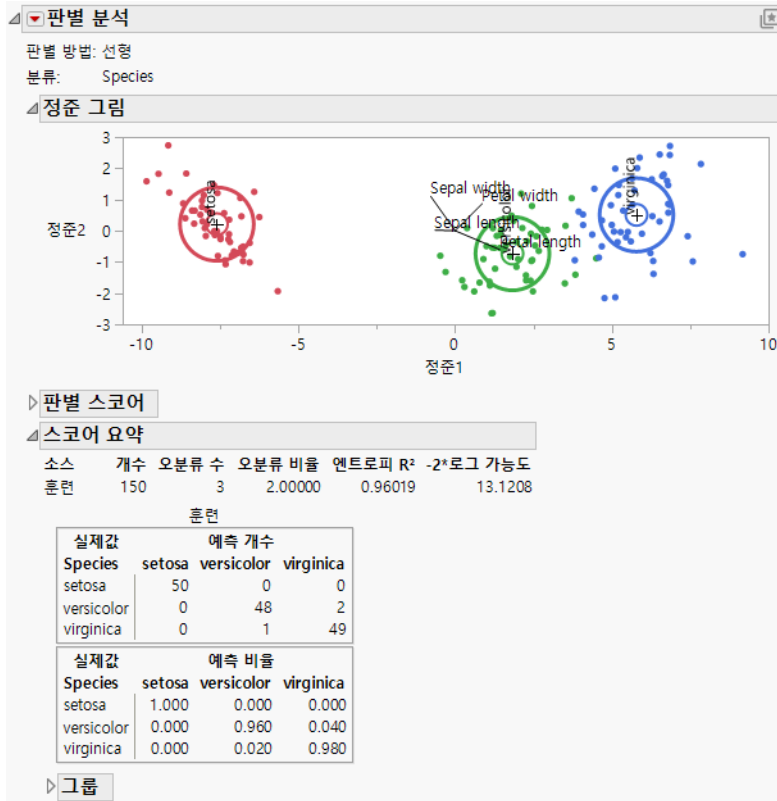
시작 창에서 선형 판별 방법과 함께 "공분산 축소" 옵션을 선택하면 람다 및 감마 값이 적절하게 지정된 정규화 판별 방법에서 제공하는 축소와 동일하게 공분산 행렬이 축소됩니다. "공분산 축소" 옵션을 선택하고 분석을 실행하면 "축소" 보고서에 "전체 축소" 값과 "전체 람다" 값이 제공됩니다. 정규화 방법을 사용하여 동일한 분석을 얻으려면 "정규화 모수" 창에서 "람다"에 1을 입력하고 "감마"에 "축소" 보고서의 "전체 람다" 값을 입력합니다.

## 판별 분석 보고서

"판별 분석" 보고서에는 선택한 판별 방법에 기반한 판별 결과가 제공됩니다. "판별 방법" 및 "분류" 변수가 보고서 맨 위에 표시됩니다. "정규화" 방법을 선택한 경우 관련 모수도 표시됩니다.

"판별 분석"의 빨간색 삼각형 메뉴에서 옵션을 선택하여 판별 방법을 변경할 수 있습니다. 그러면 보고서 결과가 업데이트되어 선택한 방법이 반영됩니다.

그림 5.6 판별 분석 보고서의 예



기본 "판별 분석" 보고서에는 다음 섹션이 포함되어 있습니다.

- "와이드 선형" 판별 방법을 선택하면 "주성분" 보고서가 나타납니다. 자세한 내용은 "주성분"에서 확인하십시오.
- "정준 그림"에는 그룹을 가장 잘 분리하는 두 개의 차원에 점과 다변량 평균이 표시됩니다. 자세한 내용은 "정준 그림 및 정준 구조"에서 확인하십시오.
- "판별 스코어" 보고서에서는 각 관측값이 어떻게 분류되는지 자세히 보여 줍니다. 자세한 내용은 "판별 스코어"에서 확인하십시오.
- "스코어 요약" 보고서에서는 관측값이 얼마나 잘 분류되었는지 간략하게 보여 줍니다. 자세한 내용은 "판별 스코어"에서 확인하십시오.

## 주성분

이 섹션은 시작 창에서 "와이드 선형"을 "판별 방법"으로 선택한 경우에만 나타납니다. 다음 표기를 고려하십시오.

- 공변량의  $n \times p$  행렬을  $\mathbf{Y}$ 로 나타냅니다. 여기서  $n$ 은 관측값 수이고  $p$ 는 공변량 수입니다.

- $Y$ 의 각 관측값에 대해 공변량 평균을 빼고 그 차이를 해당 공변량의 합동 표준편차로 나눕니다. 결과 행렬을  $Y_s$ 로 나타냅니다.

보고서에 다음과 같은 항목이 제공됩니다.

**개수** 추출된 고유값 수입니다. 고유값은 누적 백분율이 99.99% 이상이 될 때까지 추출됩니다. 이는 변동의 99.99%가 설명되었음을 나타냅니다.

**고유값**  $Y_s$ 에 대한 공분산 행렬의 고유값, 즉  $(Y_s'Y_s)/(n-p)$ 이며 내림차순으로 정렬됩니다.

**누적 백분율** 고유값의 누적합을 모든 고유값 합계의 백분율로 나타낸 값입니다. 고유값 합계는  $Y_s'Y_s$ 의 계수입니다.

**특이값**  $Y_s$ 의 특이값이며 내림차순으로 정렬됩니다.

## 정준 그림 및 정준 구조

"판별 분석" 보고서의 "정준 그림"은 변수의 정준 상관 구조를 설명하는 행렬도입니다.

### 정준 구조

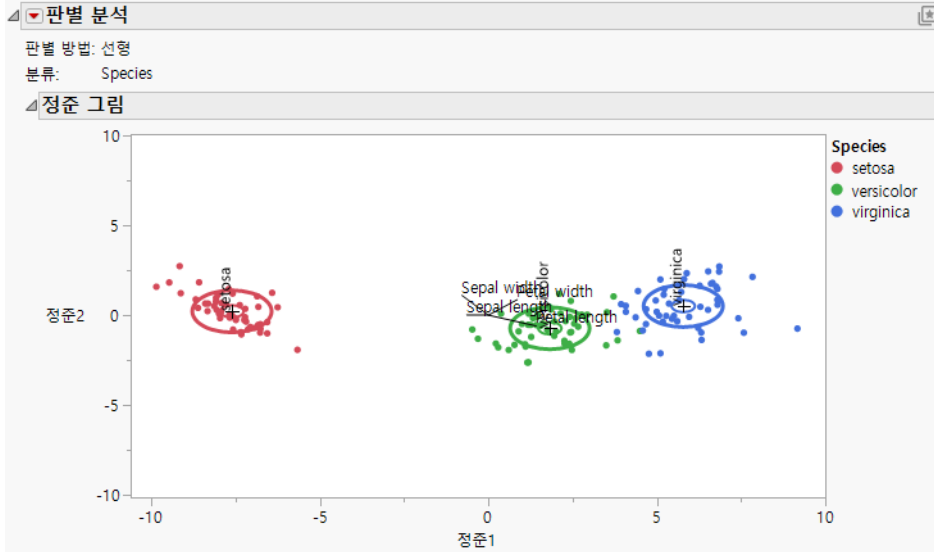
"X, 범주" 열의 각 수준은 표시 변수를 정의합니다. 범주를 나타내는 표시 변수 집합과 공변량 간에 정준 상관이 수행됩니다. 여기서 공변량의 선형 결합, 즉 정준 변수가 도출됩니다. 이러한 정준 변수는 범주 간 변동을 요약하려고 합니다.

첫 번째 정준 변수는 범주 표시 변수와 공변량 간의 다중상관을 최대화하는 공변량 선형 결합입니다. 두 번째 정준 변수는 첫 번째 정준 변수와 상관관계가 없으며 범주와의 다중상관을 최대화하는 선형 결합입니다. "X, 범주" 열에  $k$ 개의 수준이 있으면 정준 변수의 수가  $k-1$ 개입니다.

### 정준 그림

그림 5.7에서는 Iris.jmp 데이터 테이블의 선형 판별 분석에 대한 정준 그림을 보여 줍니다. Species 별로 점에 색상이 적용되어 있습니다.

그림 5.7 Iris.jmp 에 대한 정준 그림



행렬도 축은 처음 두 정준 변수입니다. 이는 그룹 간에 최대 분리를 제공하는 두 차원을 정의합니다. 각 정준 변수는 공변량의 선형 결합입니다. 자세한 내용은 "정준 구조"에서 확인하십시오. 행렬도에서는 각 관측값이 정준 변수에 따라 어떻게 표시되는지 그리고 각 공변량이 정준 변수에 어떻게 기여하는지 보여 줍니다.

- 각 그룹의 관측값 및 다변량 평균은 행렬도에 점으로 표시됩니다. 이는 처음 두 정준 변수로 표현됩니다.
  - 각 다변량 평균에 해당하는 점은 더하기 ("+") 표식으로 나타냅니다.
  - 각 평균에 대해 95% 신뢰 수준 타원이 그림에 표시됩니다. 두 그룹이 유의하게 다르면 신뢰도 타원이 교차하지 않는 경향이 있습니다.
  - 각 그룹에 대해 50% 등고선을 나타내는 타원이 그림에 표시됩니다. 이는 정규성을 가정하고 처음 두 정준 변수의 공간에서 관측값의 약 50%를 포함하는 영역을 나타냅니다.
- 그림에 표시되는 선 집합은 공변량을 나타냅니다.
  - 각 정준 변수에 대해 선형 결합된 공변량의 계수를 가중치로 해석할 수 있습니다.
  - 가중치 간에 쉽게 비교하기 위해 각 공변량이 평균이 0, 표준편차가 1이 되도록 표준화됩니다. 표준화된 공변량의 계수를 정준 가중치라고 합니다. 공변량의 정준 가중치가 클수록 정준 변수와의 연관성이 커집니다.
  - 행렬도에서 각 선의 길이와 방향은 해당 공변량과 처음 두 정준 변수의 연관성 정도를 나타냅니다. 선 길이는 정준 가중치의 배수입니다.
  - 선은 데이터의 총 평균을 정준 변수 측면에서 나타내는 점 (0,0)에서 시작됩니다.

- "판별 분석"의 빨간색 삼각형 메뉴에서 **정준 옵션 > 정준 상세 정보 표시**를 선택하여 가장 치 계수 값을 얻을 수 있습니다. "정준 상세 정보" 보고서 맨 아래에서 "표준화 스코어링 계수"를 클릭합니다. 자세한 내용은 "**표준화 스코어링 계수**"에서 확인하십시오.

### 정준 그림 수정

다음과 같이 추가 옵션을 사용하여 행렬도를 수정할 수 있습니다.

- "판별 분석"의 빨간색 삼각형 메뉴에서 **정준 옵션 > 평균 CL 타원 표시**를 선택하여 95% 신뢰도 타원을 표시하거나 숨깁니다.
- "판별 분석"의 빨간색 삼각형 메뉴에서 **정준 옵션 > 행렬도 선 표시**를 선택하여 선을 표시하거나 숨깁니다.
- 행렬도 선의 가운데를 그래프의 다른 위치로 드래그합니다. "판별 분석"의 빨간색 삼각형 메뉴에서 **정준 옵션 > 행렬도 선 위치**를 선택하여 위치와 척도를 지정합니다. 선을 표시하기 위해 조정이 필요한 경우를 제외하고 정준 그림의 기본 반지름 척도는 1.5입니다.
- "판별 분석"의 빨간색 삼각형 메뉴에서 **정준 옵션 > 정규 50% 등고선 표시**를 선택하여 50% 등고선을 표시하거나 숨깁니다.
- "판별 분석"의 빨간색 삼각형 메뉴에서 **정준 옵션 > 점에 색 적용**을 선택하여 타원과 일치하도록 점에 색상 코딩을 적용합니다.

### 세 개 이상의 범주로 분류

Iris.jmp 데이터의 경우 세 개의 Species가 있으므로 정준 변수는 두 개뿐입니다. **그림 5.7**의 그림에서는 두 정준 변수를 사용하여 세 그룹을 양호하게 분리한 상태를 보여 줍니다.

그림의 선은 다음을 나타냅니다.

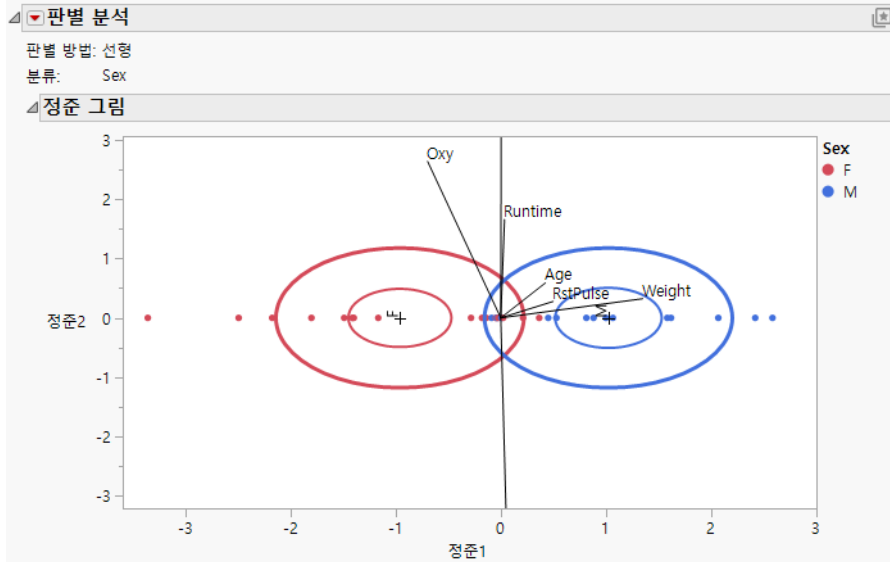
- **Petal length**는 정준 1과 양의 관계를 갖고 정준 2와 음의 관계를 갖습니다. 따라서 정준 2보다 정준 1을 정의하는 데 가중치를 더 부여합니다.
- **Petal width**는 정준 1 및 정준 2와 모두 양의 관계를 갖습니다. 따라서 두 정준 변량을 정의하는 데 동일한 가중치를 부여합니다.
- **Sepal width**는 정준 1과 음의 관계를 갖고 정준 2와 양의 관계를 갖습니다. 따라서 정준 1보다 정준 2를 정의하는 데 가중치를 더 부여합니다.
- **Sepal length**는 정준 1을 정의할 때 음의 가중치를 가지며 정준 2를 정의할 때 매우 약하게 연관됩니다.

### 두 개의 범주로 분류

분류 변수의 수준이 두 개뿐인 경우 그림에서 "정준1"로 표시되는 단일 정규 변수에 대해 점이 표시됩니다. 각 공변량의 정준 가중치는 정준1에만 관련됩니다. 선은 공변량을 분리하기 위해 연직 성분으로만 표시됩니다. 선을 정준1 축에 투영하여 단일 정준 변수와의 상대적 연관성을 비교할 수 있습니다.

그림 5.8에서는 Fitness.jmp 샘플 데이터 테이블에 대한 정준 그림을 보여 줍니다. 7개의 연속형 변량이 개인을 범주 M(남성) 또는 F(여성)로 분류하는 데 사용됩니다. 분류 변수의 범주가 두 개 뿐이므로 정준 변수가 하나만 있습니다.

그림 5.8 Fitness.jmp 에 대한 정준 그림



정준 그림의 점에 Sex별로 색상이 적용되어 있습니다. 두 그룹은 정준1의 값으로 잘 분리되어 있습니다.

7 개의 공변량에 해당하는 선에 연직 성분이 있지만 이 경우 선을 정준 1 축에 대한 투영의 관점으로만 해석해야 합니다. 다음 사항에 유의하십시오.

- MaxPulse, Runtime 및 RunPulse 는 정준 1 과 연관성이 거의 없습니다.
- Weight, RstPulse 및 Age 는 정준 1 과 양의 관계를 갖습니다. Weight 의 연관성이 가장 높습니다. RstPulse 및 Age 공변량은 연관성이 비슷하지만 더 작습니다.
- Oxy 는 정준 1 과 음의 관계를 갖습니다.

## 판별 스코어

"판별 분석" 보고서의 "판별 스코어" 섹션에는 각 관측값의 예측 분류 및 지원 정보가 제공됩니다.

**행** 데이터 테이블의 관측값 행입니다.

**실제값** 데이터 테이블에 제공된 관측값 분류입니다.

**SqDist( 실제값 )** 데이터 테이블에 제공된 관측값 분류에 대해 저장된 계산식 SqDist[<level>] 의 값입니다. 자세한 내용은 "스코어 옵션" 에서 확인하십시오.

**참고 :** 계산식의 오프셋 항 때문에 SqDist( 실제값 ) 이 음수일 수 있습니다 .

**Prob( 실제값 )** 관측값의 실제 분류에 대한 추정 확률입니다 .

**-Log( 확률 )** Prob( 실제값 ) 의 음의 로그입니다 . 이 음의 로그 가능도 값이 크면 실제 범주에 소속될 확률이 잘 예측되지 않는 관측값을 식별할 수 있습니다 .

"-Log( 확률 )" 값 오른쪽에 -Log( 확률 ) 그림이 나타납니다 . 막대가 크면 잘못된 예측을 나타냅니다 . 별표 (\*) 는 잘못 분류된 관측값을 나타냅니다 .

검증 또는 테스트 데이터 집합을 사용하는 경우 검증 데이터 집합의 관측값은 "v" 로 표시되고 테스트 데이터 집합의 관측값은 "t" 로 표시됩니다 .

**예측값** 관측값의 예측 분류입니다 . 예측 분류는 예측 소속 확률이 가장 높은 범주입니다 .

**확률 ( 예측 )** 관측값의 예측 분류에 대한 추정 확률입니다 .

**기타** 예측 확률이 0.1 을 초과하는 다른 범주를 나열합니다 ( 있는 경우 ) .

그림 5.9에서는 Iris.jmp 샘플 데이터 테이블에 대해 선형 판별 방법을 사용한 "판별 스코어" 보고서를 보여 줍니다 . **스코어 옵션 > 관심 있는 행만 표시** 옵션이 선택되어 예측 확률이 0.05에서 0.95 사이인 행 또는 잘못 분류된 행만 표시됩니다 .

**그림 5.9** 관심 있는 행만 표시

판별 스코어						
행	실제값	SqDist(실제값)	Prob(실제값)	-Log(확률)	예측값	확률(예측) 기타
71	versicolor	8.66970	0.2532	1.373	* virginica	0.7468
73	versicolor	4.87619	0.8155	0.204	versicolor	0.8155 virginica 0.18
78	versicolor	4.66698	0.6892	0.372	versicolor	0.6892 virginica 0.31
84	versicolor	8.43926	0.1434	1.942	* virginica	0.8566
120	virginica	8.19641	0.7792	0.249	virginica	0.7792 versicolor 0.22
124	virginica	3.57858	0.9029	0.102	virginica	0.9029
127	virginica	3.90184	0.8116	0.209	virginica	0.8116 versicolor 0.19
128	virginica	3.31470	0.8658	0.144	virginica	0.8658 versicolor 0.13
130	virginica	9.08495	0.8963	0.109	virginica	0.8963 versicolor 0.10
134	virginica	7.23593	0.2706	1.307	* versicolor	0.7294
135	virginica	15.83301	0.9340	0.068	virginica	0.9340
139	virginica	4.09385	0.8075	0.214	virginica	0.8075 versicolor 0.19

\*는 잘못 분류되었음을 나타냅니다.

## 스코어 요약

"판별 분석" 보고서의 "스코어 요약" 섹션에서는 판별 스코어에 대한 개요를 제공합니다 . 그림 5.10의 테이블에는 실제값 및 예측 분류가 표시됩니다 . 모든 관측값이 올바르게 분류되면 비대각 수가 0입니다 .

그림 5.10 Iris.jmp 에 대한 스코어 요약

스코어 요약						
소스	개수	오분류 수	오분류 비율	엔트로피 R <sup>2</sup>	-2*로그 가능성도	
훈련	65	1	1.53846	0.94487	7.81888	
검증	49	3	6.12245	0.86644		
테스트	36	0	0.00000	0.98410		

훈련				검증				테스트			
실제값	예측 개수			실제값	예측 개수			실제값	예측 개수		
Species	setosa	versicolor	virginica	Species	setosa	versicolor	virginica	Species	setosa	versicolor	virginica
setosa	18	0	0	setosa	16	0	0	setosa	16	0	0
versicolor	0	22	1	versicolor	0	15	2	versicolor	0	10	0
virginica	0	0	24	virginica	0	1	15	virginica	0	0	10

훈련				검증				테스트			
실제값	예측 비율			실제값	예측 비율			실제값	예측 비율		
Species	setosa	versicolor	virginica	Species	setosa	versicolor	virginica	Species	setosa	versicolor	virginica
setosa	1.000	0.000	0.000	setosa	1.000	0.000	0.000	setosa	1.000	0.000	0.000
versicolor	0.000	0.957	0.043	versicolor	0.000	0.882	0.118	versicolor	0.000	1.000	0.000
virginica	0.000	0.000	1.000	virginica	0.000	0.063	0.938	virginica	0.000	0.000	1.000

"스코어 요약" 보고서에는 다음 정보가 제공됩니다.

**열** "단계별 변수 선택"을 사용하여 모형을 생성한 경우 모형에 추가된 열이 나열됩니다.

**소스** 검증을 사용하지 않으면 모든 관측값이 훈련 데이터 집합을 구성합니다. 검증을 사용하면 훈련 및 검증 데이터 집합 또는 훈련, 검증 및 테스트 데이터 집합에 대한 행이 표시됩니다.

**오분류 수** 지정된 집합에서 잘못 분류된 관측값의 수를 제공합니다.

**오분류 비율** 지정된 집합에서 잘못 분류된 관측값의 백분율을 제공합니다.

**엔트로피 R<sup>2</sup>** 적합 척도입니다. 값이 클수록 더 나은 적합을 나타냅니다. "엔트로피 R<sup>2</sup>" 값이 1이면 분류가 완벽하게 예측되었음을 나타냅니다. 예측 확률의 불확도는 판별 모형에서 일반적으로 사용되므로 엔트로피 R<sup>2</sup> 값은 작은 경향이 있습니다.

자세한 내용은 "엔트로피 R<sup>2</sup>"에서 확인하십시오.

**참고:** 엔트로피 R<sup>2</sup> 이 음수일 수 있습니다.

**-2\* 로그 가능성도** 모형을 기반으로 훈련 데이터 집합에 있는 관측값의 음의 로그 가능성을 두 배 곱한 값입니다. 값이 작을수록 더 나은 적합을 나타냅니다. 훈련 데이터 집합에만 제공됩니다. 자세한 내용은 선형 모형 적합에서 확인하십시오.

**혼동 행렬** 범주형 X의 각 수준에 대한 실제값 대 예측값 개수 행렬을 표시합니다. JMP Pro에서 검증을 사용하는 경우 각 관측값 집합에 대해 행렬이 제공됩니다. JMP에서 제외된 행을 사용하는 경우 제외된 행이 검증 데이터 집합으로 간주되고 별도의 검증 행렬이 제공됩니다. 자세한 내용은 "JMP 및 JMP Pro에서 검증"에서 확인하십시오.

## 엔트로피 R<sup>2</sup>

엔트로피 R<sup>2</sup>은 적합 척도입니다. 이 값은 훈련 데이터 집합에 대해 계산되고, 검증을 사용하는 경우 검증 데이터 집합 및 테스트 데이터 집합에 대해 계산됩니다.

### 훈련 데이터 집합에 대한 엔트로피 R<sup>2</sup>

훈련 데이터 집합에 대한 엔트로피 R<sup>2</sup>은 다음과 같이 계산됩니다.

- 훈련 데이터 집합을 사용하여 판별 모형이 적합됩니다.
- 모형을 기반으로 한 예측 확률을 구합니다.
- 이러한 예측 확률을 사용하여 훈련 데이터 집합의 관측값에 대한 가능도가 계산됩니다. 이를  $Likelihood\_Full_{Training}$ 이라고 합니다.
- 훈련 데이터 집합을 사용하여 축소 모형 (예측 변수 없음)이 적합됩니다.
- 축소 모형의 X 수준에 대한 예측 확률을 사용하여 훈련 데이터 집합의 관측값에 대한 가능도가 계산됩니다. 이를  $Likelihood\_Reduced_{Training}$ 이라고 합니다.
- 훈련 데이터 집합에 대한 엔트로피 R<sup>2</sup>은 다음과 같습니다.

$$\text{엔트로피 } R^2_{Training} = 1 - \frac{\log(Likelihood\_Full_{Training})}{\log(Likelihood\_Reduced_{Training})}$$

### 검증 및 테스트 데이터 집합에 대한 엔트로피 R<sup>2</sup>

검증 데이터 집합에 대한 엔트로피 R<sup>2</sup>은 다음과 같이 계산됩니다.

- 훈련 데이터 집합만 사용하여 판별 모형이 적합됩니다.
- 모든 관측값에 대해 훈련 데이터 집합 모형을 기반으로 한 예측 확률을 구합니다.
- 이러한 예측 확률을 사용하여 검증 데이터 집합의 관측값에 대한 가능도가 계산됩니다. 이를  $Likelihood\_Full_{Validation}$ 이라고 합니다.
- 훈련 데이터 집합만 사용하여 축소 모형 (예측 변수 없음)이 적합됩니다.
- 축소 모형의 X 수준에 대한 예측 확률을 사용하여 검증 데이터 집합의 관측값에 대한 가능도가 계산됩니다. 이를  $Likelihood\_Reduced_{Validation}$ 이라고 합니다.
- 검증 엔트로피 R<sup>2</sup>은 다음과 같습니다.

$$\text{검증 엔트로피 } R^2 = 1 - \frac{\log(Likelihood\_Full_{Validation})}{\log(Likelihood\_Reduced_{Validation})}$$

테스트 데이터 집합의 엔트로피 R<sup>2</sup>은 검증 데이터 집합의 엔트로피 R<sup>2</sup>과 유사한 방식으로 계산됩니다.

## 판별 분석 옵션

"판별 분석"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

**단계별 변수 선택** ("와이드 선행" 방법에는 사용 불가능) "열 선택" 제어판을 표시하거나 숨깁니다. 이 제어판에는 공분산 분석 및  $p$  값을 사용하여 단계별 변수 선택을 수행할 수 있는 옵션이 포함되어 있습니다. 자세한 내용은 "**단계별 변수 선택**"에서 확인하십시오.

**판별 방법** 판별 방법을 지정합니다. "선형", "2 차", "정규화" 및 "와이드 선행" 중에서 선택합니다. 자세한 내용은 "**판별 방법**"에서 확인하십시오.

**판별 스코어** 각 행의 판별 스코어 테이블을 표시하거나 숨깁니다.

**스코어 옵션** 관측값 스코어링을 위한 옵션을 제공합니다.

**관심 있는 행만 표시** "판별 스코어" 보고서에서 잘못 분류된 행과 예측 확률이 0.05에서 0.95 사이인 행만 표시합니다.

**분류 개수 표시** "스코어 요약" 보고서에 혼동 행렬과 혼동 비율 행렬을 표시하거나 숨깁니다. 혼동 행렬은 실제 반응과 예측 반응의 이원 분류입니다. 혼동 비율 행렬은 값을 행 합계로 나눈 혼동 행렬과 동일합니다. 기본적으로 "스코어 요약" 보고서에는 범주형 X의 각 수준에 대한 혼동 행렬과 혼동 비율 행렬이 표시됩니다. JMP Pro에서 검증을 사용하는 경우 각 관측값 집합에 대해 행렬이 제공됩니다. JMP에서 제외된 행을 사용하는 경우 이러한 행이 검증 데이터 집합으로 간주되고 별도의 검증 행렬이 제공됩니다. 자세한 내용은 "**JMP 및 JMP Pro에서 검증**"에서 확인하십시오.

**각 그룹까지의 거리 표시** 각 그룹 평균에 대한 각 관측값의 Mahalanobis 거리 제곱을 포함하는 보고서를 표시하거나 숨깁니다.

**각 그룹에 대한 확률 표시** 관측값이 범주형 X로 정의된 각 그룹에 속할 확률을 포함하는 보고서를 표시하거나 숨깁니다.

**ROC 곡선** "스코어 요약" 보고서에 ROC(Receiver Operating Characteristic) 그림을 표시하거나 숨깁니다. 이 그림에는 반응 변수의 각 수준에 대한 곡선이 포함되어 있습니다. 자세한 내용은 예측 및 전문 모델링에서 확인하십시오.

**정밀도 - 재현율 곡선** "스코어 요약" 보고서에 정밀도 - 재현율 곡선 그림을 표시하거나 숨깁니다. 이 그림에는 반응 변수의 각 수준에 대한 곡선이 포함되어 있습니다. 정밀도 - 재현율 곡선은 다양한 임계값에서 정밀도 값 대 재현율 값을 표시합니다. 자세한 내용은 예측 및 전문 모델링에서 확인하십시오.

**오분류 행 선택** 행별 목록을 표시하는 보고서 창과 데이터 테이블에서 잘못 분류된 행을 선택합니다.

**불확실한 행 선택** 행별 목록을 표시하는 보고서 창과 데이터 테이블에서 분류가 불확실한 행을 선택합니다. 불확실한 행은 어떤 그룹에 대한 소속 확률이 0에 가깝지도 않고 1에 가깝지도 않은 행입니다.

이 옵션을 선택하면 불확도를 반영하는 예측 확률 범위를 지정할 수 있는 창이 열립니다. 기본적으로 확률이 0 또는 1과 0.1 넘게 차이 나는 행은 불확실한 것으로 정의됩니다. 따라서 0.1에서 0.9 사이의 확률을 갖는 행이 기본적으로 선택됩니다.

**계산식 저장** 거리, 확률 및 소속 확률 예측 계산식을 데이터 테이블에 저장합니다. 자세한 내용은 "저장된 계산식에 대한 통계 상세 정보"에서 확인하십시오.

- 거리 계산식은  $SqDist0$  및  $SqDist<level>$ 이며, 여기서  $<level>$ 은  $X$ 의 수준을 나타냅니다. 거리 계산식은 Mahalanobis 거리 계산과 관계가 있는 중간 값을 생성합니다.
- 확률 계산식은  $Prob<level>$ 이며, 여기서  $<level>$ 은  $X$ 의 수준을 나타냅니다. 각 확률 열은 해당  $X$  수준에서 관측값의 소속에 대한 사후 확률을 제공합니다. "반응 확률" 열 특성이 각 확률 열에 저장됩니다. "반응 확률" 열 특성에 대한 자세한 내용은 JMP 사용에서 확인하십시오.
- 소속 확률 예측 계산식은  $Pred <X>$ 이며 "최대 확률 분류 수준"이라는 분류 규칙을 포함합니다.
- 와이드 선형 방법은 공변량 벡터와 판별 주성분 계산식을 포함하는 판별 데이터 행렬 열도 저장합니다. 자세한 내용은 "와이드 선형 판별 방법"에서 확인하십시오.

**참고:** 와이드 선형을 제외한 다른 방법의 경우 계산식을 저장하면 "RowEdit Prob" 스크립트가 데이터 테이블에 저장됩니다. 이 스크립트는 데이터 테이블에서 불확실한 행을 선택합니다. 스크립트는 확률이 0 또는 1과 0.1 넘게 차이 나는 행을 불확실한 것으로 정의합니다. 또한 불확실한 행을 검토할 수 있도록 "행 편집기" 창도 엽니다. 와이드 선형을 제외한 새 모형을 적합시키고 "계산식 저장"을 선택하면 기존 "RowEdit Prob" 스크립트가 새 적합에 적용되는 스크립트로 대체됩니다.

**스코어링 스크립트 만들기** (JMP Standard에서만 사용 가능) "계산식 저장" 옵션으로 저장된 계산식 열을 구성하는 스크립트를 생성합니다. 이 스크립트를 저장한 후 다른 데이터 테이블과 함께 사용하여 소속 확률을 계산하고 소속 그룹을 예측하는 계산식 열을 생성할 수 있습니다.

**JMP PRO 확률 계산식 게시** (JMP Pro에서만 사용할 수 있음) 확률 계산식을 생성하여 계산식 저장소 플랫폼에 계산식 열 스크립트로 저장합니다. "계산식 저장소" 보고서가 열려 있지 않으면 "계산식 저장소" 보고서가 생성됩니다. 자세한 내용은 예측 및 전문 모델링에서 확인하십시오.

**정준 그림** 정준 그림을 표시하거나 숨깁니다. 자세한 내용은 "정준 그림 및 정준 구조"에서 확인하십시오.

**정준 옵션** 정준 그림 및 3D 정준 그림에 영향을 주는 옵션을 제공합니다.

**점 표시** 정준 그림 및 3D 정준 그림에서 점을 표시하거나 숨깁니다.

**평균 CL 타원 표시** 정규성을 가정하고 정준 그림 및 3D 정준 그림에서 각 그룹의 평균에 대한 95% 신뢰도 타원을 표시하거나 숨깁니다.

**정규 50% 등고선 표시** 각 그룹에 대해 50% 예측 타원 또는 타원체를 표시하거나 숨깁니다. 정준 그림에서 각 타원은 다변량 정규성을 가정하고 처음 두 정준 변수의 공간에서 각 그룹에 대한 관측값의 약 50%가 속할 위치를 추정하는 영역을 나타냅니다. 3D 정준 그림에서 각 타원체는 다변량 정규성을 가정하고 처음 세 정준 변수의 공간에서 관측값의 약 50%가 속할 위치를 추정하는 영역을 나타냅니다.

**행렬도 선 표시** 정준 그림 및 3D 정준 그림에서 행렬도 선을 표시하거나 숨깁니다. 라벨이 지정된 선은 정준 공간에서 공변량의 방향을 표시합니다. 이러한 선은 각 공변량과 각 정준 변수의 연관성 정도를 나타냅니다.

**행렬도 선 위치** 정준 그림 및 3D 정준 그림에서 행렬도 선의 위치와 반지름 척도를 지정할 수 있습니다.

- 기본적으로 선은 데이터의 총 평균을 정준 변수 측면에서 나타내는 점 (0,0) 에서 시작됩니다. 정준 그림에서 선을 드래그하거나 이 옵션을 사용하여 좌표를 지정할 수 있습니다.
- 선을 표시하기 위해 조정이 필요한 경우를 제외하고 정준 그림의 기본 반지름 척도는 1.5입니다. 반지름 척도는 표준화 스코어링 계수와 관련하여 완료됩니다.

**점에 색 적용** X 변수의 수준을 기반으로 정준 그림 및 3D 정준 그림의 점에 색상을 지정합니다. 데이터 테이블의 행에 색상 표식이 추가됩니다. 이 옵션은 **행 > 열 값에 따른 색상 또는 표식**을 선택하고 X 변수를 선택하는 것과 같습니다. 또한 그래프를 마우스 오른쪽 버튼으로 클릭하고 **행 범례**를 선택한 후 분류 열에 따라 색상을 적용하는 것과도 같습니다.

**정준 상세 정보 표시** "정준 상세 정보" 보고서를 표시하거나 숨깁니다. 자세한 내용은 "**정준 상세 정보 표시**"에서 확인하십시오.

**정준 구조 표시** "정준 구조" 보고서를 표시하거나 숨깁니다. 자세한 내용은 "**정준 구조 표시**"에서 확인하십시오. "와이드 선형" 판별 방법에는 사용할 수 없습니다.

**정준 스코어 저장** 데이터 테이블에 각 관측값에 대한 정준 스코어 계산식이 포함된 열을 생성합니다. k 번째 정준 스코어의 열에는 정준 <k> 라는 이름이 지정됩니다.

---

**팁 :** 스크립트에서 새 데이터 테이블에 저장 스크립트 명령을 판별 개체에 보내면 정준 변수에 대한 그룹 평균, 표준화 스코어링 계수의 반지름 척도가 1.5 인 행렬도 선 및 정준 스코어가 새 데이터 테이블에 저장됩니다. "와이드 선형" 판별 방법에는 사용할 수 없습니다.

---

**3D 정준 그림** 3 차원 정준 그림을 표시합니다. 이 옵션은 범주형 X의 수준이 네 개 이상인 경우에만 사용할 수 있습니다. 자세한 내용은 "**3D 정준 그림의 예**"에서 확인하십시오.

**사전 확률 지정** X 변수의 각 수준에 대한 사전 확률을 지정할 수 있습니다. 다음 옵션을 사용하여 사전 확률을 지정할 수 있습니다.

**동등 확률** 모든 그룹에 동일한 사전 확률을 할당합니다. 이 옵션이 기본값입니다.

**발생 횟수에 비례** 관측된 데이터에 나타나는 빈도에 비례하여 해당 그룹에 사전 확률을 할당합니다.

**기타** 사용자 사전 확률을 지정할 수 있습니다.

**새 수준 고려** 일부 점은 알려진 그룹에 적합하지 않을 수 있으며 스코어가 없는 새 그룹에 속한 것으로 간주되어야 함을 지정합니다. 자세한 내용은 "**새 수준 고려**"에서 확인하십시오.

**그룹 내 공분산 표시** 다음 보고서를 표시하거나 숨깁니다.

- 합동 그룹 내 공분산 및 상관 행렬을 제공하는 공분산 행렬 보고서
  - "2차" 및 "정규화" 방법을 선택한 경우 그룹 내 상관 행렬을 표시하는 그룹별 상관 보고서  
각 그룹에 대해 그룹 내 공분산 행렬의 행렬식 로그 (LogDetCov, 공분산 행렬식 로그)도 표시됩니다. 행렬식이 0 이면 공분산 행렬의 가역 부분을 사용하여 LogDetCov 값이 계산됩니다.
  - "2차" 판별 방법을 선택한 경우 그룹 내 공분산 행렬을 표시하는 "공분산 행렬" 보고서에 "그룹 공분산" 개요가 추가됩니다.
- "와이드 선형" 판별 방법에는 "그룹 내 공분산 표시"를 사용할 수 없습니다.

**그룹 평균 표시** 각 공변량의 평균을 제공하는 "그룹 평균" 보고서를 표시하거나 숨깁니다. X 변수의 각 수준에 대한 평균과 전체 평균이 나타납니다.

**판별 행렬 저장** 판별 분석 결과라는 스크립트를 데이터 테이블에 저장합니다. 이 스크립트는 JSL에서 사용하기 위한 다음 개체의 목록입니다.

- 공변량 (Y) 목록
- 범주형 변수 X
- X 수준 목록
- 공변량 평균 대 X 수준의 행렬
- 합동 그룹 내 공분산 행렬

"와이드 선형" 판별 방법에는 "판별 행렬 저장"을 사용할 수 없습니다. 자세한 내용은 "**판별 행렬 저장**"에서 확인하십시오.

**산점도 행렬** ("와이드 선형" 판별 방법에는 사용 불가능) 공변량에 대한 하삼각 산점도 행렬이 포함된 산점도 행렬 플랫폼을 별도의 창에서 엽니다. 산점도에는 검증이 사용된 경우에도 데이터 테이블의 모든 관측값이 포함됩니다. 범주형 변수 X의 각 수준에 대해 90% 범위의 타원이 표시됩니다. 선형 판별 방법의 경우 합동 그룹 내 공분산 행렬을 기반으로 합니다.

"산점도 행렬"의 빨간색 삼각형 메뉴에 포함된 옵션에 대한 자세한 내용은 **그래프 분석**에서 확인하십시오.

**프로파일러** 범주형 확률이 단일 프로파일러 행으로 결합된 대화식 프로파일러 보고서를 표시하거나 숨깁니다. 요인 값을 변경하면 추정된 분류 확률에 반영됩니다. 빨간색 삼각형 메뉴의 옵션에 대한 자세한 내용은 **프로파일러**에서 확인하십시오.

다음 옵션에 대한 자세한 내용은 **JMP 사용**에서 확인하십시오.

**로컬 데이터 필터** 특정 보고서에서 사용되는 데이터를 필터링할 수 있는 로컬 데이터 필터를 표시하거나 숨깁니다.

**다시 실행** 분석을 반복하거나 다시 시작할 수 있는 옵션이 포함되어 있습니다. 이 기능을 지원하는 플랫폼에서 "자동 재계산" 옵션은 해당하는 보고서 창에서 데이터 테이블에 대한 변경 사항을 즉시 반영합니다.

**플랫폼 환경 설정** 현재 플랫폼 환경 설정을 보거나, 현재 JMP 보고서의 설정과 일치하도록 플랫폼 환경 설정을 업데이트할 수 있는 옵션이 포함되어 있습니다.

**스크립트 저장** 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다.

**그룹별 스크립트 저장** 기준 변수의 모든 수준에 대한 플랫폼 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다. 시작 창에서 기준 변수를 지정한 경우에만 사용할 수 있습니다.

**참고:** 이 플랫폼의 추가 옵션은 스크립트를 통해 사용할 수 있습니다. 자세한 내용은 "도움말" 메뉴의 "스크립트 인덱스"에서 여십시오. 또한 "스크립트 인덱스"에서 이 섹션에 설명된 옵션의 스크립트 예제도 찾을 수 있습니다.

## 정준 상세 정보 표시

판별 플랫폼의 "정준 상세 정보" 보고서에서는 공변량과 그룹화 변수 X 사이의 관계를 확인하는 검정을 보여 줍니다. 보고서 아래쪽에 관련 행렬이 제공됩니다.

그림 5.11 Iris.jmp 에 대한 정준 상세 정보

정준 상세 정보								
전체 합동 그룹 내 공분산 행렬에서 계산된 정준 상세 정보입니다.								
고유값	백분율	누적 백분율	정준 상관	가능도비	근사 F	NumDF	DenDF	Prob>F
32.1919292	99.1213	99.1213	0.98482089	0.02343863	199.1453	8	288	<.0001*
0.28539104	0.8787	100.0000	0.47119702	0.77797337	13.7939	3	145	<.0001*
검정	값	근사 F	NumDF	DenDF	Prob>F			
Wilks 람다	0.0234386	199.1453	8	288	<.0001*			
Pillai 대각합	1.1918988	53.4665	8	290	<.0001*			
Hotelling-Lawley	32.47732	582.1970	8	203.4	<.0001*			
Roy 최대근	32.191929	1166.9574	4	145	<.0001*			
<ul style="list-style-type: none"> <li>▷ 그룹 내 행렬</li> <li>▷ 그룹 간 행렬</li> <li>▷ 스코어링 계수</li> <li>▷ 표준화 스코어링 계수</li> </ul>								

**참고:** 보고서의 결과를 계산하는 데 사용되는 행렬은 합동 그룹 내 공분산 행렬입니다 ("그룹 내 행렬"로 표시됨). 이 행렬은 모든 판별 방법에 대해 "정준 상세 정보" 보고서의 기초로 사용됩니다. "정준 상세 정보" 보고서의 통계량과 검정은 모든 판별 방법에 대해 동일합니다.

## 통계량 및 검정

"정준 상세 정보" 보고서에는 고유값이 나열되고 0 고유값에 대한 가능도비 검정이 제공됩니다. 정준 상관인 0이라는 귀무가설에 대해 네 가지 검정이 제공됩니다.

**고유값** 그룹 내 행렬의 역행렬과 그룹 간 행렬의 곱을 나타내는 고유값입니다. 이 값은 가장 큰 것부터 순서대로 나열됩니다. 고유값 크기는 연관된 판별 함수로 설명되는 분산의 양을 반영합니다.

**백분율** 주어진 고유값이 나타내는 고유값 합계의 비율입니다.

**누적 백분율** 비율의 누적합입니다.

**정준 상관** 공변량과 범주형  $X$  로 정의된 그룹 간의 정준 상관입니다.  $X$  로 정의된 그룹을 나타내기 위해 숫자 표시 변수를 정의하는 경우를 가정해 보겠습니다. 그런 다음 공변량을 한 변수 집합으로 사용하고  $X$  의 그룹을 나타내는 표시 변수를 다른 집합으로 사용하여 정준 상관 분석을 수행합니다. "정준 상관" 값은 이 분석의 결과로 생성된 정준 상관 값입니다.

**가능도비** 해당하는 정규 상관과 더 작은 모든 상관의 모집단 값이 0 인지 여부를 검정하기 위한 가능도비 통계량입니다. 이 비율은 주어진 정준 상관 및 더 작은 모든 정준 상관 값의 곱과 같습니다 ( $1 - \text{정준 상관}^2$ ).

**검정** 모든 그룹에서 공변량 평균이 같다는 귀무가설에 대한 네 가지 표준 검정이 나열됩니다 (Wilk 램다, Pillai 대각합, Hotelling-Lawley 및 Roy 최대근). 자세한 내용은 "[다변량 검정에 대한 통계 상세 정보](#)" 및 "[근사 F-검정에 대한 통계 상세 정보](#)"에서 확인하십시오.

**근사 F** 해당 검정과 관련된  $F$  값입니다. 특정 검정의 경우  $F$  값이 근사값 또는 상한입니다. 자세한 내용은 "[근사 F-검정에 대한 통계 상세 정보](#)"에서 확인하십시오.

**NumDF** 해당 검정에 대한 분자 자유도입니다.

**DenDF** 해당 검정에 대한 분모 자유도입니다.

**Prob>F** 해당 검정에 대한  $p$  값입니다.

## 행렬

정준 구조와 관련된 네 개의 행렬이 보고서 아래쪽에 제공됩니다. 행렬을 보려면 행렬 이름 옆의 표시 아이콘을 클릭합니다. 숨기려면 행렬 이름을 클릭합니다.

**그룹 내 행렬** 합동 그룹 내 공분산 행렬입니다.

**그룹 간 행렬** 그룹 간 공분산 행렬  $S_B$  입니다. 자세한 내용은 "[그룹 간 공분산 행렬에 대한 통계 상세 정보](#)"에서 확인하십시오.

**스코어링 계수** 원시 데이터를 기준으로 정준 스코어를 계산하는 데 사용되는 계수입니다. 이러한 계수는 [정준 옵션 > 정준 스코어 저장](#) 옵션에 사용됩니다. 계산 방법에 대한 자세한 내용은 SAS Institute Inc. (2023b) 의 "CANDISC Procedure" 장에서 확인하십시오.

**표준화 스코어링 계수** 표준화 데이터를 기준으로 정준 스코어를 계산하는 데 사용되는 계수입니다. 정준 가중치라고도 합니다. 계산 방법에 대한 자세한 내용은 SAS Institute Inc. (2023b) 의 "CANDISC Procedure" 장에서 확인하십시오.

## 정준 구조 표시

판별 플랫폼의 "정준 구조" 보고서에는 정준 변수와 공변량 간의 상관을 제공하는 세 개의 행렬이 있습니다. 다른 행렬에는 그룹 변수 수준에 대한 평균이 표시됩니다. 행렬을 보려면 행렬 이름 옆의 표시 아이콘을 클릭합니다. 숨기려면 행렬 이름을 클릭합니다.

**그림 5.12** 그룹 간 정준 구조를 보여 주는 Iris.jmp 의 정준 구조

정준 구조				
<ul style="list-style-type: none"> <li>▷ 총 정준 구조</li> <li>그룹 간 정준 구조</li> </ul>				
	Sepal length	Sepal width	Petal length	Petal width
정준1	0.9914683	-0.825658	0.99975	0.9940442
정준2	0.1303484	0.5641714	0.0223578	0.1089775
<ul style="list-style-type: none"> <li>▷ 합동 그룹 내 정준 구조</li> <li>▷ 정준 변수에 대한 클래스 평균</li> </ul>				

**총 정준 구조** 정준 변수와 공변량 간의 상관입니다. 적재라고도 합니다.

**그룹 간 정준 구조** 정준 변수에 대한 그룹 평균과 공변량에 대한 그룹 평균 간의 상관입니다.

**합동 그룹 내 정준 구조** 정준 변수와 공변량 간의 부분 상관이며 그룹 변수에 대해 조정됩니다.

**정준 변수에 대한 클래스 평균** 각 정준 변수에 대한 그룹 변수 전체 수준의 평균을 제공합니다.

## 새 수준 고려

일부 관측값이 범주형 변수의 지정된 수준에 대한 이상치라고 의심되는 경우 "판별 분석"의 빨간색 삼각형 메뉴에서 "새 수준 고려" 옵션을 사용합니다. 이 옵션을 선택하면 새 수준의 사전 확률을 지정하라는 메뉴가 나타납니다.

새 그룹을 사용하는 것이 더 적합한 관측값은 "기타"라는 새 수준에 할당됩니다. "기타" 그룹에 소속될 확률은 이러한 관측값이 그룹 구조가 가정되지 않은 전체 관측값 집합의 분포를 따른다고 가정합니다. 따라서 공분산 구조와 관련된 정규 등고선이 상대적으로 넓어집니다. 거리 계산은 지정된 사전 확률에 따라 조정됩니다.

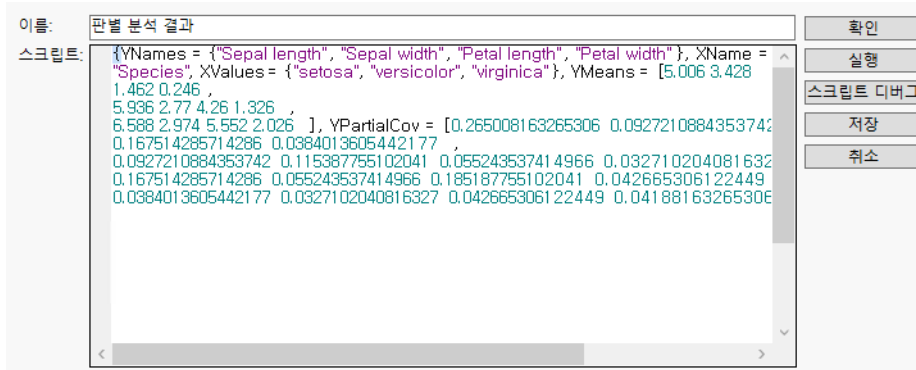
## 판별 행렬 저장

"판별 분석"의 빨간색 삼각형 메뉴에 있는 "판별 행렬 저장" 옵션은 JMP 스크립트 언어에서 사용할 전역 목록 (판별 분석 결과) 을 생성합니다. 목록에는 훈련 데이터 집합에 대해 계산된 다음 항목이 포함됩니다.

- YNames - 공변량 (Y) 목록
- XName - 범주형 변수
- XValues - X 수준 목록
- YMeans - 공변량 평균 대 X 수준의 행렬
- YPartialCov - 그룹 내 공분산 행렬

Iris.jmp 샘플 데이터 테이블의 Discriminant 스크립트를 사용하여 얻은 분석을 예로 들어보겠습니다. "판별 분석"의 빨간색 삼각형 메뉴에서 판별 행렬 저장을 선택하면 판별 분석 결과 스크립트가 데이터 테이블에 저장됩니다.

그림 5.13 Iris.jmp 에 대한 판별 분석 결과 테이블 스크립트



**참고:** 스크립트에서 "판별 행렬 가져오기" 스크립트 명령을 판별 플랫폼 개체에 보낼 수 있습니다. 이렇게 하면 "판별 행렬 저장" 과 동일한 값을 얻을 수 있지만 데이터 테이블에 저장되지는 않습니다.

## JMP 및 JMP Pro 에서 검증

판별 플랫폼에서 사용할 검증 데이터 집합은 JMP 버전에 따라 다르게 지정됩니다. JMP Standard에서는 검증 데이터 집합을 구성하는 행을 제외하여 검증 데이터 집합을 지정할 수 있습니다. 검증 데이터 집합으로 사용할 행을 선택한 후 **행 > 제외/제외 해제**를 선택합니다. 제외되지 않은 행은 훈련 데이터 집합으로 처리됩니다.

JMP Pro에서는 판별 시작 창에서 검증 열을 지정할 수 있습니다. 검증 열은 숫자 데이터 유형이어야 하고 두 개 이상의 구분되는 값을 포함해야 합니다.

다음 사항에 유의하십시오.

- 열에 두 개의 값이 포함된 경우 더 작은 값이 훈련 데이터 집합을 정의하고 더 큰 값이 검증 데이터 집합을 정의합니다.
- 열에 세 개의 값이 포함된 경우 작은 값부터 순서대로 훈련 데이터 집합, 검증 데이터 집합, 테스트 데이터 집합을 각각 정의합니다.
- 열에 네 개 이상의 구분되는 값이 포함된 경우 가장 작은 세 개의 값과 관련 관측값만 사용하여 순서대로 훈련 데이터 집합, 검증 데이터 집합 및 테스트 데이터 집합을 각각 정의합니다.

검증 데이터 집합이 지정되면 판별 플랫폼에서 다음 단계를 수행합니다.

- 훈련 데이터를 사용하여 모형이 적합됩니다.

- " 단계별 변수 선택 " 옵션을 선택하면 모형에 대한 " 검증 엔트로피  $R^2$  " 및 " 검증 오분류 비율 " 통계량이 제공됩니다. 자세한 내용은 " 통계량 " 및 " 검증 및 테스트 데이터 집합에 대한 엔트로피  $R^2$  " 에서 확인하십시오.
- " 판별 스코어 " 보고서에 검증 데이터 집합과 테스트 데이터 집합의 행을 식별하는 표시자가 나타납니다.
- " 스코어 요약 " 보고서에 훈련 데이터 집합, 검증 데이터 집합 및 테스트 데이터 집합에 대한 실제 분류 대 예측 분류가 표시됩니다.

## 판별 분석의 추가 예

이 섹션에는 판별 플랫폼을 사용하는 예가 포함되어 있습니다.

- "3D 정준 그림의 예 "
- " 단계별 변수 선택의 예 "

### 3D 정준 그림의 예

판별 플랫폼을 사용하여 여러 수준의 변수에 대한 3D 정준 그림을 생성합니다.

#### 결측값이 있는 행 숨기고 제외하기

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Owl Diet.jmp 를 엽니다.
2. **행 > 행 선택 > 선택 조건**을 선택합니다.
3. species 를 선택하고 **조건 추가**를 클릭합니다.  
이렇게 하면 species 가 결측인 모든 행이 선택됩니다. 이러한 행을 숨기고 제외합니다.
4. **행 > 숨기고 제외하기**를 선택합니다.

#### species 별로 행에 색상 적용

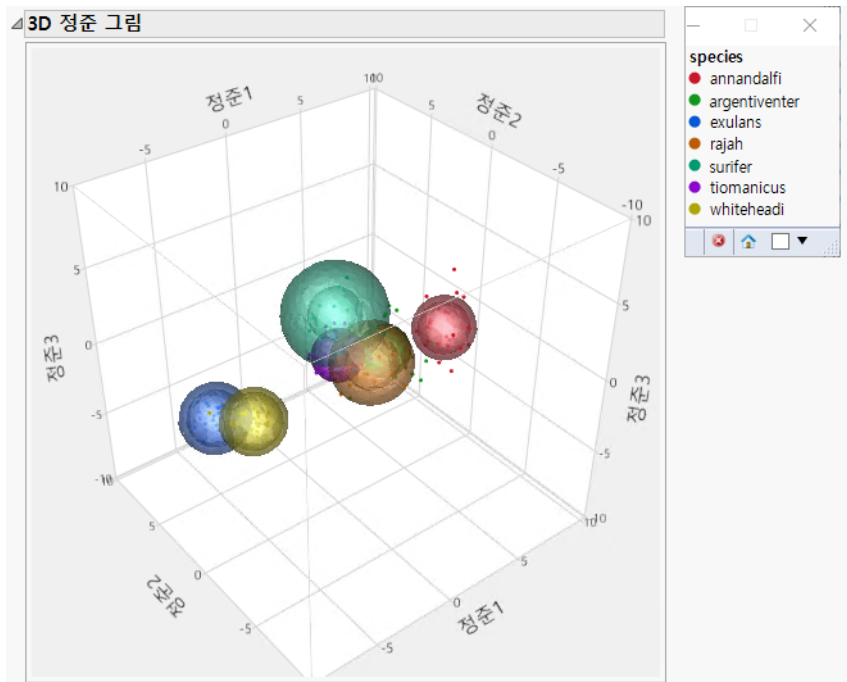
5. **행 > 열 값에 따른 색상 또는 표식**을 선택합니다.
6. species 를 선택합니다.
7. " 색상 " 메뉴에서 **JMP 진한 설정**을 선택합니다.
8. **범례가 포함된 창 만들기**를 선택합니다.
9. **확인**을 클릭합니다.  
작은 "범례" 창이 나타납니다. 데이터 테이블의 행에 species별로 색상이 지정되어 있습니다.

#### 판별 분석 수행

10. **분석 > 다변량 방법 > 판별**을 선택합니다.

11. skull length, teeth row, palatine foramen 및 jaw length 를 선택하고 **Y, 공변량**을 클릭합니다 .
12. species 를 선택하고 **X, 범주**를 클릭합니다 .
13. **확인**을 클릭합니다 .
14. " 판별 분석 " 의 빨간색 삼각형을 클릭하고 **3D 정준 그림**을 클릭합니다 .

그림 5.14 범례 차이 있는 3D 정준 그림



"범례"에서 범주를 클릭하여 "3D 정준 그림"에서 해당 점을 강조 표시할 수 있습니다. 3D 그림 내부를 클릭한 후 드래그하여 그림을 회전할 수도 있습니다.

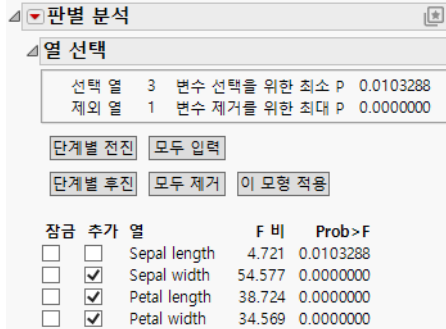
## 단계별 변수 선택의 예

이 예에서는 판별 플랫폼의 단계별 변수 선택을 사용하여 최종 모형에 포함되는 변수를 결정합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Iris.jmp 를 엽니다 .
2. **분석 > 다변량 방법 > 판별**을 선택합니다 .
3. Sepal length, Sepal width, Petal length 및 Petal width 를 선택하고 **Y, 공변량**을 클릭합니다 .
4. Species 를 선택하고 **X, 범주**를 클릭합니다 .
5. **단계별 변수 선택**을 선택합니다 .
6. **확인**을 클릭합니다 .

7. 단계별 전진을 세 번 클릭합니다.

**그림 5.15** Iris.jmp 의 변수 선별에 의한 모형



세 개의 공변량이 모형에 입력됩니다. "변수 선택을 위한 최소 P" 가 상단 패널에 나타납니다. 이 값이 0.0103288 이므로 남은 공변량 Sepal length 도 Species 에 대한 판별 분석에서 중요할 수 있음을 나타냅니다.

8. 이 모형 적용을 클릭합니다.

**그림 5.16** 선택한 공변량을 보여 주는 스코어 요약 보고서

소스	개수	오분류 수	오분류 비율	엔트로피	R <sup>2</sup>	-2*로그 가능성도
훈련	150	3	2.00000	0.95603		14.4917

훈련			
실제값	예측 개수		
Species	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	1	49

실제값	예측 비율		
Species	setosa	versicolor	virginica
setosa	1.000	0.000	0.000
versicolor	0.000	0.960	0.040
virginica	0.000	0.020	0.980

창이 업데이트되어 입력된 공변량과 선택한 판별 방법에 따라 적합 보고서가 표시됩니다. 모형에 대해 선택한 공변량이 "스코어 요약" 보고서 맨 위에 나열됩니다.

## 판별 플랫폼에 대한 통계 상세 정보

이 섹션에는 판별 분석에 사용되는 통계 상세 정보가 포함되어 있습니다.

- "와이드 선형 알고리즘에 대한 통계 상세 정보"
- "저장된 계산식에 대한 통계 상세 정보"
- "다변량 검정에 대한 통계 상세 정보"

- "근사 F-검정에 대한 통계 상세 정보"
- "그룹 간 공분산 행렬에 대한 통계 상세 정보"

## 와이드 선형 알고리즘에 대한 통계 상세 정보

와이드 선형 판별 분석은 다음과 같이 수행됩니다.

- 데이터는 그룹 평균을 빼고 합동 표준편차로 나누어 표준화됩니다.
- 특이값 분해를 사용하여 특이 벡터 집합에서 주성분 변환 행렬을 얻습니다.
- 유지된 성분 수는 특이값 제곱합의 최소 0.9999 를 나타냅니다.
- 변환된 데이터에 대해 선형 판별 분석이 수행되며, 이때 데이터가 그룹 평균에 의해 변화되지 않습니다. 합동 그룹 내 공분산 행렬이 대각 행렬이므로 빠르게 계산할 수 있습니다.

## 저장된 계산식에 대한 통계 상세 정보

이 섹션에서는 "판별 분석"의 빨간색 삼각형 메뉴에 있는 스코어 옵션 > 계산식 저장을 사용하여 저장된 계산식의 파생을 보여 줍니다. 계산식은 판별 방법에 따라 달라집니다.

범주형 변수  $X$ 에 의해 정의된 각 그룹에 대해 공변량의 관측값은  $p$  차원 다변량 정규 분포를 따른다고 가정됩니다. 여기서  $p$ 는 공변량 수입니다. 계산식에 사용되는 표기는 [표 5.2](#)에 나와 있습니다.

**표 5.2** 계산식 저장 옵션을 통해 제공된 계산식의 표기

$p$	공변량 수
$T$	총 그룹 수( $X$ 수준)
$t = 1, \dots, T$	$X$ 에 의해 정의된 그룹을 구별하기 위한 첨자
$n_t$	그룹 $t$ 의 관측값 수
$n = n_1 + n_2 + \dots + n_T$	총 관측값 수
$\mathbf{y}$	관측값에 대한 $p \times 1$ 공변량 벡터
$\mathbf{y}_{it} = (y_{i1t} \ y_{i2t} \ \dots \ y_{ipt})$	$p$ 개 공변량의 벡터로 구성된 그룹 $t$ 의 $i$ 번째 관측값
$\bar{\mathbf{y}}_t$	그룹 $t$ 의 관측값에 대한 $p \times 1$ 공변량 벡터 $\mathbf{y}$ 의 평균
$\mathbf{y}_{bar}$	모든 관측값에 대한 $p \times 1$ 공변량 벡터의 평균

표 5.2 계산식 저장 옵션을 통해 제공된 계산식의 표기 (계속)

$\mathbf{S}_t = \frac{1}{n_t - 1} \sum_{i=1}^{n_t} (\mathbf{y}_{it} - \bar{\mathbf{y}}_t)(\mathbf{y}_{it} - \bar{\mathbf{y}}_t)'$	그룹 $t$ 에 대해 추정된 $(p \times p)$ 그룹 내 공분산 행렬
$\mathbf{S}_p = \frac{1}{n - T} \sum_{t=1}^T (n_t - 1)\mathbf{S}_t$	추정된 $(p \times p)$ 합동 그룹 내 공분산 행렬
$q_t$	그룹 $t$ 에 속할 사전 확률
$p(t \mathbf{y})$	$\mathbf{y}$ 가 그룹 $t$ 에 속할 사후 확률
$ \mathbf{A} $	행렬 $\mathbf{A}$ 의 행렬식

### 선형 판별 방법

선형 판별 분석에서는 그룹 내 공분산 행렬이 모두 같다고 가정합니다. 공통 공분산 행렬은  $\mathbf{S}_p$ 로 추정됩니다. 표기는 표 5.2에서 확인하십시오.

관측값  $\mathbf{y}$ 에서 그룹  $t$ 까지의 Mahalanobis 거리는 다음과 같이 정의됩니다.

$$d_t^2 = (\mathbf{y} - \bar{\mathbf{y}}_t)' \mathbf{S}_p^{-1} (\mathbf{y} - \bar{\mathbf{y}}_t)$$

그룹  $t$ 에서 관측값  $\mathbf{y}$ 의 가능도는 다음과 같이 추정됩니다.

$$\begin{aligned} l_t(\mathbf{y}) &= (2\pi)^{-T/2} |\mathbf{S}_p|^{-1/2} \exp(-(\mathbf{y} - \bar{\mathbf{y}}_t)' \mathbf{S}_p^{-1} (\mathbf{y} - \bar{\mathbf{y}}_t)/2) \\ &= (2\pi)^{-T/2} |\mathbf{S}_p|^{-1/2} \exp(-d_t^2/2) \end{aligned}$$

합동 공분산 행렬에 대해 추정해야 할 모수의 수는  $p(p+1)/2$ 이고 평균에 대해서는  $Tp$ 입니다. 추정해야 할 모수의 총 수는  $p(p+1)/2 + Tp$ 입니다.

그룹  $t$ 에 소속될 사후 확률은 다음과 같이 정의됩니다.

$$p(t|\mathbf{y}) = \frac{q_t l_t(\mathbf{y})}{\sum_{u=1}^T q_u l_u(\mathbf{y})} = \frac{1}{1 + \sum_{u \neq t} \exp(-[(d_u^2 - 2\log(q_u)) - (d_t^2 - 2\log(q_t))]/2)}$$

관측값  $\mathbf{y}$ 는 사후 확률이 가장 큰 그룹에 할당됩니다.

선형 판별 방법에 의해 저장되는 계산식은 다음과 같이 정의됩니다.

SqDist[0]	$\mathbf{y}'\mathbf{S}_p^{-1}\mathbf{y}$
SqDist[<group $t$ >]	$d_t^2 - 2\log(q_t)$
Prob[<group $t$ >]	$p(t \mathbf{y})$
Pred <X>	$p(t \mathbf{y})$ 가 최대값이 되는 $t(t = 1, \dots, T)$

## 2차 판별 방법

2차 판별 분석에서는 그룹 내 공분산 행렬이 같다고 가정되지 않습니다. 그룹  $t$ 에 대한 그룹 내 공분산 행렬은  $\mathbf{S}_t$ 로 추정됩니다. 따라서 그룹 내 공분산 행렬에 대해 추정해야 할 모수의 수는  $Tp(p+1)/2$ 이고 평균에 대해서는  $Tp$ 입니다. 추정해야 할 모수의 총 수는  $Tp(p+3)/2$ 입니다.

그룹 표본 크기가  $p$ 에 비해 작으면 그룹 내 공분산 행렬의 추정값이 크게 변동되는 경향이 있습니다. 판별 스코어는 그룹 내 공분산 행렬에 대한 역행렬의 가장 작은 고유값에 크게 영향을 받습니다. 자세한 내용은 Friedman(1989) 연구 자료에서 확인하십시오. 따라서 그룹 표본 크기가  $p$ 에 비해 작은 경우 "정규화 판별 방법"에 설명된 정규화 방법을 고려할 수도 있습니다.

표기는 표 5.2에서 확인하십시오. 관측값  $\mathbf{y}$ 에서 그룹  $t$ 까지의 Mahalanobis 거리는 다음과 같이 정의됩니다.

$$d_t^2 = (\mathbf{y} - \bar{\mathbf{y}}_t)' \mathbf{S}_t^{-1} (\mathbf{y} - \bar{\mathbf{y}}_t)$$

그룹  $t$ 에서 관측값  $\mathbf{y}$ 의 가능도는 다음과 같이 추정됩니다.

$$\begin{aligned} l_t(\mathbf{y}) &= (2\pi)^{-T/2} |\mathbf{S}_t|^{-1/2} \exp(-(\mathbf{y} - \bar{\mathbf{y}}_t)' \mathbf{S}_t^{-1} (\mathbf{y} - \bar{\mathbf{y}}_t)/2) \\ &= (2\pi)^{-T/2} |\mathbf{S}_t|^{-1/2} \exp(-d_t^2/2) \end{aligned}$$

그룹  $t$  에 소속될 사후 확률은 다음과 같습니다.

$$p(t|\mathbf{y}) = (q_t l_t(\mathbf{y})) / \left( \sum_{u=1}^T q_u l_u(\mathbf{x}) \right)$$

$$= \frac{1}{1 + \sum_{u \neq t} \exp(-[(d_u^2 + \log|\mathbf{S}_u| - 2\log(q_u)) - (d_t^2 + \log|\mathbf{S}_t| - 2\log(q_t))]/2)}$$

판측값  $\mathbf{y}$  는 사후 확률이 가장 큰 그룹에 할당됩니다.

2 차 판별 방법에 의해 저장되는 계산식은 다음과 같이 정의됩니다.

SqDist[<group $t$ >]	$d_t^2 + \log \mathbf{S}_t  - 2\log(q_t)$
Prob[<group $t$ >]	$p(t \mathbf{y})$
Pred <X>	$p(t \mathbf{y})$ 가 최대값이 되는 $t(t = 1, \dots, T)$
<b>참고 :</b> SqDist[<group $t$ >] 는 음수일 수 있습니다.	

## 정규화 판별 방법

정규화 판별 분석에서는  $\lambda$  와  $\gamma$  라는 두 모수를 사용할 수 있습니다.

- $\lambda$  모수는 동일하다고 가정되지 않는 합동 공분산 행렬과 그룹 내 공분산 행렬에 할당된 가중치의 균형을 맞춥니다.
- $\gamma$  모수는 대각 행렬에 가깝게 축소할 크기를 결정합니다.

이 방법을 사용하면 정규화의 두 가지 측면을 활용하여 2차 판별 분석의 추정값에 안정성을 부여할 수 있습니다. 자세한 내용은 Friedman(1989) 연구 자료에서 확인하십시오. 표기는 표 5.2에서 확인하십시오.

정규화 방법의 경우 그룹  $t$  에 대한 공분산 행렬은 다음과 같습니다.

$$\Sigma_t = (1 - \gamma)(\lambda \mathbf{S}_p + (1 - \lambda)\mathbf{S}_t) + \gamma \text{Diag}((\lambda \mathbf{S}_p + (1 - \lambda)\mathbf{S}_t))$$

관측값  $\mathbf{y}$  에서 그룹  $t$  까지의 Mahalanobis 거리는 다음과 같이 정의됩니다.

$$d_t^2 = (\mathbf{y} - \bar{\mathbf{y}}_t)' \Sigma_t^{-1} (\mathbf{y} - \bar{\mathbf{y}}_t)$$

그룹  $t$  에서 관측값  $\mathbf{y}$  의 가능도는 다음과 같이 추정됩니다.

$$\begin{aligned} l_t(\mathbf{y}) &= (2\pi)^{-T/2} |\Sigma_t|^{-1/2} \exp(-(\mathbf{y} - \bar{\mathbf{y}}_t)' \Sigma_t^{-1} (\mathbf{y} - \bar{\mathbf{y}}_t)/2) \\ &= (2\pi)^{-T/2} |\Sigma_t|^{-1/2} \exp(-d_t^2/2) \end{aligned}$$

그룹  $t$  에 소속될 사후 확률은 다음과 같이 지정됩니다.

$$\begin{aligned} p(t|\mathbf{y}) &= (q_t l_t(\mathbf{y})) / \left( \sum_{u=1}^T q_u l_u(\mathbf{y}) \right) \\ &= \frac{1}{1 + \sum_{u \neq t} \exp(-[(d_u^2 + \log|\Sigma_u| - 2\log(q_u)) - (d_t^2 + \log|\Sigma_t| - 2\log(q_t))]/2)} \end{aligned}$$

관측값  $\mathbf{y}$  는 사후 확률이 가장 큰 그룹에 할당됩니다.

정규화 판별 방법에 의해 저장되는 계산식은 아래와 같이 정의됩니다.

SqDist[<group $t$ >]	$d_t^2 + \log \Sigma_t  - 2\log(q_t)$
Prob[<group $t$ >]	$p(t \mathbf{y})$
Pred <X>	$p(t \mathbf{y})$ 가 최대값이 되는 $t(t = 1, \dots, T)$

**참고 :** SqDist[<group  $t$ >] 는 음수일 수 있습니다.

### 와이드 선형 판별 방법

와이드 선형 방법은 공변량 수가 많은 경우, 특히 공변량 수가 관측값 수를 초과 ( $p > n$ ) 하는 경우에 유용합니다. 이 방법은 합동 그룹 내 공분산 행렬  $\mathbf{S}_p$  의 역행렬 또는 전치를 효율적으로 계산하는 것에 중점을 둡니다 ( $p > n$  인 경우). 여기에서는 큰 공분산 행렬에 대한 공간 할당 및 역행렬 계산이 발생하지 않도록 특이값 분해를 사용합니다.

와이드 선형 방법은 그룹 내 공분산 행렬이 같다고 가정하며, 관측값 수가 공변량 수와 같거나 이를 초과하는 경우 선형 방법과 동일합니다.

### 와이드 선형 계산

표기는 [표 5.2](#) 에서 확인하십시오. 다음 단계를 사용하여 와이드 선형 계산을 수행합니다.

1. 그룹 내 표본 평균의  $T \times p$  행렬  $\mathbf{M}$  을 계산합니다.  $\mathbf{M}$  의  $(t, j)$  번째 항목인  $m_{tj}$  는  $j$  번째 공변량에 대한 그룹  $t$  멤버의 표본 평균입니다.
2. 각 공변량  $j$  에 대해 그룹 전체의 합동 표준편차를 계산합니다. 이를  $s_{jj}$  라고 합니다.
3. 대각 항목  $s_{jj}$  가 있는 대각 행렬을  $\mathbf{S}_{diag}$  로 나타냅니다.
4. 각 공변량에 대한 값을 다음과 같이 중심화 및 척도화합니다.
  - 관측값이 속한 그룹의 평균을 뺍니다.
  - 차이를 합동 표준편차로 나눕니다.

그룹  $t$  의 관측값  $i$  에 대해  $j$  번째 공변량의 그룹 중심화 및 척도화된 값은 다음과 같이 표기합니다.

$$y_{ij}^* = \frac{y_{ij} - m_{t(i)j}}{s_{jj}}$$

$t(i)$  표기는 관측값  $i$  가 속한 그룹  $t$  를 나타냅니다.

5.  $y_{ij}^*$  값 행렬을  $\mathbf{Y}_s$  로 나타냅니다.
6. 그룹 중심화 및 척도화된 공변량의 합동 그룹 내 공분산 행렬을  $\mathbf{R}$  로 나타냅니다.  $\mathbf{R}$  행렬은 다음과 같이 지정됩니다.

$$\mathbf{R} = (\mathbf{Y}_s' \mathbf{Y}_s) / (n - T)$$

7.  $\mathbf{Y}_s$  에 특이값 분해를 적용합니다.

$$\mathbf{Y}_s = \mathbf{U} \mathbf{D} \mathbf{V}'$$

여기서  $\mathbf{U}$  와  $\mathbf{V}$  는 직교정규이며  $\mathbf{D}$  는 대각선에 양수 항목 (특이값) 이 있는 대각 행렬입니다. 자세한 내용은 "특이값 분해" 에서 확인하십시오.

그러면  $\mathbf{R}$  을 다음과 같이 쓸 수 있습니다.

$$\mathbf{R} = (\mathbf{Y}_s' \mathbf{Y}_s) / (n - T) = (\mathbf{V} \mathbf{D}^2 \mathbf{V}') / (n - T)$$

8.  $\mathbf{R}$  이 완전 계수인 경우  $\mathbf{R}^{-1/2}$  은 다음과 같습니다.

$$\mathbf{R}^{-1/2} = (\mathbf{V} \mathbf{D}^{-1} \mathbf{V}') / \sqrt{n - T}$$

여기서  $\mathbf{D}^{-1}$ 은 대각 항목이  $\mathbf{D}$ 의 대각 항목에 대한 역행렬인 대각 행렬입니다.  
 $\mathbf{R}$ 이 완전 계수가 아닌 경우  $\mathbf{R}$ 에 대한 유사역행렬을 다음과 같이 정의합니다.

$$\mathbf{R}^- = (\mathbf{V}\mathbf{D}^{-2}\mathbf{V}')/(n-T)$$

그러면  $\mathbf{R}$ 의 역제곱근을 다음과 같이 정의합니다.

$$(\mathbf{R}^-)^{1/2} = (\mathbf{V}\mathbf{D}^{-1}\mathbf{V}')/\sqrt{n-T}$$

9.  $\mathbf{R}$ 이 완전 계수이면  $\mathbf{R}^- = \mathbf{R}^{-1}$ 이 됩니다. 따라서 완비성을 위해 유사역행렬을 사용하여 계속합니다.

$p \times p$  행렬  $\mathbf{T}_s$ 를 다음과 같이 정의합니다.

$$\mathbf{T}_s = (\mathbf{S}_{diag}^{-1}\mathbf{V}\mathbf{D}^{-1})/(\sqrt{n-T})$$

그러면 다음과 같이 구성할 수 있습니다.

$$(\mathbf{T}_s\mathbf{T}_s') = \left( \mathbf{S}_{diag}^{-1}\mathbf{V}(\mathbf{D}^{-1})^2\mathbf{V}'\mathbf{S}_{diag}^{-1} \right) / (n-T) = \mathbf{S}_{diag}^{-1}\mathbf{R}^-\mathbf{S}_{diag}^{-1} = \mathbf{S}_p^-$$

여기서  $\mathbf{S}_p^-$ 는 원래 데이터에 대해 SVD를 사용하여 계산된 합동 그룹 내 공분산 행렬의 일반화 역행렬입니다.

### Mahalanobis 거리

Mahalanobis 거리, 가능도 및 사후 확률에 대한 계산식은 "선형 판별 방법"의 계산식과 동일합니다. 그러나  $\mathbf{S}_p$ 의 역행렬은 특이값 분해를 사용하여 계산된 일반화 역행렬로 대체됩니다.

계산식을 저장하면 분해 형식으로 Mahalanobis 거리가 제공됩니다. 관측값  $\mathbf{y}$ 에 대해 그룹  $t$ 까지의 제곱 거리는 다음과 같습니다. 여기서 마지막 등식의  $SqDist[0]$  및 판별 주성분은 "저장된 계산식"에 정의되어 있습니다.

$$\begin{aligned} d_t^2 &= (\mathbf{y} - \bar{\mathbf{y}}_t)' \mathbf{S}_p^- (\mathbf{y} - \bar{\mathbf{y}}_t) \\ &= (\mathbf{y} - \bar{\mathbf{y}}_t)' \mathbf{T}_s \mathbf{T}_s' (\mathbf{y} - \bar{\mathbf{y}}_t) \\ &= ((\mathbf{y} - \bar{\mathbf{y}}) - (\bar{\mathbf{y}}_t - \bar{\mathbf{y}}))' \mathbf{T}_s \mathbf{T}_s' ((\mathbf{y} - \bar{\mathbf{y}}) - (\bar{\mathbf{y}}_t - \bar{\mathbf{y}})) \\ &= (\mathbf{T}_s'(\mathbf{y} - \bar{\mathbf{y}}))' (\mathbf{T}_s'(\mathbf{y} - \bar{\mathbf{y}})) - 2(\mathbf{T}_s'(\bar{\mathbf{y}}_t - \bar{\mathbf{y}}))' (\mathbf{T}_s'(\mathbf{y} - \bar{\mathbf{y}})) + (\mathbf{T}_s'(\bar{\mathbf{y}}_t - \bar{\mathbf{y}}))' (\mathbf{T}_s'(\bar{\mathbf{y}}_t - \bar{\mathbf{y}})) \\ &= SqDist[0] - 2(\mathbf{T}_s'(\bar{\mathbf{y}}_t - \bar{\mathbf{y}}))' Discrim Prin Comp + (\mathbf{T}_s'(\bar{\mathbf{y}}_t - \bar{\mathbf{y}}))' (\mathbf{T}_s'(\bar{\mathbf{y}}_t - \bar{\mathbf{y}})) \end{aligned}$$

### 저장된 계산식

다음은 와이드 선형 판별 방법에 의해 저장되는 계산식입니다.

판별 데이터 행렬	공변량에 대한 관측값 벡터
판별 주성분	데이터를 그룹 내에서 상관관계가 없도록 렌더링하는 주 성분 스코어링 행렬에 의해 변환된 데이터입니다. $\mathbf{T}_s'(\mathbf{y} - \bar{\mathbf{y}})$ 로 지정되며, 여기서 $\bar{\mathbf{y}}$ 는 전체 평균을 포함하는 $p \times 1$ 벡터입니다.
SqDist[0]	$(\mathbf{y} - \bar{\mathbf{y}})' \mathbf{T}_s' \mathbf{T}_s (\mathbf{y} - \bar{\mathbf{y}})$
SqDist[<group t>]	관측값에서 그룹 중심까지의 Mahalanobis 거리입니다. 자세한 내용은 " <a href="#">Mahalanobis 거리</a> "에서 확인하십시오.
Prob[<group t>]	$p(t \mathbf{y})$ (" <a href="#">선형 판별 방법</a> "에 제공됨)
Pred <X>	$p(t \mathbf{y})$ 가 최대값이 되는 $t(t = 1, \dots, T)$

### 다변량 검정에 대한 통계 상세 정보

아래에서  $\mathbf{E}$ 는 잔차 교차곱 행렬이고,  $\mathbf{H}$ 는 모형 교차곱 행렬입니다.  $\mathbf{E}$ 의 대각 요소는 각 변수의 잔차 제곱합입니다.  $\mathbf{H}$ 의 대각 요소는 각 변수에 대한 모형의 제곱합입니다. 판별 분석 관련 문헌에서는 종종  $\mathbf{E}$ 를  $\mathbf{W}$ 라고 합니다( $\mathbf{W}$ 는 within을 나타냄).

다변량 결과 표의 검정 통계량은  $\mathbf{E}^{-1}\mathbf{H}$ 의 고유값  $\lambda$ 를 사용해서 구합니다. 다음 목록에서는 각 검정 통계량의 계산 방법을 설명합니다.

**참고:** 반응 설계가 지정된 후 초기  $\mathbf{E}$  및  $\mathbf{H}$  행렬의 앞에  $\mathbf{M}'$ 를 곱하고 뒤에  $\mathbf{M}$ 을 곱합니다.

- Wilks 람다

$$\Lambda = \frac{\det(\mathbf{E})}{\det(\mathbf{H} + \mathbf{E})} = \prod_{i=1}^n \left( \frac{1}{1 + \lambda_i} \right)$$

- Pillai 대각합

$$V = \text{Trace}[\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}] = \sum_{i=1}^n \frac{\lambda_i}{1 + \lambda_i}$$

- Hotelling-Lawley 대각합

$$U = \text{Trace}(\mathbf{E}^{-1}\mathbf{H}) = \sum_{i=1}^n \lambda_i$$

- Roy 최대근

$$\Theta = \lambda_1, \mathbf{E}^{-1}\mathbf{H} \text{의 최대 고유값}$$

$\mathbf{E}$ 와  $\mathbf{H}$ 는 다음과 같이 정의됩니다.

$$\mathbf{E} = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b}$$

$$\mathbf{H} = (\mathbf{Lb})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{Lb})$$

여기서  $\mathbf{b}$ 는 모형 계수의 추정 벡터이며,  $\mathbf{A}$ 는 행렬  $\mathbf{A}$ 의 일반화 역행렬을 나타냅니다.

전체 모형  $\mathbf{L}$ 은 0(절편 값)으로 구성된 열로, 행 및 열 개수가 모형의 모수 개수와 같은 단위 행렬과 연결되어 있습니다. 효과에 대한  $\mathbf{L}$  행렬은 전체 모형  $\mathbf{L}$  행렬의 일부 행으로 구성된 부분집합입니다.

## 근사 F-검정에 대한 통계 상세 정보

$F$  값과 자유도를 계산하기 위해  $p$ 의 계수가  $\mathbf{H} + \mathbf{E}$ 이고  $q$ 의 계수가  $\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'$ 이라고 하겠습니다. 여기서  $\mathbf{L}$  행렬은 검정 중인 효과와 관련된  $\mathbf{X}'\mathbf{X}$ 의 요소를 식별합니다.  $v$ 는 오차 자유도가 되도록 하고  $s$ 는  $p$  및  $q$ 의 최소값이 되도록 합니다. 또한  $m = 0.5(p - q - 1)$  및  $n = 0.5(v - p - 1)$ 이 되도록 합니다.

표 5.3에서는 해당 검정 통계량에서 구한 각 근사 F의 계산을 보여 줍니다.

표 5.3 근사 F 통계량

검정	근사 F	분자 DF	분모 DF
Wilks 람다	$F = \left( \frac{1 - \Lambda^{1/t}}{\Lambda^{1/t}} \right) \left( \frac{rt - 2u}{pq} \right)$	$pq$	$rt - 2u$
Pillai 대각합	$F = \left( \frac{V}{s - V} \right) \left( \frac{2n + s + 1}{2m + s + 1} \right)$	$s(2m + s + 1)$	$s(2n + s + 1)$

표 5.3 근사  $F$  통계량 (계속)

검정	근사 $F$	분자 DF	분모 DF
<b>Hotelling-Lawley</b> 대각합	$F = \frac{2(sn+1)U}{s^2(2m+s+1)}$	$s(2m+s+1)$	$2(sn+1)$
<b>Roy 최대근</b>	$F = \frac{\Theta(v - \max(p, q) + q)}{\max(p, q)}$	$\max(p, q)$	$v - \max(p, q) + q$

## 그룹 간 공분산 행렬에 대한 통계 상세 정보

판별 플랫폼의 그룹 간 공분산 행렬은 표 5.2의 표기를 사용하여 다음과 같이 정의됩니다.

$$S_B = \frac{1}{T-1} \sum_{t=1}^T T \binom{n_t}{n} (\bar{\mathbf{y}}_t - \mathbf{y}_{bar})(\bar{\mathbf{y}}_t - \mathbf{y}_{bar})'$$

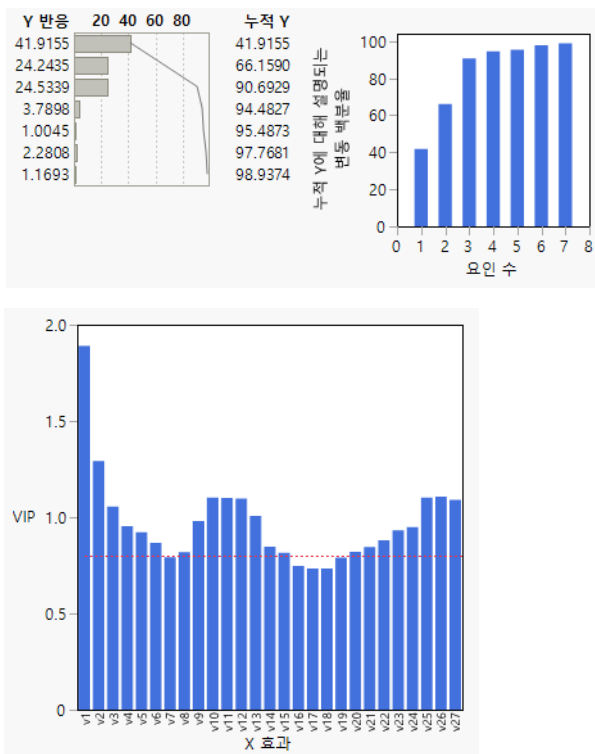
## 부분 최소 제곱 모형 Y와 X 간의 상관을 사용하여 모형 개발

PLS(부분 최소 제곱) 플랫폼에서는 요인, 즉 설명 변수(X)의 선형 결합을 기반으로 선형 모형을 적합시킵니다. 이러한 요인은 X와 하나 이상의 반응(Y) 간의 공분산을 최대화하는 방식으로 얻어집니다. PLS는 X와 Y 사이의 상관을 이용하여 기본 잠재 구조를 나타냅니다.

**JMP PRO** JMP Pro에서는 PLS-DA(PLS 판별 분석) 수행, 다양한 모형 효과 포함, 여러 검증 방법 사용, 결측 데이터 대치, 다양한 통계량 분포의 붓스트랩 추정값 계산 등의 작업이 가능한 추가 기능을 제공합니다.

부분 최소 제곱은 X 변수가 관측값보다 많은 경우, X 변수 간의 상관관계가 높은 경우, X 변수가 많은 경우, Y 변수가 몇 개 있고 X 변수가 많은 경우와 같이 보통 최소 제곱을 사용하여 만족스러운 결과를 얻을 수 없는 상황에서 효과적입니다.

그림 6.1 부분 최소 제곱 보고서 (일부)



## 목차

부분 최소 제곱 플랫폼 개요 .....	121
부분 최소 제곱의 예 .....	122
부분 최소 제곱 플랫폼 시작 .....	125
중심화 및 척도화 .....	128
X 표준화 .....	128
모형 시작 제어판 .....	129
부분 최소 제곱 옵션 .....	130
부분 최소 제곱 보고서 .....	131
모형 비교 요약 .....	131
교차 검증 보고서 .....	132
모형 적합 보고서 .....	136
모형 적합 옵션 .....	137
변수 중요도 그림 .....	141
VIP 대 계수 그림 .....	142
부분 최소 제곱의 추가 예 .....	143
부분 최소 제곱 플랫폼에 대한 통계 상세 정보 .....	145
부분 최소 제곱에 대한 통계 상세 정보 .....	145
van der Voet $T^2$ 검정에 대한 통계 상세 정보 .....	146
$T^2$ 그림에 대한 통계 상세 정보 .....	147
X 스코어 산점도 행렬의 신뢰도 타원에 대한 통계 상세 정보 .....	147
예측 및 신뢰 한계에 대한 통계 상세 정보 .....	148
표준화된 스코어 및 적재량에 대한 통계 상세 정보 .....	149
PLS 판별 분석에 대한 통계 상세 정보 .....	150

## 부분 최소 제곱 플랫폼 개요

보통 최소 제곱과 달리 PLS(부분 최소 제곱) 방법은 예측 변수가 관측값보다 많을 때 사용할 수 있습니다. PLS는 분광학, 계량화학, 유전체학, 심리학, 교육, 경제학, 정치학 및 환경 과학 등의 분야에서 고차원 데이터를 모델링하는 데 광범위하게 사용됩니다.

PLS 방법을 사용한 모형 적합은 설명 변수가 관측값보다 많거나 설명 변수의 상관관계가 높을 때 특히 유용합니다. PLS를 사용하여 단일 모형을 여러 반응에 동시에 적합시킬 수 있습니다. 자세한 내용은 Garthwaite(1994), Wold(1994), Wold et al. (2001), Eriksson et al. (2006) 및 Cox와 Gaudard(2013) 연구 자료에서 확인하십시오.

NIPALS(Nonlinear Iterative Partial Least Squares)와 SIMPLS("Statistically Inspired Modification of PLS")라는 두 가지 모형 적합 알고리즘을 사용할 수 있습니다. NIPALS에 대한 자세한 내용은 Wold(1980) 연구 자료에서 확인하십시오. SIMPLS에 대한 자세한 내용은 De Jong(1993) 연구 자료에서 확인하십시오. 두 방법에 대한 설명은 Boulesteix와 Strimmer(2007) 연구 자료에서 확인하십시오. SIMPLS 알고리즘은 특정 최적성 문제의 해를 구하는 것을 목표로 개발되었습니다. 단일 반응의 경우 두 방법 모두 동일한 모형을 제공합니다. 다중 반응의 경우 약간 차이가 있습니다.

JMP에서는 분석 > 다변량 방법 > 부분 최소 제곱을 통해서만 PLS 플랫폼에 액세스할 수 있습니다. JMP Pro에서는 분석 > 모형 적합을 통해 부분 최소 제곱 분석법에 액세스할 수도 있습니다.

**JMP PRO** JMP Pro에서는 다음을 수행할 수 있습니다.

- 모형 적합에서 부분 최소 제곱 분석법을 사용하여 명목형 모델링 유형의 반응을 적합시키는 방식으로 PLS-DA(PLS 판별 분석)를 수행합니다.
- 모형 적합에서 부분 최소 제곱 분석법을 사용하여 다항식, 교호작용 및 범주형 효과를 적합 시킵니다.
- 검증 및 교차 검증을 위한 여러 방법 중에서 선택합니다.
- 결측 데이터를 대치합니다.
- 다양한 통계량 분포의 붓스트랩 추정값을 구합니다. 관심 있는 보고서를 마우스 오른쪽 버튼으로 클릭할 수 있습니다. 붓스트랩 추정값에 대한 자세한 내용은 기본 분석에서 확인하십시오.

부분 최소 제곱 플랫폼은 추출할 최적 요인 수를 선택하는 데 도움이 되도록 van der Voet  $T^2$  검증 및 교차 검증을 사용합니다.

- JMP 플랫폼에서는 leave-one-out 교차 검증 방법을 사용합니다. 검증을 사용하지 않도록 선택할 수도 있습니다.
- **JMP PRO** JMP Pro에서는 K 폴드, Leave-One-Out 또는 랜덤 홀드백 교차 검증을 선택하거나 검증 열을 지정할 수 있습니다. 검증을 사용하지 않도록 선택할 수도 있습니다.

## 부분 최소 제곱의 예

부분 최소 제곱 플랫폼을 사용하여 발트해의 해수 표본에 존재하는 세 가지 오염 화합물의 양을 예측하기 위한 모형을 생성합니다. 세 가지 관심 화합물은 다음과 같습니다.

- 리그닌술포산염 - 펄프 산업 공해
- 부식산 - 천연 임산물
- 세제의 형광 표백제

예측 변수는  $v1 \sim v27$ 의 파장 범위로 측정된 스펙트럼 방출 강도입니다. 이 예는 부분 최소 제곱이 매우 효과적인 분광계 보정 분야에서 가져왔습니다.

모형 보정을 위해 화학적 조성이 알려진 표본을 사용합니다. 보정 데이터는 농도가 알려진 리그닌술포산염, 부식산 및 세제의 16 개 표본으로 구성됩니다. 27 개의 등거리 파장에서 방출 강도가 기록됩니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Baltic.jmp** 를 엽니다.

---

**참고 :** Baltic.jmp 데이터 테이블의 데이터는 Umetrics(1995)에 보고되어 있습니다. 원본 출처는 Lindberg, Persson 및 Wold(1983) 연구 자료입니다.

---

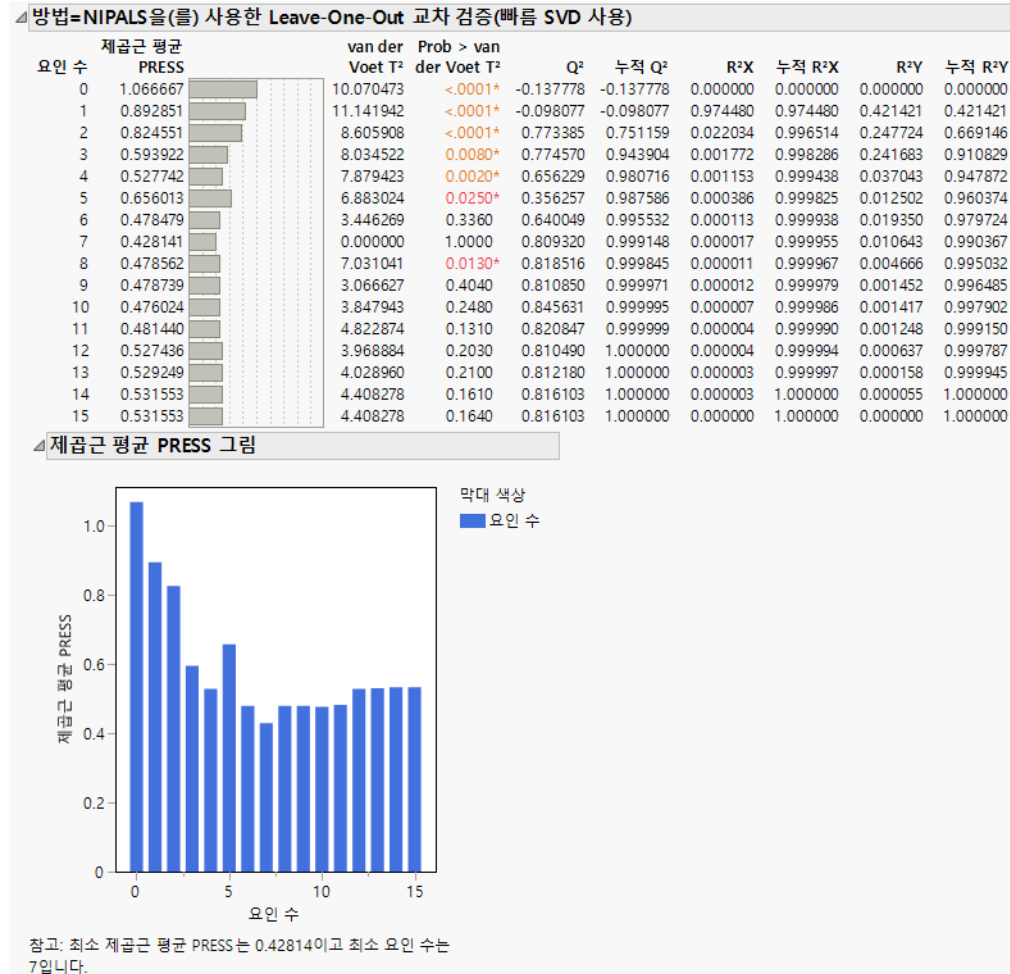
2. **분석 > 다변량 방법 > 부분 최소 제곱**을 선택합니다.
3. ls, ha 및 dt 를 **Y, 반응** 역할에 할당합니다.
4. 27 개의 강도 변수 ( $v1 \sim v27$ )가 포함된 **Intensities** 를 **X, 요인** 역할에 할당합니다.
5. **확인**을 클릭합니다.

부분 최소 제곱 모형 시작 제어판이 나타납니다.

6. "검증 방법"에서 **Leave-One-Out** 을 선택합니다.
7. **시작**을 클릭합니다.

van der Voet 검정은 난수화 검정이므로 "Prob > van der Voet T<sup>2</sup>" 값이 약간 다를 수 있습니다.

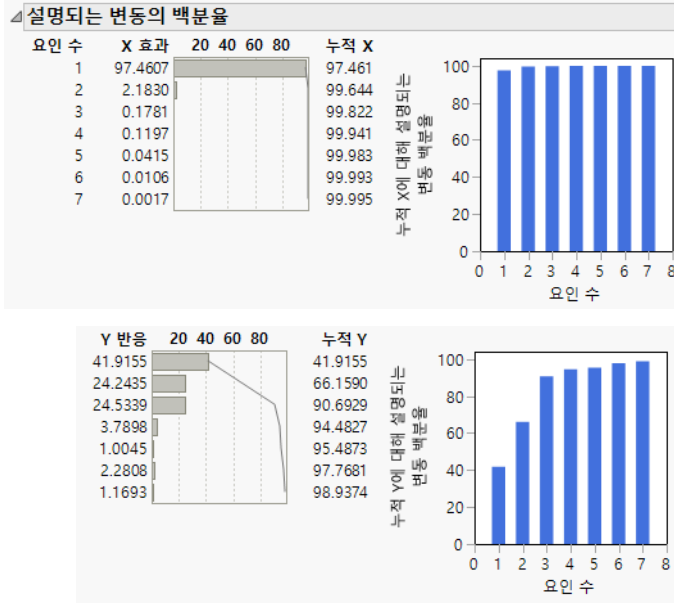
그림 6.2 부분 최소 제곱 보고서



" 제공된 평균 PRESS( 예측 잔차 제곱합 ) 그림 "에서는 요인 수가 7일 때 제공된 평균 PRESS가 최소화된다는 것을 보여 줍니다. 이 내용은 " 제공된 평균 PRESS 그림 "아래의 "참고"에 명시되어 있습니다. **7 개의 요인을 사용한 NIPALS 적합 ( 빠른 SVD 사용 )** 보고서가 생성되었습니다. 그림 6.3에 보고서의 일부가 표시되어 있습니다.

van der Voet T<sup>2</sup> 검정 통계량은 요인 수가 다른 모형이 최소 PRESS 값을 가진 모형과 유의하게 다른지 여부를 확인하는 데 사용됩니다. 일반적인 방법은 van der Voet 유의 수준이 0.10을 초과하는 가장 작은 수의 요인을 추출하는 것입니다 (SAS Institute Inc 2023f, Tobias 1995 참조). 이 방법을 여기에 적용하려면 모형 시작 패널의 요인 수에 6을 입력하여 새 모형을 적합시킵니다.

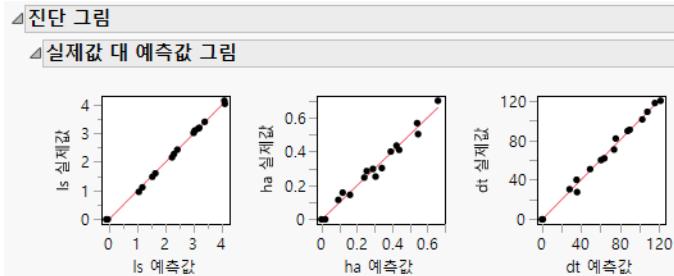
그림 6.3 추출된 7 개 요인



8. "7 개의 요인을 사용한 NIPALS 적합 ( 빠른 SVD 사용 )" 의 빨간색 삼각형을 클릭하고 **진단 그림**을 선택합니다 .

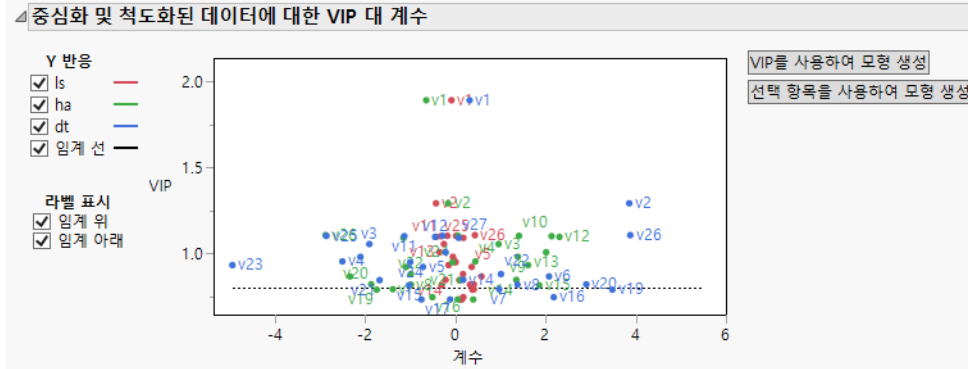
그러면 실제값 대 예측값 그림을 보여 주는 보고서와 다양한 잔차 그림을 보여 주는 세 개의 보고서가 제공됩니다 . " 실제값 대 예측값 그림 " 에서는 예측 화합물 양이 실제 양과 일치하는 정도를 보여 줍니다 .

그림 6.4 진단 그림



9. "7 개의 요인을 사용한 NIPALS 적합 ( 빠른 SVD 사용 )" 의 빨간색 삼각형을 클릭하고 **VIP 대 계수 그림**을 선택합니다 .

그림 6.5 VIP 대 계수 그림



VIP 대 계수 그림은 다양한 반응의 적합도와 관련하여 영향력 있는 변수를 식별하는 데 도움이 됩니다. 예를 들어 v23, v2 및 v26 은 모두 VIP(Variable Importance for Projection) 값이 0.8 을 초과하고 계수도 상대적으로 큼니다.

## 부분 최소 제곱 플랫폼 시작

다음과 같은 두 가지 방법으로 부분 최소 제곱 플랫폼을 시작할 수 있습니다.

- 분석 > 다변량 방법 > 부분 최소 제곱을 선택합니다.
- **JMP PRO** 분석 > 모형 적합을 선택하고 "분석법" 메뉴에서 **부분 최소 제곱**을 선택합니다. 이 방법을 사용하면 다음을 수행할 수 있습니다.
  - 범주형 변수를 Y 또는 X 로 입력합니다. 범주형 Y 를 입력하여 PLS-DA 를 수행합니다.
  - 교호작용 및 다항식 항을 모형에 추가합니다.
  - "X 표준화" 옵션을 사용하여 중심화 및 척도화된 열로 고차 항을 구성합니다.
  - 모형 규격 스크립트를 저장합니다.

모형 적합 시작 창의 일부 기능은 부분 최소 제곱 분석법에 사용할 수 없습니다.

- 가중치, 내포, 속성, 변환, 절편 없음

**팁:** "열 선택" 상자에서 변수를 마우스 오른쪽 버튼으로 클릭하고 "변환" 옵션을 선택하여 변수를 변환할 수 있습니다.

- 매크로: 혼합물 반응 표면, Scheffe 3 차

그림 6.6 JMP Pro 부분 최소 제곱 시작 창 (결측값 대체법에 EM 선택됨)

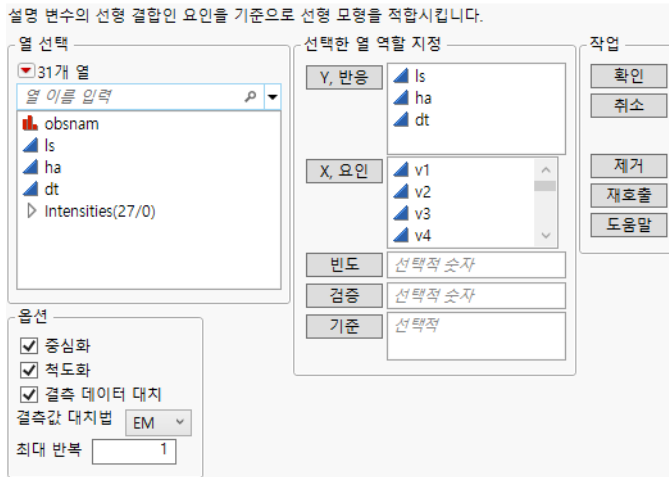
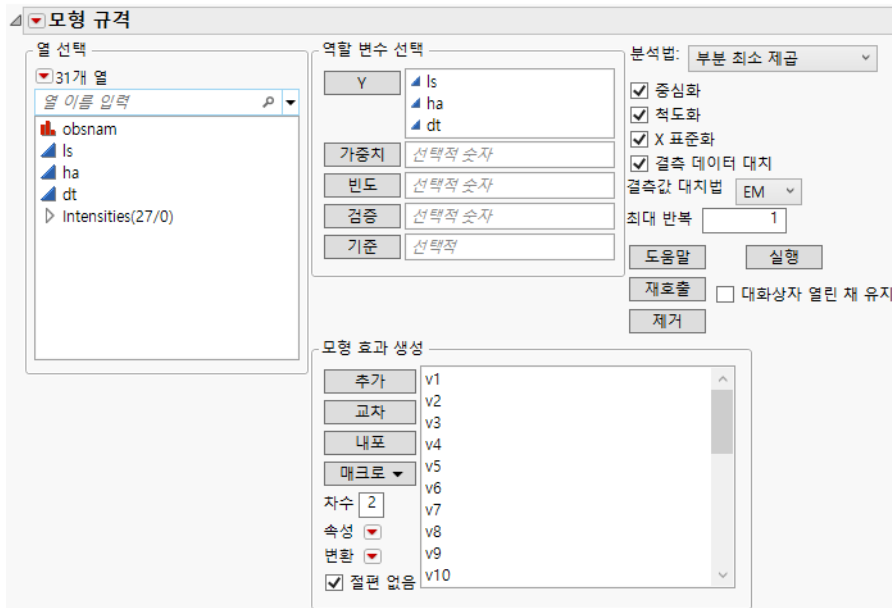


그림 6.7 JMP Pro 모형 적합 부분 최소 제곱 시작 창



모형 적합 시작 옵션에 대한 자세한 내용은 선형 모형 적합에서 확인하십시오.

"열 선택"의 빨간색 삼각형 메뉴에 포함된 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

부분 최소 제곱 시작 창에는 다음 옵션이 포함되어 있습니다.

**Y, 반응** 수치형 반응 열을 입력합니다. 열을 여러 개 입력하면 공동으로 모델링됩니다.

**JMP PRO** JMP Pro에서는 모형 적합 시작 창에 명목형 반응 열을 입력하여 PLS-DA 를 수행할 수 있습니다. 자세한 내용은 "[PLS 관별 분석에 대한 통계 상세 정보](#)"에서 확인하십시오.

**X, 요인** 예측 변수 열을 입력합니다. 부분 최소 제곱 시작 창에는 수치형 예측 변수만 허용됩니다.

**JMP PRO** JMP Pro에서는 모형 적합 시작 창에 명목형 및 순서형 모형 효과를 입력할 수 있습니다. 순서형 효과는 명목형으로 처리됩니다.

**빈도** 데이터가 요약된 경우 값에 각 행에 대한 개수가 포함된 열을 입력합니다.

**JMP PRO 검증** 검증 데이터 집합을 정의하는 숫자 열입니다. 검증 열에는 다음과 같이 연속된 정수 값만 포함되어야 합니다.

- 검증 열에 두 개의 수준이 포함된 경우 더 작은 값이 훈련 데이터 집합을 정의하고 더 큰 값이 검증 데이터 집합을 정의합니다.
- 검증 열에 세 개의 수준이 포함된 경우 작은 값부터 순서대로 훈련 데이터 집합, 검증 데이터 집합, 테스트 데이터 집합을 각각 정의합니다.
- 검증 열에 네 개 이상의 수준이 포함된 경우 K 폴드 교차 검증이 사용됩니다. 기타 검증 옵션에 대한 자세한 내용은 "[검증 방법](#)"에서 확인하십시오.

PLS 플랫폼에서는 검증 열을 사용하여 모형을 훈련 및 조정하거나 모형을 훈련, 조정 및 평가합니다. 검증에 대한 자세한 내용은 [예측 및 전문 모델링](#)에서 확인하십시오.

**참고:** "열 선택" 목록에서 아무 열도 선택하지 않은 상태로 "검증" 버튼을 클릭하면 데이터 테이블에 검증 열을 추가할 수 있습니다. "검증 열 생성" 유틸리티에 대한 자세한 내용은 [예측 및 전문 모델링](#)에서 확인하십시오.

**기준** 변수의 각 수준에 대한 개별 분석으로 구성된 보고서를 생성하는 열을 입력합니다. 기준 변수가 둘 이상 할당되면 기준 변수의 가능한 각 수준 조합에 대해 개별 분석이 생성됩니다.

**중심화** 각 열에서 평균을 빼서 모든 Y 변수와 모형 효과를 중심화합니다. 자세한 내용은 "[중심화 및 척도화](#)"에서 확인하십시오.

**척도화** 각 열을 표준편차로 나누어 모든 Y 변수와 모형 효과의 척도를 조정합니다. 자세한 내용은 "[중심화 및 척도화](#)"에서 확인하십시오.

**JMP PRO X 표준화** (모형 적합 시작 창에서만 사용 가능) 모형 효과를 구성하는 데 사용되는 모든 열을 중앙에 맞추고 척도를 조정합니다. 이 옵션을 선택하지 않으면 원래 데이터 테이블 열을 사용하여 고차 효과가 구성됩니다. 그런 다음 선택한 "중심화" 및 "척도화" 옵션에 따라 각 고차 효과가 중심화되거나 척도화됩니다. "X 표준화"는 Y 변수를 중앙에 맞추거나 척도를 조정하지 않습니다. 자세한 내용은 "[X 표준화](#)"에서 확인하십시오.

**JMP PRO 결측 데이터 대체** Y 또는 X의 결측 데이터 값을 비결측 값으로 대체합니다. **결측값 대체법** 목록에서 적절한 방법을 선택합니다.

**결측 데이터 대치**를 선택하지 않으면 X 변수의 관측값이 결측인 행이 분석에서 제외되고 해당 행에 대한 예측이 계산되지 않습니다. X 변수에는 결측 관측값이 없지만 Y 변수에 결측 관측값이 있는 행도 분석에서 제외되지만 예측이 계산됩니다.

**JMP PRO** **결측값 대치법** ( **결측 데이터 대치**를 선택한 경우에만 표시됨 ) 다음 결측값 대치 방법 중에서 선택합니다.

**평균** 각 모형 효과 또는 반응 열에 대해 결측값을 비결측 값의 평균으로 대체합니다.

**EM** 반복 EM(Expectation-Maximization) 방식을 사용하여 결측값을 대체합니다. 첫 번째 반복에서는 지정된 모형이 효과 또는 반응의 결측값을 평균으로 대체한 데이터에 적합됩니다. Y 모형의 예측값과 X 모형의 예측값을 사용하여 결측값을 대체합니다. 이후 반복의 경우 현재 추정값을 사용한 조건부 분포가 주어지면 결측값이 해당 예측값으로 대체됩니다.

결측값 대치를 위해 다항식 항은 별도의 예측 변수로 처리됩니다. 다항식 항이 지정되면 해당 항은 원래 데이터에서 계산되거나, "X 표준화"가 선택된 경우 표준화된 열 값에서 계산됩니다. 다항식 항 정의와 관련된 열의 값이 결측인 행이 있으면 해당 항목은 다항식 항에 대해 결측입니다. 이 방식으로 정의된 다항식 항을 사용하여 결측값 대치가 수행됩니다.

EM 방법에 대한 자세한 내용은 Nelson, Taylor 및 MacGregor(1996) 연구 자료에서 확인하십시오.

**JMP PRO** **최대 반복** ("결측값 대치법"에서 "EM"을 선택한 경우에만 표시됨) 알고리즘에 사용되는 최대 반복 수를 설정할 수 있습니다. 결측값의 현재 추정값과 이전 추정값 사이의 최대 차이가  $10^{-8}$ 로 제한되면 알고리즘이 종료됩니다.

시작 창을 완료하고 **확인**을 클릭하면 "모형 시작" 제어판이 나타납니다. 자세한 내용은 "**모형 시작 제어판**"에서 확인하십시오.

## 중심화 및 척도화

부분 최소 제곱에 대한 중심화 및 척도화 옵션은 기본적으로 선택됩니다. 즉, 예측 변수와 반응이 평균 0, 표준편차 1이 되도록 중심화되고 척도가 조정됩니다. 예측 변수와 반응을 중심화하면 변동에 비례하여 해당 변수와 반응이 동등하게 됩니다. 중심화하지 않으면 변수의 평균과 해당 평균에 대한 변동이 모두 연속 요인을 생성하는 데 포함됩니다. 설명을 위해 Time과 Temp가 두 개의 예측 변수라고 가정해 보겠습니다. 이 변수를 척도화하면 Time의 1 표준편차 변화가 Temp의 1 표준편차 변화와 거의 동일하다는 것을 나타냅니다.

## **JMP PRO** X 표준화

"모형 적합" 창에서 "부분 최소 제곱" 분석법을 선택하면 "X 표준화" 옵션이 기본적으로 선택됩니다. 이렇게 하면 모형 효과로 입력한 모든 열과 교호작용 또는 다항식 항에 포함된 모든 열이 표준화됩니다.

X1 과 X2 라는 두 개의 열이 있고 " 모형 적합 " 창에서 교호작용 항 X1\*X2 를 모형 효과로 입력한다고 가정해 보겠습니다. "X 표준화" 옵션을 선택하면 교호작용 항을 구성하기 전에 X1 과 X2 가 모두 중심화 및 척도화됩니다. 구성되는 교호작용 항은 다음과 같이 계산됩니다.

$$\left(\frac{X1 - \text{mean}(X1)}{\text{std}(X1)}\right) \times \left(\frac{X2 - \text{mean}(X2)}{\text{std}(X2)}\right)$$

그런 다음 모든 모형 효과가 모형에 포함되기 전에 **중심화** 및 **척도화** 옵션 선택에 따라 중심화되거나 척도화됩니다.

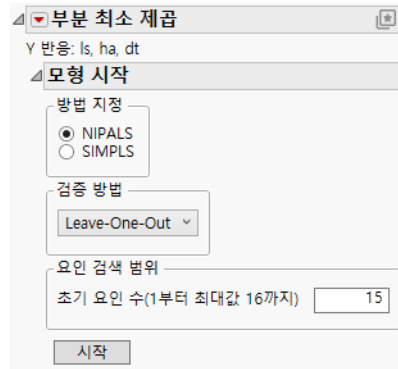
"X 표준화" 옵션을 선택하지 않고 **중심화** 및 **척도화** 옵션을 모두 선택하면 모형에 입력되는 항이 다음과 같이 계산됩니다.

$$\frac{X1 \times X2 - \text{mean}(X1 \times X2)}{\text{std}(X1 \times X2)}$$

## 모형 시작 제어판

부분 최소 제곱 시작 창에서 " 확인 " 을 클릭하거나 " 모형 적합 " 시작 창에서 " 실행 " 을 클릭하면 " 모형 시작 " 제어판이 나타납니다.

**그림 6.8** 부분 최소 제곱 모형 시작 제어판



**참고 :** JMP Pro에서는 " 모형 시작 " 제어판의 " 검증 방법 " 부분이 다르게 나타납니다.

" 모형 시작 " 제어판에는 다음과 같은 선택 옵션이 포함되어 있습니다.

**방법 지정** 모형 적합 알고리즘 유형을 선택합니다. **NIPALS** 와 **SIMPLS** 라는 두 가지 알고리즘이 있습니다. 반응 변수가 하나뿐인 경우 두 방법은 동일한 계수 추정값을 생성합니다. 두 알고리즘의 차이에 대한 자세한 내용은 " 부분 최소 제곱 플랫폼에 대한 통계 상세 정보 " 에서 확인하십시오.

**검증 방법** 검증 방법을 선택합니다. 검증은 추출할 최적 요인 수를 결정하는 데 사용됩니다. JMP Pro의 경우 플랫폼 시작 창에서 검증 열을 지정하면 이러한 옵션이 나타나지 않습니다.

**JMP PRO 홀드백** 검증 데이터 집합에 대해 지정된 비율의 데이터를 무작위로 선택하고 데이터의 다른 부분을 사용하여 모형을 적합시킵니다. 랜덤 선택은 단순 랜덤 표집을 기반으로 할 때보다 더 균형 있는 훈련 데이터 집합 및 검증 데이터 집합을 생성하기 위해 모형 요인에 대한 층화 표집을 기반으로 합니다.

**JMP PRO K 폴드** 데이터를  $K$  개의 부분집합 또는 폴드로 분할합니다. 그런 다음 각 폴드는 나머지 데이터에 적합된 모형을 검증하는 데 사용되어 총  $K$  개의 모형을 적합시킵니다. 이 방법은 제한된 양의 데이터를 효율적으로 사용하기 때문에 소규모 데이터 집합에 가장 적합합니다.

**Leave-One-Out** leave-one-out 교차 검증을 수행합니다.

**없음** 추출할 요인 수를 선택하기 위해 검증을 사용하지 않습니다. 요인 수는 "요인 검색 범위"에 지정됩니다.

**요인 검색 범위** 검증을 사용하지 않는 경우 추출할 잠재 요인 수를 지정합니다. 검증을 사용 중인 경우에는 최적 요인 수를 선택하기 전에 플랫폼에서 적합시키려고 하는 최대 요인 수입니다. 최대 요인 수는 비결측 행 수와 요인 수 중 더 작은 값입니다. 이 경우 초기 요인 수는 15와 최대 요인 수 중 더 작은 값입니다.

**요인 지정 시작**을 클릭하여 초기 모형을 적합시키면 나타납니다. 새 모형을 적합시키는 데 사용할 요인 수를 지정합니다.

**시작** 지정된 규격으로 모형 적합을 시작합니다.

## 부분 최소 제곱 옵션

"모형 시작" 제어판에서 "시작"을 클릭하기 전에 "부분 최소 제곱"의 빨간색 삼각형 메뉴에서 다음 옵션을 사용할 수 있습니다.

**JMP PRO 난수 시드값 설정**  $K$  폴드 및 홀드백 검증에 사용되는 난수화 과정을 위한 시드값을 설정합니다. 분석을 재현하려는 경우에 유용합니다. 시드를 양수 값으로 설정하고 스크립트를 저장하면 시드값이 스크립트에 자동으로 저장됩니다. 스크립트를 실행하면 항상 동일한 교차 검증 분석이 생성됩니다. "검증 방법"을 "없음"으로 설정하거나 검증 열을 사용하는 경우에는 이 옵션이 나타나지 않습니다.

**JMP PRO SVD** 모형 적합 알고리즘에서 SVD 알고리즘의 빠른 또는 전통적 방법 적용을 선택할 수 있는 하위 메뉴를 표시합니다. 기본은 빠른 SVD입니다.

### 일반 옵션

다음 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

**로컬 데이터 필터** 특정 보고서에서 사용되는 데이터를 필터링할 수 있는 로컬 데이터 필터를 표시하거나 숨깁니다.

**다시 실행** 분석을 반복하거나 다시 시작할 수 있는 옵션이 포함되어 있습니다. 이 기능을 지원하는 플랫폼에서 "자동 재계산" 옵션은 해당하는 보고서 창에서 데이터 테이블에 대한 변경 사항을 즉시 반영합니다.

**플랫폼 환경 설정** 현재 플랫폼 환경 설정을 보거나, 현재 JMP 보고서의 설정과 일치하도록 플랫폼 환경 설정을 업데이트할 수 있는 옵션이 포함되어 있습니다.

**스크립트 저장** 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다.

**그룹별 스크립트 저장** 기준 변수의 모든 수준에 대한 플랫폼 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다. 시작 창에서 기준 변수를 지정한 경우에만 사용할 수 있습니다.

---

**참고:** 이 플랫폼의 추가 옵션은 스크립트를 통해 사용할 수 있습니다. 자세한 내용은 "도움말" 메뉴의 "스크립트 인덱스"에서 여십시오. 또한 "스크립트 인덱스"에서 이 섹션에 설명된 옵션의 스크립트 예제도 찾을 수 있습니다.

---

## 부분 최소 제곱 보고서

"부분 최소 제곱" 보고서의 모양은 검증 방법을 사용하는지 여부에 따라 달라집니다. 검증 열을 지정했거나 "검증 방법" 패널에서 "홀드백"을 선택한 경우 보고서의 모든 모형 적합은 훈련 데이터를 기반으로 합니다. 그렇지 않은 경우 모든 모형 적합은 전체 데이터 집합을 기반으로 합니다.

검증을 사용한 경우 다음 세 가지 보고서가 나타납니다.

- 모형 비교 요약
- 교차 검증 보고서
- <N> 개의 요인을 사용한 NIPALS( 또는 SIMPLS) 적합

CV 방법을 **없음**으로 선택한 경우 다음 두 보고서가 나타납니다.

- 모형 비교 요약
- <N> 개의 요인을 사용한 NIPALS( 또는 SIMPLS) 적합

추가 모형을 적합시키려면 "모형 시작" 패널에서 원하는 요인 수를 지정합니다.

### 모형 비교 요약

"부분 최소 제곱" 보고서의 "모형 비교 요약" 테이블에는 각 적합 모형에 대한 요약 결과가 표시됩니다.

그림 6.9 모형 비교 요약

모형 비교 요약						
방법	SVD	행 수	요인 수	누적 X에 대해 설명되는 변동 백분율	누적 Y에 대해 설명되는 변동 백분율	VIP 수 > 0.8
NIPALS	빠름	16	7	99.995152	98.937438	22
NIPALS	빠름	16	6	99.993471	97.768092	22

이 보고서에는 다음과 같은 요약 정보가 포함됩니다.

**방법** "모형 시작" 제어판에서 지정한 분석 방법을 표시합니다.

**SVD** 지정한 SVD 알고리즘 구현 방법을 표시합니다.

**행 수** 훈련 데이터 집합에 사용된 관측값 수를 표시합니다.

**요인 수** 추출된 요인 수를 표시합니다.

**누적 X에 대해 설명되는 변동 백분율** 모형에 의해 설명되는 X의 변동 백분율을 표시합니다.

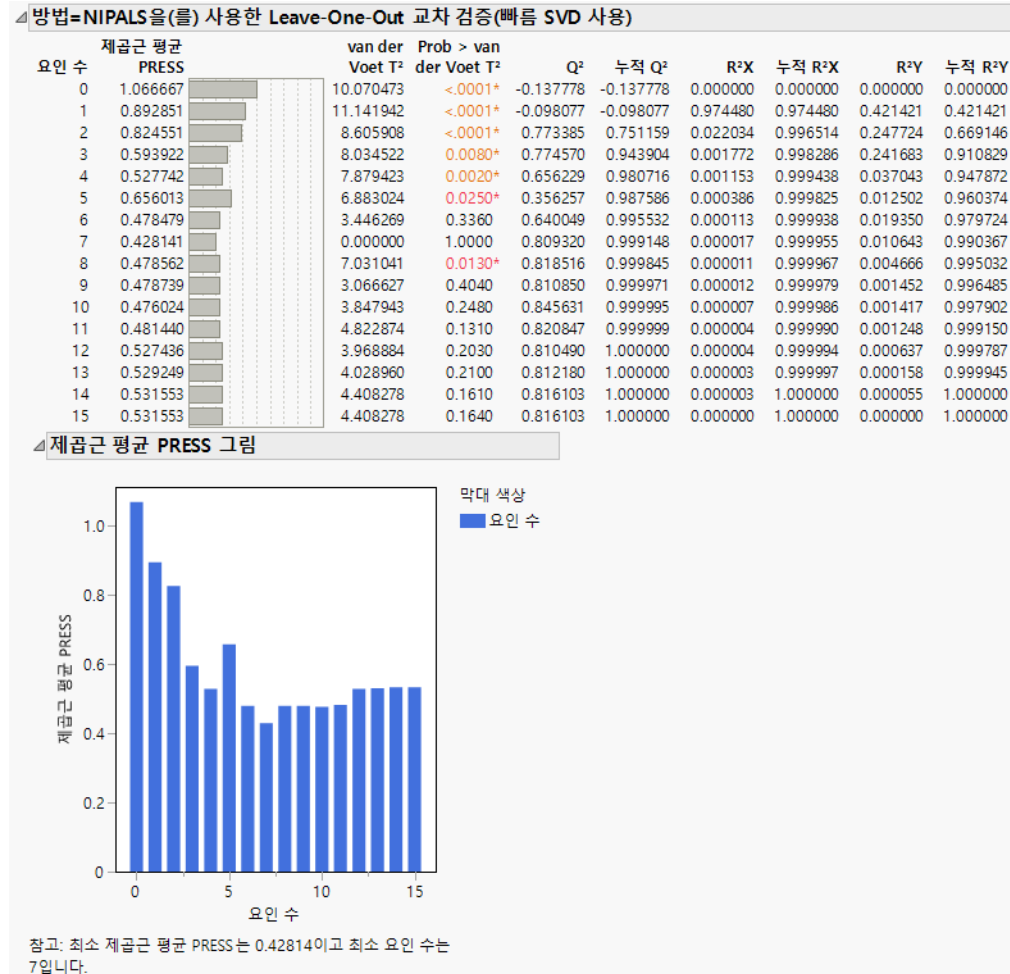
**누적 Y에 대해 설명되는 변동 백분율** 모형에 의해 설명되는 Y의 변동 백분율을 표시합니다.

**VIP 수 > 0.8** VIP(Variable Importance for Projection) 값이 0.8 보다 큰 모형 효과의 수를 표시합니다. VIP 스코어는 X 및 Y 모델링에 대한 변수의 상대적 중요도를 나타내는 척도입니다 (Wold 1994, Eriksson et al. 2006 참조).

## 교차 검증 보고서

이 보고서는 부분 최소 제곱에 대한 "모형 시작" 제어판에서 교차 검증 형식을 검증 방법으로 선택한 경우에만 나타납니다. 보고서 제목은 제어판에서 선택한 교차 검증 및 방법 옵션에 따라 동적으로 "<방법 = 방법 지정>"을 사용한 <교차 검증 방법>"으로 지정됩니다. 여기에는 "모형 시작" 제어판에 지정한 대로 0 ~ 추출된 최대 요인 수를 사용하여 모형 적합에 대한 요약 통계량이 표시됩니다. 또한 제공된 평균 PRESS 값을 보여 주는 그림도 제공됩니다. 자세한 내용은 "[제공된 평균 PRESS 그림](#)"에서 확인하십시오. 최적 요인 수는 최소 제곱근 평균 PRESS 통계량을 사용하여 식별됩니다.

그림 6.10 교차 검증 보고서



**JMP<sup>PRO</sup> X 표준화** 옵션을 선택하면 표준화가 전체 데이터 테이블에 한 번 적용됩니다. 개별 훈련 데이터 집합에 다시 적용되지는 않습니다. 그러나 **중심화** 또는 **척도화** 옵션 조합을 선택하면 이 선택 조합이 각 교차 검증 훈련 데이터 집합에 적용됩니다. 이러한 옵션을 선택한 경우 개별적으로 중심화 및 척도화되는 훈련 데이터 집합을 사용하여 교차 검증이 진행됩니다.

다음 통계량이 보고서에 표시됩니다. 검증 또는 교차 검증이 사용되는 경우 보고된 결과는 훈련 데이터 집합 통계량의 요약입니다.

**요인 수** 모형 적합에 사용된 요인 수입니다.

**제공근 평균 PRESS** 모든 반응에 대한 PRESS 값 평균의 제공근입니다. 자세한 내용은 "**제공근 평균 PRESS**"에서 확인하십시오.

**van der Voet T<sup>2</sup>** 추출된 요인 수가 다른 모형이 최적 모형과 유의하게 다른지 여부를 검정하는 van der Voet 검정의 검정 통계량입니다. 각 van der Voet T<sup>2</sup> 검정에 대한 귀무가설은 해당

요인 수를 기반으로 하는 모형이 최적 모형과 다르지 않다는 것입니다. 대립가설은 모형이 최적 모형과 다르다는 것입니다. 자세한 내용은 "[van der Voet T<sup>2</sup> 검정에 대한 통계 상세 정보](#)"에서 확인하십시오.

**Prob > van der Voet T<sup>2</sup>** van der Voet T<sup>2</sup> 검정의  $p$  값입니다. 자세한 내용은 "[van der Voet T<sup>2</sup> 검정에 대한 통계 상세 정보](#)"에서 확인하십시오.

**Q<sup>2</sup>** 다음과 같이  $Y$ 의 총 제곱합으로 나눈 PRESS 값 비율을 1에서 뺀 값으로 정의되는 무차원 예측 능력 척도입니다.

$$1 - \text{PRESS}/\text{SSY}$$

자세한 내용은 "[Q<sup>2</sup> 계산](#)"에서 확인하십시오.

**누적 Q<sup>2</sup>** 주어진 수 이하의 요인을 사용한 모형의 예측 능력을 나타내는 지표입니다. 주어진 요인 수  $f$ 에 대해 누적 Q<sup>2</sup>은 다음과 같이 정의됩니다.

$$1 - \prod_{i=1}^f (\text{Press}_i/\text{SSY}_i)$$

여기서  $\text{PRESS}_i$ 와  $\text{SSY}_i$ 는  $i$ 개 요인에 대한 값에 해당합니다.

**R<sup>2</sup>X** 지정된 요인에 의해 설명되는  $X$ 변동의 백분율입니다. R<sup>2</sup>X가 큰 성분이  $X$ 변수의 변동을 많이 설명합니다. 자세한 내용은 "[검증이 사용되는 경우 R<sup>2</sup>X 및 R<sup>2</sup>Y 계산](#)"에서 확인하십시오.

**누적 R<sup>2</sup>X** 주어진 요인 수를 사용한 모형에 의해 설명되는  $X$ 변동의 백분율입니다. 이 값은 R<sup>2</sup>X 값의 합계입니다 ( $i=1 \sim$  주어진 요인 수).

**R<sup>2</sup>Y** 지정된 요인에 의해 설명되는  $Y$ 변동의 백분율입니다. R<sup>2</sup>Y가 큰 성분이  $Y$ 변수의 변동을 많이 설명합니다. 자세한 내용은 "[검증이 사용되는 경우 R<sup>2</sup>X 및 R<sup>2</sup>Y 계산](#)"에서 확인하십시오.

**누적 R<sup>2</sup>Y** 주어진 요인 수를 사용한 모형에 의해 설명되는  $Y$ 변동의 백분율입니다. 이 값은 R<sup>2</sup>Y 값의 합계입니다 ( $i=1 \sim$  주어진 요인 수).

## Q<sup>2</sup> 및 누적 R<sup>2</sup>Y 해석

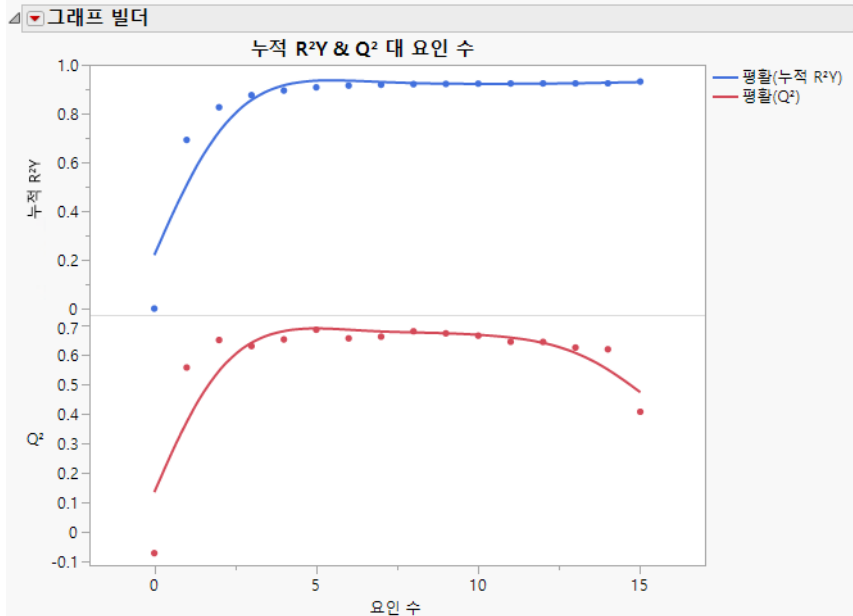
Q<sup>2</sup> 및 누적 R<sup>2</sup>Y 통계량은 둘 다 예측 능력을 측정하지만 방식이 다릅니다.

- 누적 R<sup>2</sup>Y는 요인 수가 증가함에 따라 함께 증가합니다. 모형에 요인이 추가될수록 더 많은 변동이 설명되기 때문입니다.
- Q<sup>2</sup>은 요인 수가 증가함에 따라 증가했다가 감소하거나, 적어도 증가를 중단하는 경향이 있습니다. 요인이 추가될수록 모형이 훈련 데이터 집합에 맞게 조정되고 새 데이터에 대해 잘 일반화되지 않아 PRESS 통계량이 감소하기 때문입니다.

Q<sup>2</sup> 및 누적 R<sup>2</sup>Y 분석은 모형에 포함할 요인 수를 결정하기 위해 van der Voet 검정 대신 사용할 수 있습니다. Q<sup>2</sup>이 크고 감소가 시작되지 않은 요인 수를 선택합니다. 누적 R<sup>2</sup>Y도 커야 합니다.

[그림 6.11](#)에서는 Leave-One-Out 검증 방법을 사용하는 Penta.jmp 데이터 테이블의 요인 수에 대한 누적 R<sup>2</sup>Y 및 Q<sup>2</sup> 그림을 보여 줍니다. 누적 R<sup>2</sup>Y는 약 네 개 요인에 대해 증가 후 안정됩니다. Q<sup>2</sup> 통계량은 두 요인에 대해 가장 크고 그 이후에 안정되기 시작합니다. 이 그림은 요인이 두 개인 모형이 데이터를 과대적합하지 않고  $Y$ 변동의 많은 부분을 설명한다는 것을 나타냅니다.

그림 6.11 Penta.jmp 에 대한 누적  $R^2Y$  및  $Q^2$



### 제공된 평균 PRESS 그림

이 막대 차트에는 요인 수가 가로 축에 표시되고 제공된 평균 PRESS 값이 세로 축에 표시됩니다. 이는 교차 검증 보고서의 "제공된 평균 PRESS" 열 오른쪽에 나타나는 가로 막대 차트와 동일합니다(그림 6.10 참조).

### 제공된 평균 PRESS

지정된 요인 수  $a$  에 대한 제공된 평균 PRESS 는 다음 단계를 사용하여 계산됩니다.

1. 요인이  $a$  개인 모형을 각 훈련 데이터 집합에 적합시킵니다.
2. 결과 예측 계산식을 검증 데이터 집합의 관측값에 적용합니다.
3. 각  $Y$  에 대해 다음을 수행합니다.
  - 각 검증 데이터 집합에 대해 관측된 각 검증 데이터 집합 값과 예측값 간의 차이 제곱을 계산합니다 (제곱 예측 오차).
  - 각 검증 데이터 집합에 대해 이러한 차이 제곱의 평균을 구하고 결과를 반응에 대한 분산 추정값으로 나눕니다.  $K$  폴드 및 Leave-One-Out 검증 방법의 경우 전체 반응 열의 분산으로 나눕니다. 홀드백 검증의 경우 훈련 데이터 집합에 있는 반응 값의 분산으로 나눕니다.
  - 이러한 평균을 합산하고 검증 데이터 집합이 두 개 이상인 경우 합계를 검증 데이터 집합 수 - 1로 나눕니다. 이 값이 주어진  $Y$  에 대한 PRESS 통계량입니다.
4.  $a$  개 요인의 제공된 평균 PRESS 는 모든 반응에 대한 PRESS 값 평균의 제공된 값입니다.

5. 여러  $Y$ 의 PRESS 통계량은 모든 반응에 대해 3 단계에서 구한 PRESS 통계량의 평균을 계산한 값입니다.

## Q<sup>2</sup> 계산

Q<sup>2</sup> 통계량은  $1 - PRESS/SSY$ 로 정의됩니다. PRESS 통계량은 훈련 데이터 집합을 기반으로 개발되었지만 검증 데이터 집합에서 평가된 모형의 모든 반응에 대해 평균을 구한 예측 오차 제곱합입니다. SSY 값은 모든 반응에 대해 평균을 구하고 검증 데이터 집합의 관측값을 기반으로 하는  $Y$ 의 제곱합입니다.

교차 검증 보고서의 Q<sup>2</sup> 통계량은 선택한 검증 방법에 따라 다음과 같은 방식으로 계산됩니다.

**Leave-One-Out** Q<sup>2</sup>은 1 - 평균 (PRESS) / 평균 (SSY) 값입니다. 각  $Y$ 에 대한 PRESS는 관측값을 한 번에 하나씩 제외하여 생성된 모형에서 계산됩니다. SSY는 각  $Y$ 열의 모든 값에서 계산됩니다.

**K 폴드** Q<sup>2</sup>은 각  $K$  폴드를 제외하여 생성된  $K$  개 모형을 기반으로 한 검증 데이터 집합에 대해 계산되는  $1 - PRESS/SSY$  값의 평균입니다.

**홀드백 또는 검증 데이터 집합** Q<sup>2</sup>은 단일 훈련 데이터 집합을 사용하여 생성된 모형을 기반으로 한 검증 데이터 집합에 대해 계산되는  $1 - PRESS/SSY$  값입니다.

## 검증이 사용되는 경우 R<sup>2</sup>X 및 R<sup>2</sup>Y 계산

교차 검증 보고서의 R<sup>2</sup>X 및 R<sup>2</sup>Y 통계량은 선택한 검증 방법에 따라 다음과 같은 방식으로 계산됩니다.

---

**참고 :** 이러한 모든 계산에서 R<sup>2</sup>Y는 유사하게 계산됩니다.

---

**Leave-One-Out** R<sup>2</sup>X는 관측값을 한 번에 하나씩 제외하여 생성된 모형에 대한 "X 효과에 대해 설명되는 변동 백분율" 값의 평균입니다.

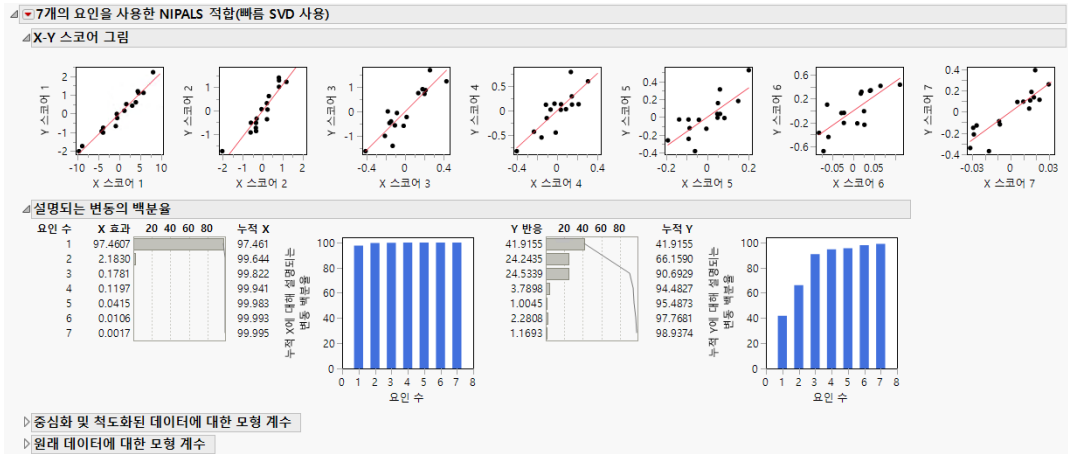
**K 폴드** R<sup>2</sup>X는 각 폴드를 제외하여 생성된  $K$  개 모형에 대한 "X 효과에 대해 설명되는 변동 백분율" 값의 평균입니다.

**홀드백 또는 검증 데이터 집합** R<sup>2</sup>X는 훈련 데이터를 사용하여 생성된 모형에 대한 "X 효과에 대해 설명되는 변동 백분율" 값입니다.

## 모형 적합 보고서

부분 최소 제곱 플랫폼의 모형 적합 보고서에는 각 적합 모형에 대한 상세 결과가 표시됩니다. 적합에는 교차 검증에 기반한 최적 요인 수가 사용되거나, 교차 검증 방법이 지정되지 않은 경우 지정된 요인 수가 사용됩니다. 보고서 제목은 NIPALS 또는 SIMPLS 중 사용된 방법과 빠른 SVD 또는 전통적 SVD 중 사용된 방법을 나타내고 추출된 요인 수를 제공합니다.

그림 6.12 모형 적합 보고서



모형 적합 보고서에는 다음과 같은 요약 정보가 포함됩니다.

**X-Y 스코어 그림** 추출된 각 요인에 대한 X 및 Y 스코어의 산점도입니다.

**설명되는 변동의 백분율** X와 Y 둘 다에 대해 설명되는 변동 백분율 및 누적 변동 백분율을 표시합니다. 추출된 각 요인에 대한 결과가 제공됩니다.

**중심화 및 척도화된 데이터에 대한 모형 계수** 각 Y를 대상으로 중심화 및 척도화된 데이터를 기반으로 하는 모형에 대한 X의 계수를 표시합니다.

## 모형 적합 옵션

부분 최소 제곱 플랫폼에서 "모형 적합"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

**변동 백분율 그림** "X 효과에 대해 설명되는 변동 백분율"과 "Y 효과에 대해 설명되는 변동 백분율"이라는 두 개의 그림을 추가합니다. 여기에는 X 및 Y에 대해 추출된 각 요인에 의해 설명되는 변동 백분율을 나타내는 누적 막대 차트가 표시됩니다.

**변수 중요도 그림** 각 X 변수의 VIP(Variable Importance for Projection) 값을 그림에 표시합니다. VIP 스코어가 "변수 중요도 테이블"에 나타납니다. 자세한 내용은 "변수 중요도 그림"에서 확인하십시오.

**VIP 대 계수 그림** 모형 계수에 대한 VIP(Variable Importance for Projection) 통계량을 그림에 표시합니다. 선택한 Y에 해당하는 점만 표시할 수 있습니다. 추가 라벨 옵션이 제공됩니다. 중심화 및 척도화된 데이터에 대한 그림과 원래 데이터에 대한 그림이 둘 다 있습니다. 자세한 내용은 "VIP 대 계수 그림"에서 확인하십시오.

**VIP 임계 설정** 변수 중요도 그림, 변수 중요도 테이블 및 VIP 대 계수 그림에 사용할 임계 수준을 설정합니다.

**계수 그림** 전체 X 변수에서 각 반응에 대한 모형 계수를 그림에 표시합니다. 선택한 Y에 해당하는 점만 표시할 수 있습니다. 중심화 및 척도화된 데이터에 대한 그림과 원래 데이터에 대한 그림이 둘 다 있습니다.

**적재 그림** 추출된 각 요인에 대한 X 및 Y 적재를 그림에 표시합니다. X와 Y에 대해 각각 별도의 그림이 있습니다.

**적재 산점도 행렬** X 적재 및 Y 적재의 산점도 행렬을 표시합니다.

**상관 적재 그림** 동일한 그림에 중첩된 X 및 Y 적재의 단일 산점도 또는 산점도 행렬을 표시합니다. 이 옵션을 선택할 때 그림에 표시할 요인 수를 지정합니다.

- 두 개의 요인을 지정하면 단일 상관 적재 산점도가 나타납니다. 그림 아래에서 축을 정의하는 두 요인을 선택합니다. 오른쪽 화살표 버튼을 클릭하면 각 요인 조합이 연속해서 그림에 표시됩니다.

- 세 개 이상의 요인을 지정하면 선택한 개수까지의 요인 쌍에 대한 셀이 포함된 산점도 행렬이 나타납니다.

두 경우 모두 체크박스를 사용하여 라벨 지정을 제어합니다.

**X-Y 스코어 그림** 다음 옵션을 포함합니다.

**선형 적합** "X-Y 스코어 그림"에 점을 통과하는 적합선을 표시하거나 숨깁니다.

**신뢰 대역 표시** "X-Y 스코어 그림"에 적합선에 대한 95% 신뢰 대역을 표시하거나 숨깁니다.

**스코어 산점도 행렬** X 스코어의 산점도 행렬과 Y 스코어의 산점도 행렬을 표시합니다. 각 X 스코어 산점도에는 이상치 감지에 사용할 수 있는 95% 신뢰도 타원이 표시됩니다. 신뢰도 타원에 대한 통계 상세 정보는 "X 스코어 산점도 행렬의 신뢰도 타원에 대한 통계 상세 정보"에서 확인하십시오.

**거리 그림** 다음을 나타내는 그림을 표시합니다.

- 각 관측값에서 X 모형까지의 거리
- 각 관측값에서 Y 모형까지의 거리
- X 모형까지의 거리와 Y 모형까지의 거리를 둘 다 보여 주는 산점도

적합한 모형에서는 X 거리와 Y 거리가 모두 작으므로 점이 원점 (0,0)에 가깝습니다. 그림을 사용하여 X 또는 Y에 대한 이상치를 찾을 수 있습니다. 점 그룹이 함께 군집화되면 공통 특징을 가질 수 있으며 개별적으로 분석할 수 있습니다. 검증 데이터 집합 또는 검증 및 테스트 데이터 집합이 사용 중인 경우 이러한 데이터 집합과 훈련 데이터 집합에 대해 별도의 보고서가 제공됩니다.

**T 제곱 그림** 각 관측값에 대한  $T^2$  통계량 그림과 관리 한계를 함께 표시합니다. 관측값의  $T^2$  통계량은 추출된 요인에 대한 해당 관측값의 스코어를 기반으로 계산됩니다. 관리 한계 및  $T^2$  계산에 대한 자세한 내용은 " $T^2$  그림에 대한 통계 상세 정보"에서 확인하십시오.

**진단 그림** 모형 적합을 평가하기 위한 진단 그림을 표시합니다. "실제값 대 예측값 그림", "잔차 대 예측값 그림", "잔차 대 행 번호 그림" 및 "잔차 정규 분위수 그림"이라는 네 가지 유형의 그림이 있습니다. 각 반응에 대한 그림이 제공됩니다. 검증 데이터 집합 또는 검증 및

테스트 데이터 집합이 사용 중인 경우 이러한 데이터 집합과 훈련 데이터 집합에 대해 별도의 보고서가 제공됩니다.

**프로파일러** 각 Y 변수에 대한 프로파일러를 표시합니다.

**스펙트럼 프로파일러** 모든 반응 변수가 그림의 첫 번째 셀에 나타나는 단일 프로파일러를 표시합니다. 이 프로파일러는 X 변수의 변화가 Y 변수에 미치는 영향을 동시에 시각화하는 데 유용합니다.

**열 저장** 다양한 계산식 및 결과를 저장하기 위한 옵션이 포함되어 있습니다.

**예측 계산식 저장** 새 계산식 열을 원래 데이터 테이블에 저장합니다. 각 Y 변수마다 X 변수의 함수 형태인 예측 계산식을 포함하는 예측 계산식 <반응> 열이 있습니다.

**예측을 X 스코어 계산식으로 저장** 새 계산식 열을 원래 데이터 테이블에 저장합니다. 각 Y 변수마다 X 스코어 계산식의 함수 형태인 예측 계산식을 포함하는 예측 계산식 <반응> 열이 있습니다. 각 X 스코어에 대한 계산식 열도 데이터 테이블에 저장됩니다.

**예측 계산식 표준 오차 저장** 새 계산식 열을 원래 데이터 테이블에 저장합니다. 각 Y 변수마다 X 변수의 함수 형태인 예측 평균 표준 오차 계산식을 포함하는 예측값 표준 오차 <반응> 열이 있습니다. 자세한 내용은 "예측 및 신뢰 한계에 대한 통계 상세 정보"에서 확인하십시오.

**평균 신뢰 한계 계산식 저장** 새 계산식 열을 원래 데이터 테이블에 저장합니다. 각 Y 변수마다 X 스코어 계산식의 함수 형태인 평균 반응 신뢰 하한 및 상한 열이 있습니다. 새 열 이름은 평균 <반응> 의 95% 하한과 평균 <반응> 의 95% 상한입니다. 이 열에는 반응 평균에 대한 95% 신뢰 한계가 포함됩니다. 자세한 내용은 "예측 및 신뢰 한계에 대한 통계 상세 정보"에서 확인하십시오.

**개별값 신뢰 한계 계산식 저장** 새 계산식 열을 원래 데이터 테이블에 저장합니다. 각 Y 변수마다 X 스코어 계산식의 함수 형태인 개별 예측 신뢰 하한 및 상한 열이 있습니다. 새 열 이름은 개별 <반응> 의 95% 하한과 개별 <반응> 의 95% 상한입니다. 이 열에는 개별값에 대한 95% 예측 한계가 포함됩니다. 자세한 내용은 "예측 및 신뢰 한계에 대한 통계 상세 정보"에서 확인하십시오.

**스코어 계산식 저장** 새 계산식 열을 원래 데이터 테이블에 저장합니다. 추출된 각 요인마다 X 스코어 계산식이 포함된 X 스코어 <N> 계산식 열과 Y 스코어 계산식이 포함된 Y 스코어 <N> 계산식 열이 있습니다. X 스코어 계산식은 X 변수의 함수이고 Y 스코어 계산식은 X 스코어 계산식의 함수입니다. 각 X 스코어 계산식 열에는 MDMCC 열 특성이 있어 MDMCC(모형 기반 다변량 관리도) 플랫폼에서 해당 열을 사용할 수 있습니다. 계산식은 "부분 최소 제곱에 대한 통계 상세 정보"에서 확인하십시오.

**Y 예측값 저장** 새 열을 원래 데이터 테이블에 저장합니다. 각 Y 변수마다 예측 Y 값을 포함하는 열이 있습니다.

**Y 잔차 저장** 새 열을 원래 데이터 테이블에 저장합니다. 각 Y 변수마다 Y 잔차 값을 포함하는 열이 있습니다.

- X 예측값 저장** 새 열을 원래 데이터 테이블에 저장합니다. 각 X 변수마다 예측 X 값을 포함하는 열이 있습니다.
- X 잔차 저장** 새 열을 원래 데이터 테이블에 저장합니다. 각 X 변수마다 X 잔차 값을 포함하는 열이 있습니다.
- X 효과에 대해 설명되는 변동 백분율 저장** 열을 새 데이터 테이블에 저장합니다. 각 X 변수마다 추출된 모든 요인에 대해 설명되는 변동률을 포함하는 열이 있습니다.
- Y 효과에 대해 설명되는 변동 백분율 저장** 열을 새 데이터 테이블에 저장합니다. 각 Y 변수마다 추출된 모든 요인에 대해 설명되는 변동률을 포함하는 열이 있습니다.
- 스코어 저장** 새 열을 원래 데이터 테이블에 저장합니다. 추출된 각 요인마다 X 스코어를 포함하는 열과 Y 스코어를 포함하는 열이 있습니다.
- 적재 저장** 열을 두 개의 새 데이터 테이블에 저장합니다. X 변수에 대한 적재를 포함하는 데이터 테이블과 Y 변수에 대한 적재를 포함하는 데이터 테이블이 있습니다.
- 표준화된 스코어 저장** 새 열을 원래 데이터 테이블에 저장합니다. 새 열에는 추출된 각 요인에 대한 표준화된 X 및 Y 스코어가 포함됩니다. 계산식은 "[표준화된 스코어 및 적재량에 대한 통계 상세 정보](#)"에서 확인하십시오.
- 표준화된 적재량 저장** 열을 두 개의 새 데이터 테이블에 저장합니다. X 변수에 대한 표준화된 적재량을 포함하는 데이터 테이블과 Y 변수에 대한 표준화된 적재량을 포함하는 데이터 테이블이 있습니다. 계산식은 "[표준화된 스코어 및 적재량에 대한 통계 상세 정보](#)"에서 확인하십시오.
- T 제곱 저장** 새 계산식 열을 원래 데이터 테이블에 저장합니다. 새 열에는  $T^2$  계산식이 X 변수의 함수 형태로 포함됩니다. 이 열의 값은 T 제곱 그림에도 사용됩니다.
- T 제곱을 X 스코어 계산식으로 저장** 새 계산식 열을 원래 데이터 테이블에 저장합니다. 새 열에는  $T^2$  계산식이 X 스코어 계산식의 함수 형태로 포함됩니다.
- 거리 저장** 새 열을 원래 데이터 테이블에 저장합니다. 새 열에는 X 모형까지의 거리(DModX) 및 Y 모형까지의 거리(DModY) 값이 포함됩니다. 이러한 값은 거리 그림에 사용됩니다.
- 거리를 X 스코어 계산식으로 저장** 새 계산식 열을 원래 데이터 테이블에 저장합니다. 새 열에는 X 모형까지의 거리(DModX) 및 Y 모형까지의 거리(DModY) 계산식이 X 스코어 계산식의 함수 형태로 포함됩니다.
- X 가중치 저장** 열을 새 데이터 테이블에 저장합니다. 추출된 각 요인에 대해 X 변수의 가중치를 포함하는 열이 있습니다.
- JMP PRO 검증 저장** 새 열을 원래 데이터 테이블에 저장합니다. 새 열에는 각 관측값이 검증에 사용된 방법을 나타내는 숫자가 포함됩니다. 홀드백 검증의 경우 행이 혼란 또는 검증에 사용되었는지 여부를 식별합니다. K 폴드 검증의 경우 행이 할당된 부분군의 번호를 식별합니다.

**JMP PRO** **결측값 대치 저장** 열을 새 데이터 테이블에 저장합니다. 각 X 및 Y 변수에 대해 결측값이 대치 값으로 대체된 원래 데이터 열을 포함하는 열이 있습니다. 검증 열을 지정한 경우 검증 열도 포함됩니다.

**JMP PRO** **예측 계산식 게시** 예측 계산식을 생성하여 계산식 저장소 플랫폼에 계산식 열 스크립트로 저장합니다. "계산식 저장소" 보고서가 열려 있지 않으면 "계산식 저장소" 보고서가 생성됩니다. 자세한 내용은 **예측 및 전문 모델링**에서 확인하십시오.

**JMP PRO** **스코어 계산식 게시** X 및 Y 스코어 계산식을 생성하여 계산식 저장소 플랫폼에 계산식 열 스크립트로 저장합니다. "계산식 저장소" 보고서가 열려 있지 않으면 "계산식 저장소" 보고서가 생성됩니다. 자세한 내용은 **예측 및 전문 모델링**에서 확인하십시오.

**적합 제거** 주 플랫폼 보고서에서 모형 보고서를 제거합니다.

**VIP 를 사용하여 모형 생성** 시작 창을 연 후, 적절한 반응을 Y 로 입력하고 VIP(Variable Importance for Projection) 값이 지정된 임계를 초과하는 변수를 X 로 입력하여 창을 채웁니다. "중심화 및 척도화된 데이터에 대한 VIP 대 계수" 보고서의 버튼과 동일한 기능을 수행합니다. 자세한 내용은 "**VIP 대 계수 그림**"에서 확인하십시오.

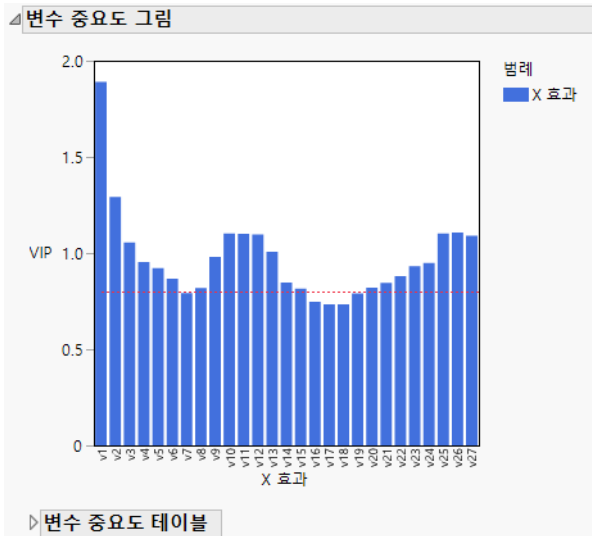
**저장된 X 스코어에 대한 모형 기반 다변량 관리도** ( 범주형 변수 또는 교호작용 효과를 포함하지 않는 모형에만 사용 가능) 각 X 스코어에 대한 계산식을 저장하고 MDMCC( 모형 기반 다변량 관리도) 시작 창을 실행합니다. MDMCC 시작 창에서 스코어 계산식이 공정 열로 할당됩니다. 공정을 추가 또는 제거하고, 시간 ID 를 추가하거나, 과거 데이터 끝 위치를 설정한 후 "확인" 을 클릭할 수 있습니다. 자세한 내용은 **품질 및 공정 방법**에서 확인하십시오.

**예측값 프로파일러** "프로파일러" 시작 창을 실행합니다. "프로파일러" 시작 창에서 예측 계산식은 지정된 모형 적합의 각 변수에 대한 예측 계산식입니다. "확인" 을 클릭하기 전에 잡음 요인 또는 다른 예측 계산식을 추가할 수 있습니다. 자세한 내용은 **프로파일러**에서 확인하십시오.

## 변수 중요도 그림

부분 최소 제곱 플랫폼의 "변수 중요도 그림" 옵션은 각 X 변수에 대한 VIP(Variable Importance for Projection) 값 그래프를 표시합니다. "변수 중요도 테이블"에는 VIP 스코어가 표시됩니다. VIP 스코어는 X와 Y를 모델링할 때 변수의 중요도를 나타내는 척도입니다. 변수의 계수 및 VIP 가 작으면 해당 변수는 모형에서 삭제할 후보가 됩니다(Wold 1994 참조). 값이 0.8이면 일반적으로 VIP가 작은 것으로 간주되며(Eriksson et al. 2006 참조) 그림의 0.8 위치에 빨간색 파선이 표시됩니다.

그림 6.13 변수 중요도 그림



### VIP 대 계수 그림

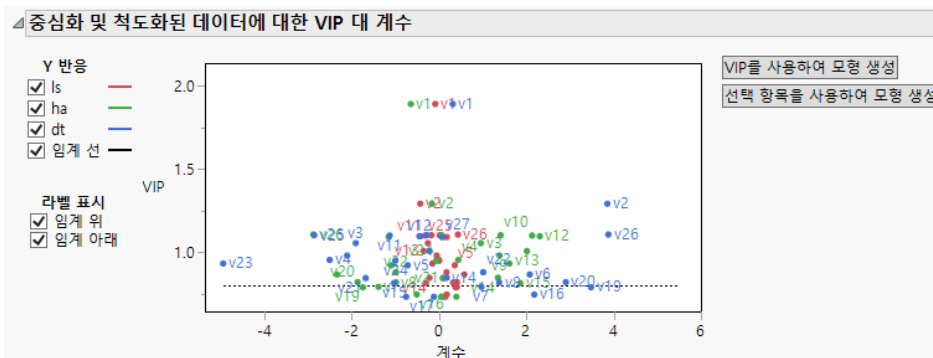
부분 최소 제곱 플랫폼의 "VIP 대 계수 그림" 옵션은 모형 계수에 대한 VIP(Variable Importance for Projection) 통계량 그래프를 표시합니다. 그림 오른쪽에는 변수 축소 및 모형 생성을 쉽게 할 수 있는 두 가지 옵션이 있습니다.

**VIP를 사용하여 모형 생성** 시작 창을 연 후, 적절한 반응을 Y로 입력하고 VIP 값이 특정 임계를 초과하는 변수를 X로 입력하여 창을 채웁니다.

**선택 항목을 사용하여 모형 생성** 그림에서 직접 X를 선택할 수 있으며 그러면 Y와 선택한 X만 시작 창에 입력됩니다.

현재 열 선택을 기반으로 다른 플랫폼을 사용하려면 원하는 플랫폼을 엽니다. 선택 항목이 시작 창에 그대로 유지됩니다. 역할 버튼을 클릭하면 선택한 열이 채워집니다.

그림 6.14 중심화 및 척도화된 데이터에 대한 VIP 대 계수 그림

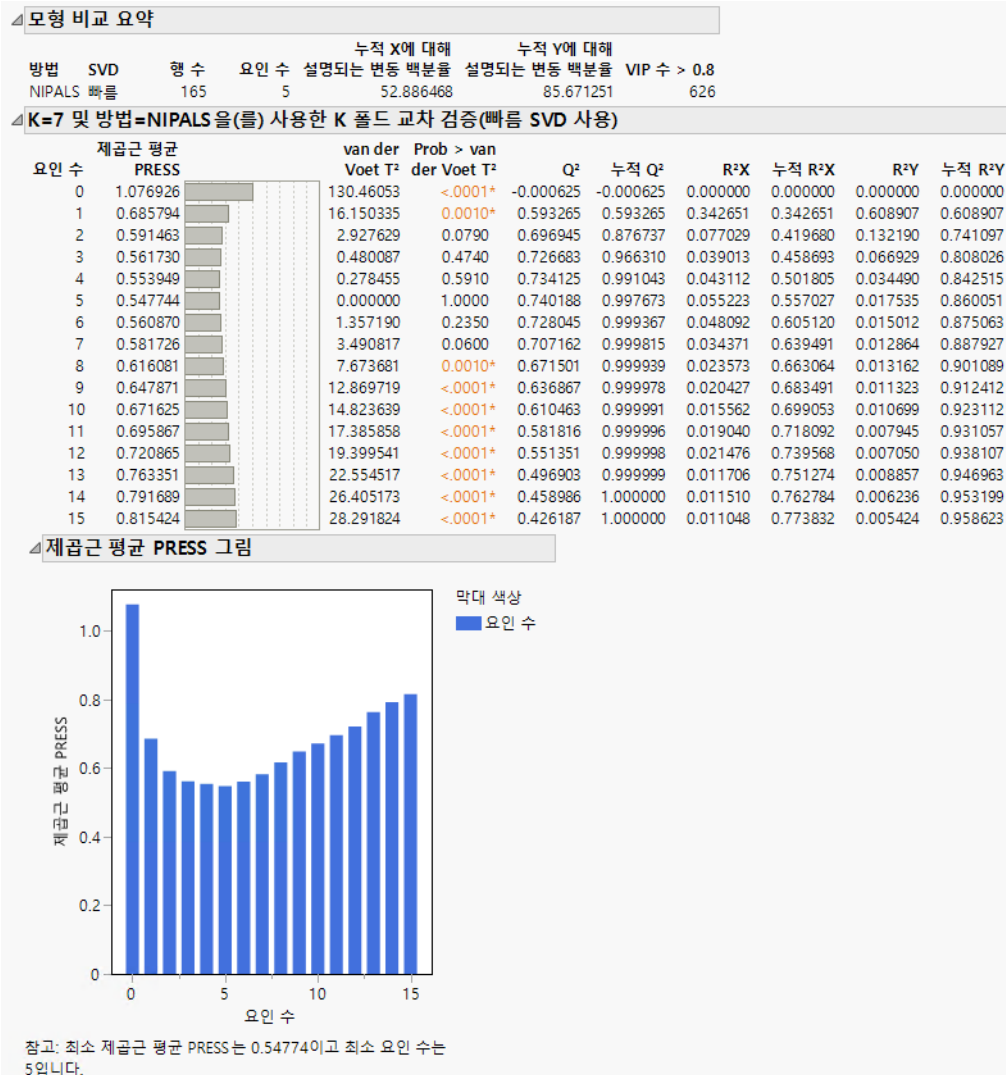


## JMP<sup>PRO</sup> 부분 최소 제곱의 추가 예

이 예에서는 모형 적합의 부분 최소 제곱 분석법을 사용하여 범주형 반응과 다수의 변수에 대한 분석을 수행합니다. 예제 데이터 집합은 165명의 남성에게서 수집한 혈청 표본으로 구성되어 있으며, 그 중 84명은 전립선암을 앓고 있고 81명은 그렇지 않습니다. 분석 목표는 혈청 표본을 기반으로 전립선암에 걸린 남성을 판별하는 것입니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Prostate Cancer.jmp 를 엽니다 .
2. **분석 > 모형 적합**을 선택합니다 .
3. **Status** 를 선택하고 **Y** 를 클릭합니다 .
4. **Proteins** 열 그룹을 선택하고 **추가**를 클릭합니다 .
5. " 분석법 " 목록에서 **부분 최소 제곱**을 선택합니다 .
6. **실행**을 클릭합니다 .
7. ( 선택 사항 )" 부분 최소 제곱 "의 빨간색 삼각형을 클릭하고 **난수 시드값 설정**을 선택합니다 .
8. ( 선택 사항 )" 난수 시드값 지정 " 옆에 "1234" 를 입력합니다 .  
난수 시드값을 지정하면 이 예에 표시된 결과를 재현할 수 있습니다 .
9. ( 선택 사항 ) **확인**을 클릭합니다 .
10. " 모형 시작 "에서 **시작**을 클릭합니다 .

그림 6.15 부분 최소 제곱 보고서



" 제공된 평균 PRESS 그림 "에서는 요인 수가5 일 때 제공된 평균 PRESS가 최소화된다는 것을 보여 줍니다 . 이 내용은 " 제공된 평균 PRESS 그림 "아래의 "참고"에 명시되어 있습니다 . 5개의 요인을 사용한 NIPALS 적합 ( 빠른 SVD 사용 ) 보고서가 생성되었습니다 .

## 부분 최소 제곱 플랫폼에 대한 통계 상세 정보

이 섹션에는 부분 최소 제곱 플랫폼에 사용되는 몇 가지 방법에 대한 통계 상세 정보가 포함되어 있습니다. 자세한 내용은 Hoskuldsson(1988), Garthwaite(1994) 또는 Cox와 Gaudard(2013) 연구 자료에서 확인하십시오.

- "부분 최소 제곱에 대한 통계 상세 정보"
- "van der Voet T<sup>2</sup> 검정에 대한 통계 상세 정보"
- "T<sup>2</sup> 그림에 대한 통계 상세 정보"
- "X 스코어 산점도 행렬의 신뢰도 타원에 대한 통계 상세 정보"
- "예측 및 신뢰 한계에 대한 통계 상세 정보"
- "표준화된 스코어 및 적재량에 대한 통계 상세 정보"
- "PLS 판별 분석에 대한 통계 상세 정보"

### 부분 최소 제곱에 대한 통계 상세 정보

부분 최소 제곱은 요인, 즉 설명 변수(X)의 선형 결합을 기반으로 선형 모형을 적합시킵니다. 이러한 요인은 X와 하나 이상의 반응(Y) 간의 공분산을 최대화하는 방식으로 얻어집니다. 이런 식으로 PLS는 X와 Y 사이의 상관을 이용하여 기본 잠재 구조를 나타냅니다. 요인은 반응 변동과 예측 변수 변동을 설명해야 하는 합동 목표를 해결합니다. 부분 최소 제곱은 X 변수가 관측값보다 많거나 X 변수의 상관관계가 높을 때 특히 유용합니다.

### NIPALS

NIPALS(Nonlinear Iterative Partial Least Squares) 방법은 한 번에 하나의 요인을 추출하는 방식으로 작동합니다.  $\mathbf{X} = \mathbf{X}_0$  은 중심화 및 척도화된 예측 변수 행렬이고,  $\mathbf{Y} = \mathbf{Y}_0$  은 중심화 및 척도화된 반응 값 행렬이라고 하겠습니다. PLS 방법은 예측 변수의 선형 결합인  $\mathbf{t} = \mathbf{X}_0\mathbf{w}$  로 시작합니다. 여기서  $\mathbf{t}$  는 스코어 벡터이고  $\mathbf{w}$  는 연관된 가중치 벡터입니다. PLS 방법은  $\mathbf{t}$  에 대한 회귀 분석을 통해  $\mathbf{X}_0$  과  $\mathbf{Y}_0$  을 모두 예측합니다.

$$\hat{\mathbf{X}}_0 = \mathbf{t}\mathbf{p}', \text{ 여기서 } \mathbf{p}' = (\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{X}_0$$

$$\hat{\mathbf{Y}}_0 = \mathbf{t}\mathbf{c}', \text{ 여기서 } \mathbf{c}' = (\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{Y}_0$$

벡터  $\mathbf{p}$  와  $\mathbf{c}$  를 각각 X 및 Y 적재라고 합니다.

특정 선형 결합  $\mathbf{t} = \mathbf{X}_0\mathbf{w}$  는 일부 반응 선형 결합  $\mathbf{u} = \mathbf{Y}_0\mathbf{q}$  에서 최대 공분산  $\mathbf{t}'\mathbf{u}$  를 갖는 결합입니다. 또 다른 특성은 X 및 Y 가중치를 나타내는  $\mathbf{w}$  와  $\mathbf{q}$  가 공분산 행렬  $\mathbf{X}_0'\mathbf{Y}_0$  의 첫 번째 왼쪽 및 오른쪽 특이 벡터에 비례한다는 것입니다. 이는  $\mathbf{X}_0'\mathbf{Y}_0\mathbf{Y}_0'\mathbf{X}_0$  과  $\mathbf{Y}_0'\mathbf{X}_0\mathbf{X}_0'\mathbf{Y}_0$  의 첫 번째 고유 벡터에 각각 해당합니다.

이것은 첫 번째 PLS 요인이 추출되는 방법을 설명합니다. 두 번째 요인은  $\mathbf{X}_0$  과  $\mathbf{Y}_0$  을 첫 번째 요인의  $X$  및  $Y$  잔차로 대체하여 동일한 방식으로 추출됩니다.

$$\mathbf{X}_1 = \mathbf{X}_0 - \hat{\mathbf{X}}_0$$

$$\mathbf{Y}_1 = \mathbf{Y}_0 - \hat{\mathbf{Y}}_0$$

이러한 잔차를 수축된  $X$  및  $Y$  블록이라고도 합니다. 스코어 벡터 추출 및 데이터 행렬 수축 과정이 추출된 요인 수만큼 반복됩니다.

## SIMPLS

SIMPLS(Statistically Inspired Modification of Partial Least Squares) 알고리즘은 통계 기준을 최적화하기 위해 개발되었습니다. 이 방법은  $X$  스코어가 직교해야 한다는 요건에 따라  $X$ 와  $Y$ 의 선형 결합 간에 공분산을 최대화하는 스코어 벡터를 찾습니다.  $\mathbf{X}_0$  및  $\mathbf{Y}_0$  행렬이 수축되는 NIPALS와 달리 SIMPLS는 교차곱 행렬  $\mathbf{X}_0' \mathbf{Y}_0$ 을 수축시킵니다.

단일  $Y$  변수의 경우 이 두 알고리즘이 동일합니다. 그러나 다변량  $Y$ 의 경우 모형이 다릅니다. SIMPLS는 De Jong(1993)에 의해 제안되었습니다.

## van der Voet $T^2$ 검정에 대한 통계 상세 정보

부분 최소 제곱 플랫폼의 van der Voet  $T^2$  검정은 지정된 수만큼 추출된 요인이 있는 모형이 제안된 최적 모형과 유의하게 다른지 여부를 판별하는 데 도움이 됩니다. 이 검정은 두 모형의 잔차 제곱이 동일한 분포를 따른다는 귀무가설에 기반한 난수화 검정입니다. 귀무가설은 두 모형의 예측 능력이 동일하다는 것을 나타낸다고 직관적으로 생각할 수 있습니다.

교차 검증 보고서에 제공된 van der Voet  $T^2$  통계량을 얻기 위해 각 검증 데이터 집합에 대해 아래 계산이 수행됩니다. 단일 검증 데이터 집합의 경우 결과는 보고된 값입니다. Leave-One-Out 및  $K$  폴드 검증의 경우 각 검증 데이터 집합의 결과를 구하여 평균을 계산합니다.

추출된 요인이  $i$  개 있는 모형의 반응  $k$ 에 대한  $j$  번째 예측 잔차를  $R_{i,jk}$ 로 나타냅니다. 제안된 최적 요인 수  $opt$ 에 기반한 모형의 해당 값을  $R_{opt,jk}$ 로 나타냅니다. 검정 통계량은 다음 차이를 기반으로 합니다.

$$D_{i,jk} = R_{i,jk}^2 - R_{opt,jk}^2$$

$K$  개의 반응이 있다고 가정해 보겠습니다. 다음 표기를 고려하십시오.

$$\mathbf{d}_{i,j} = (D_{i,j1}, D_{i,j2}, \dots, D_{i,jK})'$$

$$\mathbf{d}_{i,.} = \sum_j \mathbf{d}_{i,j}$$

$$\mathbf{S}_i = \sum_j \mathbf{d}_{i,j} \mathbf{d}_{i,j}'$$

추출된  $i$  개의 요인에 대한 van der Voet 통계량은 다음과 같이 정의됩니다.

$$C_i = \mathbf{d}_{i,.}' \mathbf{S}_i^{-1} \mathbf{d}_{i,.}$$

$R^2_{i,jk}$  및  $R^2_{opt,jk}$ 를 무작위로 교환한 결과 값의 분포와  $C_i$ 를 비교하여 유의 수준을 구합니다. 이러한 값의 몬테카를로 표본이 시뮬레이션되고, 유의 수준은  $C_i$ 보다 크거나 같은 시뮬레이션된 임계값의 비율로 근사값이 산출됩니다.

## T<sup>2</sup> 그림에 대한 통계 상세 정보

부분 최소 제곱 플랫폼에서  $i$  번째 관측값의 T<sup>2</sup> 값은 다음과 같이 계산됩니다.

$$T_i^2 = (n-1) \sum_{j=1}^p \left( t_{ij}^2 / \sum_{k=1}^n t_{kj}^2 \right)$$

여기서  $t_{ij}$  =  $i$  번째 행과  $j$  번째 추출된 요인의 X 스코어,  $p$  = 추출된 요인 수,  $n$  = 모형을 훈련하는 데 사용된 관측값 수입니다. 검증을 사용하지 않는 경우  $n$  = 총 관측값 수입니다.

T<sup>2</sup> 그림의 관리 한계는 다음과 같이 계산됩니다.

$$((n-1)^2/n) * \text{BetaQuantile}(0.95, p/2, (n-p-1)/2)$$

여기서  $p$  = 추출된 요인 수,  $n$  = 모형을 훈련하는 데 사용된 관측값 수입니다. 검증을 사용하지 않는 경우  $n$  = 총 관측값 수입니다. 자세한 내용은 Tracy et al. (1992)에서 확인하십시오.

## X 스코어 산점도 행렬의 신뢰도 타원에 대한 통계 상세 정보

부분 최소 제곱 플랫폼의 "스코어 산점도 행렬" 옵션은 X 스코어 산점도에 95% 신뢰도 타원을 추가합니다. NIPALS 알고리즘과 SIMPLE 알고리즘 모두 직교 스코어 벡터를 생성하므로 X 스코어에 상관관계가 없습니다. 타원은 X 스코어의 각 쌍이 상관계수가 0인 이변량 정규 분포를 따른다고 가정합니다.

세로 축의 스코어  $i$  와 가로 축의 스코어  $j$  에 대한 산점도를 예로 들어보겠습니다. 타원의 위쪽, 아래쪽, 왼쪽 및 오른쪽 극단에 대한 좌표는 다음과 같이 정의됩니다.

- 위쪽 / 아래쪽 극단 =  $\pm \sqrt{\text{var}(\text{score } i) * z}$

- 왼쪽 / 오른쪽 극단 =  $\pm \sqrt{\text{var}(\text{score } j) * z}$

여기서  $z = ((n-1)*(n-1)/n) * \text{BetaQuantile}(0.95, 1, (n-3)/2)$ 입니다. Z 값에 대한 배경 정보는 Tracy et al. (1992)에서 확인하십시오.

## 예측 및 신뢰 한계에 대한 통계 상세 정보

이 섹션에서는 부분 최소 제곱 플랫폼에서 예측 및 신뢰 한계의 표준 오차 계산에 대해 설명합니다.  $\mathbf{X}$ 를 예측 변수 행렬로 나타내고  $\mathbf{Y}$ 를 반응 값 행렬로 나타냅니다. 이러한 값은 시작 창에서 선택한 항목에 따라 중심화 및 척도화될 수 있습니다.  $\mathbf{Y}$ 의 성분은 공통 분산이  $\sigma^2$ 인 독립 정규 분포를 따른다고 가정합니다.

Hoskuldsson(1988)은  $\mathbf{Y}$ 에 대한 PLS 모형을 스코어 관점으로 보면 다중 선형 회귀 모형과 형식적으로 유사하다는 것을 발견했습니다. 또한 이 유사성을 사용하여 예측값의 분산에 대한 근사 계산식을 도출했습니다. Umetrics(1995) 연구 자료도 확인하십시오. 그러나 Denham(1997)은 PLS에 의해 예측된 값은  $\mathbf{Y}$ 의 비선형 함수임을 지적했습니다. 또한 예측 구간을 얻기 위해 붓스트랩 및 교차 검증 기법을 제안했습니다. PLS 플랫폼에서는 Umetrics(1995)에 설명된 정규성 기반 방식을 사용합니다.

열에 스코어가 포함된 행렬을  $\mathbf{T}$ 로 나타내고  $\mathbf{X}$ 의 새 관측값  $\mathbf{x}_0$ 을 예로 들어보겠습니다.  $\mathbf{Y}$ 에 대한 예측 모형은  $\mathbf{T}$ 를 기반으로  $\mathbf{Y}$ 를 모델링하는 회귀를 통해 얻습니다.  $\mathbf{x}_0$ 과 관련된 스코어 벡터를  $\mathbf{t}_0$ 으로 나타냅니다.

$a$ 가 요인 수를 나타낸다고 하겠습니다.  $s^2$ 을  $df = n - a - 1$ (데이터가 중심화된 경우) 및  $df = n - a$ (데이터가 중심화되지 않은 경우)로 나눈 잔차 제곱합으로 정의합니다.  $s^2$  값은  $\sigma^2$ 의 추정값입니다.

### 예측 계산식의 표준 오차

$\mathbf{x}_0$ 에서 예측 평균의 표준 오차는 다음과 같이 추정됩니다.

$$SE(\bar{Y}_{x_0}) = s \sqrt{\left(\frac{1}{n} + \mathbf{t}_0(\mathbf{T}'\mathbf{T})^{-1}\mathbf{t}_0'\right)}$$

### 평균 신뢰 한계 계산식

$t_{0.975, df}$ 를 자유도가  $df = n - a - 1$ (데이터가 중심화된 경우) 및  $df = n - a$ (데이터가 중심화되지 않은 경우)인  $t$  분포의 0.975 분위수를 나타낸다고 하겠습니다.

평균에 대한 95% 신뢰 구간은 다음과 같이 계산됩니다.

$$\bar{Y}_{x_0} \pm t_{0.975, df} SE(\bar{Y}_{x_0})$$

### 개별값 신뢰 한계 계산식

$x_0$  에서 예측된 개별 반응의 표준 오차는 다음과 같이 추정됩니다.

$$SE(\hat{Y}_{x_0}) = s \sqrt{\left(\frac{1}{n} + 1 + \mathbf{t}_0(\mathbf{T}'\mathbf{T})^{-1}\mathbf{t}_0'\right)}$$

$t_{0.975, df}$  를 자유도가  $df = n - a - 1$  (데이터가 중심화된 경우) 및  $df = n - a$  (데이터가 중심화되지 않은 경우)인  $t$  분포의 0.975 분위수를 나타낸다고 하였습니다.

개별 반응에 대한 95% 신뢰 구간은 다음과 같이 계산됩니다.

$$\bar{Y}_{x_0} \pm t_{0.975, df} SE(\hat{Y}_{x_0})$$

### 표준화된 스코어 및 적재량에 대한 통계 상세 정보

이 섹션에서는 부분 최소 제곱 플랫폼의 표준화된 스코어 및 적재량 계산에 대해 설명합니다.

다음 표기를 고려하십시오.

- $n_{tr}$  = 훈련 데이터 집합의 관측값 수
- $m = X$  의 효과 수
- $k = Y$  의 반응 수
- $VarX_i = i$  번째 요인에 의해 설명되는  $X$  의 변동 백분율
- $VarY_i = i$  번째 요인에 의해 설명되는  $Y$  의 변동 백분율
- $XScore_i = i$  번째 요인에 대한  $X$  스코어 벡터
- $YScore_i = i$  번째 요인에 대한  $Y$  스코어 벡터
- $XLoad_i = i$  번째 요인에 대한  $X$  적재량 벡터
- $YLoad_i = i$  번째 요인에 대한  $Y$  적재량 벡터

### 표준화된 스코어

$i$  번째 표준화된  $X$  스코어의 벡터는 다음과 같이 정의됩니다.

$$\frac{XScore_i}{(n_{tr} - 1) \sqrt{m VarX_i / n_{tr}}}$$

$i$  번째 표준화된  $Y$  스코어의 벡터는 다음과 같이 정의됩니다.

$$\frac{\mathbf{YScore}_i}{(n_{tr} - 1)\sqrt{k\text{Var}Y_i/n_{tr}}}$$

### 표준화된 적재량

$i$  번째 표준화된  $X$  적재량의 벡터는 다음과 같이 정의됩니다.

$$\mathbf{XLoad}_i\sqrt{m\text{Var}X_i}$$

$i$  번째 표준화된  $Y$  적재량의 벡터는 다음과 같이 정의됩니다.

$$\mathbf{YLoad}_i\sqrt{k\text{Var}Y_i}$$

## **JMP** PRO PLS 판별 분석에 대한 통계 상세 정보

모형 적합 플랫폼에서 부분 최소 제곱 분석법을 사용하여 PLS-DA(부분 최소 제곱 판별 분석)를 수행할 수 있습니다. 시작 창에서 범주형 변수를  $Y$ 로 입력하면 표시자 코딩을 사용하여 변수가 코드화됩니다. 수준이  $k$ 개인 경우 각 수준은 해당 수준의 행이면 값이 1이고 그렇지 않으면 0인 표시 변수로 나타냅니다. 결과로 생성된  $k$ 개의 표시 변수가 연속형으로 처리되고, 연속형  $Y$ 를 사용하는 경우와 동일하게 PLS 분석이 진행됩니다.

# 7 장

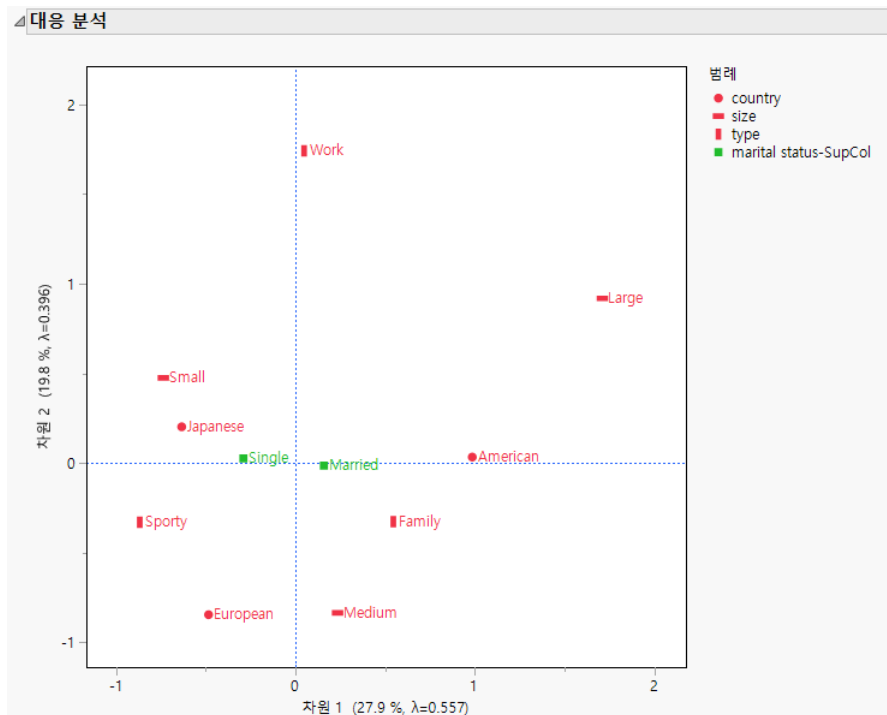
## 다중 대응 분석 범주형 변수 수준 간의 연관성 식별

MCA(다중 대응 분석)에서는 여러 범주형 변수를 사용하여 해당 변수 수준 간의 연관성을 식별합니다. MCA는 두 개의 변수에서 다수로 대응 분석을 확장합니다. 이는 양적 변수에 대한 주성분 분석과 유사한 것으로 간주될 수 있습니다. 또한 다른 다변량 방법과 마찬가지로 차원 축소 방법이며, 데이터를 2차원 또는 3차원 공간의 점으로 나타냅니다.

다중 대응 분석은 사회 과학 분야에 주로 사용됩니다. 설문 조사 분석에서 질문 함치도를 식별하는 데 사용할 수 있습니다. 또한 상품의 잠재 시장을 파악하기 위해 소비자 조사에도 사용됩니다.

다중 대응 분석에 대한 자세한 내용은 LeRoux와 Rouanet(2010) 연구 자료에서 확인하십시오.

그림 7.1 다중 대응 분석



## 목차

다중 대응 분석의 예.....	153
다중 대응 분석 플랫폼 시작.....	156
다중 대응 분석 보고서.....	157
다중 대응 분석 플랫폼 옵션.....	158
그림 표시.....	161
상세 정보 표시.....	161
조정 관성 표시.....	161
좌표 표시.....	162
요약 통계량 표시.....	162
관성에 대한 부분 기여도 표시.....	163
제공코사인 표시.....	163
Cochran Q 검정.....	164
교차표.....	164
다중 대응 분석의 추가 예.....	164
보조 변수를 사용한 예.....	165
보조 ID를 사용한 예.....	166
Cochran Q 검정의 예.....	167
다중 대응 분석 플랫폼에 대한 통계 상세 정보.....	168
상세 정보 보고서에 대한 통계 상세 정보.....	169
조정 관성에 대한 통계 상세 정보.....	170
요약 통계량에 대한 통계 상세 정보.....	170
Cochran Q 통계량에 대한 통계 상세 정보.....	171

---

## 다중 대응 분석의 예

이 예에서는 다양한 그룹의 직원을 이해하기 위해 네 가지 질문에 대한 직원 응답 간의 연관성을 탐색합니다. 이 예에서는 55명의 JMP 직원을 대상으로 다음 영역에서 개인의 선호도 또는 취향에 대해 수집한 데이터를 사용합니다.

- 선호하는 TV 프로그램 (8 개 범주): 뉴스, 코미디, 경찰, 자연, 스포츠, 영화, 드라마 또는 연속극
- 선호하는 영화 (8 개 범주): 액션, 코미디, 시대극, 다큐멘터리, 공포, 뮤지컬, 로맨스 또는 SF
- 선호하는 미술 유형 (7 개 범주): 행위 미술, 풍경화, 르네상스, 정물화, 초상화, 현대 미술 또는 인상파
- 선호하는 외식 장소 (6 개 범주): 버거 & 감자튀김, 피자, 인도 식당, 이탈리아 식당, 프랑스 식당 또는 스테이크 전문 식당

---

**참고:** 선호도 측도는 LeRoux 와 Rouanet(2010) 의 질문을 기반으로 합니다.

---

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Employee Taste.jmp 를 엽니다.
2. **분석 > 다변량 방법 > 다중 대응 분석**을 선택합니다.
3. TV, Film, Art 및 Restaurant 를 선택하고 **Y, 반응**을 클릭합니다.  
주성분 분석에서처럼 MCA에서 요인은 일반적으로 반응과 설명 변수가 아니라 반응으로 간주됩니다.
4. **확인**을 클릭합니다.

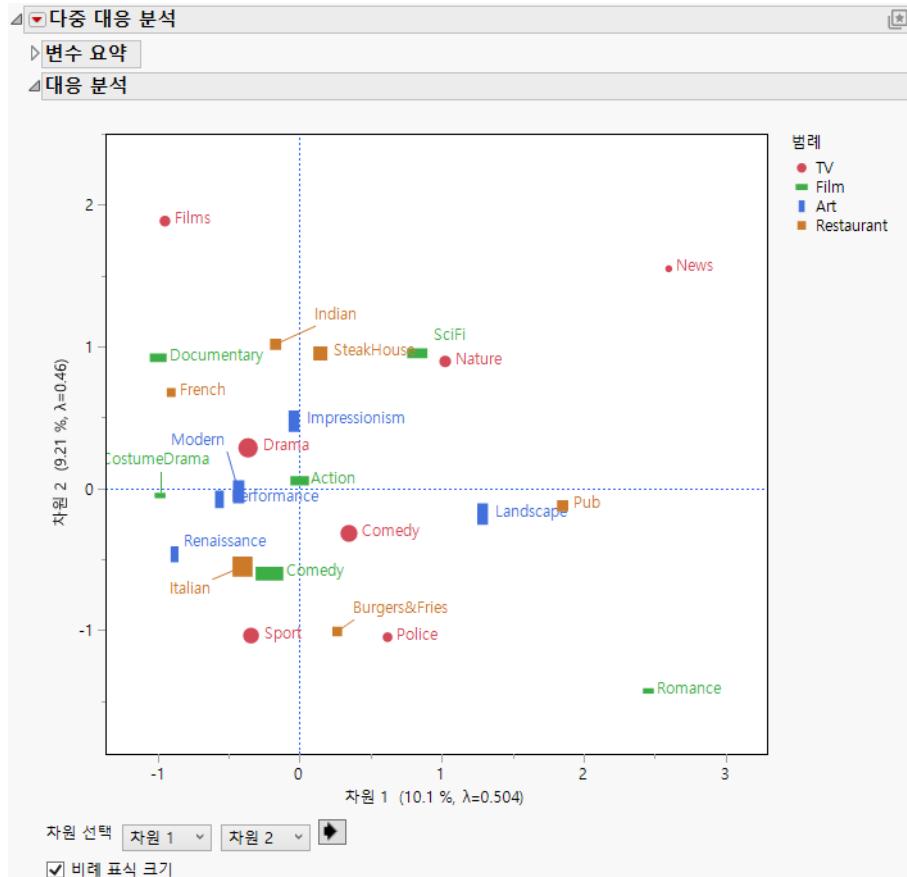
---

**팁:** 중첩 라벨을 클릭한 후 드래그하여 재배열할 수 있습니다.

---

5. 그림 아래에서 **비례 표시 크기**를 선택합니다.  
표시 크기는 해당 범주에서 반응의 상대적 비율을 나타냅니다.

그림 7.2 초기 다중 대응 분석 보고서 (일부)



"대응 분석" 보고서에는 처음 두 개의 주축(차원)에 투영된 네 가지 변수의 범주가 표시됩니다. 컨트롤을 사용하여 그림에 표시된 차원을 변경할 수 있습니다. 점 사이의 거리는 직원의 반응 패턴 간 비유사성을 나타냅니다.

이 그래프에서 주제에 대한 지식을 활용하여 결과를 해석할 수 있습니다.

- "Sport" 와 "Police" TV 프로그램 선호 군집이 "Burgers&Fries" 식당 선호 항목 근처에 그룹화되어 있습니다. 이 군집을 "대중문화" 그룹이라고 할 수 있습니다.
- 차원 2 를 보면 개인의 취향이 대중문화 그룹에서 다큐멘터리 영화를 좋아하고, TV 드라마를 선호하며, 프랑스와 인도 요리 같은 전문 식당을 선호하는 사람처럼 좀 더 "정교한" 취향으로 분류될 수 있는 것으로 옮겨갑니다.

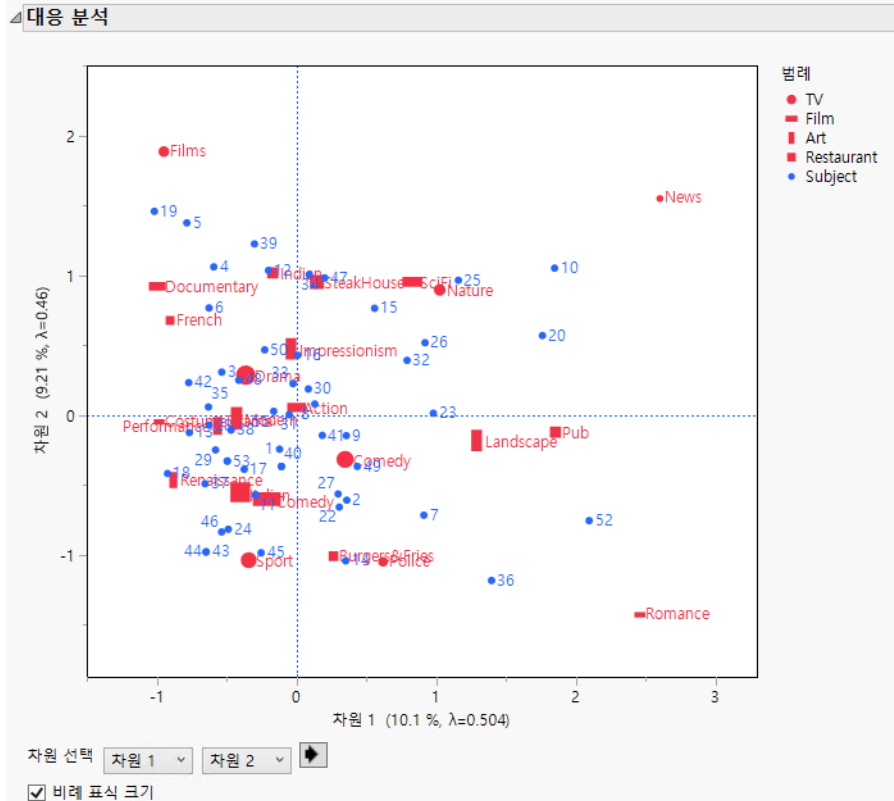
개별값에 대한 스코어를 얻으려면 다음과 같이 Subject 를 **X**, **요인**으로 추가합니다.

#### 6. "변수 요약" 개요를 엽니다.

"변수 요약" 패널을 사용하면 플랫폼을 다시 시작하지 않고 분석을 수정할 수 있습니다. 또한 완료된 분석에 대한 간략한 보기를 제공합니다.

7. Subject 를 선택하고 **X 추가**를 클릭합니다. 그러면 분석이 자동으로 업데이트됩니다.

그림 7.3 Subject 가 포함된 다중 대응 분석 보고서



Subject 를 포함하면 그림의 왼쪽 아래 사분면에 취향이 비슷한 직원 군집이 강조 표시됩니다. 해당 직원들은 대중문화 그룹으로 식별한 영역에 속합니다.

## 다중 대응 분석 플랫폼 시작

분석 > 다변량 방법 > 다중 대응 분석을 선택하여 다중 대응 분석 플랫폼을 시작합니다.

그림 7.4 다중 대응 분석 시작 창

명목형(범주형) 변수의 데이터 집합에서 기본 구조를 탐지하고 나타냅니다.

The screenshot shows the following interface elements:

- 열 선택 (Column Selection):** A list of 8 variables: Subject, TV, Film, Art, Restaurant, Gender, Age, and Job. The '8개 열' (8 columns) label is highlighted in red.
- 선택한 열 역할 지정 (Designate Selected Column Roles):** A table where roles are assigned to selected variables.
 

Role	Selected Variables
Y, 반응 (Y, Response)	필수 선택적 (Mandatory/Optional)
X, 요인 (X, Factor)	선택적 (Optional)
Z, 보조 변수 (Z, Auxiliary Variable)	선택적 (Optional)
보조 ID (Auxiliary ID)	선택적 숫자 (Optional Number)
빈도 (Frequency)	선택적 숫자 (Optional Number)
기준 (Reference)	선택적 (Optional)
- 작업 (Actions):** A set of buttons including '확인' (OK), '취소' (Cancel), '제거' (Remove), '재호출' (Re-call), and '도움말' (Help).

"열 선택"의 빨간색 삼각형 메뉴에 포함된 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

**Y, 반응** 분석할 범주형 열을 할당합니다. MCA에서는 일반적으로 변수 간의 연관성에 관심이 있지만 명시적 "설명" 변수와 "반응" 변수가 없습니다.

**X, 요인** 요인 또는 설명 변수로 사용할 하나 이상의 범주형 열을 할당합니다.

**참고:** 개별값에 대한 스코어를 얻으려면 개체 ID 열을 단일 X로 사용합니다.

**Z, 보조 변수** 보조 변수로 사용할 열을 할당합니다. 이러한 변수는 연관성을 식별하는 데 관심이 있지만 계산에 포함되지 않습니다. 보조 변수는 데이터 해석을 향상시키는 데 사용됩니다.

**보조 ID** 보조로 사용할 행을 식별하는 열을 할당합니다. 일반적으로 보조 ID 열의 값은 1과 0입니다. ID 0과 관련된 행은 보조 행으로 처리됩니다. 비보조 행에 없는 X 또는 Y 변수의 수준이 보조 행에 있으면 보조 ID 열이 무시됩니다.

**참고:** "보조 ID" 역할과 "Z, 보조 변수" 역할은 한 번에 하나만 지정할 수 있습니다.

**빈도** 이 역할에 빈도 변수를 할당합니다. 이 기능은 데이터를 요약할 경우에 유용합니다.

**기준** 기준 변수의 각 수준에 대해 개별 보고서를 생성합니다. 기준 변수가 둘 이상 할당되면 기준 변수의 가능한 각 수준 조합에 대해 개별 보고서가 생성됩니다.

---

**참고 :** 다중 대응 분석 플랫폼에서는 결측값을 많은 다른 JMP 플랫폼과 다르게 처리합니다. 분석할 때 모든 비결측 셀 쌍을 하나의 행에 사용합니다. 계산에서 전체 행이 제거되지 않습니다.

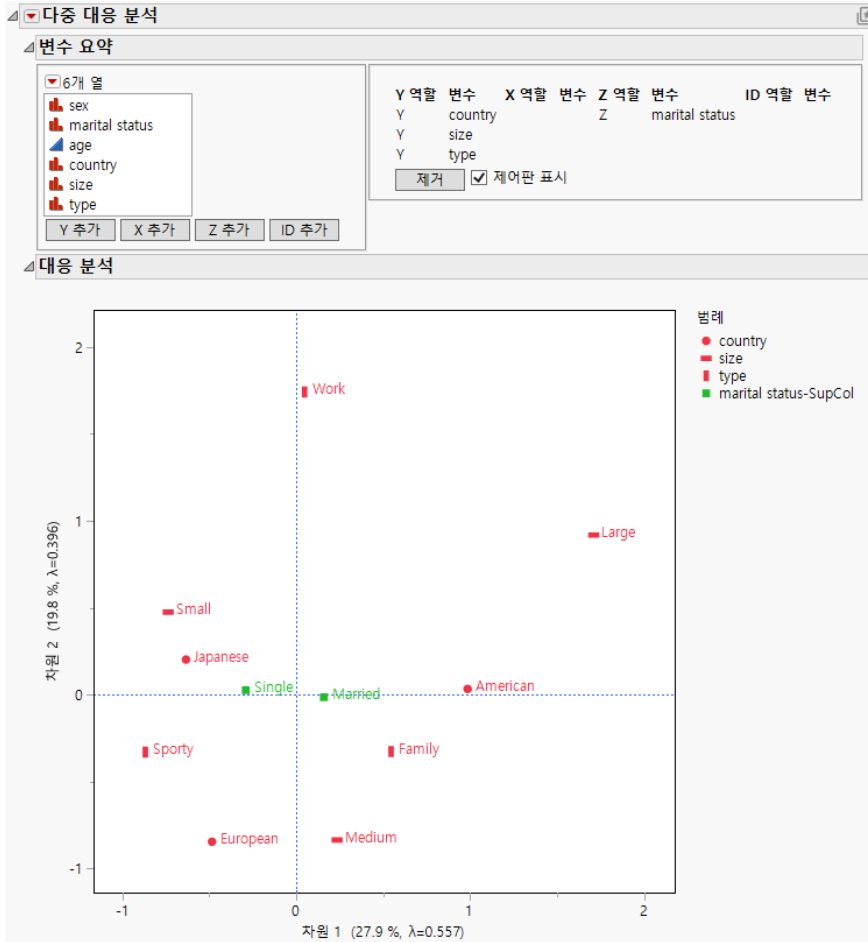
---

## 다중 대응 분석 보고서

초기 MCA 보고서에는 변수 요약, 대응 분석 그림 및 중요도 순으로 나타낸 데이터 차원 상세 정보가 표시됩니다. 범주 또는 개별값 클라우드는 그림을 통해 데이터 내에 존재하는 연관성을 식별할 수 있습니다. 상세 정보에는 그림에 표시된 두 차원이 데이터 내의 관계를 해석하기에 충분한지 여부에 대한 정보가 제공됩니다.

대화식 "변수 요약" 패널을 사용하면 플랫폼을 다시 시작하지 않고 분석을 수정할 수 있습니다. 이 패널에는 분석에 사용된 열과 역할이 표시됩니다. "제어판 표시" 체크박스를 선택하면 데이터 테이블의 열 목록이 왼쪽에 나타납니다. 열을 선택하고 "Y 추가", "X 추가", "Z 추가" 또는 "ID 추가"를 클릭하여 분석의 열을 변경할 수 있습니다. 또는 열을 변수 요약 테이블의 머리글로 드래그할 수 있습니다. 이렇게 하면 시작 창으로 돌아가지 않고 분석을 수정할 수 있습니다.

그림 7.5 제어판 표시가 선택된 다중 대응 분석 보고서



## 다중 대응 분석 플랫폼 옵션

MCA의 빨간색 삼각형 메뉴에는 필요에 따라 보고서를 사용자 정의할 수 있는 옵션이 포함되어 있습니다.

**교차표** 선택한 변수 역할에 맞게 Burt 표 또는 분할표를 표시하거나 숨깁니다. 자세한 내용은 "[교차표](#)"에서 확인하십시오.

**대응 분석** 대응 분석 보고서의 하위 메뉴를 표시합니다. 사용할 수 있는 보고서는 사용자가 수행하는 분석 유형에 따라 결정됩니다. 이러한 보고서에는 그림, 상세 정보, 좌표 및 요약 통계량이 제공됩니다.

**그림 표시** 처음 두 개의 주 축에 의해 설명되는 평면에 2 차원 범주 클라우드를 표시하거나 숨깁니다. 이 그림은 기본적으로 나타납니다. 이 그림에는 등간격 척도가 사용됩니다. 자세한 내용은 "**그림 표시**"에서 확인하십시오.

**상세 정보 표시** 특이값, 관성, 카이제곱 통계량, 백분율 및 누적 백분율을 포함하는 분석 상세 정보를 표시하거나 숨깁니다. 이 보고서는 기본적으로 나타납니다. 자세한 내용은 "**상세 정보 표시**"에서 확인하십시오.

**조정 관성 표시** Benzécri 조정 관성 보고서와 Greenacre 조정 관성 보고서를 표시하거나 숨깁니다. 자세한 내용은 Benzécri(1979) 및 Greenacre(1984) 연구 자료에서 확인하십시오. 하나 이상의 X 변수가 있는 경우 이 옵션을 사용할 수 없습니다. 자세한 내용은 "**조정 관성 표시**"에서 확인하십시오.

**좌표 표시** 필요한 경우 분석의 범주에 대해 최대 처음 세 개의 주 좌표 보고서를 표시하거나 숨깁니다. 자세한 내용은 "**좌표 표시**"에서 확인하십시오.

**요약 통계량 표시** 분석의 각 범주에 대한 품질, 질량 및 관성 요약 통계량 보고서를 표시하거나 숨깁니다. 자세한 내용은 "**요약 통계량 표시**"에서 확인하십시오.

**관성에 대한 부분 기여도 표시** 최대 처음 세 개의 각 차원에서 관성에 대한 각 범주의 기여도를 보여 주는 보고서를 표시하거나 숨깁니다. 자세한 내용은 "**관성에 대한 부분 기여도 표시**"에서 확인하십시오.

**제공코사인 표시** 최대 처음 세 개의 각 차원에서 각 범주의 제공코사인을 보여 주는 보고서를 표시하거나 숨깁니다. 이 보고서에는 각 Y 변수의 각 수준에서 처음 세 개의 각 차원에 대한 제공코사인 값을 표시하는 막대 차트가 포함되어 있습니다. 자세한 내용은 "**제공코사인 표시**"에서 확인하십시오.

**Cochran Q 검정** (모든 Y 변수에 두 수준으로 구성된 동일한 집합이 있고 X 변수에 각 행에 대한 고유 값이 있는 경우에만 사용 가능) 특정 반응의 주변 확률이 Y 변수에 따라 변하지 않음을 검정하는 Cochran Q 통계량을 제공합니다. Cochran Q 통계량은 세 개 이상의 반응 변수에 대한 McNemar 통계량을 일반화한 것입니다. Agresti(2013) 연구 자료를 참조할 수 있습니다. 자세한 내용은 "**Cochran Q 검정**"에서 확인하십시오.

**3D 대응 분석** 처음 세 개의 주 축에 의해 설명되는 공간에 Y, X 및 Z 변수의 3 차원 범주 클라우드를 표시하거나 숨깁니다. 차원이 세 개 미만인 경우 이 옵션을 사용할 수 없습니다.

**좌표 저장** 주 좌표를 하나 이상의 JMP 데이터 테이블에 저장합니다. 열 좌표, 행 좌표, 보조 열 좌표 및 보조 행 좌표가 별도의 JMP 데이터 테이블에 저장됩니다. 저장할 열 수를 선택할 수 있습니다.

**좌표 계산식 저장** 여러 차원의 주 좌표에 대한 계산식 열을 데이터 테이블에 저장합니다. 각 관측값에 대한 값은 각 차원에 대한 특이값으로 척도화된 Y 변수의 좌표 평균입니다. 저장할 열 수를 선택할 수 있습니다.

**보조 행 교차표** ( 하나 이상의 보조 Z 및 X 변수가 지정되거나 ID 가 지정된 경우 사용 가능 ) 다음 규칙에 따라 Burt 표 또는 분할표를 표시하거나 숨깁니다 .

- 하나 이상의 X 및 Z 변수가 지정된 경우 X 변수 대 반응 변수의 분할표를 표시하거나 숨깁니다 .
- ID 가 지정되고 X 변수가 지정되지 않은 경우 추가 관측값에 대한 Burt 표를 표시하거나 숨깁니다 . 추가 관측값은 ID 변수가 0 인 관측값입니다 .
- ID 와 하나 이상의 X 변수가 지정된 경우 추가 관측값에 대한 분할표를 표시하거나 숨깁니다 . 추가 관측값은 ID 변수가 0 인 관측값입니다 .

**보조 열 교차표** ( 하나 이상의 보조 Z 변수가 지정된 경우 사용 가능 ) 다음 규칙에 따라 분할표를 표시하거나 숨깁니다 .

- X 변수가 지정되지 않은 경우 반응 변수 대 보조 변수의 분할표를 표시하거나 숨깁니다 .
- 하나 이상의 X 변수가 지정된 경우 X 변수 대 보조 변수의 분할표를 표시하거나 숨깁니다 .

**모자이크 그림** ( X 변수와 Y 변수가 각각 하나씩 있는 경우에만 사용 가능 ) Y 대 X 모자이크 그림을 표시하거나 숨깁니다 . 모자이크 그림은 각 세그먼트가 해당 그룹의 빈도 수에 비례하는 누적 막대 차트입니다 .

**독립성에 대한 검정** ( X 변수와 Y 변수가 각각 하나씩 있는 경우에만 사용 가능 ) X 변수와 Y 변수 간의 독립성 검정을 표시하거나 숨깁니다 . 이 검정에는 Pearson 형식과 가능도비 형식의 두 가지 버전이 있으며 둘 다 카이제곱 통계량을 사용합니다 .

다음 옵션에 대한 자세한 내용은 **JMP 사용**에서 확인하십시오 .

**로컬 데이터 필터** 특정 보고서에서 사용되는 데이터를 필터링할 수 있는 로컬 데이터 필터를 표시하거나 숨깁니다 .

**다시 실행** 분석을 반복하거나 다시 시작할 수 있는 옵션이 포함되어 있습니다 . 이 기능을 지원하는 플랫폼에서 " 자동 재계산 " 옵션은 해당하는 보고서 창에서 데이터 테이블에 대한 변경 사항을 즉시 반영합니다 .

**플랫폼 환경 설정** 현재 플랫폼 환경 설정을 보거나 , 현재 JMP 보고서의 설정과 일치하도록 플랫폼 환경 설정을 업데이트할 수 있는 옵션이 포함되어 있습니다 .

**스크립트 저장** 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다 .

**그룹별 스크립트 저장** 기준 변수의 모든 수준에 대한 플랫폼 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다 . 시작 창에서 기준 변수를 지정한 경우에만 사용할 수 있습니다 .

---

**참고 :** 이 플랫폼의 추가 옵션은 스크립트를 통해 사용할 수 있습니다 . 자세한 내용은 " 도움말 " 메뉴의 " 스크립트 인덱스 " 에서 여십시오 . 또한 " 스크립트 인덱스 " 에서 이 섹션에 설명된 옵션의 스크립트 예제도 찾을 수 있습니다 .

---

## 그림 표시

다중 대응 분석 플랫폼에서 이 옵션은 처음 두 개의 주 축에 의해 설명되는 평면에 범주 또는 개별값이 투영된 투영 그래프를 표시합니다. 이 그림에는 등간격 척도가 사용됩니다. 그림 아래의 "차원 선택" 컨트롤을 사용하여 그림에 표시되는 차원을 전환할 수 있습니다. 첫 번째 컨트롤은 그림의 가로 축을 정의하고 두 번째 컨트롤은 그림의 세로 축을 정의합니다. 화살표 버튼을 클릭하여 그림에 표시되는 차원을 순환할 수 있습니다. 그림의 점 크기를 각 점에 해당하는 관측값 수에 비례하여 표시할지 여부를 지정하려면 **비례 표시 크기** 체크박스를 사용합니다.

---

**참고:** 대응 분석 그림에서 점을 선택하면 보고서 창의 다른 테이블에 있는 해당 행도 선택됩니다. 그러나 데이터 테이블의 행은 선택되지 않습니다. 그림에서 특정 변수와 관련된 모든 점을 선택하려면 그림 범례에서 변수 이름을 선택합니다.

---

## 상세 정보 표시

다중 대응 분석 플랫폼에서 이 옵션은 특이값 테이블을 표시합니다.

**특이값** 분할표 또는 Burt 표의 특이값 분해에서 특이값을 표시합니다. 계산식은 "[상세 정보 보고서에 대한 통계 상세 정보](#)"에서 확인하십시오.

**관성** 정준 차원의 측면을 고려한 상대적인 변동을 반영하는 특이값 제곱을 나열합니다.

**카이제곱** Burt 표 또는 분할표에 대한 전체 카이제곱에서 차원이 나타내는 부분을 나열합니다.

**백분율** 총 관성에 대한 관성 비율입니다.

**누적 백분율** 관성의 누적 비율을 보여 줍니다. 처음 두 개의 특이값에서 큰 관성이 발견되는 경우에는 2D 대응 분석 그림으로 테이블에서의 관계를 충분히 나타낼 수 있습니다.

## 조정 관성 표시

다중 대응 분석에서 Burt 표의 주 관성은 고유값입니다. 이러한 관성을 사용하면 비관적인 적합 표시가 제공된다는 문제가 있습니다. Benzécri 는 관성 조정을 제안했습니다. Greenacre 는 Benzécri 의 조정이 적합 품질을 과대추정한다고 주장하며 대체 조정 방법을 제안했습니다. 참조를 위해 두 조정 값이 모두 계산됩니다. 자세한 내용은 "[조정 관성에 대한 통계 상세 정보](#)"에서 확인하십시오.

**관성** 정준 차원의 측면을 고려한 상대적인 변동을 반영하는 특이값 제곱을 나열합니다.

**조정 관성** Benzécri 또는 Greenacre 조정에 따라 조정된 관성을 나열합니다.

**백분율** 총 관성에 대한 조정 관성의 비율입니다.

**누적 백분율** 조정 관성의 누적 비율을 보여 줍니다. 처음 두 개의 특이값에서 큰 관성이 발견되는 경우에는 2D 대응 분석 그림으로 테이블에서의 관계를 충분히 나타낼 수 있습니다.

## 좌표 표시

다중 대응 분석 플랫폼에서 이 옵션은 "열 좌표" 테이블 또는 "행 및 열 좌표" 테이블을 표시합니다.

**X** X, 요인 변수로 지정된 열을 나열합니다.

**Y** Y, 반응 변수로 지정된 열을 나열합니다.

**Z** Z, 보조 변수로 지정된 열을 나열합니다.

**범주** X, Y 또는 Z 변수의 수준을 나열합니다.

**차원 1, 차원 2, 차원 3** 표시된 주 축에서 각 수준 또는 각 반응의 좌표를 나열합니다. 기본적으로 처음 세 개 차원에 대한 좌표가 테이블에 표시됩니다. 추가 차원에 대한 좌표 열은 숨겨집니다. 이러한 선택적 열을 표시하려면 테이블에서 마우스 오른쪽 버튼을 클릭하고 **열** 하위 메뉴에서 차원 열을 선택합니다.

**참고:** X, 요인 변수로 지정된 열이 있는 경우 동일한 머리글을 사용하는 좌표 보고서에 X 테이블과 Y 테이블이 모두 표시됩니다. Z, 보조 변수가 지정된 경우에는 X 및 Y 좌표 아래에 해당 좌표가 나열됩니다.

## 요약 통계량 표시

다중 대응 분석 플랫폼에서 이 옵션은 "열 점에 대한 요약 통계량" 테이블 또는 "행 및 열 점에 대한 요약 통계량" 테이블을 표시합니다. Y 테이블에는 열 점이라고 하는 각 반응의 각 수준에 대한 품질, 질량 및 관성이 제공됩니다. X 테이블에는 X, 요인 변수의 각 수준에 대한 품질, 질량 및 관성이 제공됩니다. 자세한 내용은 "요약 통계량에 대한 통계 상세 정보"에서 확인하십시오.

**X** X, 요인 변수로 지정된 열을 나열합니다.

**Y** Y, 반응 변수로 지정된 열을 나열합니다.

**범주** X 또는 Y 변수의 수준을 나열합니다.

**품질 (차원 =2)** 해별로 수준 표현의 품질을 나열합니다.

**질량** 반응 수준에 대한 행 빈도를 총 빈도로 나눈 값을 나열합니다. Burt 표에서 이 값은 각 행의 총 %입니다.

**관성** 반응 수준에 의해 설명되는 총 관성의 비율을 나열합니다. 전체 수준 및 해당 반응에 대해 관성 값의 합은 1입니다.

**참고:** X, 요인 변수로 지정된 열이 있는 경우 동일한 머리글을 사용하는 요약 통계량 보고서에 X 테이블과 Y 테이블이 모두 표시됩니다.

## 관성에 대한 부분 기여도 표시

다중 대응 분석 플랫폼에서 이 옵션은 "열 점의 관성에 대한 부분 기여도" 테이블 또는 "행 및 열 점의 관성에 대한 부분 기여도" 테이블을 표시합니다. "열 점의 관성에 대한 부분 기여도 그림"도 표시됩니다. 이 그림은 각 Y 변수의 각 수준을 대상으로 테이블에 표시된 각 차원에 대한 부분 기여도를 보여 주는 막대 차트입니다.

**X** X, 요인 변수로 지정된 열을 나열합니다.

**Y** Y, 반응 변수로 지정된 열을 나열합니다.

**범주** X 또는 Y 변수의 수준을 나열합니다.

**차원 1, 차원 2, 차원 3** 표시된 차원의 관성에 대한 반응 또는 요인 수준의 기여도를 나열합니다. 기본적으로 처음 세 개 차원에 대한 열이 테이블에 표시됩니다. 추가 열은 숨겨집니다. 이러한 선택적 열을 표시하려면 테이블을 마우스 오른쪽 버튼으로 클릭하고 **열** 하위 메뉴에서 차원 열을 선택합니다.

각 반응의 각 수준은 각 차원의 관성에 기여합니다. 각 차원 내에서 부분 기여도의 합은 1입니다.

차원의 관성에 대한 행 또는 열의 기여도는 다음과 같이 계산됩니다.

$$\text{기여도} = (\text{질량})(\text{좌표})^2 / (\text{차원 관성})$$

**참고:** X, 요인 변수로 지정된 열이 있는 경우 동일한 머리글을 사용하는 관성에 대한 부분 기여도 보고서에 X 테이블과 Y 테이블이 모두 표시됩니다.

## 제공코사인 표시

다중 대응 분석 플랫폼에서 이 옵션은 "열 점에 대한 제공코사인" 테이블 또는 "행 및 열 점에 대한 제공코사인" 테이블을 표시합니다. "열 점의 제공코사인 그림"도 표시됩니다. 이 그림은 각 Y 변수의 각 수준을 대상으로 표시된 처음 세 개의 각 차원에 대한 제공코사인 값을 보여 주는 막대 차트입니다.

**X** X, 요인 변수로 지정된 열을 나열합니다.

**Y** Y, 반응 변수로 지정된 열을 나열합니다.

**범주** X 또는 Y 변수의 수준을 나열합니다.

**차원 1, 차원 2, 차원 3** 표시된 차원별로 수준 표현의 품질을 나열합니다. 기본적으로 처음 세 개 차원에 대한 결과가 테이블에 표시됩니다. 추가 열은 숨겨집니다. 이러한 선택적 열을 표시하려면 테이블을 마우스 오른쪽 버튼으로 클릭하고 **열** 하위 메뉴에서 차원 열을 선택합니다.

값은 표시된 차원에 대한 각 점의 품질을 나타냅니다. 제공코사인을 점과 차원의 상관 제곱으로 해석할 수 있습니다. 처음 두 차원의 제공코사인 합은 요약 통계량 보고서의 "품질 (차원 =2)" 값과 같습니다. 자세한 내용은 "요약 통계량에 대한 통계 상세 정보"에서 확인하십시오.

**참고 :** X, 요인 변수로 지정된 열이 있는 경우 동일한 머리글을 사용하는 제곱코사인 보고서에 X 테이블과 Y 테이블이 모두 표시됩니다.

## Cochran Q 검정

다중 대응 분석 플랫폼에서 Cochran Q 검정은 세 개 이상의 이분 결과에 대한 전체 매칭 표본의 비모수 동질성 검정입니다. 이 검정을 사용하여 매칭 쌍의 비율 차이를 판별할 수 있습니다. Cochran Q 검정은 결과가 두 개인 경우에 사용되는 McNemar 검정을 확장한 것입니다.

## 교차표

다중 대응 분석 플랫폼에서 "교차표" 옵션은 Burt 표 또는 분할표를 표시하거나 숨깁니다. X, 요인 열 없이 Y, 반응 열을 여러 개 선택하면 Burt 표가 생성됩니다. X, 요인 열을 선택하면 Burt 표 대신 기존의 분할표가 생성됩니다. 개요 노트 제목은 교차표의 구조를 반영합니다.

Burt 표는 다중 대응 분석의 기반입니다. 이 표는 모든 범주형 변수 쌍에 대한 분할된 대칭 테이블입니다. 대각 분할은 대각 행렬입니다(변수 자체의 교차표). 비대각 분할은 일반적인 분할표입니다.

"Burt 표" 또는 "분할표"의 빨간색 삼각형 메뉴에는 표에 표시할 다음 통계량 옵션이 포함되어 있습니다.

**개수** 셀 빈도, 빈도 소계 및 총 합계 (총 표본 크기)입니다. 이 값은 기본적으로 표시됩니다.

**총 %** 총 합계 대비 셀 개수 및 소계의 백분율입니다. 이 값은 기본적으로 표시됩니다.

**셀 카이제곱** 각 셀에 대해  $(O - E)^2 / E$  로 계산된 카이제곱 값입니다.

**열 %** 열 합계 대비 각 셀 개수의 백분율입니다.

**행 %** 행 합계 대비 각 셀 개수의 백분율입니다.

**기대값** 독립성을 가정한 상태에서 각 셀의 기대 빈도 (E)입니다. 해당 행 합계와 열 합계의 곱을 총 합계로 나눠서 계산됩니다.

**편차** 관측된 셀 빈도 (O)에서 기대 셀 빈도 (E)를 뺀 값입니다.

**열 누적** 누적 열 합계입니다.

**열 누적 %** 누적 열 백분율입니다.

**행 누적** 누적 행 합계입니다.

**행 누적 %** 누적 행 백분율입니다.

**데이터 테이블로 만들기** 표에 표시된 각 통계량에 대해 하나씩 데이터 테이블을 생성합니다.

## 다중 대응 분석의 추가 예

이 섹션에는 다중 대응 분석 플랫폼을 사용하는 예가 포함되어 있습니다.

- " 보조 변수를 사용한 예 "
- " 보조 ID 를 사용한 예 "
- "Cochran Q 검정의 예 "

## 보조 변수를 사용한 예

다중 대응 분석 플랫폼에서 보조 변수를 사용하여 직원의 직무 유형과 선호도 사이의 관계를 탐색합니다.

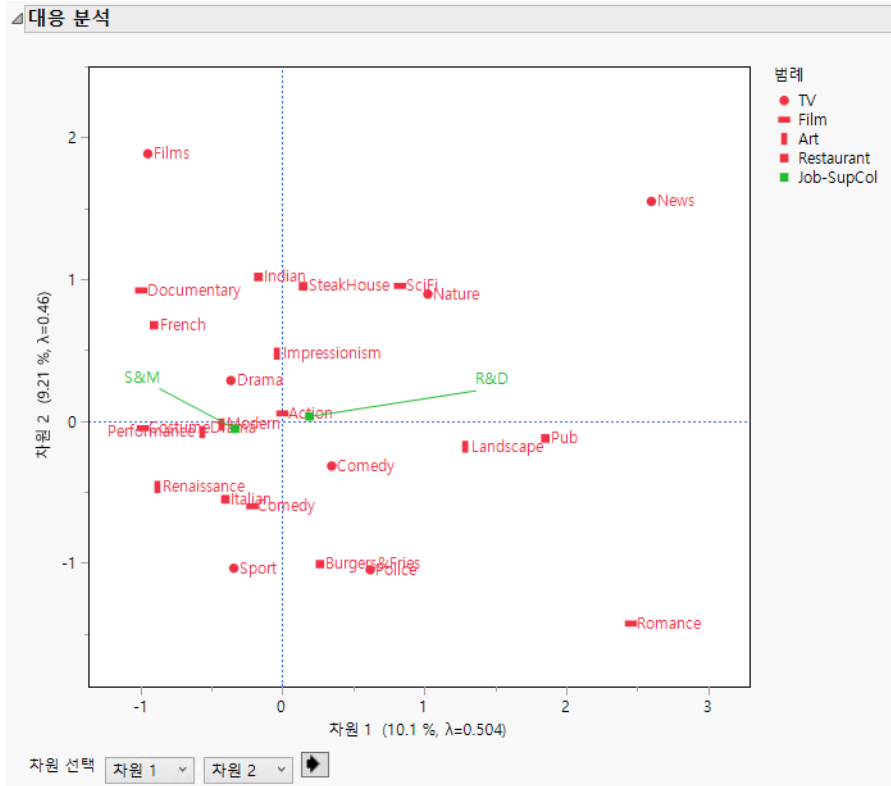
1. **도움말 > 샘플 데이터 폴더**를 선택하고 Employee Taste.jmp 를 엽니다.
2. **분석 > 다변량 방법 > 다중 대응 분석**을 선택합니다.
3. TV, Film, Art 및 Restaurant 를 선택하고 **Y, 반응**을 클릭합니다.
4. Job 을 선택하고 **Z, 보조 변수**를 클릭합니다.
5. **확인**을 클릭합니다.

---

**참고 :** Job 은 보조 변수이므로 계산에 사용되지 않습니다. 계산이 완료된 후 Job 유형이 그림에 표시됩니다.

---

그림 7.6 보조 변수를 사용한 MCA 보고서



그림에서 R&D(연구 개발) 직원과 S&M(영업 및 마케팅) 직원에 대한 데이터 점이 서로 가까이 있다는 것을 알 수 있습니다. 이는 직원의 업무 기능으로 인한 선호도에 큰 차이가 없다는 의미로 해석됩니다. 즉, 직원의 업무 역할은 선호도에 유의한 요인이 아닙니다.

## 보조 ID 를 사용한 예

다중 대응 분석을 사용하여 Alaska 와 Hawaii 의 인구 증가가 미국의 다른 지역과 다른지 여부를 판별합니다. 두 주는 전체 데이터 수집 기간 중에 주가 아니었고 미국 본토에 속하지 않으므로 보조 지역으로 처리됩니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 US Regional Population.jmp 를 엽니다.
2. **분석 > 다변량 방법 > 다중 대응 분석**을 선택합니다.
3. Year 를 선택하고 **Y, 반응**을 클릭합니다.
4. Region 을 선택하고 **X, 요인**을 클릭합니다.
5. ID 를 선택하고 **보조 ID** 를 클릭합니다.
6. Population 을 선택하고 **빈도**를 클릭합니다.

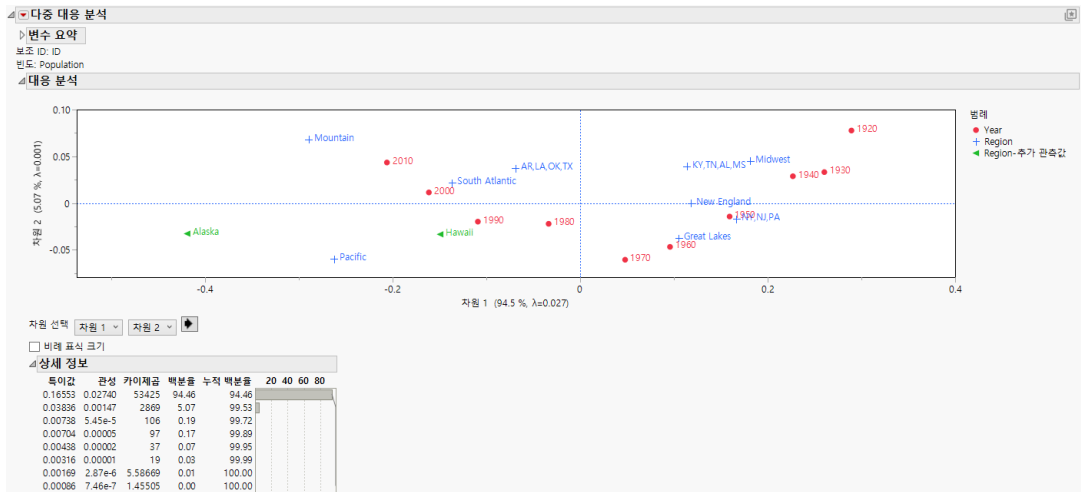
7. **확인**을 클릭합니다.

"상세 정보" 보고서에서는 연도와 지역 간의 연관성이 첫 번째 차원에 의해 거의 전부 설명되는 것을 보여 줍니다. 그림을 보면 첫 번째 차원에서 연도가 정확한 순서로 되어 있음을 알 수 있습니다. 이 순서는 대응 분석을 통해 자연스럽게 일어났으며 순서 정보가 분석에 제공되지 않았습니다. 데이터를 표시하는 데 사용된 등간격 척도가 그림에서 강조 표시됩니다.

지역 순서는 Midwest에서 시작하여 Northeast, South, Mountain West로 이어지는 인구 이동을 반영합니다.

Alaska와 Hawaii는 분석 계산에 사용되지 않았지만 결과를 기반으로 그림에 표시되었습니다. 증가 패턴은 Pacific 주와 가장 유사합니다. Alaska의 증가는 Pacific 지역보다 훨씬 더 극단적입니다.

그림 7.7 보조 ID 를 사용한 MCA 보고서



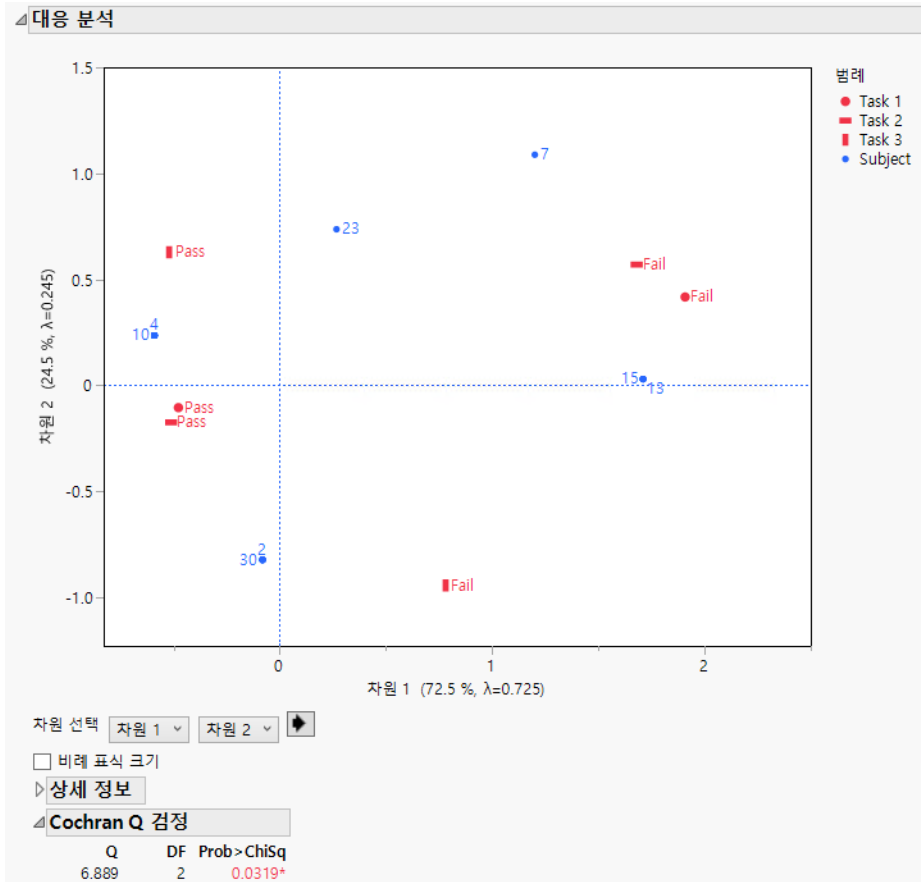
Cochran Q 검정의 예

다중 대응 분석을 사용하여 30명의 피실험자를 대상으로 세 가지 작업의 완료 용이성을 평가합니다. Task 1은 완료율 80%로 시뮬레이션되었습니다. Tasks 2와 3은 각각 피실험자의 95%와 80%에 대해 Task 1과 동일한 결과를 갖도록 시뮬레이션되었습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Cochrans Q.jmp** 를 엽니다.
2. **분석 > 다변량 방법 > 다중 대응 분석**을 선택합니다.
3. Task 1, Task 2, Task 3 을 선택하고 **Y, 반응**을 클릭합니다.
4. Subject 를 선택하고 **X, 요인**을 클릭합니다.
5. **확인**을 클릭합니다.

6. "다중 대응 분석"의 빨간색 삼각형 메뉴를 클릭하고 **대응 분석 > Cochran Q 검정**을 선택합니다.

그림 7.8 Cochran Q 검정



대응 분석 그림에서 Pass 및 Fail 결과가 함께 군집화되어 있으므로 Task 1과 Task 2 간의 유사성을 보여 줍니다. 그러나 Task 3은 Tasks 1과 2에서 멀리 떨어져 있습니다. Cochran Q 검정 통계량 6.889의 관련  $p$  값은 0.0319입니다. 이  $p$  값은 0.05 유의 수준에서 모든 작업에 대한 통과율이 동일하다는 가설이 기각됨을 뒷받침합니다.

## 다중 대응 분석 플랫폼에 대한 통계 상세 정보

이 섹션에는 다중 대응 분석 플랫폼에 대한 통계 상세 정보가 포함되어 있습니다.

- "상세 정보 보고서에 대한 통계 상세 정보"
- "조정 관성에 대한 통계 상세 정보"

- " 요약 통계량에 대한 통계 상세 정보 "
- "Cochran Q 통계량에 대한 통계 상세 정보 "

## 상세 정보 보고서에 대한 통계 상세 정보

다중 대응 분석 플랫폼의 상세 정보 보고서에는 특이값이 나열됩니다. 이러한 값을 얻기 위해 사용되는 특이값 분해는 단순 대응 분석에 사용되는 특이값 분해의 확장입니다.

단순 대응 분석이 수행되면 다음과 같은 특이값 분해의 특이값이 보고서에 나열됩니다.

$$\mathbf{D}_r^{-0.5}(\mathbf{P} - r\mathbf{c}')\mathbf{D}_c^{-0.5} = \mathbf{U}\mathbf{Diag}(\Lambda)\mathbf{V}'$$

다음은 각 요소에 대한 설명입니다.

- $\mathbf{P}$  = 개수를 총 빈도로 나눈 값의 행렬
- $r, c$  =  $\mathbf{P}$ 의 행 및 열 합계
- $\mathbf{D}$  행렬 =  $r$  및  $c$  값의 대각 행렬
- $\Lambda$  = 특이값의 열 벡터

다중 대응 분석이 수행되면 특이값 분해가 다음 방정식으로 확장됩니다.

$$\mathbf{D}^{-0.5}(\mathbf{C} - d\mathbf{d}')\mathbf{D}^{-0.5} = \mathbf{U}\mathbf{Diag}(\Lambda)\mathbf{V}'$$

다음은 각 요소에 대한 설명입니다.

- $\mathbf{C}$  = Burt 표
- $d$  =  $\mathbf{C}$ 의 열 합계에 대한 열 벡터 ( $\mathbf{C}$ 가 대칭이므로  $d$  = 행 합계)
- $\mathbf{D}$  =  $d$  값의 대각 행렬

"상세 정보" 보고서에서 관성은 열 벡터  $\Lambda$ 입니다. 특이값은 관성 벡터의 제곱근입니다. 열 좌표는 다음과 같이 계산됩니다.

$$\mathbf{D}^{-0.5}\mathbf{V}\mathbf{Diag}(\Lambda^{0.5})$$

## 조정 관성에 대한 통계 상세 정보

다중 대응 분석에서  $m$  개의 범주형 변수로 생성된 Burt 표의 일반적인 주 관성은  $\Lambda^2$ 의 고유값  $\lambda_k$ 입니다. 이러한 관성은 비관적인 적합 표시를 제공합니다. Benzécri(1979)는 다음과 같은 관성 조정을 제안했습니다. 이 내용은 Greenacre(1984, p. 145) 연구 자료에도 설명되어 있습니다.

$$\left(\frac{m}{m-1}\right)^2 \times \left(\lambda_k - \frac{1}{m}\right)^2 \quad \text{단, } \lambda_k > \frac{1}{m}$$

이 조정은  $1/m$ 보다 큰 모든 관성에 대해 조정 관성의 합을 기준으로 조정 관성의 백분율을 계산합니다.

Greenacre(1984, p. 156) 연구 자료에는 Benzécri 조정이 적합 품질을 과대추정한다는 내용이 나옵니다. 대신 Greenacre는 다음을 기준으로 조정 관성의 백분율을 계산할 것을 제안합니다.

$$\frac{m}{m-1} \left( \text{tr}(\text{Diag}(\Lambda^4)) - \frac{n_c - m}{m^2} \right)$$

이 계산은  $1/m$ 보다 큰 모든 관성을 대상으로 합니다. 여기서  $\text{tr}(\text{Diag}(\Lambda^4))$ 은 관성 제곱합이고,  $n_c$ 는  $m$ 개의 변수에 대한 총 범주 수입니다.

## 요약 통계량에 대한 통계 상세 정보

이 섹션에서는 다중 대응 분석 플랫폼에 보고되는 요약 통계량에 대해 설명합니다.

품질은 지정된 수의 차원으로 정의되는 공간에서 원점과 점 사이의 거리 제곱과 최대 차원 수를 갖는 공간에서 원점과의 거리에 대한 비율입니다. 카이제곱 측정 기준의 경우 주어진 차원의 점 품질은 차원을 정의하는 벡터를 사용하여 계산된 코사인을 통해 얻을 수 있습니다. 또한 품질은 지정된 차원의 관성 합 대 모든 차원의 관성 합 비율과 같습니다. 품질은 저차원 공간에서 점이 얼마나 잘 표현되는지를 나타냅니다.

질량은 총 빈도에 대한 행 또는 열 빈도의 비율입니다.

관성은 주성분 분석의 분산과 유사합니다. 전체 관성은 테이블에 있는 모든 관측값의 합으로 나눈 이원 빈도 테이블의 총 Pearson 카이제곱입니다.

상대 관성은 전체 관성에 대한 점의 기여 비율입니다. 요약 통계량 테이블의 "관성" 열에 상대 관성이 나열됩니다.

## Cochran Q 통계량에 대한 통계 상세 정보

Cochran Q 검정은 다중 대응 분석 플랫폼에서 사용할 수 있습니다. Cochran Q 검정 통계량은 다음과 같이 계산됩니다.

$$Q = k(k-1) \frac{\sum_{i=1}^k \left( \sum_{j=1}^b X_{ij} - \frac{N}{k} \right)^2}{\sum_{i=1}^b \left( \sum_{j=1}^k X_{ij} \left( k - \sum_{j=1}^k X_{ij} \right) \right)}$$

다음은 각 요소에 대한 설명입니다.

$k$  = 처리 수

$b$  = 블록 수

$X_{ij}$  =  $k$  번째 처리에 대한  $i$  번째 개체의 반응 (0 또는 1)

$N$  = 모든 개체와 처리에 대한 양의 반응의 총 합계

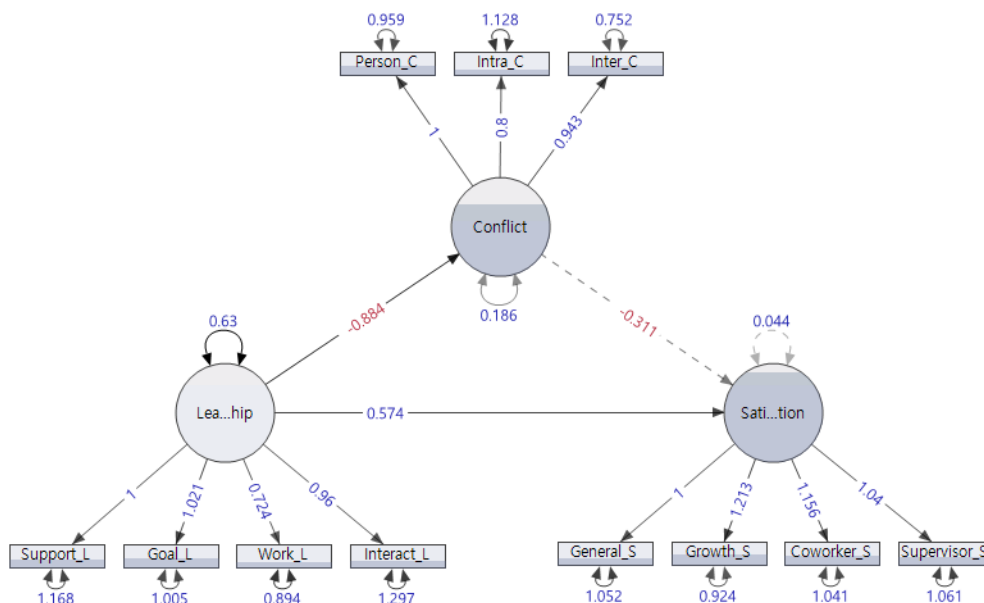


구조 방정식 모형 플랫폼은 JMP Pro 에서만 사용할 수 있습니다.

구조 방정식 모형 플랫폼을 사용하면 다양한 모형을 적합시킬 수 있습니다. 이러한 모형에는 확증적 요인 분석, 잠재 변수를 사용하거나 사용하지 않는 경로 모형, 측정 오차 모형, 중재 분석 및 잠재 성장 곡선 모형 등이 포함됩니다. 시작점 및 끝점 변수를 선택하고 두 변수의 연결 방법을 결정하여 모형을 지정할 수 있습니다. 연결은 다이어그램에서 단방향 또는 양방향 화살표로 표시됩니다. 이러한 선택 항목으로 모형의 두 가지 대체 보기(동적 경로 다이어그램, 모형 엣지 목록)를 채웁니다. 또한 이 플랫폼에서는 모형을 지정할 때 모형 식별을 확인합니다.

다중 그룹 분석 모형을 적합시킬 수도 있으므로 그룹 간의 통계적 효과를 체계적으로 비교할 수 있습니다. 이 방법은 모집단 간의 차이를 조사하는 데 유용할 수 있습니다.

**그림 8.1** 구조 방정식 모형 경로 다이어그램



## 목차

구조 방정식 모형 개요.....	175
구조 방정식 모형의 유형.....	175
구조 방정식 모형의 예.....	179
구조 방정식 모형 플랫폼 시작.....	182
구조 방정식 모형 보고서.....	183
모형 규격 보고서.....	184
모형 비교 보고서.....	192
카이제곱 차이 검정 보고서.....	193
구조 방정식 모형 적합 보고서.....	194
구조 방정식 모형 플랫폼 옵션.....	195
모형 옵션.....	197
경로 다이어그램 사용자 정의.....	202
경로 다이어그램 팝업 메뉴 옵션.....	202
경로 다이어그램 사용자 정의 옵션.....	204
구조 방정식 모형의 추가 예.....	205
잠재 경로 변수 모형의 예.....	205
잠재 성장 곡선 모형의 예.....	211
측정 모형 평가 보고서의 예.....	213
다중 그룹 분석의 예.....	216
중재 분석의 예.....	218
구조 방정식 모형 플랫폼에 대한 통계 상세 정보.....	222
추정 방법에 대한 통계 상세 정보.....	223
적합 요약 측도에 대한 통계 상세 정보.....	223
중재 분석에 대한 통계 상세 정보.....	228

## JMP PRO 구조 방정식 모형 개요

SEM(구조 방정식 모형) 플랫폼을 사용하면 변수 간의 관계 이론을 검증하는 데 사용할 수 있는 다양한 모형을 적합시킬 수 있습니다. 모형의 변수는 관측된 명시 변수이거나 관측되지 않은 잠재 변수일 수 있습니다. 구조 방정식 모델링은 사회 과학 및 행동 과학 분야에서 많이 사용됩니다.

기본적으로 이 플랫폼에서는 모든 변수에 대한 평균과 분산이 있는 모형을 지정합니다. 그런 다음 모형이 생성되는 동안 여러 보기를 확인할 수 있는 모형 생성 인터페이스를 제공합니다. 또한 모형 생성 과정 중에 몇 가지 모형 상세 정보를 제공하여 모형을 실행하기 전에 유지할 수 없는 모형에 대해 알려 줍니다.

하나 이상의 모형을 적합시킨 후 "모형 비교" 보고서에서 적합 모형과 두 개의 기준 모형을 비교할 수 있습니다. 기준 모형은 제약 없는 모형과 독립 모형입니다. 제약 없는 모형은 데이터에 구조를 적용하지 않고 지정된 모형 변수의 모든 평균, 분산 및 공분산을 적합시키는 완전 포화 모형입니다. 독립 모형은 지정된 모형 변수의 모든 평균과 분산을 적합시킵니다. 지정된 모형 변수에서 모든 공분산이 0으로 고정되므로 매우 제한적인 모형이 됩니다.

SEM 플랫폼에서는 완전 정보 최대 가능성(Finkelstein 1979) 방법을 사용합니다. 이렇게 하면 랜덤 결측값이 있는 관측값 비율이 높은 경우에도 데이터에서 사용 가능한 모든 정보를 완전히 사용할 수 있습니다.

구조 방정식 모델링에 대한 자세한 내용은 SAS Institute Inc. (2023a)의 "CALIS Procedure" 장, Bollen(1989) 및 Kline(2016)에서 확인하십시오.

**참고:** 구조 방정식 모형 플랫폼의 모든 모형은 평균 구조로 추정되므로 Constant 항이 포함됩니다. 관측 변수의 평균에 구조를 적용하지 않으려면 기본 모형 규격에서와 같이 평균이 자유롭게 추정되어야 합니다.

## JMP PRO 구조 방정식 모형의 유형

이 섹션에서는 구조 방정식 모형 플랫폼에서 적합시킬 수 있는 다양한 모형 유형 중 일부에 대해 설명합니다.

### 경로 분석

경로 분석을 사용하면 관측 변수 간의 연관성에 대한 대안 설명 모형을 검증할 수 있습니다. 이 기법은 연구에서 관심 구성당 하나의 변수만 사용할 수 있을 때 주로 사용됩니다. 대개 가장 단순한 경로 분석 모형은  $X$ 가  $Y$ 를 예측하는 표준 회귀 모형입니다. SEM 플랫폼을 사용하면 이 단순 회귀 모형을 적합시킬 수 있지만 더 복잡한 모형을 지정할 수도 있습니다. 예를 들어 이론이나 이전 조사에 기초하여  $X \Rightarrow Y$  관계의 중재 변수로 간주되는  $Z$  변수가 있을 수 있습니다. 즉,  $X$ 가  $Z$ 를 예측한 후  $Y$ 를 예측합니다. 따라서 원래  $X \Rightarrow Y$  관계는  $Z$ 가 원래 모형에서 제외되어야만 존재할 수 있습니다. 자세한 내용은 "중재 분석"에서 확인하십시오. 경로 분석 모형의 변수 수가 많아질수록 이러한 모형의 복잡성도 증가합니다. 복잡성이 추가되면 연구원들이 더 정

교한 이론을 테스트할 수 있습니다 . SEM 플랫폼에서 다음 단계에 따라 경로 분석을 수행할 수 있습니다 .

1. 시작 창에서 모든 관측 변수를 선택하고 " 모형 변수 " 를 클릭한 후 " 확인 " 을 클릭합니다 .
2. " 모형 규격 " 보고서의 " 시작점 목록 " 에서 예측 변수를 선택하고 " 끝점 목록 " 에서 해당하는 결과를 선택한 후 단방향 화살표 버튼을 클릭합니다 .

**참고 :** 상관계수가 0 이라는 가설을 검정하는 경우 외에는 모든 외생 변수 ( 단방향 화살표가 가리키지 않는 변수 ) 가 모형에서 자유롭게 상관되어야 합니다 . 이러한 공분산은 양방향 화살표 버튼으로 지정됩니다 .

### 확증적 요인 분석

CFA( 확증적 요인 분석 ) 를 사용하면 대안 측정 모형을 검정할 수 있습니다 . CFA 는 조사 개발에서 주로 사용되며 구조 회귀 모형을 적합시키기 전의 초기 단계로 사용됩니다 . SEM 플랫폼을 사용하면 다음 단계를 수행하여 확증적 요인 분석 모형을 적합시킬 수 있습니다 .

1. 시작 창에서 모든 관측 변수를 선택하고 " 모형 변수 " 를 클릭한 후 " 확인 " 을 클릭합니다 .
2. " 모형 규격 " 의 " 끝점 목록 " 을 사용하여 잠재 변수에 적재될 것으로 간주되는 변수를 선택합니다 .
3. " 끝점 목록 " 아래의 상자에 잠재 변수 이름을 입력하고 잠재 변수 추가 버튼 **+** 을 클릭합니다 .
4. 모형에 대한 모든 잠재 변수가 지정될 때까지 이 과정을 반복합니다 .

SEM 플랫폼에는 항상 평균 구조가 포함되므로 모든 관측 변수가 "평균/절편" 목록에 **Constant** 항 결과로 나열됩니다 . 또한 모든 잠재 변수는 첫 번째 표시자의 적재를 1로 설정하거나(기본값), 시작 창에서 "잠재 변수 표준화" 옵션을 선택한 경우 분산을 1로 설정하여 자동으로 식별됩니다 . 마지막으로 기존의 CFA 모형에서는 모든 잠재 변수가 공변할 수 있습니다 . "시작점 목록"과 "끝점 목록"에서 모든 잠재 변수를 선택한 후 양방향 화살표 버튼을 클릭하여 이러한 공분산을 지정할 수 있습니다 .

### 중재 분석

중재 분석을 사용하면 하나 이상의 중재 변수가 원인 예측 변수의 효과를 결과로 전달한다고 가정하는 인과 이론을 검정할 수 있습니다 . 중재 모형은 중재 변수라고도 하는 하나 이상의 중간 변수를 통해 예측 변수와 결과 변수 사이에 간접 효과가 존재하는지 여부 및 효과의 강도를 탐색합니다 . SEM 플랫폼에서 원인, 중재 및 결과 역할의 변수를 지정하여 단순 중재 모형을 지정할 수 있습니다 . 체크박스를 사용하면 중재 변수를 두 개 이상 선택하여 다중 중재 모형을 지정할 수 있습니다 .

1. 시작 창에서 예측 변수, 결과 변수, 중재 변수를 선택하고 " 모형 변수 " 를 클릭한 후 " 확인 " 을 클릭합니다 .
2. " 모형 바로 가기 " 옵션을 사용하여 횡단면 ( 전통적 ) > 중재 분석 옵션을 선택합니다 .
3. " 중재 모형 지정 " 창에서 원인 예측 변수, 중재 변수, 결과 변수를 선택합니다 .

이 모형에서는 예측 변수의 효과가 중재 변수를 거쳐 결과 변수로 전달됩니다. 예를 들어 특정 교수법(독립 변수)과 학생의 학업 성취도(종속 변수) 사이의 관계를 파악하는 데 관심이 있다고 가정해 보겠습니다. 학생의 수업 참여도(중재 변수)가 이 관계를 설명할 수 있다고 생각됩니다. 중재 분석 바로 가기를 사용하면 간접 효과의 유의성을 평가하고, 학생의 참여도가 교수법이 학업 성취도에 미치는 영향을 매개하는지 여부 및 그 정도를 파악할 수 있습니다. 자세한 내용은 "**중재 분석의 예**"에서 확인하십시오.

### 구조 회귀 모형

SR(구조 회귀) 모형은 잠재 변수가 있는 경로 분석이라고도 합니다. 이러한 모형은 주로 CFA(확증적 요인 분석)를 통해 적절한 측정 모형을 식별한 후에 사용됩니다. SR 모형을 사용하면 잠재 변수 간의 특정 관계 패턴을 검정할 수 있습니다. 즉, CFA는 잠재 변수 사이의 효과에 방향성을 적용하지 않지만(모든 잠재 변수가 자유롭게 공변할 수 있음) SR 모형은 방향성을 적용합니다. 경영진의 리더십(Leadership)이 직장 내 팀 갈등(Conflict)을 줄이고 직원 만족도(Satisfaction)를 높일 수 있다고 가정한 예에서 Leadership 잠재 변수는 Conflict 및 Satisfaction 잠재 변수를 예측할 수 있습니다. 다음 단계에 따라 CFA를 수행한 후 이러한 방향 효과(회귀)를 지정할 수 있습니다.

1. 시작 창에서 모든 관측 변수를 선택하고 "모형 변수"를 클릭한 후 "확인"을 클릭합니다.
2. "모형 규격"의 "끝점 목록"을 사용하여 잠재 변수에 적재될 것으로 간주되는 변수를 선택합니다.
3. "끝점 목록" 아래의 상자 옆에 잠재 변수 이름을 입력하고 잠재 변수 추가 버튼 **+**을 클릭합니다.
4. 모형에 대한 모든 잠재 변수가 지정될 때까지 이 과정을 반복합니다.
5. "모형 규격" 보고서의 "시작점 목록"에서 예측 변수를 선택하고 "끝점 목록"에서 해당하는 결과를 선택한 후 단방향 화살표 버튼을 클릭합니다.

### 잠재 성장 곡선 모형

LGC(잠재 성장 곡선) 모형을 사용하면 반복 측정 데이터에 대한 잠재 궤적을 적합시키고 검정할 수 있습니다. 이러한 모형은 혼합 모형 프레임워크의 임의 효과 모형과 매우 유사합니다. 무성장 모형을 선형 모형과 비교하려는 경우가 종종 있습니다. 무성장 모형에서는 개체의 시작점이 다를 수 있지만 궤적이 평평합니다. 선형 모형에서는 개체의 시작점과 시간에 따른 선형 기울기가 모두 다를 수 있습니다. 사용 가능한 데이터가 충분한 경우에는 이러한 모형을 개체의 시작점과 시간에 따라 선형 및 2차 비율이 달라질 수 있는 2차 모형과 비교할 수도 있습니다. SEM 플랫폼 모형 규격을 사용하여 LGC 모형을 적합시키거나, 다음 단계를 수행하여 모형 바로 가기를 통해 LGC 모형 적합을 간소화할 수 있습니다.

1. 시작 창에서 모든 관측 변수(반복 측정)를 선택하고 "모형 변수"를 클릭한 후 "확인"을 클릭합니다.

**참고:** 잠재 성장 모형 바로 가기에서 모형을 올바르게 지정하려면 관측 변수가 시간 오름차순으로 나열되고 시간 간격이 동일해야 합니다.

2. " 모형 바로 가기 " 옵션을 사용하여 중단 분석 > 절편만 있는 잠재 성장 곡선 모형을 선택한 후 " 실행 " 을 클릭합니다 .
3. " 모형 바로 가기 " 옵션을 사용하여 중단 분석 > 선형 잠재 성장 곡선 모형을 선택한 후 " 실행 " 을 클릭합니다 .
4. " 모형 바로 가기 " 옵션을 사용하여 중단 분석 > 2 차 잠재 성장 곡선 모형을 선택한 후 " 실행 " 을 클릭합니다 .

그러면 "모형 비교" 테이블에 대안적 적합 지수가 표시되고 최적 모형을 선택할 수 있습니다.

### 조건부 잠재 성장 곡선 모형

조건부 잠재 성장 곡선 모형은 위의 단계에 따라 이상적인 성장 궤적을 식별한 후 사용할 수 있습니다. 이때 변경 요인 및 절편의 예측 변수를 모형에 추가할 수 있습니다. 이러한 예측 변수는 성장 과정의 초기 스코어와 이후의 변화를 결정하는 데 중요한 요인일 수 있습니다. 조건부 LGC 를 적합시키려면 시작 창에서 잠재 변수의 가설 예측 변수를 포함하여 모든 관측 변수 ( 반복 측정 ) 를 선택합니다. 다음 단계를 쉽게 수행하려면 예측 변수가 " 모형 변수 " 목록의 마지막 변수 여야 합니다.

1. " 모형 바로 가기 " 옵션을 사용하여 적절한 성장 궤적을 선택합니다. 이 옵션은 예측 변수를 포함하여 LGC 모형의 모든 변수를 지정합니다. 따라서 예측 변수를 성장 과정에서 제외하고 예측 변수로 올바르게 지정해야 합니다.
2. " 적재 " 목록에서 예측 변수를 찾고 해당 변수를 포함하는 모든 효과를 선택한 후 " 제거 " 를 클릭합니다.
3. " 시작점 목록 " 에서 예측 변수를 선택하고 " 끝점 목록 " 에서 " 절편 " 또는 " 기울기 " 를 선택합니다.
4. 단방향 화살표를 클릭하여 조건부 LGC 를 지정합니다.

**참고:** 예측 변수가 두 개 이상인 경우 " 시작점 목록 " 과 " 끝점 목록 " 에서 예측 변수를 선택하고 양방향 화살표 버튼을 클릭하여 공분산을 지정해야 합니다.

### 다중 그룹 분석

다중 그룹 분석 모형을 사용하면 SEM 프레임워크의 모형에 대해 그룹화 변수를 지정할 수 있습니다. 그러면 그룹 전체에서 모형이 추정되므로 여러 모집단에 대해 추론할 수 있습니다. 다음 단계를 사용하여 다중 그룹 분석 모형을 지정할 수 있습니다.

1. 시작 창에서 모델링할 관측 변수를 선택하고 " 모형 변수 " 를 클릭합니다.
2. 범주형 그룹화 변수 ( 주로 수준 수가 적은 변수 ) 를 선택하고 " 그룹 " 을 클릭한 후 " 확인 " 을 클릭합니다.
3. " 모형 규격 " 보고서를 사용하여 선택 모형을 지정합니다. 회귀 경로를 추가하려면 " 시작점 목록 " 에서 예측 변수를 선택하고 " 끝점 목록 " 에서 해당 결과를 선택한 후 단방향 화살표 → 버튼을 클릭합니다. 공분산 경로를 추가하려면 동일한 단계를 수행하지만 대신 양방향

화살표 ↔ 버튼을 클릭합니다. 잠재 변수를 추가하려면 "끝점 목록"에서 해당 표시자를 선택한 후 "끝점 목록" 아래의 잠재 변수 추가 + 버튼을 클릭합니다.

4. "결합" 탭 경로 다이어그램을 사용하여 엷지를 선택하고 "등가 설정"을 클릭하여 그룹 간에 등식 제약 조건을 적용합니다. "결합" 탭에서 변경한 모형 규격은 모든 그룹에 적용됩니다. 그룹별 제약 조건 또는 규격 변경은 그룹별 탭을 사용하여 적용할 수 있습니다. 기본적으로 모형의 경로는 그룹 간에 자유롭게 추정됩니다.

## JMP PRO 구조 방정식 모형의 예

예를 들어 인사과 직원이 주요 직장 구성 요인을 측정하기 위한 설문 조사를 개발하려고 합니다. 이 예에서는 구조 방정식 모형 플랫폼을 사용하여 확증적 요인 분석 모형을 생성합니다. 이 모형을 사용하여 200 명을 대상으로 실시한 직장 관련 설문 조사의 응답을 분석할 수 있습니다. 설문 조사는 직무 만족도와 관련된 11 개 질문에 대한 응답을 포함합니다. 확증적 요인 분석 모형은 설문 조사 질문에 대한 답변을 리더십 특성, 역할 갈등 및 전반적인 직무 만족도의 잠재 변수와 연관시킵니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Job Satisfaction.jmp 를 엷니다.
2. **분석 > 다변량 방법 > 구조 방정식 모형**을 선택합니다.
3. Support\_L ~ Supervisor\_S 를 선택하고 **모형 변수**를 클릭합니다.
4. **확인**을 클릭합니다.  
"구조 방정식 모형" 보고서의 "모형 규격" 개요가 나타납니다.

### 잠재 변수 생성

5. "끝점 목록"에서 Support\_L ~ Interact\_L 을 선택합니다. "끝점 목록" 아래의 상자에 "Leadership"을 입력한 후 잠재 변수 추가 + 버튼을 클릭합니다.
6. "끝점 목록"에서 Person\_C ~ Inter\_C 를 선택합니다. "끝점 목록" 아래의 상자에 "Conflict"를 입력한 후 잠재 변수 추가 + 버튼을 클릭합니다.
7. "끝점 목록"에서 General\_S ~ Supervisor\_S 를 선택합니다. "끝점 목록" 아래의 상자에 "Satisfaction"을 입력한 후 잠재 변수 추가 + 버튼을 클릭합니다.

### 공분산 추가 및 모형 적합

8. "시작점 목록"과 "끝점 목록"에서 모두 Leadership, Conflict 및 Satisfaction 을 선택합니다. 양방향 화살표 ↔ 버튼을 클릭합니다.
9. "모형 규격" 보고서의 왼쪽 위에 있는 "모형 이름" 아래 텍스트 상자에 "3-Factor CFA"를 입력합니다.
10. **실행**을 클릭합니다.
11. (선택 사항) "구조 방정식 모형 : 3-Factor CFA" 옆의 빨간색 삼각형을 클릭하고 **경로 다이어그램 설정 > 레이아웃 > 위쪽에서 아래쪽으로**를 선택합니다.

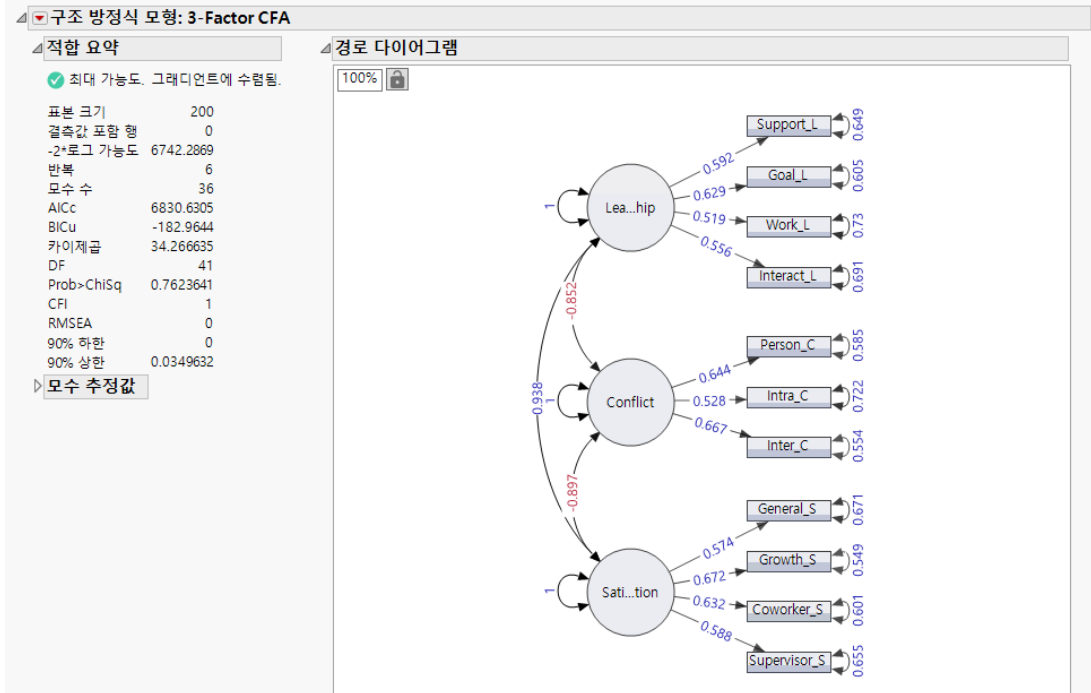
12. (선택 사항) "모수 추정값" 옆의 회색 표시 아이콘을 클릭합니다.

"모수 추정값" 보고서를 닫으면 전체 경로 다이어그램을 볼 수 있습니다.

13. 경로 다이어그램을 마우스 오른쪽 버튼으로 클릭하고 **표시 > 추정값 표시 > 표준화**를 선택합니다.

이제 "경로 다이어그램"의 숫자는 모형에 대한 표준화 모수 추정값을 나타냅니다.

그림 8.2 구조 방정식 모형 보고서

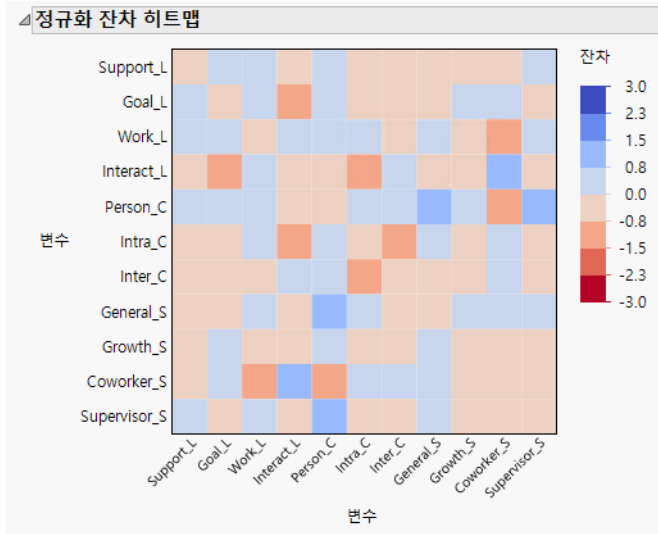


"적합 요약" 보고서에 나열된 이 모형의 카이제곱 통계량은 34.27 이고 자유도는 41 입니다. 해당  $p$  값은 0.7624 로, 유의하지 않습니다. 이는 모형이 잘 적합된다는 귀무가설을 기각할 증거가 없음을 나타냅니다. 따라서 이 모형이 데이터를 적절하게 적합시킨다는 결론을 내릴 수 있습니다.

카이제곱 값은 표본 크기에 따라 달라지므로 일부 적합 모형에서 유의한 카이제곱 값을 생성할 수 있습니다. CFI(비교 적합 지수)와 RMSEA(근사의 제곱근 평균 제곱 오차)는 모형 적합을 결정하는 추가 지표를 제공합니다. 이러한 지수는 0과 1 사이로 제한됩니다. CFI 값이 0.90 보다 크고 RMSEA 값이 0.10 보다 작으면 좋습니다(Browne와 Cudeck 1993, Hu와 Bentler 1999 참조). 여기서 CFI 1과 RMSEA 0은 적합도가 우수하다는 것을 나타냅니다. 이 설문 조사는 리더십, 갈등 및 만족도에 대한 잠재 변수를 측정하는 데 좋은 도구라는 결론을 내릴 수 있습니다.

14. "구조 방정식 모형 : 3-Factor CFA" 옆의 빨간색 삼각형을 클릭하고 **히트맵 > 정규화 잔차 히트맵**을 선택합니다.

그림 8.3 정규화 잔차 히트맵



"정규화 잔차 히트맵"에 양의 방향 또는 음의 방향으로 2.0 단위를 초과하는 값이 없습니다. 이는 모형이 데이터를 잘 적합시킨다는 추가적인 증거입니다. 잔차는 더 세분화된 수준에서 부적합 모형을 진단할 수도 있습니다. 이 모형의 정규화 잔차는 국소적 부적합 증거를 나타내지 않습니다.

## JMP PRO 구조 방정식 모형 플랫폼 시작

분석 > 다변량 방법 > 구조 방정식 모형을 선택하여 구조 방정식 모형 플랫폼을 시작합니다.

그림 8.4 구조 방정식 모형 시작 창

가로형 데이터 형식   요약 데이터 형식

가로형 데이터를 사용하여 구조 방정식 모형을 적합시키기 위한 프레임워크를 제공합니다.

열 선택

- 15개 열
- ID
- Support\_L
- Goal\_L
- Work\_L
- Interact\_L
- Person\_C
- Intra\_C
- Inter\_C
- General\_S
- Growth\_S
- Coworker\_S
- Supervisor\_S
- Leadership\_Avg
- Conflict\_Avg
- Satisfaction\_Avg

선택한 열 역할 지정

모형 변수	필수 숫자 선택적 숫자
그룹	선택적
빈도	선택적 숫자
기준	선택적

작업

확인  
취소  
제거  
재호출  
도움말

잠재 변수 표준화  
 제약 없는(포화) 모형 적합  
 독립 모형 적합

"열 선택"의 빨간색 삼각형 메뉴에 포함된 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

시작 창에는 두 가지 데이터 형식에 대한 탭이 있습니다.

**가로형 데이터 형식** 데이터 테이블의 각 행이 단일 관측값에 해당하고 열에 모형에 사용할 변수가 포함된 경우에 선택합니다. 결측값만 포함된 행은 분석에서 제외됩니다.

**요약 데이터 형식** 데이터 테이블의 데이터가 상관 행렬 또는 공분산 행렬로 요약된 경우에 선택합니다. 평균과 표준편차도 열로 지정할 수 있습니다. 이러한 값을 지정하지 않으면 평균은 0으로 간주되고 표준편차는 입력 행렬 대각선의 제곱근으로 간주됩니다.

### JMP PRO 시작 창 옵션

**모형 변수** 모형에 포함할 열입니다. 하나 이상의 열을 지정해야 합니다. 모든 열에는 숫자 데이터 유형과 연속형 모델링 유형이 있어야 합니다.

요약 데이터 형식의 경우 모형 변수 열은 요약 데이터에 대한 상관 행렬 또는 공분산 행렬을 포함하는 열입니다.

**그룹** ("가로형 데이터 형식"에만 사용 가능) 다중 그룹 분석을 위한 그룹화 변수를 지정하는 열이며, 그룹 간 효과에 대한 등식 제약 조건을 테스트할 수 있습니다. 이 변수는 수준 수가 몇 개인 범주형 변수여야 합니다.

**빈도** 분석의 각 행에 빈도를 할당하는 숫자 값이 포함된 열입니다.

**기준** ("가로형 데이터 형식"에만 사용 가능) 변수의 각 수준에 대한 개별 분석으로 구성된 보고서를 생성하는 열입니다. 기준 변수가 둘 이상 할당되면 기준 변수의 가능한 각 수준 조합에 대해 개별 분석이 생성됩니다.

---

**주의:** 기준 변수를 사용하면 다중 그룹 분석 모형이 적합되지 않습니다. 다중 그룹 분석을 수행하려면 그룹 변수를 지정해야 합니다.

---

**평균** ("요약 데이터 형식"에만 사용 가능) 상관 행렬 또는 공분산 행렬의 각 행에 있는 변수에 해당하는 평균 열입니다. 지정하지 않으면 평균이 0으로 간주됩니다.

**표준편차** ("요약 데이터 형식"에만 사용 가능) 상관 행렬 또는 공분산 행렬의 각 행에 있는 변수에 해당하는 표준편차 열입니다. 지정하지 않으면 표준편차가 입력 행렬 대각선의 제곱근으로 간주됩니다.

**라벨** ("요약 데이터 형식"에만 사용 가능) 상관 행렬 또는 공분산 행렬의 각 행에 있는 변수에 해당하는 라벨 열입니다. 지정하지 않으면 입력 행렬이 포함된 열의 이름이 "모형 규격" 보고서의 변수에 사용됩니다.

**잠재 변수 표준화** 이 옵션을 선택하면 분산을 1로 고정하고 모든 적재를 자유롭게 추정할 수 있도록 하여 잠재 변수의 척도를 설정합니다.

**제약 없는 모형 적합** 이 옵션을 선택하면 첫 번째 모형을 적합시킬 때 제약 없는 모형이 자동으로 적합됩니다. 그런 다음 "모형 비교" 보고서에서 적합 모형을 제약 없는 모형과 비교할 수 있습니다. 제약 없는 모형은 데이터에 구조를 적용하지 않고 지정된 모형 변수의 모든 평균, 분산 및 공분산을 적합시키는 완전 포화 모형입니다.

**독립 모형 적합** 이 옵션을 선택하면 첫 번째 모형을 적합시킬 때 독립 모형이 자동으로 적합됩니다. 그런 다음 "모형 비교" 보고서에서 적합 모형을 독립 모형과 비교할 수 있습니다. 독립 모형은 지정된 모형 변수의 모든 평균과 분산을 적합시킵니다. 지정된 모형 변수에서 모든 공분산이 0으로 고정되므로 매우 제한적인 모형이 됩니다.

**표본 크기** ("요약 데이터 형식"에만 사용 가능) 요약 데이터로 나타내는 관측값 수를 지정합니다.

---

## JMP PRO 구조 방정식 모형 보고서

"구조 방정식 모형" 보고서에는 다음 항목이 포함됩니다.

- "모형 규격 보고서"
- "모형 비교 보고서"
- "카이제곱 차이 검정 보고서"
- "구조 방정식 모형 적합 보고서"

## JMP PRO 모형 규격 보고서

구조 방정식 모형 플랫폼의 "모형 규격" 보고서에는 모형을 지정하기 위한 컨트롤이 포함되어 있습니다. 시작 창에서 "확인" 을 클릭하면 "모형 규격" 보고서의 "다이어그램" 및 "목록" 탭에 기본 독립 모형이 나타납니다. 독립 모형은 지정된 모형 변수의 모든 평균과 분산을 포함합니다.

## JMP PRO 규격 패널

"규격" 패널에는 모형을 생성하기 위한 컨트롤이 있습니다.

**모형 이름** 모형 이름을 지정할 수 있습니다. "작업" 패널에서 "실행" 을 클릭하면 "모형 이름" 의 텍스트를 사용하는 모형 보고서가 생성됩니다.

**시작점 목록** 시작 창에서 지정한 모형 변수와 Constant 항을 나열합니다. Constant 항은 상수에서 각 변수를 회귀하여 관측 변수 또는 잠재 변수의 평균을 추정합니다. 이 목록에서 하나 이상의 변수를 선택하고 "끝점 목록" 에서 하나 이상의 변수를 선택한 후 화살표 버튼 중 하나를 클릭하면 모형에 항이 추가됩니다. 잠재 변수를 생성하면 "시작점 목록" 및 "끝점 목록" 에 추가됩니다. 목록 맨 위에 있는 검색 필터 컨트롤을 사용하여 이 목록의 항목을 필터링할 수 있습니다. 자세한 내용은 "검색 필터 옵션" 에서 확인하십시오.

**끝점 목록** 시작 창에서 지정한 모형 변수와 Constant 항을 나열합니다. Constant 항은 상수에서 각 변수를 회귀하여 관측 변수 또는 잠재 변수의 평균을 추정합니다. "시작점 목록" 에서 하나 이상의 변수를 선택하고 이 목록에서 하나 이상의 변수를 선택한 후 화살표 버튼 중 하나를 클릭하면 모형에 항이 추가됩니다. 잠재 변수를 생성하면 "시작점 목록" 및 "끝점 목록" 에 추가됩니다. 모형에 잠재 변수를 추가하려면 "끝점 목록" 의 변수를 선택해야 합니다. 목록 맨 위에 있는 검색 필터 컨트롤을 사용하여 이 목록의 항목을 필터링할 수 있습니다. 자세한 내용은 "검색 필터 옵션" 에서 확인하십시오.

**단방향 화살표** → "시작점 목록" 과 "끝점 목록" 에서 선택한 변수 간의 관계 유형을 지정합니다. 단방향 화살표는 회귀 효과와 동일합니다.

**양방향 화살표** ↔ "시작점 목록" 과 "끝점 목록" 에서 선택한 변수 간의 관계 유형을 지정합니다. 양방향 화살표는 공분산 효과와 동일합니다.

**잠재 변수 추가** + "시작점 목록" 과 "끝점 목록" 에 잠재 변수를 추가합니다. 변수 이름은 + 버튼 왼쪽에 있는 상자의 텍스트를 기반으로 합니다.

**참고:** + 버튼을 클릭하기 전에 "끝점 목록" 에서 잠재 변수에 대한 표시자를 선택해야 합니다.

**잠재 변수 제거** — "시작점 목록" 또는 "끝점 목록" 에서 현재 선택된 잠재 변수를 제거합니다. "시작점 목록" 과 "끝점 목록" 에서 잠재 변수를 제거하면 모형에서도 잠재 변수가 제거됩니다.

**모형 바로 가기** 공통 모형을 빠르게 생성하기 위한 세 가지 범주의 옵션이 포함되어 있습니다. 이러한 옵션 중 하나를 선택하면 현재 모형 규격이 모두 지워집니다.

**규격 지원** 외생 변수 또는 내생 변수에 대한 공분산 추가와 같은 모형 지정 과정을 단축하는 옵션이 포함되어 있습니다. "잠재 변수 척도 전환" 옵션을 사용하면 잠재 변수의 척도를 변경할 수도 있으며, 이는 확증적 요인 분석을 적합시킬 때 유용합니다.

**공변 외생 변수** 기존 외생 변수 간에 사용 가능한 모든 공분산을 빠르게 지정할 수 있습니다. 외생 변수는 모형의 다른 변수에 의해 예측되지 않는 변수입니다. 대부분의 경우 이러한 변수는 적절한 모형 규격을 위해 다른 변수와의 공변동이 0이 아닐 것으로 예상됩니다.

**공변 내생 변수** 기존 내생 변수 간에 사용 가능한 모든 공분산을 빠르게 지정할 수 있습니다. 내생 변수는 모형에 있는 하나 이상의 다른 변수에 의해 예측되는 변수입니다. 경우에 따라 내생 변수의 잔차 분산은 적절한 모형 규격을 위해 서로 공변할 것으로 예상됩니다. 이 바로 가기는 주로 하나의 결과 집합에 대한 하나의 예측 변수 집합이 있지만 중재 변수나 잠재 변수가 없는 경로 모형에 가장 유용합니다.

**잠재 변수 척도 전환** 모형에서 잠재 변수의 척도를 설정하기 위한 두 가지 기본 방식 간을 전환합니다. 잠재 변수가 있는 모형을 추정하려면 잠재 변수의 척도를 설정해야 합니다. 첫 번째 적재를 1로 고정하거나, 잠재 변수의 분산을 1로 고정하는 두 가지 옵션이 가장 일반적인 척도 설정 방법입니다. 첫 번째 방법은 잠재 변수의 척도를 적재가 1로 고정된 표시자의 척도로 설정합니다. 두 번째 방법은 잠재 변수의 분산이 1이 되도록 하고, 잠재 변수가 z 스코어 측정 기준을 갖도록 평균 0과 쌍을 이룰 수 있습니다.

**측정 오차 수정** 관측된 변수에 대한 오차 분산 추정값을 지정하여 오차 없는 잠재 변수를 모델링할 수 있습니다. 신뢰도 측도 (예: 오메가 계수), 비신뢰도 (1 - 신뢰도) 또는 측정 오차 분산을 사용하여 알려진 오차 분산을 지정할 수 있습니다. 이 옵션을 사용하여 선형 측정 오차 모형을 지정할 수 있습니다.

**횡단면 (전통적)** 확증적 요인 분석, 중재 분석 및 다중 그룹 측정 불변성을 위한 옵션이 포함되어 있습니다.

**확증적 요인 분석** 잠재 변수의 표시자를 선택하고 이름을 사용자 정의할 수 있습니다. 직교 (상관되지 않음) 또는 사각 (상관됨) 규격 중에서 선택할 수 있는 옵션이 제공됩니다.

**중재 분석** 하나 이상의 중재 변수가 예측 변수와 결과 변수 간의 연결에 미치는 영향을 조사할 수 있습니다. 이 바로 가기를 사용하여 원인, 중재, 결과 역할을 지정합니다. 고급 분석의 경우 체크박스를 사용하여 여러 중재 변수를 포함할 수 있습니다.

**다중 그룹 측정 불변성** 기존 잠재 변수를 사용하거나 필요에 따라 재정의하여 여러 그룹에 대한 측정 특성을 테스트합니다. 기본적으로 이 바로 가기를 통해 모든 측정 불변성 모형을 적합시키고 비교할 수 있습니다. 특정한 측정 불변성 모형을 선택할 수도 있습니다. 기본 설정은 모든 모형을 포괄적으로 분석하므로 데이터 표현에 가장 적절한 모형을 식별하는 데 도움이 됩니다.

**중단 분석** 일반적인 잠재 성장 곡선 모형을 위한 옵션이 포함되어 있습니다. 잠재 성장 곡선 모형은 시간에 따른 관측 단위의 궤적을 특성화하는 데 사용되는 중단 모형의 한 유형입니다. 이러한 모형의 표준 규격을 사용하면 모집단의 평균 궤적 추정값과 해당 평균으로부터의 산포 측도를 얻을 수 있습니다. 절편만 있는 잠재 성장 곡선, 선형 잠재 성장 곡선, 2차 잠재 성장 곡선 및 잠재 변수 기반 성장 곡선이라는 네 가지 공통 성장 곡선 바로 가기가 제공됩니다.

**참고:** 잠재 성장 곡선 옵션 중 하나를 선택할 때 "끝점 목록"의 변수를 선택하면 선택된 변수에만 모형이 적용됩니다. 그렇지 않으면 "끝점 목록"의 모든 변수에 모형이 적용됩니다.

**시간 값 정의** 수동으로 또는 여러 옵션을 사용하여 중단 모형에서 각 반복 측정에 대한 시간 값을 입력할 수 있습니다. "이름으로 추측" 옵션은 변수 이름에 나올 수 있는 숫자를 기반으로 시간 값을 할당합니다. "모두 증가" 및 "모두 감소" 옵션은 모든 시간 값을 1씩 증분하여 균일하게 조정합니다. "오름차순" 또는 "내림차순" 옵션은 0에서 시작하거나 끝나는 값이 오름차순 또는 내림차순 시간 순서를 따르도록 지정합니다. "모두 지우기" 옵션은 이전에 입력한 시간 값을 제거합니다. 시간 값을 정의하는 것은 중단 모형 규격에 반드시 필요하며, 정확한 분석을 위해 계열이 오름차순 시간으로 올바르게 정렬되고 적절하게 그룹화되게 합니다.

**절편만 있는 잠재 성장 곡선** 관측 단위에 대한 평평한 궤적을 나타내는 모형을 지정합니다. 이 모형에는 주어진 과정에 대한 절편으로 해석할 수 있는 하나의 잠재 변수가 있습니다. 절편에는 추정된 평균과 분산이 있으며, 이는 각각 모집단의 평균 과정 수준과 해당 평균으로부터 개별 단위의 이탈을 나타냅니다. 이 모형은 주로 다른 잠재 성장 곡선 모형과의 비교를 위한 기준으로 사용됩니다. 예를 들어 선형 잠재 성장 곡선 모형이 절편만 있는 모형보다 더 잘 적합되는 경우 선형 추세를 설정할 수 있습니다.

**선형 잠재 성장 곡선** 관측 단위에 대한 선형 궤적을 나타내는 모형을 지정합니다. 이 모형에는 절편과 선형 기울기라는 두 개의 잠재 변수가 있습니다. 이러한 각 잠재 변수에는 평균 수준을 해석할 수 있는 평균과 전체 관측 단위에 대한 변화 및 해당 평균으로부터의 이탈을 해석할 수 있는 분산이 있습니다. 모든 잠재 변수 또는 성장 요인은 공변할 수 있습니다. 기울기의 첫 번째 적재는 0으로 고정되어 첫 번째 시점에서 절편을 중앙에 맞춥니다. 이전에 "시간 값 정의" 모형 바로 가기를 사용한 경우를 제외하고 이 바로 가기에서는 변수가 시간 오름차순으로 나열된다고 가정합니다. 즉, 기울기 요인에 대한 적재가 0에서 시작하고 각 변수에 대해 1단위씩 증가합니다.

**2차 잠재 성장 곡선** 관측 단위에 대한 2차 궤적을 나타내는 모형을 지정합니다. 이 모형에는 절편, 선형 기울기 및 2차 기울기라는 세 개의 잠재 변수가 있습니다. 이러한 각 잠재 변수에는 평균 2차 궤적을 해석할 수 있는 평균과 해당 평균으로부터의 개별 이탈을 해석할 수 있는 분산이 있습니다. 모든 잠재 변수 또는 성장 요인은 공변할 수 있습니다. 선형 및 2차 기울기의 첫 번째 적재는 0으로 고정되어 첫 번째 시점에서 절편을 중앙에 맞춥니다. 이 바로 가기에서는 변수가 시간 오름차순으로 나열된다고 가정합니다. 즉, 선형 기울기 요인에 대한 적재가 0에서 시작하고 각 변수에 대해 1단위씩 증가합니다." 시간 값 정의 " 모형 바로 가기를 사용하여 시간 값을 지정한 경우 해당 값이 기울기 지정에 사용됩니다. 2차 기울기 요인에 대한 적재는 선형 기울기 요인의 적재를 제공한 값입니다.

**잠재 변수 기반 성장 곡선** 알려진 함수 형태가 없는 성장 곡선을 지정합니다. 이 모형은 주로 비선형 궤적에 사용됩니다. 이 모형에는 절편과 비선형 기울기라는 두 개의 잠재 변수가 있으며, 각각 평균 궤적을 해석할 수 있는 평균과 해당 평균으로부터의 개별 이탈을 해석할 수 있는 분산을 갖습니다. 이러한 잠재 변수 또는 성장 요인은 공변할 수 있습니다. 비선형 기울기의 첫 번째 적재는 0으로 고정되어 첫 번째 시점에서 절편을 중앙에 맞춥니다. 비선형 기울기의 마지막 적재는 1로 고정되며, 모든 중간 적재가 자유롭게 추정됩니다. 이 규격을 사용하면 적재 추정값을 변화의 비율로 해석할 수 있습니다. 이전에 " 시간 값 정의 " 모형 바로 가기를 사용한 경우를 제외하고 이 바로 가기에서는 변수가 시간 오름차순으로 나열된다고 가정합니다.

**성장 모형 적합 및 비교** (다중 그룹 분석에는 사용 불가능) 카이제곱 차이 검정을 사용하여 절편만 있는 모형, 선형 성장 모형 및 2차 성장 곡선 모형의 적합을 비교할 수 있습니다. 종단 분석의 공통된 목표는 데이터의 변화 패턴을 특성화하는 것입니다. 이 세 모형의 적합을 비교하면 데이터에 가장 적합한 궤적 유형을 식별하여 목표를 이룰 수 있습니다. 특정 모형 결과를 검토하기 전에 최량 적합 모형을 식별할 수 있도록 모든 적합 모형 보고서가 기본적으로 축소되어 있습니다.

**다변량 잠재 성장 곡선** 다변량 성장 곡선 모형을 빠르게 지정할 수 있습니다. 주어진 과정에 대한 변수 부분집합과 해당 과정에 지정할 성장 곡선 유형을 선택할 수 있습니다. + 버튼을 사용하면 지정된 각 단변량 모형을 추가할 수 있습니다. 기본적으로 모든 잠재 변수는 공변할 수 있으므로 성장 요인 간의 연관성을 추론할 수 있습니다. 이전에 " 시간 값 정의 " 모형 바로 가기를 사용한 경우를 제외하고 이 바로 가기에서는 각 과정의 변수가 시간 오름차순으로 나열된다고 가정합니다.

**교차 시차 패널 모형** 패널 데이터를 분석하고, 수동 지정을 통해 복잡하고 시간이 많이 걸리는 선행 - 후행 관계를 풀 수 있습니다. 이 바로 가기는 열 이름을 기반으로 자동으로 과정을 식별하고 필요에 따라 수정할 수 있습니다. 이는 과정의 시간 안정성을 시험하기 위해 등식 제약 조건 적용을 간소화합니다. 이 바로 가기는 효율적이고 정확한 모형 적합 과정에 기여하는 자기회귀 및 교차 시차 효과와 잔차 분산, 공분산에 대해 등식 제약 조건이 필요한지 판단하기 위해 카이제곱 차이 검정을 사용하여 모형의 지정, 적합, 비교를 지원합니다.

**측정 불변성** 기준 형태 모형을 적합시키고, 잠재 변수의 측정 특성을 테스트하기 위해 제약 조건을 점진적으로 추가하여 여러 시점에 걸쳐 측정의 일관성을 확인할 수 있습니다. 시간 값을 정의하는 것은 측정 불변성 바로 가기의 사용과 정확한 모형 규격을 위해 매우 중요합니다. 자세한 내용은 "**시간 값 정의**"에서 확인하십시오. 단일 모형을 지정하거나, 포괄적 분석을 위해 모든 모형 (형태 동일성, 약한 동일성, 강한 동일성, 엄격한 동일성)을 적합시키고 비교하도록 선택할 수 있습니다. "수동 계열 정의" 옵션으로 열의 시간 순서를 사용자 정의할 수 있습니다. 이 바로 가기를 사용하면 집중적인 모형 비교와 측정 불변성에 대한 상세한 평가가 가능합니다.

## JMP PRO 작업 패널

"작업" 패널의 버튼은 "목록" 탭의 지정된 모형 변수 목록에서 작동합니다. 다음과 같은 작업을 수행할 수 있습니다.

**모수 고정** 현재 선택된 효과의 모수 값을 상수 값으로 고정할 수 있습니다. 변수의 모수 값을 고정하면 항 이름 뒤의 괄호 안에 고정 값이 표시됩니다.

---

**참고:** 기본값 0을 유지하고 "확인"을 클릭하면 선택한 효과가 모형에서 제거됩니다.

---

**등가 설정** 두 개 이상 선택한 효과의 모수 값이 같도록 제한할 수 있습니다. 둘 이상의 효과에 대한 모수 값이 같도록 설정하면 항 이름 뒤의 괄호 안에 임의의 영숫자 라벨 ("c1")이 나타납니다. 모수 값이 서로 같도록 설정된 효과 집합이 여러 개인 경우 번호가 순차적으로 지정됩니다 ("c1", "c2" 등).

---

**참고:** 동등 제약 조건은 유형이 동일한 모수 내에서만 허용됩니다.

---

**제약 해제** 모형에서 선택한 효과에 적용된 제한을 제거할 수 있습니다. 제한에는 고정 값으로 설정된 효과 또는 다른 효과와 동일하게 설정된 효과가 포함됩니다.

**제거** 선택한 효과를 모형에서 제거합니다.

**실행 취소** 모형에 대한 마지막 수정 작업을 실행 취소합니다.

**다시 실행** 모형에 대해 마지막으로 실행 취소한 수정 작업을 다시 실행합니다.

**실행** 현재 지정된 모형을 적합시키고 보고서를 생성합니다.

**재설정** 모형 규격을 독립 모형 (기본값)으로 재설정합니다.

## JMP PRO 다이어그램 패널

"다이어그램" 패널에는 모형 다이어그램 레이아웃을 조정할 수 있는 버튼이 포함되어 있습니다. 첫 번째 버튼을 사용하면 다이어그램에서 직사각형으로 표시된 명시 변수를 회전할 수 있습니다. 두 번째 버튼을 사용하면 다이어그램의 두 가지 다른 배열을 순환할 수 있습니다. 세 번째 버튼을 사용하여 다이어그램을 사용자 정의할 수 있습니다. 자세한 내용은 "**경로 다이어그램 사용자 정의**"에서 확인하십시오.

## JMP PRO 상세 정보 패널

"상세 정보" 패널에는 현재 "모형 규격" 보고서에 지정된 모형에 대한 정보가 포함되어 있습니다. 모형에 대한 명시 변수의 수, 잠재 변수의 수, 자유 추정 모수의 수 및 자유도 정보를 제공합니다. 최대 반복 수를 조정할 수도 있습니다.

## JMP PRO 다이어그램 탭

"다이어그램" 탭에는 현재 지정된 모형을 시각화할 수 있는 모형 다이어그램이 포함됩니다. 잠재 변수는 원으로 표시되고 명시 변수는 직사각형으로 표시됩니다. 단방향 화살표는 적재 및 회귀를 나타냅니다. 양방향 화살표는 분산 및 공분산을 나타냅니다. 분산은 해당 변수 자신에게 돌아오는 양방향 화살표로 지정됩니다. 공분산은 두 변수 사이의 양방향 화살표로 지정됩니다.

경로 다이어그램은 대화식이므로 항목을 드래그하여 배열할 수 있습니다. 화살표 키를 사용하여 경로 다이어그램에서 선택한 항목을 이동할 수도 있습니다. "다이어그램" 패널의 버튼을 사용하여 경로 다이어그램의 항목을 재배열할 수 있습니다. 경로 다이어그램 팝업 메뉴의 옵션을 사용하여 경로 다이어그램에서 항목을 추가하거나 제거할 수 있습니다. 경로 다이어그램 자체를 마우스 오른쪽 버튼으로 클릭하면 경로 다이어그램 팝업 메뉴가 나타납니다. 팝업 메뉴 내용은 빈 캔버스 부분을 클릭했는지 아니면 경로 다이어그램 요소를 클릭했는지에 따라 다릅니다. 경로 다이어그램 왼쪽 위에 있는 확대/축소 설정을 조정하여 경로 다이어그램을 확대하거나 축소할 수도 있습니다. 또한 경로 다이어그램에는 경로 다이어그램에서 항목 위치를 잠글 수 있는 "잠금" 버튼이 왼쪽 위에 있습니다. 경로 다이어그램이 잠기면 "잠금" 버튼이 파란색입니다. 경로 다이어그램에서 선택한 후 팝업 메뉴에서 **선택 > 선택 항목 숨기기** 옵션을 선택하여 해당 항목을 숨길 수 있습니다. 경로 다이어그램의 항목이 숨겨지면 숨겨진 모든 항목을 표시할 수 있는 "모두 숨기기 해제" 버튼이 경로 다이어그램 맨 위에 나타납니다.

**참고 :** 시작 창에서 "그룹" 변수를 지정하면 "다이어그램" 탭이 여러 그룹 탭 안에 내포됩니다. "다이어그램" 탭 왼쪽의 패널 탭을 사용하여 이러한 탭 간에 이동할 수 있습니다. 그룹 모형 집합을 표시하는 "결합" 탭과 "그룹" 변수의 각 수준에 대한 패널 탭이 있습니다.

경로 다이어그램 사용자 정의 옵션에 대한 자세한 내용은 "경로 다이어그램 사용자 정의"에서 확인하십시오. 자세한 내용은 "다이어그램 패널", "상세 정보 패널" 및 "경로 다이어그램 설정"에서 확인하십시오.

**팁 :** 경로 다이어그램을 이미지로 복사하려면 다이어그램을 마우스 오른쪽 버튼으로 클릭하고 **다이어그램 복사**를 선택합니다. 최상의 품질을 유지하려면 클립보드 이미지를 벡터 그래픽으로 붙여 넣습니다.

## JMP PRO 목록 탭

"목록" 탭에는 모형의 각 변수 유형에 대한 목록이 포함됩니다. 이러한 목록은 모형 다이어그램에 나타내는 데 사용되는 화살표로 분류됩니다. 단방향 화살표는 평균이나 절편, 적재 및 회귀를

지정하는 데 사용됩니다. 양방향 화살표는 분산 및 공분산을 지정하는 데 사용됩니다. "목록" 탭의 "분산" 및 "공분산" 섹션에서 양방향 화살표를 커서로 가리키면 잔차에 분산 또는 공분산이 있는지 여부를 나타내는 톨팁이 표시됩니다.

**참고:** 시작 창에서 "그룹" 변수를 지정하면 "목록" 탭이 여러 그룹 탭 안에 내포됩니다. "목록" 탭 왼쪽의 패널 탭을 사용하여 이러한 탭 간에 이동할 수 있습니다. 그룹 모형 집합을 표시하는 "결합" 탭과 "그룹" 변수의 각 수준에 대한 패널 탭이 있습니다.

각 목록 맨 위에 있는 검색 필터 컨트롤을 사용하여 각 목록의 항목을 필터링할 수 있습니다. 자세한 내용은 "[검색 필터 옵션](#)"에서 확인하십시오.

## JMP PRO 상태 탭

"상태" 탭에는 모형 식별성 확인이 포함됩니다. 신뢰할 수 있는 추정값을 얻으려면 구조 방정식 모형을 식별해야 합니다. SEM 컨텍스트에서 식별은 입력 데이터의 모집단 공분산 행렬을 기반으로 모형의 모든 모수에 대한 고유 해를 도출할 수 있음을 의미합니다. SEM 프레임워크에서 다양한 모형을 지정할 수 있으므로 식별 규칙도 다양합니다. "상태" 탭은 "식별 규칙", "모형 상세 정보" 및 "데이터 상세 정보"라는 세 가지 정보 패널로 구성됩니다.

- "식별 규칙" 패널에는 지정된 모형에 적합한 규칙 (최대 8 개) 목록이 포함됩니다. 모형 식별을 보장하기 위한 필요 충분 규칙이 있습니다. 필요 규칙이 실패할 경우 모형을 적합시키기 전에 규칙을 수정하는 단계를 수행해야 합니다. 충분 규칙이 실패할 경우 문제를 반드시 해결할 필요는 없습니다. 경우에 따라 충분 규칙이 모형에 문제가 있다는 증거 없이 실패할 수 있습니다. 모든 규칙은 양정치 공분산 행렬을 가정합니다. 공분산 행렬이 양정치가 아니면 "모형 상세 정보" 패널 아래에 경고가 나타납니다.

**팁:** 특정 식별 규칙에 대한 자세한 내용을 보려면 규칙 테이블의 행을 클릭하십시오. 그러면 테이블 오른쪽에 규칙 설명이 나타납니다.

- "모형 상세 정보" 패널에는 현재 지정된 모형을 기술하는 값 목록이 포함됩니다. 이러한 값을 사용하여 모형의 잠재적 문제를 감지할 수 있습니다.
- "데이터 상세 정보" 패널에는 입력 데이터를 기술하는 값 목록이 포함됩니다. 이러한 값을 사용하여 데이터의 잠재적 문제를 감지할 수 있습니다.

지정된 데이터 열에 특이성이 있으면 보고서에 "특이성 상세 정보" 테이블이 포함됩니다.

**참고:** 시작 창에서 "그룹" 변수를 지정하면 "상태" 탭이 여러 그룹 탭 안에 내포됩니다. "상태" 탭 왼쪽의 패널 탭을 사용하여 이러한 탭 간에 이동할 수 있습니다. 그룹 모형 집합을 표시하는 "결합" 탭과 "그룹" 변수의 각 수준에 대한 패널 탭이 있습니다.

"상태" 탭 자체에는 지정된 모형의 현재 상태를 보여 주는 동적 아이콘이 있습니다. 아이콘은 다음과 같이 상태를 지정합니다.

- ✔ 적용 가능한 모든 식별 규칙을 통과합니다.
- ✘ 하나 이상의 필수 식별 규칙이 충족되지 않으며 모형을 적합시키기 전에 문제 해결 단계를 수행해야 합니다.
- ⓘ 필수가 아닌 식별 규칙 중 하나 이상이 충족되지 않으며 모형이 올바르게 지정되었는지 여부를 판별하기 위해 추가 조사가 필요합니다. SEM 고급 응용 프로그램에서 이러한 상황이 종종 발생하지만 모형에 문제가 있다는 의미는 아닙니다. 이는 식별 규칙이 모형 식별을 보장할 수 없음을 나타냅니다.

그림 8.5 상태 탭의 예



**JMP PRO** 검색 필터 옵션

"모형 규격" 보고서의 "시작점 목록", "끝점 목록" 및 "" 목록" 탭의 목록 상자에는 특정 목록의 항목을 필터링할 수 있는 검색 필터가 포함되어 있습니다.

검색 상자 옆의 아래쪽 화살표 버튼을 클릭하여 검색을 구체화할 수 있습니다.

**용어 포함** 검색 기준의 일부가 포함된 항목을 반환합니다. "ease oom" 을 검색하면 "Release Zoom" 과 같은 메시지가 반환됩니다.

**구 포함** 검색 기준이 정확히 포함된 항목을 반환합니다. "text box" 를 검색하면 "text" 와 "box" 가 바로 연이어 포함된 항목 ( 예 : "Context Box" 및 "Text Box") 이 반환됩니다.

**구로 시작** 검색 기준으로 시작하는 항목을 반환합니다.

**구로 끝남** 검색 기준으로 끝나는 항목을 반환합니다.

**전체 구** 전체 문자열로 구성된 항목을 반환됩니다. "text box" 를 검색하면 "text box" 만 포함된 항목이 반환됩니다.

**정규 표현식** 검색 상자에서 와일드카드 (\*) 와 마침표 (.) 를 사용할 수 있습니다. "get.\*name" 을 검색하면 "get" 다음에 하나 이상의 단어가 포함된 항목을 찾을 수 있습니다. 즉, "Get Color Theme Names", "Get Name Info" 및 "Get Effect Names" 등이 반환됩니다.

**결과 반전** 검색 기준과 매칭되지 않는 항목을 반환합니다.

**모든 용어 일치** 문자열이 모두 포함된 항목을 반환합니다. "t test" 를 검색하면 검색 문자열 중 하나 또는 둘 모두가 포함된 요소 ( 예 : "Pat Test", "Shortest Edit Script" 및 "Paired t test") 가 반환됩니다.

**대 / 소문자 무시** 검색 기준의 대 / 소문자를 무시합니다.

**전체 단어 일치** "모든 용어 일치" 설정에 따라 문자열의 각 단어가 포함된 항목을 반환합니다. "모든 용어 일치" 옵션이 선택되어 있는 경우 "data filter" 를 검색하면 "data" 와 "filter" 가 반환됩니다.

## JMP PRO 모형 비교 보고서

구조 방정식 모형 플랫폼의 "모형 비교" 보고서에는 적합된 모든 모형에 대한 테이블이 포함됩니다. 두 번째 열의 아이콘을 사용하여 "모형 비교" 보고서 아래에 나타나는 모형 보고서를 제어할 수 있습니다. "모형 이름" 오른쪽의 나머지 열을 사용하면 다양한 기준에 따라 모형을 비교할 수 있습니다.

모형의 "AICc 가중치" 값은 적합 모형 중 하나가 사실일 때 특정 모형이 참모형일 확률로 해석할 수 있습니다. 따라서 AICc 가중치가 1에 가장 가까운 모형이 최량 적합 모형입니다. AICc 가중치는 비결측 AICc 값만 사용하여 계산되고 다음과 같이 정의됩니다.

$$\text{AICcWeight} = \exp[-0.5(\text{AICc} - \min(\text{AICc}))] / \sum(\exp[-0.5(\text{AICc} - \min(\text{AICc}))])$$

여기서  $\min(\text{AICc})$  는 적합 모형 중 가장 작은 AICc 값입니다.

"모형 비교" 보고서의 다른 기준에 대한 자세한 내용은 "[구조 방정식 모형 적합 보고서](#)"에서 확인하십시오.

**참고:** 모형이 수렴되지 않으면 "모형 비교" 보고서의 모형 이름 시작 부분에 별표가 표시됩니다.

적합 모형 성능에 대한 컨텍스트를 제공하기 위해 다음 두 모형이 "모형 비교" 보고서에 기본적으로 표시됩니다.

**제약 없음 (포화)** 데이터에 구조를 적용하지 않고 지정된 모형 변수의 모든 평균, 분산 및 공분산을 적합시킵니다.

**독립** 지정된 모형 변수의 모든 평균과 분산을 적합시키고 모든 공분산을 0 으로 고정합니다.

### 모형 비교 테이블 옵션

테이블의 행을 선택한 후 "모형 비교" 테이블의 팝업 메뉴에서 다음 옵션을 사용할 수 있습니다.

**모형 이름 바꾸기** (테이블에서 " 제약 없음 (포화)" 또는 " 독립 " 이외의 모형을 선택한 경우에만 해당) 이미 적합된 모형의 이름을 변경할 수 있습니다. "모형 비교" 테이블에서 모형 이름을 변경하면 보고서의 해당 개요에 있는 모형 이름도 업데이트됩니다.

**독립 모형으로 설정** (테이블에서 " 제약 없음 (포화)" 또는 " 독립 " 이외의 모형을 선택한 경우에만 해당) 보고서에서 독립 모형을 선택한 모형으로 바꿉니다. 이 옵션은 특히 종단 모형에 유용합니다. 독립 모형을 변경하면 "모형 비교" 테이블의 비교 적합 지수가 업데이트됩니다.

---

**주의 :** 잘못된 독립 모형을 지정하면 적합 지수가 유효하지 않을 수 있습니다.

---

**기본 독립 모형 재설정** (기본 독립 모형이 아닌 모형이 독립 모형으로 설정된 경우에만 사용 가능) 모형 비교를 위해 기본 독립 모형을 독립 모형으로 복원합니다. 이 옵션을 사용하면 "모형 비교" 테이블의 비교 적합 지수도 원래 값으로 업데이트됩니다.

### 모형 비교 보고서 옵션

"모형 비교" 테이블 아래에 두 가지 옵션이 나타납니다. 이러한 옵션은 "모형 비교" 테이블에서 행을 클릭하여 강조 표시한 후 사용할 수 있습니다.

**선택한 모형 비교** (테이블에서 두 개 이상의 행이 강조 표시된 경우에만 사용 가능) 선택한 행의 모든 내포 모형 조합 사이의 내포 카이제곱 차이 검정을 계산합니다.

---

**팁 :** 선택한 모형이 내포되지 않은 경우 경고가 나타나고 내포되지 않은 모형 조합은 "카이제곱 차이 검정" 보고서에 표시되지 않습니다.

---

**선택 해제** (테이블에서 하나 이상의 행이 강조 표시된 경우에만 사용 가능) 테이블 행 선택을 모두 지웁니다.

## JMP PRO 카이제곱 차이 검정 보고서

구조 방정식 모형 플랫폼의 "카이제곱 차이 검정" 보고서에는 내포 카이제곱 검정 테이블이 포함됩니다. 이 테이블에는 모형을 정의하는 두 열이 있습니다. 첫 번째 열은 두 모형 중 더 제한된 모형을 포함하고 두 번째 열은 두 모형 중 덜 제한된 모형을 포함합니다. 더 작은 모형이 더 큰 모형 안에 내포됩니다. 나머지 열에는 카이제곱 값, 자유도, CFI 및 RMSEA의 차이와 내포 카이제곱 검정의  $p$  값이 표시됩니다. 열 이름의  $\Delta$  표기는 차이를 나타냅니다. 유의한  $\Delta$  카이제곱 값은 내포 모형의 추가 제약 조건으로 인해 통계적으로 유의한 부적합 증가가 발생하므로 덜 제한된 모형을 유지해야 함을 나타냅니다. 카이제곱 검정은 표본 크기의 영향을 받습니다. 즉, 표본 크기가 증가할수록 더 유의할 가능성이 높기 때문에 " $\Delta$ CFI"와 " $\Delta$ RMSEA"도 고려해야 합니다. 이상적으로는  $\Delta$ CFI가  $-0.01$ 을 초과하지 않고  $\Delta$ RMSEA가  $0.015$ 를 초과하지 않아야 합니다(Chen 2007 참조).

**주의:** 이 보고서의 차이 검정은 내포 모형에서만 의미가 있습니다.

빨간색 X 버튼을 클릭하여 보고서의 행을 제거할 수 있습니다. 보고서를 완전히 제거하려면 테이블의 모든 행을 제거해야 합니다. 주 보고서에서 "구조 방정식 모형" 노드를 제거하면 해당 모형이 포함된 모든 차이 검정이 테이블에서 제거됩니다.

## JMP PRO 구조 방정식 모형 적합 보고서

"모형 규격" 보고서에서 "실행" 을 클릭할 때마다 지정된 모형에 대한 "구조 방정식 모형" 보고서가 나타납니다. 기본적으로 이 보고서에는 "적합 요약" 보고서, "모수 추정값" 보고서 및 "경로 다이어그램" 이 포함됩니다.

**참고:** 시작 창에서 "그룹" 변수를 지정하면 그룹 변수의 각 수준에 대한 개별 모형 적합 보고서가 있습니다. "구조 방정식 모형" 적합 보고서 개요 상자 위의 그룹 탭을 사용하여 이러한 보고서 간에 이동할 수 있습니다.

**적합 요약** 수렴 상태 및 추정 방법을 포함한 모형 적합 정보 테이블입니다. 시작 창에서 "그룹" 변수를 지정하면 이 테이블에 현재 모형 적합 탭의 그룹 변수 수준만 참조하는 통계량에 대한 두 번째 열이 포함됩니다. 다음 통계량이 이 테이블에 보고됩니다.

**표본 크기** 모형을 적합시키는 데 사용된 관측값 (행) 수입니다.

**결측값 포함 행** 결측값이 하나 이상 포함된 관측값 (행) 수입니다. 모든 결측값은 완전 정보 최대 가능도를 사용하여 처리됩니다 (Finkelstein 1979 참조).

**-2\* 로그 가능도** 적합 모형의 로그 가능도에 -2 를 곱한 값입니다. 이 값은 내포 모형을 비교하는 데 사용할 수 있습니다. 두 모형의 -2\* 로그 가능도 값 차이는 자유도가 모형 간의 자유도 차이와 동일한 카이제곱 분포입니다. 자세한 내용은 **선형 모형 적합**에서 확인하십시오.

**반복** 모형을 적합시키는 데 사용된 반복 수입니다.

**모수 수** 모형에서 자유 추정 모수의 수입니다.

**AICc** 수정 Akaike 정보 기준입니다. 이 값은 모형을 비교하는 데 사용할 수 있으며 값이 작을수록 더 나은 모형 적합을 나타냅니다. 자세한 내용은 "**AICc, BIC 및 BICu**" 에서 확인하십시오.

**BICu** BICu (제약 없는 모형에 대한 BIC) 는 베이지안 정보 기준을 재구성한 것입니다. BICu 는 제약 없는 모형과의 비교로 정의됩니다. 음수 값은 적합 모형을 뒷받침하고 양수 값은 제약 없는 모형을 뒷받침합니다. 다른 정보 기준과 마찬가지로 이 값을 사용하여 모형을 비교할 수 있으며, 두 모형 사이에서 값이 작을수록 모형이 더 적합하다는 것을 나타냅니다. 자세한 내용은 "**AICc, BIC 및 BICu**" 에서 확인하십시오.

**카이제곱** 모형에 대한 카이제곱 통계량입니다.

**DF** 모형 적합에 대한 카이제곱 검정의 자유도입니다.

**Prob>ChiSq** 모형에 대한 카이제곱 통계량의  $p$  값입니다.

**CFI** Bentler CFI(비교 적합 지수)는 모형 적합을 결정하는 추가 지표를 제공합니다. CFI는 0 과 1 사이로 제한됩니다. 0.90 보다 큰 값이 좋습니다 (Browne 과 Cudeck 1993, Hu 와 Bentler 1999 참조). 자세한 내용은 "CFI" 에서 확인하십시오.

**RMSEA** RMSEA(근사의 제공근 평균 제공 오차)는 모형 적합을 결정하는 추가 지표를 제공합니다. RMSEA는 0 과 1 사이로 제한됩니다. 0.10 보다 작은 값이 좋습니다 (Browne 과 Cudeck 1993, Hu 와 Bentler 1999 참조). 자세한 내용은 "RMSEA" 에서 확인하십시오.

**90% 하한** RMSEA 에 대한 90% 신뢰 하한입니다. 자세한 내용은 "RMSEA" 에서 확인하십시오.

**90% 상한** RMSEA 에 대한 90% 신뢰 상한입니다. 자세한 내용은 "RMSEA" 에서 확인하십시오.

**모수 추정값** 모형 모수에 대한 추정값 테이블입니다. 이 테이블은 "평균 / 절편", "적재", "회귀" 및 "분산" 섹션으로 구성됩니다. 각 추정값에 대해 표준 오차 ("표준 오차"), Wald 검정 통계량 ("Wald Z") 및 해당  $p$  값 ("Prob>|Z|") 이 제공됩니다. 시작 창에서 "그룹" 변수를 지정하면 이 테이블에 현재 모형 적합 탭의 그룹 변수 수준에 대한 모수 추정값만 포함됩니다.

**팁:** "모수 추정값" 테이블에는 숨겨진 열이 더 포함되어 있습니다. 이러한 열을 표시하려면 테이블을 마우스 오른쪽 버튼으로 클릭하고 열 하위 메뉴에서 추가 열을 선택합니다.

**경로 다이어그램** 적합 모형의 경로 다이어그램 표현을 표시합니다. 자세한 내용은 "다이어그램 탭" 에서 확인하십시오. 시작 창에서 "그룹" 변수를 지정하면 이 다이어그램에 현재 모형 적합 탭의 그룹 변수 수준만 나타납니다.

---

## JMP PRO 구조 방정식 모형 플랫폼 옵션

"구조 방정식 모형" 의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

**경로 다이어그램 설정** 경로 다이어그램 표시를 수정하기 위한 다음 옵션을 포함합니다.

**다이어그램 사용자 정의** 경로 다이어그램의 여러 측면을 사용자 정의할 수 있습니다. 자세한 내용은 "경로 다이어그램 사용자 정의 옵션" 에서 확인하십시오.

**레이아웃** 경로 다이어그램의 전체 모양을 변경하는 두 가지 옵션을 포함합니다. "왼쪽에서 오른쪽으로" 레이아웃 또는 "위쪽에서 아래쪽으로" 레이아웃 중에서 선택할 수 있습니다.

**팁:** 경로 다이어그램에서 항목을 드래그하여 특정 항목의 배열을 변경할 수도 있습니다.

**다이어그램 복사** 경로 다이어그램 이미지를 클립보드에 저장합니다. 최상의 품질을 유지하려면 클립보드 이미지를 벡터 그래픽으로 붙여 넣습니다.

**다이어그램 특성 복사** 현재 경로 다이어그램 특성을 클립보드에 복사합니다. 그런 다음 특성을 다른 SEM 경로 다이어그램에 붙여 넣을 수 있습니다.

**다이어그램 특성 붙여넣기** 클립보드의 경로 다이어그램 특성을 현재 SEM 경로 다이어그램에 붙여 넣습니다.

**기술 통계량** 기술 통계량을 생성하기 위한 다음 옵션을 포함합니다.

**다변량 단순 통계량** 다변량 단순 통계량 보고서를 표시하거나 숨깁니다. 이 보고서의 통계량은 결측 데이터가 있을 수 있는 다른 열과 독립적으로 각 열에 대해 추정됩니다.

**완전 정보 다변량 통계량** 다변량 단순 통계량 보고서를 표시하거나 숨깁니다. 이 보고서의 통계량은 결측 데이터를 설명하기 위해 완전 정보 최대 가능도를 사용하여 추정됩니다.

**추론** 비정규 데이터를 모델링하기 위한 로버스트 추론 및 붓스트랩 추론을 생성하는 다음 옵션을 포함합니다.

**로버스트 추론** ("가로형 데이터 형식"에만 사용 가능) 로버스트 적합 통계량과 ML(최대 가능도) 및 FIML(완전 정보 최대 가능도) 모수 추정값에 대한 샌드위치 표준 오차를 계산합니다. 이 옵션은 설문 조사 데이터나 순서형 또는 비정규 분포의 데이터 집합과 같이 정규 분포가 아닌 연속형 데이터에 특히 유용합니다.

**붓스트랩 추론** 비정규 데이터 및 중재 모형의 간접 효과와 같이 일반적으로 정규 분포를 따르지 않는 효과 검정을 위한 붓스트랩 추정값을 계산합니다. 보고서에서 붓스트랩 추정값을 구할 통계량을 선택할 수 있습니다. 카이제곱 통계량에 대한 편향 수정 신뢰 구간 및 Bollen-Stine 변환을 사용할 수 있습니다. 붓스트랩 신뢰 구간의 유의 수준, 붓스트랩 표본 수 및 단순 시드값(선택적)을 사용자 정의할 수 있습니다. "붓스트랩 추론" 보고서에는 붓스트랩 추정값, 표준 오차, 모수 추정값의 신뢰 구간이 포함됩니다.

**이상치 탐색 시작** 이상치 탐색 플랫폼을 시작합니다. 자세한 내용은 **예측 및 전문 모델링**에서 확인하십시오.

**결측값 탐색 시작** (결측값이 없는 경우에는 사용 불가능) 결측값 탐색 플랫폼을 시작합니다. 자세한 내용은 **예측 및 전문 모델링**에서 확인하십시오.

**명시 변수 추가** 현재 모형에 새 명시 변수를 추가할 수 있습니다. 추가할 새 명시 변수를 선택하면 새 "구조 방정식 모형" 보고서가 시작됩니다. 새 보고서에는 명시 변수가 새로 추가된 기존 모형 규격이 사용됩니다.

**명시 변수 제거** 지정된 명시 변수가 제거된 새 "구조 방정식 모형" 보고서를 시작할 수 있습니다. 제거할 명시 변수를 선택하면 새 보고서가 시작됩니다. 새 보고서에는 지정된 명시 변수가 제거된 기존 모형 규격이 사용됩니다.

**모형 규격 복사** 현재 구조 방정식 모형 규격을 클립보드에 복사합니다. 그런 다음 모형 규격을 다른 SEM 플랫폼 보고서에 붙여 넣을 수 있습니다.

**모형 규격 붙여넣기** 클립보드의 모형 규격을 현재 모형 규격에 붙여 넣습니다.

다음 옵션에 대한 자세한 내용은 **JMP 사용**에서 확인하십시오.

**로컬 데이터 필터** 특정 보고서에서 사용되는 데이터를 필터링할 수 있는 로컬 데이터 필터를 표시하거나 숨깁니다.

**다시 실행** 분석을 반복하거나 다시 시작할 수 있는 옵션이 포함되어 있습니다. 이 기능을 지원하는 플랫폼에서 "자동 재계산" 옵션은 해당하는 보고서 창에서 데이터 테이블에 대한 변경 사항을 즉시 반영합니다.

**플랫폼 환경 설정** 현재 플랫폼 환경 설정을 보거나, 현재 JMP 보고서의 설정과 일치하도록 플랫폼 환경 설정을 업데이트할 수 있는 옵션이 포함되어 있습니다.

**스크립트 저장** 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다.

**그룹별 스크립트 저장** 기준 변수의 모든 수준에 대한 플랫폼 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다. 시작 창에서 기준 변수를 지정한 경우에만 사용할 수 있습니다.

---

**참고:** 이 플랫폼의 추가 옵션은 스크립트를 통해 사용할 수 있습니다. 자세한 내용은 "도움말" 메뉴의 "스크립트 인덱스"에서 여십시오. 또한 "스크립트 인덱스"에서 이 섹션에 설명된 옵션의 스크립트 예제도 찾을 수 있습니다.

---

## JMP PRO 모형 옵션

"모형 규격" 보고서에서 "실행"을 클릭하면 지정된 모형에 대한 "구조 방정식 모형" 보고서가 나타납니다. 시작 창에서 "그룹" 변수를 지정한 경우 모형 옵션은 그룹 변수의 모든 수준에 대한 보고서에서 요소를 표시하거나 숨깁니다. 이 보고서에는 다음 옵션이 포함된 빨간색 삼각형 메뉴가 있습니다.

**경로 다이어그램 표시** 모형 보고서에 경로 다이어그램을 표시하거나 숨깁니다.

**경로 다이어그램 설정** 모형의 경로 다이어그램을 수정하기 위한 다음 옵션을 포함합니다.

**다이어그램 사용자 정의** 경로 다이어그램의 여러 측면을 사용자 정의할 수 있습니다. 자세한 내용은 "경로 다이어그램 사용자 정의 옵션"에서 확인하십시오.

**레이아웃** 경로 다이어그램의 전체 모양을 변경하는 두 가지 옵션을 포함합니다. "왼쪽에서 오른쪽으로" 레이아웃 또는 "위쪽에서 아래쪽으로" 레이아웃 중에서 선택할 수 있습니다.

---

**팁:** 경로 다이어그램에서 항목을 드래그하여 특정 항목의 배열을 변경할 수도 있습니다.

---

**다이어그램 복사** 경로 다이어그램 이미지를 클립보드에 저장합니다. 최상의 품질을 유지하려면 클립보드 이미지를 벡터 그래픽으로 붙여 넣습니다.

**다이어그램 특성 복사** 현재 경로 다이어그램 특성을 클립보드에 복사합니다. 그런 다음 특성을 다른 SEM 경로 다이어그램에 붙여 넣을 수 있습니다.

**다이어그램 특성 붙여넣기** 클립보드의 경로 다이어그램 특성을 현재 SEM 경로 다이어그램에 붙여 넣습니다.

**지수 적합** 적합 모형을 평가할 수 있는 다양한 지수 값이 포함된 보고서를 표시하거나 숨깁니다. "지수 적합" 보고서에는 "적합 요약" 보고서에 나오는 값 ("구조 방정식 모형 적합 보고서" 참조) 외에 다음과 같은 지수 값도 포함됩니다.

**BIC** 베이저안 정보 기준입니다. 이 값은 모형을 비교하는 데 사용할 수 있으며 값이 작을수록 더 나은 모형 적합을 나타냅니다. 자세한 내용은 "AICc, BIC 및 BICu" 에서 확인하십시오.

**RNI** RNI(상대 비중심성 지수)는 모형 적합을 결정하기 위한 추가 지표를 제공합니다. 이 값은 CFI와 같지만 1로 제한되지 않습니다. 0.90보다 큰 값이 좋습니다. 자세한 내용은 "RNI" 에서 확인하십시오.

**TLI** TLI(Tucker-Lewis 지수)는 모형 적합을 결정하는 추가 지표를 제공합니다. 이 지수는 NNFI(비표준 적합 지수)라고도 합니다. TLI는 0과 1 사이로 제한됩니다. 0.95보다 큰 값이 좋습니다 (West et al. 2012 참조). 자세한 내용은 "TLI" 에서 확인하십시오.

**NFI** NFI(표준 적합 지수)는 모형 적합을 결정하는 추가 지표를 제공합니다. NFI는 0과 1 사이로 제한됩니다. 0.95보다 큰 값이 좋습니다 (West et al. 2012 참조). 자세한 내용은 "NFI" 에서 확인하십시오.

**수정 GFI** 수정 적합도 지수는 모형 적합을 결정하는 추가 지표를 제공합니다. 수정 GFI는 0과 1 사이로 제한됩니다. 0.95보다 큰 값이 좋습니다 (West et al. 2012 참조). 자세한 내용은 "수정 GFI 및 수정 AGFI" 에서 확인하십시오.

**수정 AGFI** 수정 조정 적합도 지수는 모형 적합을 결정하는 추가 지표를 제공합니다. 수정 AGFI는 0과 1 사이로 제한됩니다 (West et al. 2012 참조). 자세한 내용은 "수정 GFI 및 수정 AGFI" 에서 확인하십시오.

**RMR** RMR(제공된 평균 제공 잔차)은 모형 적합을 결정하는 추가 지표를 제공합니다. RMR의 잔차는 관측 공분산과 모형 내재 공분산 간의 차이에서 가져옵니다. RMR은 양수이고 값이 작을수록 좋습니다 (West et al. 2012 참조). 자세한 내용은 "RMR 및 SRMR" 에서 확인하십시오.

**SRMR** SRMR(표준화된 제공된 평균 제공 잔차)은 모형 적합을 결정하는 추가 지표를 제공합니다. SRMR의 잔차는 관측 공분산과 모형 내재 공분산 간의 표준화된 차이에서 가져옵니다. SRMR은 양수이고 값이 작을수록 좋습니다 (West et al. 2012 참조). 자세한 내용은 "RMR 및 SRMR" 에서 확인하십시오.

**참고 :** "지수 적합" 보고서의 다른 값에 대한 설명은 "구조 방정식 모형 적합 보고서" 에서 확인하십시오.

**적합 요약** 모형 적합에 대한 상세 정보가 포함된 보고서를 표시하거나 숨깁니다.

**모수 추정값** 모형에 대한 표준화되지 않은 모수 추정값이 포함된 보고서를 표시하거나 숨깁니다.

**표준화 모수 추정값** 모형에 대한 표준화된 모수 추정값이 포함된 보고서를 표시하거나 숨깁니다.

**신뢰 구간** "모수 추정값" 및 "표준화 모수 추정값" 보고서에 신뢰 구간을 표시하거나 숨깁니다.

**총 효과** (모형에 하나 이상의 회귀 또는 적재 변수가 포함되고 효과가 수렴하는 경우에만 사용 가능) 모형에 있는 총 효과의 표준화되지 않은 추정값 및 표준화된 추정값 테이블을 표시하거나 숨깁니다. 표준 오차도 포함됩니다. 효과 수렴에 대한 검정은 Bentler와 Freeman(1983) 연구 자료에 설명되어 있습니다. 테이블 오른쪽에는 표준화 추정값의 막대 차트가 포함되어 있습니다.

**간접 효과** (모형에 중재 변수가 포함되고 효과가 수렴하는 경우에만 사용 가능) 모형에 있는 간접 효과의 표준화되지 않은 추정값 및 표준화된 추정값 테이블을 표시하거나 숨깁니다. 표준 오차도 포함됩니다. 효과 수렴에 대한 검정은 Bentler와 Freeman(1983) 연구 자료에 설명되어 있습니다. 테이블 오른쪽에는 표준화 추정값의 막대 차트가 포함되어 있습니다.

---

**팁:** "간접 효과" 테이블에서 값의 붓스트랩 추정값을 얻을 수 있습니다. 붓스트랩 분석을 실행하려면 붓스트랩을 수행할 통계량이 포함된 테이블 열을 마우스 오른쪽 버튼으로 클릭하고 "붓스트랩" 을 선택합니다. 자세한 내용은 기본 분석에서 확인하십시오.

---

**예측 프로파일러** 결과 변수 집합의 조건부 기대값에 대한 예측 변수 집합의 효과를 볼 수 있습니다. 이 옵션을 선택하면 하나 이상의 예측 변수와 하나 이상의 결과를 선택해야 하는 창이 나타납니다. 예측 및 95% 신뢰 구간은 모형 내재 공분산 행렬을 기반으로 합니다. 예측 프로파일러에 대한 자세한 내용은 프로파일러에서 확인하십시오.

---

**참고:** 설정 창의 초기 변수 목록은 모형과 일치하는 변수로 제한됩니다. 예를 들어 "예측 변수 선택" 목록에는 모형에서 특정 항목을 예측하는 변수만 포함되고 "결과 선택" 목록에는 모형의 다른 변수에 의해 예측되는 변수만 포함됩니다. 두 목록의 모든 모형 변수를 보려면 **모든 변수 표시** 상자를 선택합니다.

---

**모형 내재 공분산** 모형에 내재된 공분산 행렬이 포함된 보고서를 표시하거나 숨깁니다.

**모형 내재 상관계수** 모형에 내재된 상관 행렬이 포함된 보고서를 표시하거나 숨깁니다.

**모형 내재 평균** 모형에 내재된 각 변수에 대한 평균이 포함된 보고서를 표시하거나 숨깁니다.

**잔차** 모형의 잔차 행렬이 포함된 보고서를 표시하거나 숨깁니다. 이 행렬은 모형 내재 공분산 행렬과 표본 공분산 행렬 간의 차이입니다.

**정규화 잔차** 모형의 정규화 잔차 행렬이 포함된 보고서를 표시하거나 숨깁니다.

**RAM 행렬** RAM(Reticular Action Model) 표기법에 사용되는 모형 행렬이 포함된 보고서를 표시하거나 숨깁니다.

**추정값 공분산** 모형의 모수 추정값 공분산 행렬이 포함된 보고서를 표시하거나 숨깁니다.

**추정값 상관계수** 모형의 모수 추정값 상관 행렬이 포함된 보고서를 표시하거나 숨깁니다.

**내생 변수의  $R^2$**  (모형이 재귀적이고 내생 변수가 포함된 경우에만 사용 가능) 모형의 각 내생 변수에 대한  $R^2$  값이 포함된 보고서를 표시하거나 숨깁니다. 이 값은 1에서 각 내생 변수에 대한 잔차 분산과 모형 내재 분산의 비율을 뺀 값으로 계산됩니다.  $R^2$  값은 내생 변수에서 모형에 의해 설명되는 분산 크기를 나타냅니다. 내생 변수는 경로 다이어그램에서 해당 변수를 가리키는 경로가 있는 변수입니다.

**히트맵** 모형의 잔차, 공분산 및 상관계수를 시각화할 수 있습니다.

**정규화 잔차 히트맵** 모형의 정규화 잔차 히트맵이 포함된 보고서를 표시하거나 숨깁니다.

**모형 내재 공분산 히트맵** 모형에 내재된 공분산 행렬의 히트맵이 포함된 보고서를 표시하거나 숨깁니다.

**모형 내재 상관계수 히트맵** 모형에 내재된 상관 행렬의 히트맵이 포함된 보고서를 표시하거나 숨깁니다.

**추정값 공분산 히트맵** 모형에 대한 모수 추정값 공분산 행렬의 히트맵이 포함된 보고서를 표시하거나 숨깁니다.

**추정값 상관계수 히트맵** 모형에 대한 모수 추정값 상관 행렬의 히트맵이 포함된 보고서를 표시하거나 숨깁니다.

**수정 지수** 모형 수정 지수의 모든 추정값 또는 일부를 표시할 수 있습니다. 이러한 값은 모형 적합을 개선하기 위해 모형에 추가할 수 있는 모수를 결정하는 데 사용될 수 있습니다. 각 테이블은 카이제곱 열을 기준으로 내림차순 정렬됩니다.

**모든 수정 지수** 모든 모형 수정 지수의 추정값을 포함하는 테이블을 표시하거나 숨깁니다. 이 테이블에는 각 추정값의 모수 유형을 나타내는 열이 있습니다.

**평균에 대한 수정 지수** 평균 및 절편에 대한 모형 수정 지수 추정값을 포함하는 테이블을 표시하거나 숨깁니다.

**적재에 대한 수정 지수** 적재 모수에 대한 모형 수정 지수 추정값을 포함하는 테이블을 표시하거나 숨깁니다.

**회귀에 대한 수정 지수** 회귀 모수에 대한 모형 수정 지수 추정값을 포함하는 테이블을 표시하거나 숨깁니다.

**공분산에 대한 수정 지수** 공분산 모수에 대한 모형 수정 지수 추정값을 포함하는 테이블을 표시하거나 숨깁니다.

**측정 모형 평가** (고유 요인 간의 공분산이 없는 확증적 요인 모형에만 사용 가능) 표시자 신뢰도, 오메가 및 H 계수, 구성 타당도 행렬 등 검정과 측도의 신뢰도 및 타당도를 정량화하기 위한 다양한 통계량과 그래프를 표시하거나 숨깁니다.

"표시자 신뢰도" 그림에는 잠재 변수의 표준화된 적재량 제곱과 제안된 최소 허용 가능 신뢰도 임계값 (0.25) 이 표시됩니다. 변수의 값이 낮으면 변수가 해당하는 잠재 변수의 변동을 포착하는 데 그다지 유용하지 않음을 나타냅니다.

"종합 신뢰도" 및 "구성 최대 신뢰도" 보고서에는 각 잠재 변수에 대한 오메가 (McDonald 1999) 계수와 H(Hancock 과 Mueller 2001) 계수가 각각 표시됩니다. 이러한 값의 범위는 0~1이며 약 0.70 이상을 권장합니다. 오메가는 관측된 종합 스코어에서 잠재 변수의 분산 비율을 나타냅니다. H는 표시자가 나타내는 잠재 변수의 분산 비율을 나타냅니다. 이러한 추정값은 모형에 따라 다릅니다. 한 요인 모형이 적합되면 결과 오메가는 일반 오메가입니다. 둘 이상의 잠재 변수가 있는 요인 모형이 적합되면 결과 오메가 추정값을 하위 척도 오메가라고 합니다. 그러나 이중 요인 모형이 적합되는 경우 일반 요인 추정값의 오메가는 계층적 오메가이고 그룹 요인은 계층적 하위 척도 오메가라고 합니다 (Rodriguez et al. 2015 참조). 제안

된 임계값은 설문 조사 목표의 컨텍스트에서 사용되어야 합니다. 종합 스코어를 사용하여 개별값에 대한 결정을 내리려면 제안된 임계값보다 신뢰도가 높아야 하지만 (약 0.90 이상) 조사 목적으로 종합 스코어를 사용하려면 임계값의 아래쪽 끝이 허용됩니다 (Nunnally 1978 참조).

"구성 타당도 행렬" 보고서는 사용자가 생각하는 대상을 잠재 변수가 측정하고 있는지 여부를 판별하는 데 도움이 됩니다.

- 하삼각 항목은 잠재 변수 상관을 포함합니다. 이러한 항목을 사용하면 잠재 변수가 서로 어느 정도 상관되어 있는지 확인하고 가설 상관 강도와 비교할 수 있습니다.
- 상삼각 항목은 잠재 변수의 상관 계수입니다. 이러한 항목을 사용하면 잠재 변수 간의 분산 중첩에 초점을 맞출 수 있습니다. 이 통계량은 행렬의 대각 항목과 비교할 때 특히 유용합니다.
- 대각 항목은 각 잠재 변수에 의해 추출된 평균 분산 크기를 포함합니다. 이 값은 각 잠재 변수에 대한 표시자 신뢰도의 평균과 동일합니다. 대각 항목은 표시자가 잠재 변수를 올바르게 정의하는 데 충분한 계통적 분산을 가지므로 대각 항목의 값이 높은 잠재 변수가 좋습니다. 이상적으로는 각 잠재 변수의 대각 항목이 위쪽 및 오른쪽 항목보다 높아야 합니다.

구성 타당도 행렬을 시각화하면 대각 항목을 상삼각 항목과 비교할 수 있습니다.

자세한 내용은 "측정 모형 평가 보고서의 예"에서 확인하십시오.

**예측값 그림** 모형의 내생 변수에 대한 예측값 그림을 표시하거나 숨깁니다. 종단 데이터의 경우 이 그림은 시간에 따른 모형 내재 성장 궤적을 보여 줍니다. 기본적으로 예측값은 상자 그림으로 표시됩니다. **데이터 점 연결** 체크박스를 선택하면 표시가 선 그림으로 전환됩니다.

**열 저장** 적합 구조 방정식 모형을 기반으로 한 열을 데이터 테이블에 저장할 수 있습니다.

**요인 스코어 저장** (모형에 잠재 변수가 있는 경우에만 사용 가능) 각 잠재 변수에 대해 회귀 방법을 사용하여 계산된 요인 스코어를 포함하는 열을 데이터 테이블에 저장합니다. 요인 스코어는 데이터 테이블에도 추가되는 숨겨진 열에서 계산됩니다. 이 숨겨진 열은 Estimate Factor Score() JSL 함수를 사용합니다. 이 함수에 대한 자세한 내용은 도움말 > 스크립트 인덱스에서 확인하십시오.

**Bartlett 요인 스코어 저장** (모형에 잠재 변수가 있는 경우에만 사용 가능) 각 잠재 변수에 대해 Bartlett 방법을 사용하여 계산된 요인 스코어를 포함하는 열을 데이터 테이블에 저장합니다. 요인 스코어는 데이터 테이블에도 추가되는 숨겨진 열에서 계산됩니다. 이 숨겨진 열은 Estimate Bartlett Factor Score() JSL 함수를 사용합니다. 이 함수에 대한 자세한 내용은 도움말 > 스크립트 인덱스에서 확인하십시오.

**예측 계산식 저장** (모형에 하나 이상의 내생 변수 또는 종속 변수가 있는 경우에만 사용 가능) 각 변수에 대한 관측 결과의 예측값 계산식이 포함된 열을 데이터 테이블에 저장합니다. 모형에 잠재 변수가 있으면 Bartlett 방법을 사용하여 계산된 요인 스코어도 데이터 테이블에 저장됩니다.

**관측 잔차 저장** (모형에 하나 이상의 내생 변수 또는 종속 변수가 있는 경우에만 사용 가능) 각 변수에 대한 관측 결과의 잔차 값이 포함된 열을 데이터 테이블에 저장합니다. 모형에 잠재 변수가 있으면 Bartlett 방법을 사용하여 계산된 요인 스코어도 데이터 테이블에 저장됩니다.

**모형 규격 복사** 현재 구조 방정식 모형 규격을 클립보드에 복사합니다. 그런 다음 모형 규격을 다른 SEM 플랫폼 보고서에 붙여 넣을 수 있습니다.

**모형 규격으로 재호출** 모형 규격 보고서의 모형을 지정된 모형으로 설정합니다.

**적합 제거** 보고서 창에서 지정된 모형 보고서를 제거합니다.

## JMP PRO 경로 다이어그램 사용자 정의

다양한 방법으로 구조 방정식 모형 경로 다이어그램을 사용자 정의할 수 있습니다. 대부분의 사용자 정의 옵션은 경로 다이어그램의 팝업 메뉴와 "다이어그램 모양 사용자 정의" 창 중 하나 또는 둘 다에 있습니다.

- "경로 다이어그램 팝업 메뉴 옵션"
- "경로 다이어그램 사용자 정의 옵션"

## JMP PRO 경로 다이어그램 팝업 메뉴 옵션

경로 다이어그램 자체를 마우스 오른쪽 버튼으로 클릭하면 경로 다이어그램 팝업 메뉴가 나타납니다. 메뉴에서 사용할 수 있는 특정 옵션 집합은 다이어그램에서 마우스 오른쪽 버튼을 클릭한 위치와 마우스 오른쪽 버튼을 클릭할 때 경로 다이어그램 요소가 선택되었는지 여부에 따라 달라집니다.

**선택** "시작점 목록" 및 "끝점 목록"에서 선택, 다이어그램에 단방향 및 양방향 화살표 추가, 새 잠재 변수 생성, 선택한 항목 숨기기 및 숨겨진 모든 항목 표시를 위한 옵션을 포함합니다. 경로 다이어그램에서 항목을 선택하지 않으면 이 하위 메뉴에 있는 대부분의 옵션을 사용할 수 없습니다.

**팁 :** 경로 다이어그램의 항목이 숨겨지면 숨겨진 모든 항목을 표시할 수 있는 "모두 숨기기 해제" 버튼이 경로 다이어그램 맨 위에 나타납니다.

**표시** 경로 다이어그램의 다양한 요소를 표시하거나 숨기는 옵션을 포함합니다. 표시된 요소의 원래 집합으로 돌아가려면 **기본값 표시**를 선택합니다.

**참고 :** "기본값 표시" 옵션은 "평균 / 절편 표시" 옵션 설정에 영향을 주지 않습니다.

적합 모형 보고서에 있는 경로 다이어그램의 경우 이 하위 메뉴에는 경로 다이어그램의 화살표에 표시되는 추정값을 변경하는 옵션도 포함되어 있습니다. 표준화되지 않은 모수 추정값 또는 표준화된 모수 추정값을 표시하거나 추정값을 표시하지 않도록 선택할 수 있습니다.

**이 유형 모두 선택** ( 경로 다이어그램에서 항목을 마우스 오른쪽 버튼으로 클릭한 경우에만 사용 가능 ) 경로 다이어그램에서 마우스 오른쪽 버튼으로 클릭한 항목과 유형이 같은 다른 모든 항목을 선택합니다.

**잠재 변수 그룹 선택** ( 경로 다이어그램에서 잠재 변수 항목을 마우스 오른쪽 버튼으로 클릭한 경우에만 사용 가능 ) 선택한 잠재 변수와 연결된 전체 항목 그룹을 선택합니다. 잠재 변수를 선택하지 않은 경우 이 옵션은 마우스 오른쪽 버튼으로 클릭한 잠재 변수와 연결된 항목 그룹만 선택합니다.

**변수 이름 바꾸기** ( 경로 다이어그램에서 잠재 변수 항목을 마우스 오른쪽 버튼으로 클릭한 경우에만 사용 가능 ) 경로 다이어그램에서 잠재 변수 이름을 변경할 수 있습니다.

---

**참고 :** 경로 다이어그램에서 잠재 변수 이름을 바꾸면 " 시작점 목록 " 및 " 끝점 목록 " 에서 이름이 업데이트됩니다.

---

**회귀 추가** ( 경로 다이어그램에서 변수를 마우스 오른쪽 버튼으로 클릭한 경우에만 사용 가능 ) 마우스 오른쪽 버튼으로 클릭한 변수와 경로 다이어그램에 있는 다른 변수 사이의 회귀를 나타내는 단방향 화살표를 추가할 수 있습니다. 이 옵션을 선택하는 경우 화살표로 가리킬 변수를 클릭해야 합니다.

**공분산 추가** ( 경로 다이어그램에서 변수를 마우스 오른쪽 버튼으로 클릭한 경우에만 사용 가능 ) 마우스 오른쪽 버튼으로 클릭한 변수와 경로 다이어그램에 있는 다른 변수 사이의 공분산을 나타내는 양방향 화살표를 추가할 수 있습니다. 이 옵션을 선택하는 경우 화살표로 가리킬 변수를 클릭해야 합니다.

**다이어그램 사용자 정의** " 다이어그램 모양 사용자 정의 " 창을 시작합니다. 자세한 내용은 " [경로 다이어그램 사용자 정의 옵션](#) " 에서 확인하십시오.

**잠재 변수 그룹 회전** 경로 다이어그램에서 선택한 잠재 변수에 대해 시계 방향으로 90 썸 증분하여 잠재 변수 표시자를 회전합니다. 잠재 변수를 선택하지 않은 경우 이 옵션은 경로 다이어그램의 모든 잠재 변수 표시자를 회전합니다.

**레이아웃** 경로 다이어그램의 항목에 대한 두 가지 배열 옵션 중 하나를 선택할 수 있습니다.

**다이어그램 복사** 경로 다이어그램 이미지를 클립보드에 저장합니다. 최상의 품질을 유지하려면 클립보드 이미지를 벡터 그래픽으로 붙여 넣습니다.

**다이어그램 특성 복사** 현재 경로 다이어그램 특성을 클립보드에 복사합니다. 그런 다음 특성을 다른 SEM 경로 다이어그램에 붙여 넣을 수 있습니다.

**다이어그램 특성 붙여넣기** 클립보드의 경로 다이어그램 특성을 현재 SEM 경로 다이어그램에 붙여 넣습니다.

**실행 취소** 경로 다이어그램에 대한 마지막 변경 내용을 실행 취소합니다.

**다시 실행** 경로 다이어그램에 대해 마지막으로 실행 취소한 변경 내용을 다시 실행합니다.

**편집** JMP 의 그래프에 대한 표준 옵션을 포함합니다. 이러한 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

**레이아웃 재설정** 경로 다이어그램을 원래대로 재설정합니다.

**JMP PRO** 경로 다이어그램 사용자 정의 옵션

"다이어그램" 패널에서 사용자 정의 버튼을 클릭하거나, 빨간색 삼각형 메뉴에서 "다이어그램 사용자 정의"를 선택하거나, 경로 다이어그램 팝업 메뉴에서 "다이어그램 사용자 정의"를 선택하면 "다이어그램 모양 사용자 정의" 창이 나타납니다. "다이어그램 모양 사용자 정의" 창에는 사전 설정된 두 개의 색상 테마 중에서 선택할 수 있는 드롭다운 메뉴와 네 개의 패널이 있습니다.

**테마 사전 설정** 사전 설정된 색상 테마 옵션을 포함합니다. "흑백" 테마 또는 "파란색" 테마 중에서 선택할 수 있습니다. 두 테마 중 하나를 선택하면 해당 테마에 맞게 설정이 빠르게 업데이트됩니다.

**변수 모양** 경로 다이어그램에 있는 항목의 채우기 색상, 테두리 색상, 텍스트 색상 및 크기를 설정하는 옵션을 포함합니다. 명시 변수, 잠재 변수 및 상수 변수의 모양과 이러한 각 항목의 글꼴을 설정할 수 있습니다.

**경로 설정** 화살표 모양에 대한 옵션을 포함합니다. 각 경로 화살표의 두께 및 투명도를 고정 값으로 설정하거나 표준화 추정값을 기반으로 할 수 있습니다. 유의하지 않은  $p$  값을 파선으로 나타내고, 비유의성에 대한  $\alpha$  수준을 설정하고, 화살표와 관련된 색상과 글꼴을 변경할 수도 있습니다. 기본적으로 화살표 두께는 고정 값으로 설정되고 화살표 투명도는 표준화 모수 추정값에 비례합니다.

**기타 설정** 다음 옵션이 포함되어 있습니다.

**격자 활성화** 다이어그램의 항목을 정렬하는 데 사용할 수 있는 격자를 경로 다이어그램에 표시하거나 숨깁니다.

**다이어그램 잠금** 경로 다이어그램에서 항목 배치를 잠그거나 잠금 해제합니다. 경로 다이어그램이 잠겨 있으면 항목을 드래그할 수 없으며 경로 다이어그램 맨 위의 잠금 표시자가 강조 표시됩니다.

**평균 / 절편 표시** 경로 다이어그램에 평균과 절편을 표시하거나 숨깁니다. 기본적으로 평균 구조는 경로 다이어그램에 표시되지 않습니다.

**잠재 변수 표시자 표시** 경로 다이어그램에 잠재 변수 표시자를 표시하거나 숨깁니다.

**회귀 표시** 경로 다이어그램에 회귀 화살표를 표시하거나 숨깁니다.

**분산 표시** 경로 다이어그램에 분산 화살표를 표시하거나 숨깁니다.

**공분산 표시** 경로 다이어그램에 공분산 화살표를 표시하거나 숨깁니다.

**등식 제약 조건 표시** 경로 다이어그램에 등식 제약 조건을 표시하거나 숨깁니다.

**R 제공 값 표시** 경로 다이어그램의 각 노드에 대한  $R^2$  값을 표시하거나 숨깁니다.

**상수 평균 제공 표시** 경로 다이어그램에 상수 화살표를 표시하거나 숨깁니다.

**R 제공으로 노드 채우기** 경로 다이어그램에서 각 노드에 대한  $R^2$  값에 해당하는 각 노드의 채우기 색상을 표시하거나 숨깁니다.

**미리보기** 창의 현재 설정을 기반으로 하는 경로 다이어그램 모양 미리보기를 포함합니다.

**팁:** 이후 경로 다이어그램에 동일한 설정을 사용할 수 있도록 창의 현재 설정을 플랫폼 환경 설정에 저장하려면 "다이어그램 모양 사용자 정의" 창 아래쪽에 있는 **환경 설정에 저장** 버튼을 사용합니다.

## JMP PRO 구조 방정식 모형의 추가 예

이 섹션에는 구조 방정식 모형 플랫폼을 사용하는 예가 포함되어 있습니다.

- "잠재 경로 변수 모형의 예"
- "잠재 성장 곡선 모형의 예"
- "측정 모형 평가 보고서의 예"
- "다중 그룹 분석의 예"
- "중재 분석의 예"

### 잠재 경로 변수 모형의 예

이 예에서는 75 개 개발도상국의 데이터를 사용하는 Bollen(1989) 연구 자료에 설명된 산업화 및 정치적 민주주의에 대한 구조 회귀 모형을 생성합니다. 데이터 테이블의 변수에는 1960 년과 1965 년의 네 가지 민주주의 척도와 1960 년의 세 가지 산업화 척도가 포함됩니다. 이러한 변수는 데이터 테이블의 각 열에서 "노트" 열 특성에 설명되어 있습니다. "노트" 열 특성을 보려면 열 이름을 마우스 오른쪽 버튼으로 클릭하고 "열 정보"를 선택한 후 "열 특성" 아래의 "노트"를 선택합니다.

모형 규격 과정은 잠재 변수 생성, 적재 및 회귀 변수 추가, 공분산 항 추가 및 적재 변수에 제약 조건 적용이라는 네 가지 주 단계로 구성됩니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Political Democracy.jmp 를 엽니다.
2. **분석 > 다변량 방법 > 구조 방정식 모형**을 선택합니다.
3. Prod60 ~ Legis65 를 선택하고 **모형 변수**를 클릭합니다.
4. **확인**을 클릭합니다.  
"구조 방정식 모형" 보고서의 "모형 규격" 개요가 나타납니다.
5. 보기 패널 상자에서 **목록** 탭을 클릭합니다.

#### 잠재 변수 생성

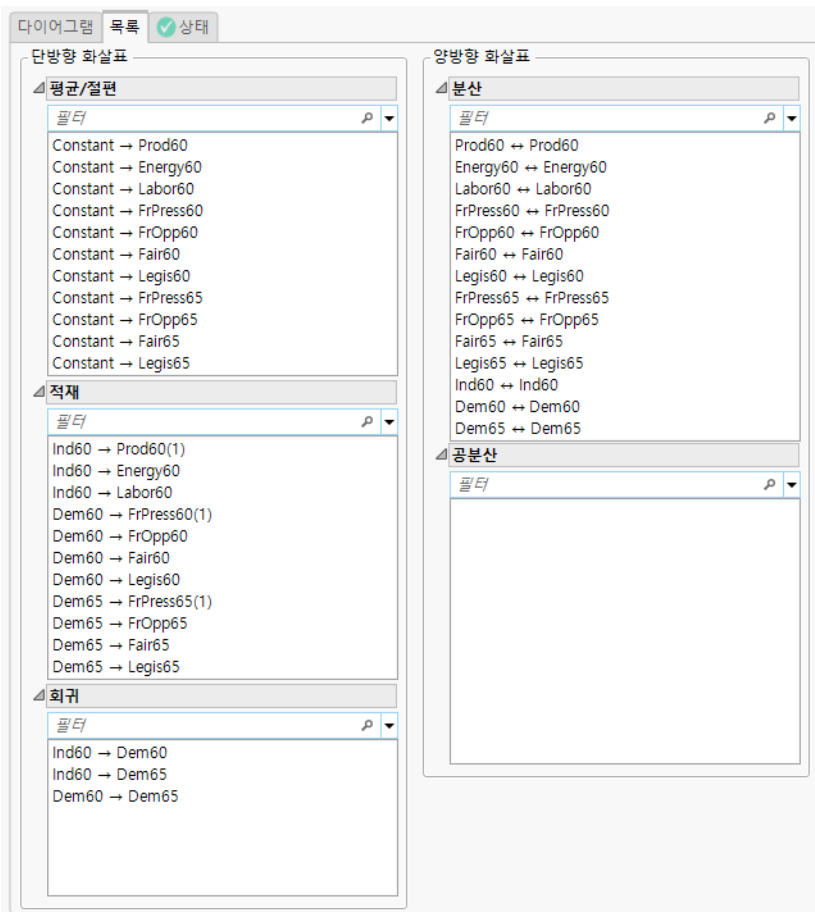
6. "끝점 목록"에서 Prod60 ~ Labor60 을 선택하고 "끝점 목록" 아래의 상자에 "Ind60"을 입력한 후 잠재 변수 추가 버튼 **+**을 클릭합니다.
7. "끝점 목록"에서 FrPress60 ~ Legis60 을 선택하고 "끝점 목록" 아래의 상자에 "Dem60"을 입력한 후 잠재 변수 추가 버튼 **+**을 클릭합니다.

8. " 끝점 목록 " 에서 FrPress65 ~ Legis65 를 선택하고 " 끝점 목록 " 아래의 상자에 "Dem65" 를 입력한 후 잠재 변수 추가 버튼  $\oplus$  을 클릭합니다 .

### 적재 및 회귀 변수 추가

9. " 시작점 목록 " 에서 Ind60 을 선택하고 " 끝점 목록 " 에서 Dem60 을 선택한 후 단방향 화살표 버튼  $\rightarrow$  을 클릭합니다 .
10. " 시작점 목록 " 에서 Ind60 을 선택하고 " 끝점 목록 " 에서 Dem65 를 선택한 후 단방향 화살표 버튼  $\rightarrow$  을 클릭합니다 .
11. " 시작점 목록 " 에서 Dem60 을 선택하고 " 끝점 목록 " 에서 Dem65 를 선택한 후 단방향 화살표 버튼  $\rightarrow$  을 클릭합니다 .

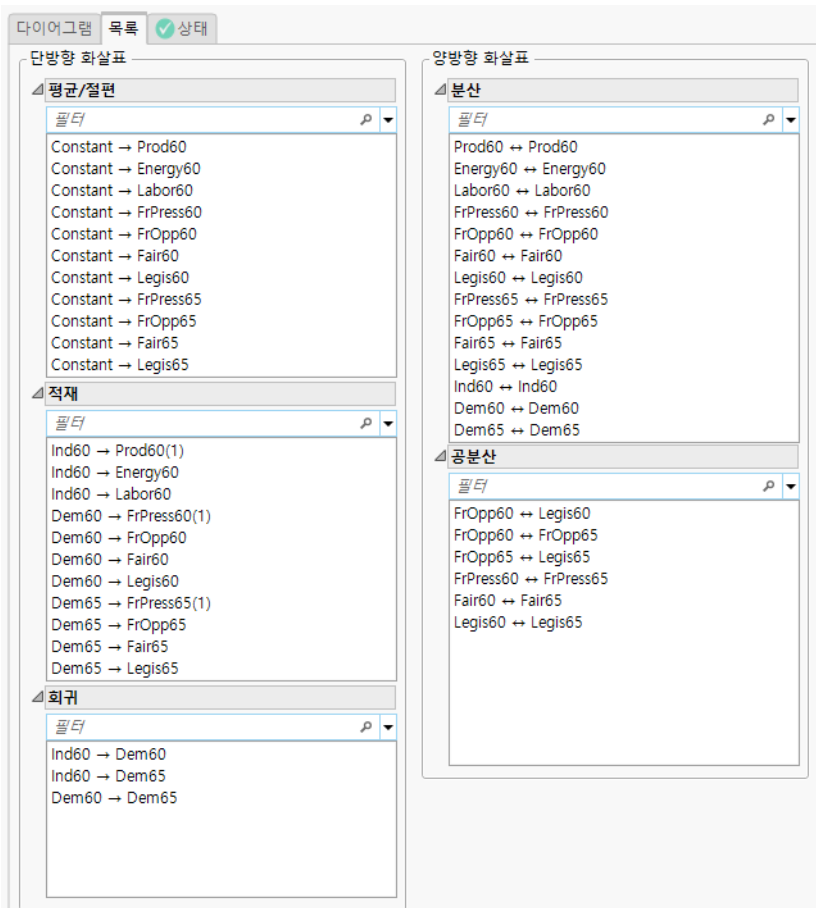
그림 8.6 적재 및 회귀



### 공분산 추가

12. "시작점 목록"에서 FrOpp60 을 선택하고 "끝점 목록"에서 Legis60 과 FrOpp65 를 선택한 후 양방향 화살표 버튼 ↔ 을 클릭합니다.
13. "시작점 목록"에서 FrOpp65 를 선택하고 "끝점 목록"에서 Legis65 를 선택한 후 양방향 화살표 버튼 ↔ 을 클릭합니다.
14. "시작점 목록"에서 FrPress60 을 선택하고 "끝점 목록"에서 FrPress65 를 선택한 후 양방향 화살표 버튼 ↔ 을 클릭합니다.
15. "시작점 목록"에서 Fair60 을 선택하고 "끝점 목록"에서 Fair65 를 선택한 후 양방향 화살표 버튼 ↔ 을 클릭합니다.
16. "시작점 목록"에서 Legis60 을 선택하고 "끝점 목록"에서 Legis65 를 선택한 후 양방향 화살표 버튼 ↔ 을 클릭합니다.

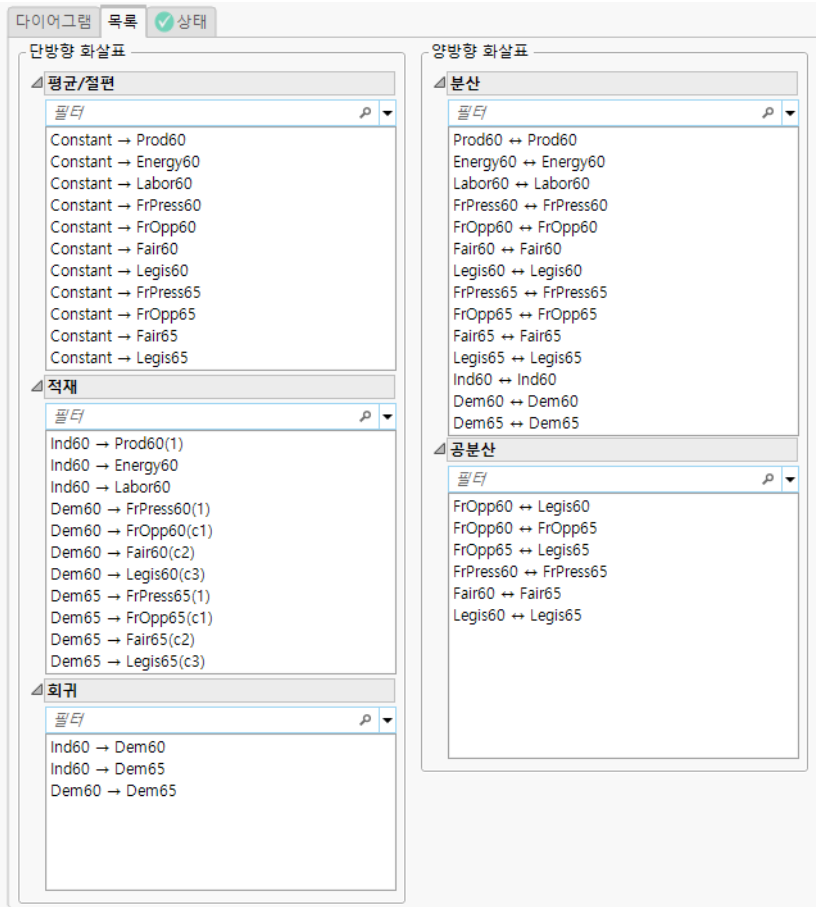
그림 8.7 공분산



## 적재에 제약 조건 추가

17. "적재" 목록에서 Dem60->FrOpp60 과 Dem65->FrOpp65 를 선택하고 **등가 설정**을 클릭합니다.
18. "적재" 목록에서 Dem60->Fair60 과 Dem65->Fair65 를 선택하고 **등가 설정**을 클릭합니다.
19. "적재" 목록에서 Dem60->Legis60 과 Dem65->Legis65 를 선택하고 **등가 설정**을 클릭합니다.

그림 8.8 완료된 모형 규격



적재에 대한 제약 조건은 영숫자 라벨로 지정됩니다 . 예를 들어 Dem60->FrOpp60 과 Dem65->FrOpp65 는 둘 다 "c1" 이라는 라벨이 지정되어 있으므로 등가 설정된 것을 알 수 있습니다 .

20. "모형 이름" 아래의 텍스트 상자에 "Industrialization and Political Democracy"를 입력합니다.
21. **실행**을 클릭합니다 .

그림 8.9 구조 방정식 모형의 적합 요약 보고서

적합 요약	
✔ 최대 가능도, 그래디언트에 수렴됨.	
표본 크기	75
결측값 포함 행	0
-2*log 가능도	3097.6362
반복	7
모수 수	39
AICc	3264.779
BICu	-123.8851
카이제곱	40.17949
DF	38
Prob>ChiSq	0.3738824
CFI	0.9967743
RMSEA	0.0276538
90% 하한	0
90% 상한	0.0870678

" 적합 요약 " 보고서에 나열된 이 모형의 카이제곱 통계량은 40.18 이고 자유도는 38 입니다 . 해당  $p$  값은 0.3739 로 , 유의하지 않습니다 . 이는 모형이 잘 적합된다는 귀무가설을 기각할 증거가 없음을 나타냅니다 . 따라서 이 모형이 데이터를 적절하게 적합시킨다는 결론을 내릴 수 있습니다 .

카이제곱 값은 표본 크기에 따라 달라지므로 일부 적합 모형에서 유의한 카이제곱 값을 생성할 수 있습니다 . CFI( 비교 적합 지수 ) 와 RMSEA( 근사의 제곱근 평균 제곱 오차 ) 는 모형 적합을 결정하는 추가 지표를 제공합니다 . 이러한 지수는 0 과 1 사이로 제한됩니다 . CFI 값이 0.90보다 크고 RMSEA 값이 0.10보다 작으면 좋습니다(Browne과 Cudeck 1993, Hu와 Bentler 1999 참조) . 여기서 CFI 0.9968 과 RMSEA 0.0277 은 적합도가 매우 좋음을 나타냅니다 .

그림 8.10 구조 방정식 모형의 모수 추정값 보고서

구조 방정식 모형: Industrialization and Political Democracy				
모수 추정값				
평균/절편	추정값	표준 오차	Wald Z	Prob> Z
Constant → Prod60	5.0543838	0.0840624	60.126553	<.0001*
Constant → Energy60	4.7921946	0.1732697	27.657433	<.0001*
Constant → Labor60	3.5576898	0.1612318	22.065683	<.0001*
Constant → FrPress60	5.4646667	0.2989108	18.28193	<.0001*
Constant → FrOpp60	4.2564429	0.43899	9.695907	<.0001*
Constant → Fair60	6.5631103	0.3939806	16.658459	<.0001*
Constant → Legis60	4.452533	0.37963	11.728613	<.0001*
Constant → FrPress65	5.1362519	0.304444	16.870925	<.0001*
Constant → FrOpp65	2.9780741	0.3923495	7.590361	<.0001*
Constant → Fair65	6.1962639	0.3643986	17.004083	<.0001*
Constant → Legis65	4.0433897	0.3753514	10.772279	<.0001*
적재	추정값	표준 오차	Wald Z	Prob> Z
Ind60 → Prod60	1	0	.	.
Ind60 → Energy60	2.1796563	0.1389144	15.690643	<.0001*
Ind60 → Labor60	1.8182091	0.1521279	11.951848	<.0001*
Dem60 → FrPress60	1	0	.	.
Dem60 → FrOpp60	1.1907892	0.1416452	8.4068422	<.0001*
Dem60 → Fair60	1.1745429	0.1197994	9.8042513	<.0001*
Dem60 → Legis60	1.2509852	0.1229265	10.17669	<.0001*
Dem65 → FrPress65	1	0	.	.
Dem65 → FrOpp65	1.1907892	0.1416452	8.4068422	<.0001*
Dem65 → Fair65	1.1745429	0.1197994	9.8042513	<.0001*
Dem65 → Legis65	1.2509852	0.1229265	10.17669	<.0001*
회귀	추정값	표준 오차	Wald Z	Prob> Z
Ind60 → Dem60	1.4713298	0.3914705	3.7584694	0.0002*
Ind60 → Dem65	0.6004651	0.2382793	2.520005	0.0117*
Dem60 → Dem65	0.8650429	0.0756829	11.429833	<.0001*
분산	추정값	표준 오차	Wald Z	Prob> Z
Prod60 ↔ Prod60	0.0813876	0.0196996	4.1314429	<.0001*
Energy60 ↔ Energy60	0.1204279	0.0699023	1.7228042	0.0849
Labor60 ↔ Labor60	0.4666598	0.0891232	5.2361204	<.0001*
FrPress60 ↔ FrPress60	1.8546635	0.4569832	4.0584942	<.0001*
FrOpp60 ↔ FrOpp60	7.5813057	1.3449558	5.6368437	<.0001*
Fair60 ↔ Fair60	4.9556824	0.9612174	5.1556312	<.0001*
Legis60 ↔ Legis60	3.2244586	0.7417042	4.3473645	<.0001*
FrPress65 ↔ FrPress65	2.3130445	0.4834091	4.7848595	<.0001*
FrOpp65 ↔ FrOpp65	4.968181	0.8945069	5.5541001	<.0001*
Fair65 ↔ Fair65	3.5600424	0.7379151	4.8244607	<.0001*
Legis65 ↔ Legis65	3.3076929	0.7128664	4.63999	<.0001*
Ind60 ↔ Ind60	0.4485992	0.0867473	5.1713357	<.0001*
Dem60 ↔ Dem60	3.8752798	0.888605	4.3610824	<.0001*
Dem65 ↔ Dem65	0.1644203	0.2333327	0.7046604	0.4810
공분산	추정값	표준 오차	Wald Z	Prob> Z
FrOpp60 → Legis60	1.4400946	0.6909749	2.084149	0.0371*
FrOpp60 ↔ FrOpp65	2.1830127	0.731106	2.9859044	0.0028*
FrOpp65 → Legis65	1.3717826	0.5781705	2.3726262	0.0177*
FrPress60 ↔ FrPress65	0.5825441	0.3644288	1.5985127	0.1099
Fair60 ↔ Fair65	0.711572	0.6194126	1.1487851	0.2506
Legis60 ↔ Legis65	0.3628	0.4607878	0.7873472	0.4311

다음으로, " 회귀 " 아래의 모수 추정값은 Dem60 과 Dem65 에 대한 Ind60 의 양의 효과 및 Dem65 에 대한 Dem60 의 양의 효과를 보여 줍니다 . 따라서 Ind60 의 스코어가 높을수록 Dem60 과 Dem65 의 스코어가 높고 Dem60 의 스코어가 높을수록 Dem65 의 스코어가 높습니다 . 모수 추정값에 대한 해당  $p$  값은 " 회귀 "아래에 표시됩니다 . 3 개의 회귀 모수는  $\alpha = 0.05$  수준에서 모두 유의합니다 . 따라서 잠재 변수 사이에 0 이 아닌 관계가 있다는 결론을 내릴 수 있습니다 .

## 잠재 성장 곡선 모형의 예

LGC(잠재 성장 곡선) 모형을 사용하면 분석자가 종단 데이터를 모델링하고 시간에 따른 궤적을 특성화할 수 있습니다. 가장 일반적인 잠재 성장 곡선 모형은 과정의 전체 궤적과 전체 궤적에서 개별 관측값의 편차를 포착하기 위해 절편 및 기울기 잠재 변수를 지정하는 선형 성장 모형입니다. 이 모형은 랜덤 계수 모형과 매우 유사합니다. LGC 모형의 주요 특징은 데이터의 평균 구조에 제약 조건을 적용하는 것입니다. 평균 구조를 모델링하면 성장과 관련된 가설을 검정할 수 있습니다.

이 예에서는 선다형 시험 스코어를 사용하여 4년 동안 학업 프로그램을 이수한 학생들의 성취도를 모델링합니다. 선형 궤적이 이 과정의 특징이라는 가설을 검정하려고 합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Academic Achievement.jmp 를 엽니다.
2. **분석 > 다변량 방법 > 구조 방정식 모형**을 선택합니다.
3. Multiple Choice Year1 ~ Multiple Choice Year4 를 선택하고 **모형 변수**를 클릭합니다.
4. **확인**을 클릭합니다.

"구조 방정식 모형" 보고서의 "모형 규격" 개요가 나타납니다.

5. **모형 바로 가기 > 종단 분석 > 선형 잠재 성장 곡선**을 선택합니다.

이제 모형 규격 다이어그램에 선형 잠재 성장 곡선 모형이 표시됩니다.

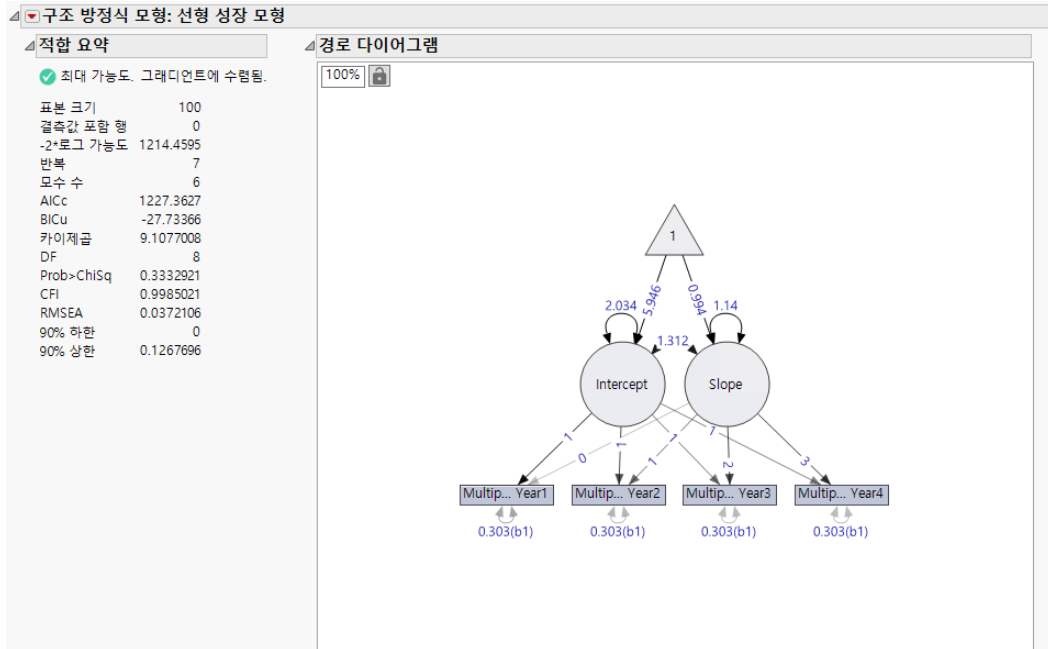
6. 보기 패널 상자에서 **목록** 탭을 클릭합니다.

7. "분산" 목록에서 처음 4개 항목을 선택하고 **등가 설정**을 클릭합니다.

이렇게 하면 각 변수의 잔차 분산이 동일한 값으로 제한되며 이는 분산 가정의 동질성과 유사합니다. SEM에서는 이 가정을 검정할 수 있으므로 등식 제약 조건 없이 다른 모형을 적합시키고 카이제곱 차이 검정과 비교할 수 있습니다.

8. "모형 이름" 아래의 텍스트 상자에 "Linear LGC with Equal Variances" 를 입력합니다.
9. **실행**을 클릭합니다.

그림 8.11 선형 LGC 모형에 대한 적합 요약 및 경로 다이어그램



"적합 요약" 보고서에 나열된 이 모형의 카이제곱 통계량은 4.18 이고 자유도는 8 입니다. 해당  $p$  값은 0.3739 로, 유의하지 않습니다. 이는 모형이 잘 적합된다는 귀무가설을 기각할 증거가 없음을 나타냅니다. 따라서 이 모형이 데이터를 적절하게 적합시킨다는 결론을 내릴 수 있습니다. 또한 CFI 및 RMSEA 적합 지수가 각각 0.9 보다 크고 0.1 보다 작으므로 적합도가 우수하다는 것을 나타냅니다.

그림 8.12 선형 LGC 모형에 대한 모수 추정값

구조 방정식 모형: Linear Growth Model					
모수 추정값					
평균/절편	추정값	표준 오차	Wald Z	Prob> Z	
Constant → Intercept	5.9461	0.1498668	39.675887	<.0001*	
Constant → Slope	0.9942	0.1095503	9.075282	<.0001*	
적재	추정값	표준 오차	Wald Z	Prob> Z	
Intercept → Multiple Choice Year1	1	0	.	.	
Intercept → Multiple Choice Year2	1	0	.	.	
Intercept → Multiple Choice Year3	1	0	.	.	
Intercept → Multiple Choice Year4	1	0	.	.	
Slope → Multiple Choice Year1	0	0	.	.	
Slope → Multiple Choice Year2	1	0	.	.	
Slope → Multiple Choice Year3	2	0	.	.	
Slope → Multiple Choice Year4	3	0	.	.	
분산	추정값	표준 오차	Wald Z	Prob> Z	
Multiple Choice Year1 ↔ Multiple Choice Year1	0.3030585	0.0303059	10	<.0001*	
Multiple Choice Year2 ↔ Multiple Choice Year2	0.3030585	0.0303059	10	<.0001*	
Multiple Choice Year3 ↔ Multiple Choice Year3	0.3030585	0.0303059	10	<.0001*	
Multiple Choice Year4 ↔ Multiple Choice Year4	0.3030585	0.0303059	10	<.0001*	
Intercept ↔ Intercept	2.0338661	0.318341	6.3889543	<.0001*	
Slope ↔ Slope	1.1395155	0.1698318	6.7096706	<.0001*	
공분산	추정값	표준 오차	Wald Z	Prob> Z	
Intercept ↔ Slope	1.3124985	0.2048419	6.4073734	<.0001*	

절편의 평균 추정값은 5.96이며, 이는 첫 해의 전체 성취도를 나타냅니다. 즉, 첫 번째 측정에서 학생들의 학업 성취도가 평균 5.96입니다. 첫 번째 측정에 해당하는 변수에 대한 기울기 요인의 적재가 0 이므로 절편의 평균은 이 변수에서 중심화됩니다.

기울기의 평균 추정값은 0.95이며, 이는 전체 성취도가 매년 0.95씩 증가한다는 것을 나타냅니다. 절편 및 기울기 요인의 분산 추정값이 유의하므로 평균 궤적의 유의한 변동을 나타냅니다. 학생들마다 시작 지점도 다르고 증가 비율도 다릅니다. 외부 변수를 절편 및 기울기의 예측 변수로 사용하여 개인의 궤적 차이를 발생시키는 요인을 파악할 수 있습니다.

마지막으로, 절편과 기울기 사이의 양의 공분산은 첫 해에 더 높은 수준의 학업 성취도로 시작하는 학생이 시간이 지남에 따라 더 빠르게 증가하는 경향이 있음을 나타냅니다.

LGC 모형은 시간에 따른 과정 변화를 조사하기 위한 유연한 접근 방식을 제공합니다. "모형 바로 가기" 메뉴는 무성장 궤적, 2차 궤적 및 잠재 변수 기반 비선형 궤적을 포함하여 최량 모형 적합을 확인하기 위해 서로 적합시키고 검정할 수 있는 대안 궤적을 제공합니다.

## 측정 모형 평가 보고서의 예

CFA(확증적 요인 분석) 모형을 사용하면 대안 측정 모형을 검정할 수 있습니다. "측정 모형 평가" 보고서는 검정과 측도의 신뢰도 및 타당도를 정량화하기 위한 도구를 제공합니다. 결과에는 표시자 신뢰도, 오메가 및 H 계수, 구성 타당도 행렬이 포함됩니다.

이 예에서는 소비자 데이터 조사의 타당도와 신뢰도를 평가합니다. Privacy, Security, Reputation, Trust 및 Purchase Intent 라는 다섯 가지 잠재 변수에 대한 확증적 요인 분석 모형을 적합시킵니다. 그런 다음 "측정 모형 평가" 옵션을 사용하여 설문 조사 신뢰도를 검정합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Online Consumer Data.jmp 를 엽니다.
2. **SEM: CFA** 스크립트 옆의 녹색 삼각형을 클릭합니다.

이 스크립트는 설문 조사 데이터에 대한 확증적 요인 분석 모형을 실행합니다.

3. "모형 규격" 및 "모형 비교" 옆의 회색 표시 아이콘을 클릭하여 해당 보고서 부분을 숨깁니다.

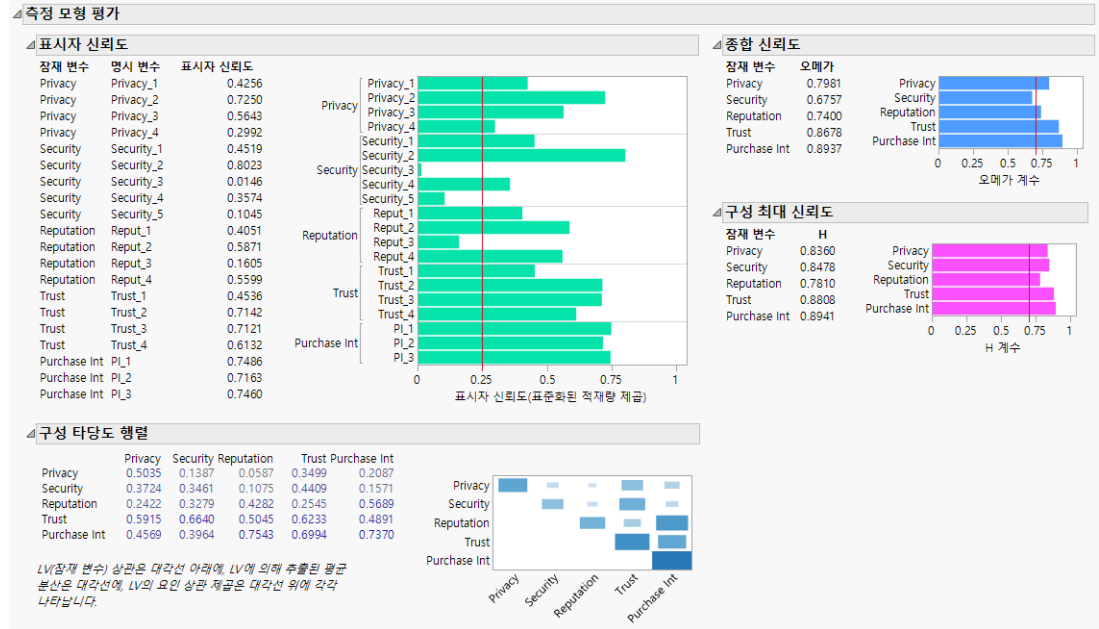
그림 8.13 CFA 모형에 대한 적합 요약

적합 요약	
✓ 최대 가능도, 그래디언트에 수렴됨.	
표본 크기	843
결측값 포함 행	0
-2*로그 가능도	45106.777
반복	6
모수 수	70
AICc	45259.653
BICu	-269.8055
카이제곱	808.10917
DF	160
Prob>ChiSq	3.738e-87
CFI	0.9160319
RMSEA	0.0693187
90% 하한	0.0646117
90% 상한	0.0741036

"적합 요약" 보고서에 나열된 이 모형의 카이제곱 통계량은 808.11 이고 자유도는 160 입니다. 해당  $p$  값은 유의합니다. 이는 모형이 잘 적합된다는 귀무가설을 기각할 몇 가지 증거가 있음을 나타냅니다. 그러나 카이제곱 통계량은 표본 크기에 큰 영향을 받으며, 관측값 수가 200 개 ~ 300 개를 초과하는 표본의 데이터에 모형이 잘 적합되는 경우에도 유의한 것으로 나타났습니다. 이 예에서 표본 크기는 843 입니다. 따라서 CFI 및 RMSEA 적합 지수로도 모형 적합을 평가해야 합니다. 이 예에서는 두 지수가 각각 0.9 보다 크고 0.1 보다 작으므로 모두 양호한 적합을 나타냅니다.

4. "구조 방정식 모형 : CFA" 옆의 빨간색 삼각형을 클릭하고 **측정 모형 평가**를 선택합니다.

그림 8.14 CFA 모형에 대한 측정 모형 평가 보고서



" 표시자 신뢰도 " 그림에는 잠재 변수의 표준화된 적재량 제곱과 제안된 최소 허용 가능 신뢰도 임계값 (0.25) 이 표시됩니다 . Security 관련 질문 두 개와 Reputation 관련 질문 하나가 해당 잠재 변수의 변동을 포착하는 데 그다지 유용하지 않다는 것을 알 수 있습니다 . Security\_3 질문은 값이 매우 낮으므로 구성의 신뢰도에 거의 영향을 주지 않고 제거할 수 있는 반면 , Security\_5 및 Reput\_3 질문은 신뢰도 향상을 위해 수정할 수 있습니다 .

" 종합 신뢰도 " 및 " 구성 최대 신뢰도 " 보고서에는 각 잠재 변수에 대한 오메가 계수와 H 계수가 각각 표시됩니다 . 이러한 값의 범위는 0 ~ 1 이며 약 0.70 이상을 권장합니다 . 오메가는 관측된 종합 스코어에서 잠재 변수의 분산 비율을 나타냅니다 . H 는 표시자가 나타내는 잠재 변수의 분산 비율을 나타냅니다 . 이러한 측도에 따르면 Security 의 오메가 계수는 제안된 임계값 0.70 보다 아주 약간 낮지만 조사 목적으로는 이 종합 신뢰도가 충분하다고 할 수 있습니다 . 제안된 임계값은 설문 조사 목표의 컨텍스트에서 사용되어야 합니다 . 종합 스코어를 사용하여 개별값에 대한 결정을 내리려면 제안된 임계값보다 신뢰도가 높아야 하지만 ( 약 0.90 이상 ) 조사 목적으로 종합 스코어를 사용하려면 임계값의 아래쪽 끝이 허용됩니다 (Nunnally 1978 참조 ) . Security\_3, Security\_5 및 Reput\_4 질문에 대한 표시자 신뢰도 향상에 초점을 맞추면 Security 및 Reputation 의 종합 신뢰도를 향상시킬 수 있습니다 .

" 구성 타당도 행렬 " 보고서는 사용자가 생각하는 대상을 잠재 변수가 측정하고 있는지 여부를 판별할 수 있는 방법을 제공합니다 . 행렬 시각화 그림에서 Privacy, Trust 및 Purchase Intent 의 대각 값이 위쪽 및 오른쪽 값보다 큼니다 . 그러나 Security 와 Reputation 의 경우에는 그렇지 않습니다 . 이는 Security 및 Reputation 관련 질문을 개선하면 Security 및 Reputation 잠재 변수를 측정할 수 있다는 추가적인 증거입니다 .

Security 및 Reputation 관련 질문 중 일부를 삭제하거나 수정하여 설문 조사를 개선할 수 있다는 결론을 내릴 수 있습니다 .

보고서 구성 요소에 대한 자세한 내용은 "[측정 모형 평가](#)"에서 확인하십시오.

## 다중 그룹 분석의 예

구조 방정식 모형 플랫폼에서 그룹 변수 역할을 사용하면 여러 그룹 간에 통계적 효과에 대해 모형을 비교할 수 있습니다. 이 예에서는 두 그룹의 어린이에 대한 학업 성취도의 성장 궤적을 비교하려고 합니다. 선다형 시험에서 학생들의 스코어를 네 번 반복 측정합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Academic Achievement.jmp 를 엽니다.
2. **분석 > 다변량 방법 > 구조 방정식 모형**을 선택합니다.
3. Multiple Choice Year1 ~ Multiple Choice Year4 를 선택하고 **모형 변수**를 클릭합니다.
4. **Sex** 를 선택하고 **그룹**을 클릭합니다.
5. **확인**을 클릭합니다.  
"구조 방정식 모형" 보고서의 "모형 규격" 개요가 나타납니다.
6. **모형 바로 가기 > 종단 분석 > 선형 잠재 성장 곡선**을 선택합니다.  
이제 모형 규격 다이어그램에 선형 잠재 성장 곡선 모형이 표시됩니다.
7. "모형 이름" 아래의 텍스트 상자에 "Grouped Linear LGC Model (Equality Constraint)" 을 입력합니다.
8. 경로 다이어그램에서 삼각형 아이콘과 "절편" 및 "기울기" 사이의 화살표를 각각 선택하고 **등가 설정**을 클릭합니다.  
이렇게 하면 절편 및 기울기 잠재 변수의 평균이 두 학생 그룹 간에 같아집니다.
9. "등가 설정" 창에서 **확인**을 클릭하여 두 그룹 간에 등식 제약 조건을 적용합니다.
10. **실행**을 클릭합니다.
11. 보고서 창의 "모형 규격" 섹션이 나올 때까지 위로 스크롤합니다.
12. "모형 이름" 아래의 텍스트 상자에 "Grouped Linear LGC Model" 을 입력합니다.
13. 경로 다이어그램에서 a1 및 a2 라벨이 지정된 화살표를 각각 선택하고 **제약 해제**를 클릭합니다.
14. **실행**을 클릭합니다.
15. "모형 규격" 옆의 회색 표시 아이콘을 클릭합니다.
16. "모형 비교" 테이블에서 Grouped LGC Model 을 둘 다 선택하고 **선택한 모형 비교**를 클릭합니다.

그림 8.15 모형 비교 보고서

모형 비교																	
모형 이름	-2* 로그 가능도	모수 수	AICc	AICc 기준치	.2	.4	.6	.8	BICu	카이제곱	DF	Prob>ChiSq	CFI	RMSEA	90% 하한	90% 상한	
1 제약 없음(포화)	1170.3253	28	1249.1985	0.0000					0.0000	0.0000	0		1.0000	0.0000	0.0000	0.0000	
2 독립	1852.9631	16	1891.5173	0.0000					627.3758	682.6378	12	<.0001*	0.0000	0.7476	0.7004	0.7958	
3 Grouped Linear LGC Model (Equality Constraint)	1204.3683	16	1242.9225	0.0005					-21.2190	34.0430	12	0.0007*	0.9671	0.1917	0.1176	0.2690	
4 Grouped Linear LGC Model	1183.2528	18	1227.6972	0.9995					-33.1242	12.9275	10		0.2278	0.9956	0.0765	0.0000	0.1812

선택한 모형 비교 선택 해제

카이제곱 차이 검정						
내포 모형	모형	Δ카이제곱	ΔDF	Prob>ChiSq	ΔCFI	ΔRMSEA
Grouped Linear LGC Model (Equality Constraint)	Grouped Linear LGC Model	21.1155	2	<.0001*	-0.029	0.1152

차이 값들은 내포 모형에서만 의미가 있습니다.

카이제곱 차이 검정에 따르면 등식 제약 조건으로 인해 통계적으로 유의한 모형 부적합 증가가 발생하므로 적절하지 않음을 알 수 있습니다. 이와 관련하여 두 Grouped LGC Model의 BICu 값이 모두 음수이며, 이는 두 모형이 모두 제약 없는 모형에 우호적임을 나타냅니다. 카이제곱 차이 검정에서 뒷받침되고 BICu 값이 더 작은 제약 없는 Grouped LGC Model을 계속 사용하도록 선택합니다.

17. "Grouped Linear LGC Model (Equality Constraint)" 옆의 버튼을 클릭합니다.

이렇게 하면 등식 제약 조건이 있는 모형과 관련된 보고서 섹션이 숨겨집니다.

18. Alt 키를 누르고 "구조 방정식 모형 : Grouped Linear LGC Model" 옆의 빨간색 삼각형을 클릭합니다.

**팁 :** Alt 키를 사용하면 여러 개의 빨간색 삼각형 메뉴를 동시에 선택할 수 있습니다.

19. 경로 다이어그램 표시, 적합 요약 및 모수 추정값 옵션을 선택 취소합니다.

20. 예측값 그림 옵션을 선택합니다.

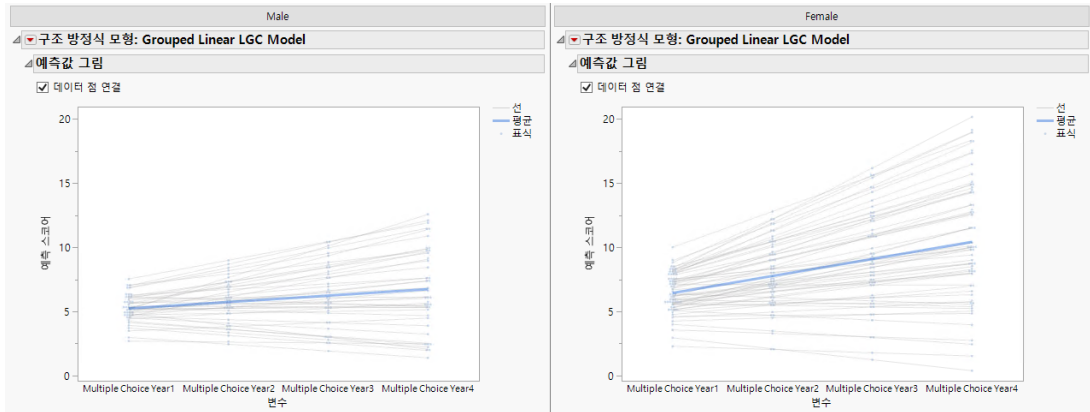
21. 확인을 클릭합니다.

**참고 :** "구조 방정식 모형 : Grouped Linear LGC Model"의 빨간색 삼각형 메뉴에서 선택한 내용이 "Male" 탭 보고서와 "Female" 탭 보고서에 모두 적용됩니다.

22. "Male" 또는 "Female" 탭 중 하나를 마우스 오른쪽 버튼으로 클릭하고 스타일 설정 > 수평 퍼짐을 선택합니다.

23. 데이터 점 연결 상자를 선택합니다.

그림 8.16 예측값 그림



4년 동안의 남학생에 대한 예측값이 여학생에 대한 예측값보다 기울기가 더 평평하고 변동이 적습니다.

## 중재 분석의 예

이 예에서는 구조 방정식 모형 플랫폼에서 중재 분석을 수행하는 방법을 보여 줍니다. 데이터는 다양한 배경을 가진 중등학교 학생들의 인종 간 분위기를 조사하기 위한 교육 변수로 구성되어 있습니다. 이 예의 목표는 중등학교 학생들에게 학교에서의 소속감이 학교에 대한 흥미에 미치는 영향을 중재하는 자율성의 역할을 조사하는 것입니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Secondary Students.jmp** 를 엽니다.
2. **분석 > 다변량 방법 > 구조 방정식 모형**을 선택합니다.
3. **Interest, Belonging, Autonomy** 를 선택하고 **모형 변수**를 클릭합니다.
4. **확인**을 클릭합니다.
5. "모형 바로 가기" 의 빨간색 삼각형을 클릭하고 **횡단면 (전통적) > 중재 분석**을 선택합니다.
6. "중재 모형 지정" 창에서 "원인 변수", "중재 변수", "결과 변수" 옵션에 대해 **Belonging, Autonomy, Interest** 를 각각 선택합니다.
7. **확인**을 클릭합니다.

그림 8.17 구조 방정식 모형 규격

8. 실행을 클릭합니다.

그림 8.18 구조 방정식 모형 보고서

평균/절편	추정값	표준 오차	Wald Z	Prob> Z
Constant → Interest	0.2588694	0.150868	1.7158667	0.0862
Constant → Belonging	3.6848073	0.0266537	138.24733	<.0001*
Constant → Autonomy	1.5091022	0.1431389	10.542918	<.0001*
회귀				
Belonging → Autonomy	0.2982896	0.0380392	7.8416299	<.0001*
Belonging → Interest	0.6235978	0.0390082	15.986311	<.0001*
Autonomy → Interest	0.3368475	0.0345591	9.747006	<.0001*
분산				
Interest ↔ Interest	0.6744588	0.0333295	20.236106	<.0001*
Belonging ↔ Belonging	0.5818352	0.0287523	20.236106	<.0001*
Autonomy ↔ Autonomy	0.6895207	0.0340738	20.236106	<.0001*

표시된 보고서에는 중재 분석 모형에 대한 "적합 요약" 및 "모수 추정값" 테이블이 포함되어 있습니다. 이 모형은 종속 변수와 독립 변수 간의 직접 효과를 강조합니다. "모수 추정값" 테이블에는 각 변수에 대한 추정 평균과 절편 값, 예측 변수 (Belonging, Autonomy) 와 결과 변수 (Interest, Autonomy) 간의 회귀 계수 추정값, 추정 분산 및 잔차 분산이 포함되어 있습니다. 보고서에 따르면 이러한 추정값이 나타내는 효과가 통계적으로 유의합니다. 이러한 추정값을 기반으로 다음과 같은 결론을 내릴 수 있습니다.

- Belonging  $\Rightarrow$  Autonomy 경로 : Belonging 에서 1 단위 차이가 나는 두 학생은 Autonomy 에서 0.298 단위 차이가 나는 것으로 추정됩니다. 계수의 양수 부호는 Belonging 이 올라갈 수록 Autonomy 도 올라갈 가능성이 있음을 나타냅니다.
- Autonomy  $\Rightarrow$  Interest 경로 : Belonging 스코어는 같지만 Autonomy 에서 1 단위 차이가 나는 두 학생은 Interest 에서 0.337 단위 차이가 나는 것으로 추정됩니다. 계수의 양수 부호는 Autonomy 가 많으면 Interest 가 많다는 것과 관련이 있음을 나타냅니다.
- Belonging  $\Rightarrow$  Interest 경로 : Belonging 에서 1 단위 차이가 나고 Autonomy 에서 차이가 없는 두 학생은 Interest 에서 0.624 단위 차이가 나는 것으로 추정됩니다. 직접 효과는 양수 값을 나타내며 이는 Belonging 이 증가하면 Interest 도 증가할 것이라는 의미입니다.

이제 총 효과와 간접 효과를 계산합니다.

1. "구조 방정식 모형: 중재 분석" 옆의 빨간색 삼각형을 클릭하고 **총 효과** 옵션을 선택합니다.
2. "구조 방정식 모형: 중재 분석" 옆의 빨간색 삼각형을 클릭하고 **간접 효과** 옵션을 선택합니다.

**그림 8.19** 총 효과 및 간접 효과

총 효과							
예측 변수	결과	추정값	표준 오차	Wald Z	Prob> Z	표준화 추정값	.2 .4 .6 .8
Belonging	Interest	0.7240759	0.0397437	18.218651	<.0001*	0.5370241	
Belonging	Autonomy	0.2982896	0.0380392	7.8416299	<.0001*	0.2642676	
Autonomy	Interest	0.3368475	0.0345591	9.747006	<.0001*	0.2819923	

간접 효과							
예측 변수	결과	추정값	표준 오차	Wald Z	Prob> Z	표준화 추정값	.2 .4 .6 .8
Belonging	Interest	0.1004781	0.0164454	6.1097993	<.0001*	0.0745214	

"총 효과" 테이블에는 예측 변수 (Belonging, Autonomy) 와 결과 변수 (Interest, Autonomy) 간의 총 효과가 포함되어 있습니다. 총 효과는 예측 변수가 결과 변수에 미치는 직간접 영향이 결합된 것입니다. Belonging 에서 Autonomy 로, Autonomy 에서 Interest 로의 직접 효과만 있으므로 총 효과가 "모수 추정값" 테이블의 직접 효과와 동일합니다. 이 테이블에 따르면 Belonging 이 Interest 에 강력하고 긍정적인 총 효과 (0.724) 를 미치고 있음을 알 수 있습니다. 유의 효과는 Interest 에 대한 Belonging 의 의미 있는 기여를 나타냅니다. 총 효과는 다른 변수를 일정하게 유지한 상태에서 예측 변수의 1 단위 변화에 따른 결과 변수의 평균 변화를 정량화합니다. 예를 들어 Belonging 이 1 단위 증가할 때마다 Interest 가 0.724 단위 증가하는 것으로 추정됩니다.

"간접 효과" 테이블에는 Belonging 이 Autonomy 를 거쳐 Interest 에 미치는 효과를 강조하는 간접 효과 추정값이 포함되어 있습니다. 이 테이블에 따르면 Belonging 이 중재 변수 Autonomy 를 통해 Interest 에 미치는 간접 효과 추정값 (0.100) 이 통계적으로 유의한 것으로 나타나며, 이는 Belonging 의 변화가 Autonomy 중재 변수의 변화를 통해 Interest 에 영향을 미친다는 의미입니다.

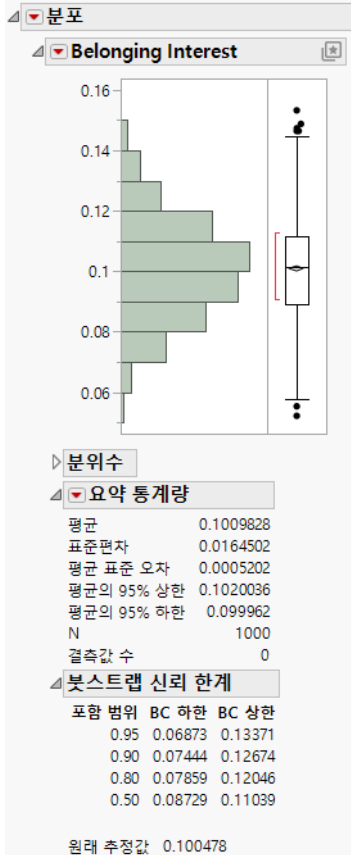
Belonging에서 1 단위 차이가 나는 두 학생의 경우 Belonging이 높은 학생의 Autonomy가 높은 경향으로 인해 Interest에서 0.100 단위 차이가 나는 것으로 추정된다는 결론을 내릴 수 있으며, 실제로 Interest가 더 높은 결과를 보입니다.

중재 분석의 과제 중 하나는 간접 효과를 추정하는 것입니다. 이 추정값은 특히 표본 크기가 작을 때 비정규 분포를 따르는 경향이 있습니다. 붓스트랩 방법이 이에 대한 해결책을 제공합니다. 붓스트랩은 관측된 데이터 집합에서 반복적으로 복원 재표집을 수행하여 간접 효과의 표집 분포를 경험적으로 추정합니다. 이 방식은 간접 효과에 대해 더 정확한 표준 오차와 신뢰 구간을 제공합니다.

모수 추정값, 총 효과 및 간접 효과에 대한 붓스트랩 신뢰 구간을 구하려면 다음 단계를 수행하십시오.

1. 관심 테이블에서 **추정값** 열을 마우스 오른쪽 버튼으로 클릭하고 **붓스트랩**을 선택합니다.
2. **붓스트랩 표본 수**에 "1000" 을 입력합니다.
3. (선택 사항) **그림 8.20**의 결과와 일치시키려면 **난수 시드값**에 "1234" 를 입력합니다.
4. **확인**을 클릭합니다.
5. "붓스트랩 결과" 테이블에서 **분포 스크립트**를 실행합니다.

그림 8.20 붓스트랩 보고서



표시된 보고서에는 붓스트랩 추정값 분포에 대한 정보가 포함되어 있습니다. "요약 통계량" 테이블에 따르면 간접 효과의 붓스트랩 추정값이 0.101입니다. 이 추정값은 그림 8.19에 표시된 간접 효과 추정값과 거의 동일합니다. "붓스트랩 신뢰 한계" 보고서에는 추정된 간접 효과에 대한 편향 수정 신뢰 한계가 포함되어 있습니다. 95% 신뢰 구간에 0이 포함되지 않으므로 중재 분석에서 간접 효과가 유의함을 나타냅니다.

**팁:** "구조 방정식 모형"의 빨간색 삼각형 메뉴에서 추론 > 붓스트랩 추론 옵션을 사용하여 완전 모형에 대한 붓스트랩 추정값을 얻을 수 있습니다.

## JMP PRO 구조 방정식 모형 플랫폼에 대한 통계 상세 정보

이 섹션에는 구조 방정식 모형 플랫폼에 대한 통계 상세 정보가 포함되어 있습니다.

- "추정 방법에 대한 통계 상세 정보"

- "적합 요약 측도에 대한 통계 상세 정보"
- "중재 분석에 대한 통계 상세 정보"

## 추정 방법에 대한 통계 상세 정보

구조 방정식 모형 플랫폼에서 사용되는 기본 추정 방법은 결측 데이터가 있는지 여부에 따라 달라 집니다. 결측 데이터가 없는 경우 구조 방정식 모형 플랫폼에서는 기본적으로 ML(최대 가능도) 추정을 사용합니다. 결측 데이터가 감지되는 경우 플랫폼에서는 기본적으로 FIML(완전 정보 최대 가능도) 추정을 사용합니다. 두 경우 모두 관측된 정보 행렬을 사용하여 표준 오차를 얻습니다. 로버스트 최대 가능도 표준 오차는 "구조 방정식 모형"의 빨간색 삼각형 메뉴에서 추론 > 로버스트 추론 옵션을 사용하여 얻을 수 있습니다. "로버스트 추론" 옵션을 사용하면 비정규 데이터가 있을 때 효율성이 떨어지는 문제를 보완하기 위해 표준 오차에 샌드위치 수정이 적용됩니다. 샌드위치의 빵은 사용자 지정 모형의 관측된 정보 행렬이고, 샌드위치의 고기는 개별 관측값 그래디언트의 외적 합입니다. 이 플랫폼에서는 Newton-Raphson, Quasi-Newton 및 Fisher 스코어링을 포함한 최적화 알고리즘 조합을 추정 프로세스에 사용합니다.

**참고:** 구조 방정식 모형 플랫폼의 ML 방법은  $N - 1$ 이 아니라 표본 크기  $N$ 을 추정에 사용합니다.

## 적합 요약 측도에 대한 통계 상세 정보

이 섹션에서는 구조 방정식 모형 플랫폼에 보고되는 적합 요약 측도에 대해 설명합니다.

### AICc, BIC 및 BICu

AICc 및 BIC는 다음과 같이 정의됩니다.

$$AICc = -2\log L + 2k + \frac{2k(k+1)}{n - (k+1)}$$

$$BIC = -2\log L + k \ln(n)$$

다음은 각 요소에 대한 설명입니다.

$-2\log L$  = 음의 로그 가능도의 두 배

$n$  = 표본 크기

$k$  = 모수 수

"모형 비교" 보고서의 가능도 기반 측도에 대한 자세한 내용은 선형 모형 적합에서 확인하십시오.

BICu( 제약 없는 모형에 대한 BIC)는 다음과 같이 정의됩니다.

$$BICu = \chi_{min}^2 - df_{min} \log(n)$$

다음은 각 요소에 대한 설명입니다.

$\chi_{min}^2$  = 적합 모형의 카이제곱 통계량

$df_{min}$  = 적합 모형의 자유도

$n$  = 표본 크기

적합 모형의 BICu는 적합 모형의 BIC에서 제약 없는 모형의 BIC를 뺀 값과 같습니다. 제약 없는 모형의 BIC에 대한 자세한 내용은 Bollen et al. (2014)에서 확인하십시오.

### CFI

CFI(비교 적합 지수)는 다음과 같이 정의됩니다.

$$CFI = \frac{\max(\chi_0^2 - df_0, 0) - \max(\chi_{min}^2 - df_{min}, 0)}{\max(\chi_0^2 - df_0, 0)}$$

다음은 각 요소에 대한 설명입니다.

$\chi_0^2$  = 독립 모형의 카이제곱 통계량

$df_0$  = 독립 모형의 자유도

$\chi_{min}^2$  = 적합 모형의 카이제곱 통계량

$df_{min}$  = 적합 모형의 자유도

CFI에 대한 자세한 내용은 Bentler(1990) 연구 자료에서 확인하십시오.

### RNI

RNI(상대 비중심성 지수)는 다음과 같이 정의됩니다.

$$RNI = \frac{(\chi_0^2 - df_0) - (\chi_{min}^2 - df_{min})}{\chi_0^2 - df_0}$$

다음은 각 요소에 대한 설명입니다.

$\chi_0^2$  = 독립 모형의 카이제곱 통계량

$df_0$  = 독립 모형의 자유도

$\chi_{min}^2$  = 적합 모형의 카이제곱 통계량

$df_{min}$  = 적합 모형의 자유도

RNI에 대한 자세한 내용은 McDonald & Marsh(1990)에서 확인하십시오.

### TLI

TLI(Tucker-Lewis 지수)는 다음과 같이 정의됩니다.

$$TLI = \frac{\frac{\chi_0^2}{df_0} - \frac{\chi_{min}^2}{df_{min}}}{\frac{\chi_0^2}{df_0} - 1}$$

다음은 각 요소에 대한 설명입니다.

$\chi_0^2$  = 독립 모형의 카이제곱 통계량

$df_0$  = 독립 모형의 자유도

$\chi_{min}^2$  = 적합 모형의 카이제곱 통계량

$df_{min}$  = 적합 모형의 자유도

자세한 내용은 West et al. (2012) 연구 자료에서 확인하십시오.

### NFI

Bentler-Bonett NFI(표준 적합 지수)는 다음과 같이 정의됩니다.

$$NFI = \frac{\chi_0^2 - \chi_{min}^2}{\chi_0^2}$$

다음은 각 요소에 대한 설명입니다.

$\chi_0^2$  = 독립 모형의 카이제곱 통계량

$$\chi_{min}^2 = \text{적합 모형의 카이제곱 통계량}$$

자세한 내용은 West et al. (2012) 연구 자료에서 확인하십시오.

### 수정 GFI 및 수정 AGFI

수정 GFI(수정 적합도 지수)는 다음과 같이 정의됩니다.

$$\text{수정 GFI} = \frac{p}{p + 2 \left( \frac{\chi_{min}^2 - df_{min}}{n - 1} \right)}$$

다음은 각 요소에 대한 설명입니다.

$$\chi_{min}^2 = \text{적합 모형의 카이제곱 통계량}$$

$$df_{min} = \text{적합 모형의 자유도}$$

$$p = \text{적합 모형에서 관측 변수의 수}$$

$$n = \text{표본 크기}$$

수정 AGFI(수정 조정 적합도 지수)는 다음과 같이 정의됩니다.

$$\text{수정 AGFI} = 1 - \frac{p^*}{df_{min}} (1 - \text{RevisedGFI})$$

다음은 각 요소에 대한 설명입니다.

$$p^* = \text{공분산 행렬 및 관측 변수 평균 벡터의 고유 항목 수}$$

$$df_{min} = \text{적합 모형의 자유도}$$

자세한 내용은 Maiti & Mukherjee(1991) 및 West et al. (2012)에서 확인하십시오.

### RMSEA

RMSEA(근사의 제공근 평균 제공 오차)는 다음과 같이 정의됩니다.

$$\text{RMSEA} = \sqrt{\frac{\max(\chi_{min}^2 - df_{min}, 0)}{n \times df_{min}}}$$

다음은 각 요소에 대한 설명입니다.

$n$  = 표본 크기

$df_{min}$  = 적합 모형의 자유도

$\chi_{min}^2$  = 적합 모형의 카이제곱 통계량

RMSEA 에 대한 신뢰 한계는 비중심 카이제곱 분포  $\Phi(x|\lambda, d)$  의 누적 분포 함수를 사용하여 계산됩니다. 90% 신뢰 한계는 다음과 같이 계산됩니다.

$$\text{하한} = \sqrt{\frac{\lambda_L}{n \times df_{min}}}$$

$$\text{상한} = \sqrt{\frac{\lambda_U}{n \times df_{min}}}$$

다음은 각 요소에 대한 설명입니다.

$\lambda_L$  은  $\Phi(\chi_{min}^2 | \lambda_L, df_{min}) = 0.95$  를 충족합니다.

$\lambda_U$  는  $\Phi(\chi_{min}^2 | \lambda_U, df_{min}) = 0.05$  를 충족합니다.

자세한 내용은 Maydeu-Olivares et al. (2017)에서 확인하십시오.

### RMR 및 SRMR

RMR 및 SRMR 계산식은 다음과 같이 정의됩니다.

$$\text{RMR} = \sqrt{\frac{1}{b} \left[ \sum_i \sum_j^p (s_{ij} - \hat{\sigma}_{ij})^2 + \sum_i^p (\bar{x}_i - \hat{\mu}_i)^2 \right]}$$

$$\text{SRMR} = \sqrt{\frac{1}{b} \left[ \sum_i \sum_j^p \frac{(s_{ij} - \hat{\sigma}_{ij})^2}{s_{ii}s_{jj}} + \sum_i^p \frac{(\bar{x}_i - \hat{\mu}_i)^2}{s_{ii}} \right]}$$

다음은 각 요소에 대한 설명입니다.

$p$  = 명시 변수의 수

$b$  = 공분산 행렬 및 관측 변수 평균 벡터의 고유 항목 수

$$b = \frac{p(p+1)}{2} + p$$

$s_{ij}$  = 입력 공분산 행렬의  $(i, j)$  번째 요소

$\hat{\sigma}_{ij}$  = 예측 공분산 행렬의  $(i, j)$  번째 요소

$\bar{x}_i$  = 표본 평균 벡터의  $i$  번째 요소

$\hat{\mu}_i$  = 예측 평균 벡터의  $i$  번째 요소

자세한 내용은 SAS Institute Inc. (2023a)의 "CALIS Procedure" 장에서 확인하십시오.

## 중재 분석에 대한 통계 상세 정보

이 섹션에서는 결과 ( $y$ ) 및 중재 ( $M$ ) 변수에 대해 중재 분석에서 사용되는 선형 방정식 시스템을 설명합니다. 예측 변수 ( $x$ ) 와 중재 변수 ( $M$ ), 예측 변수 ( $x$ ) 와 결과 변수 ( $y$ ) 간의 관계를 다음과 같이 표현할 수 있습니다.

$$M = \beta_{0M} + ax + e_M$$

$$y = \beta_{0y} + c'x + bM + e_y$$

첫 번째 방정식은 예측 변수( $x$ )가 중재 변수( $M$ )에 미치는 효과를 보여 줍니다. 예측 변수( $x$ )가 1 단위 증가할 때마다 중재 변수( $M$ )는  $a$  단위씩 증가할 것으로 예상됩니다. 두 번째 방정식은 중재 변수( $M$ )가 동일한 두 관측값의 경우 한 관측값의 예측 변수( $x$ ) 값이 클수록 결과 변수( $y$ )에  $c'$  단위의 변화가 있을 것으로 예상됨을 나타냅니다. 직접 효과( $c'$ )가 양수 값을 나타내는 경우 예측 변수( $x$ )가 증가할수록 결과 변수( $y$ )도 증가할 가능성이 높습니다. 반면에 효과가 음수인 경우 예측 변수( $x$ )가 증가하면 결과 변수( $y$ )가 감소할 가능성이 높습니다.

이러한 방정식은 중재 분석 컨텍스트에서 직접 효과와 간접 효과를 측정하기 위한 기본 프레임워크를 제공합니다. 다음과 같이 방정식을 조합할 수 있습니다.

$$y = \beta_{0y} + c'x + b(\beta_{0M} + ax + e_M) + e_y$$

$$y = \beta_{0y} + c'x + b\beta_{0M} + abx + be_M + e_y$$

$$y = \beta_{0y} + b\beta_{0M} + (c' + ab)x + be_M + e_y$$

이 시나리오에서는 다음 사항이 적용됩니다.

$\beta_{0y} + b\beta_{0M}$ 은 새 절편입니다.

$c' + ab$ 는 새 기울기입니다. 이 새 기울기를 총 효과( $c$ )라고 하며 직접 효과( $c'$ )와 간접 효과( $ab$ )로 구성됩니다.

$be_M + e_y$ 는 새 오차 항입니다.

총 효과는 다음과 같이 정의할 수 있습니다.

총 효과 ( $c$ ) = 직접 효과 ( $c'$ ) + 간접 효과 ( $ab$ )

$$c = c' + ab = \{\beta_{0y} + c'x\} + \{(\beta_{0M} + ax)(\beta_{0y} + bM)\}$$



## 요인 분석

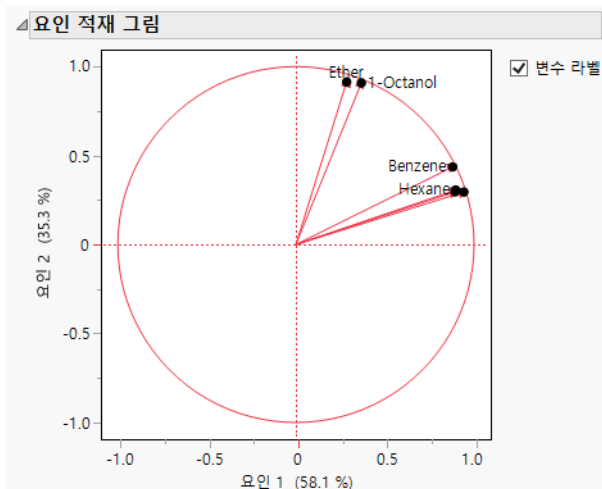
### 데이터의 잠재 변수 식별

요인 분석은 관측된 변수를 더 적은 수의 관측 불가능한 잠재 변수 또는 요인의 관점에서 설명하려고 합니다. 요인 분석은 공통 요인 분석 및 탐색 요인 분석이라고도 합니다. 요인은 관측 변수 (오차 포함)의 선형 결합으로 정의됩니다. 이러한 요인은 관측 변수에 공통된 변동을 설명하기 위해 구성됩니다. 요인 분석의 목적은 관측되지 않은 요인의 관점에서 관측 변수에 대한 의미 있는 해석을 찾는 것입니다. 변수 수를 줄이는 목적도 있습니다.

요인 분석은 심리학, 사회학 및 교육에 기원을 둔 여러 분야에 사용됩니다. 이러한 분야에서 요인 분석은 관측된 행동이 기본 패턴 및 구조 측면에서 어떻게 해석될 수 있는지 이해하는 데 사용됩니다. 예를 들어 외부 활동 참여, 취미, 운동 및 여행 측도는 모두 "적극적인 성격 유형과 소극적인 성격 유형"으로 설명될 수 있는 요인과 관련이 있을 수 있습니다.

데이터의 기본 패턴과 구조를 탐색하거나 해석해야 하는 경우 요인 분석을 사용합니다. 또한 더 적은 수의 잠재 변수를 사용하여 변수의 정보를 요약할 때도 사용해 보십시오.

**그림 9.1** 회전된 요인 적재



**목차**

요인 분석 플랫폼 개요.....	233
요인 분석 플랫폼의 예.....	233
요인 분석 플랫폼 시작.....	236
요인 분석 보고서 .....	236
모형 시작.....	238
회전 방법.....	239
요인 분석 플랫폼 옵션.....	240
요인 분석 모형 적합 옵션 .....	241

## 요인 분석 플랫폼 개요

요인 분석은 더 적은 수의 관측 불가능한 요인의 관점에서 관측 가능 변수 집합을 모델링합니다. 요인은 관측 변수 간의 상관 또는 공분산을 설명하기 위해 구성됩니다. 요인의 해석력을 높이기 위해 요인 회전을 사용하여 요인의 기준 축을 변경할 수 있습니다.

10 개의 관측 변수  $X_1, X_2, \dots, X_{10}$  이 있는 경우를 가정해 보겠습니다. 이러한 10 개의 변수를  $F_1$  과  $F_2$  라는 두 잠재 요인의 관점에서 모델링하려고 합니다. 편의상 요인은 상관관계가 없으며 각 요인의 평균은 0, 분산은 1 이라고 가정합니다. 도출하려는 모형의 형식은 다음과 같습니다.

$$X_i = \beta_{i0} + \beta_{i1}F_1 + \beta_{i2}F_2 + \varepsilon_i$$

따라서  $Var(X_i) = \beta_{i1}^2 + \beta_{i2}^2 + Var(\varepsilon_i)$ 입니다.  $X_i$ 의 분산 중 요인, 즉 공통 분산 또는 공통분으로 인한 부분은  $\beta_{i1}^2 + \beta_{i2}^2$ 입니다. 나머지 분산  $Var(\varepsilon_i)$ 는 고유 분산이며  $X_i$ 에 고유한 특정 분산과 오차 분산의 조합으로 간주됩니다.

이 플랫폼에서는 상관 행렬 또는 공분산 행렬의 고유값에 대한 스크리 그림을 제공합니다. 이 그림을 참고로 사용하여 추출할 요인 수를 결정할 수 있습니다. 플랫폼의 기본 요인 수는 1을 초과하는 고유값 수입니다.

이 플랫폼에서는 이 모형의 모수를 추정하기 위해 "주 축"과 "최대 가능도"라는 두 가지 요인화 방법을 제공합니다. 각 변수의 공통 요인이 기여하는 분산 비율을 추정하기 위한 두 가지 "사전 공통분" 옵션이 있습니다. 이러한 옵션은 상관 행렬 또는 공분산 행렬의 대각에 대한 가정을 설정합니다. "주성분" 옵션은 대각의 값이 1인 상관 행렬 또는 대각의 값이 분산인 공분산 행렬을 분석할 구조로 처리합니다. "공통 요인 분석" 옵션은 대각 항목을 SMC(다중상관제곱)로 설정합니다. 이러한 값은 다른 변수와 공유하는 변동 비율을 반영합니다.

요인 회전은 추출된 요인의 해석력을 뒷받침하는 데 사용됩니다. 요인 분석 플랫폼에서는 직교 회전과 사각 회전을 모두 포함하는 다양한 회전 방법을 제공합니다.

공통 분산을 고려하는 요인 분석과 달리 주성분 분석은 관측 변수의 총 분산을 설명합니다. 자세한 내용은 "주성분"에서 확인하십시오.

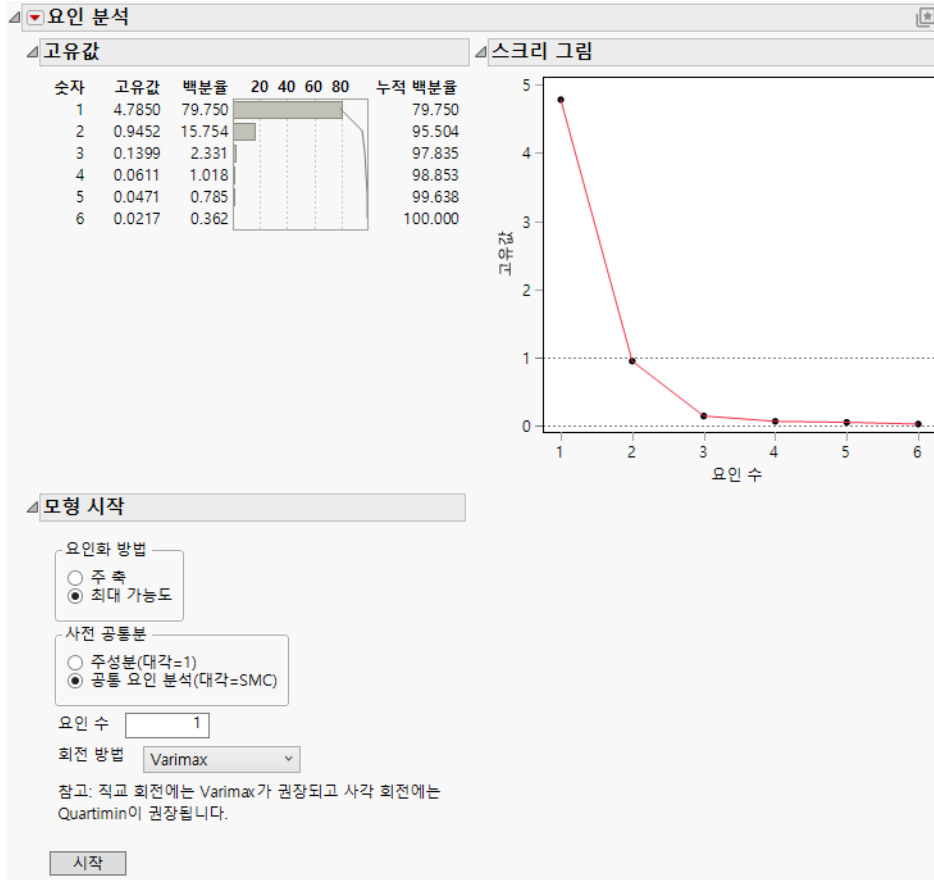
요인 분석에 대한 자세한 내용은 Jöreskog(1977) 또는 Cudeck 과 MacCallum(2007) 연구 자료에서 확인하십시오.

## 요인 분석 플랫폼의 예

요인 분석 플랫폼을 사용하여 6 개의 용매로 설명되는 두 가지 요인을 추출합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Solubility.jmp 를 엽니다.
2. **분석 > 다변량 방법 > 요인 분석**을 선택합니다.
3. 1-Octanol ~ Hexane 을 선택하고 **Y, 열**을 클릭합니다.
4. **확인**을 클릭합니다.

그림 9.2 초기 요인 분석 보고서



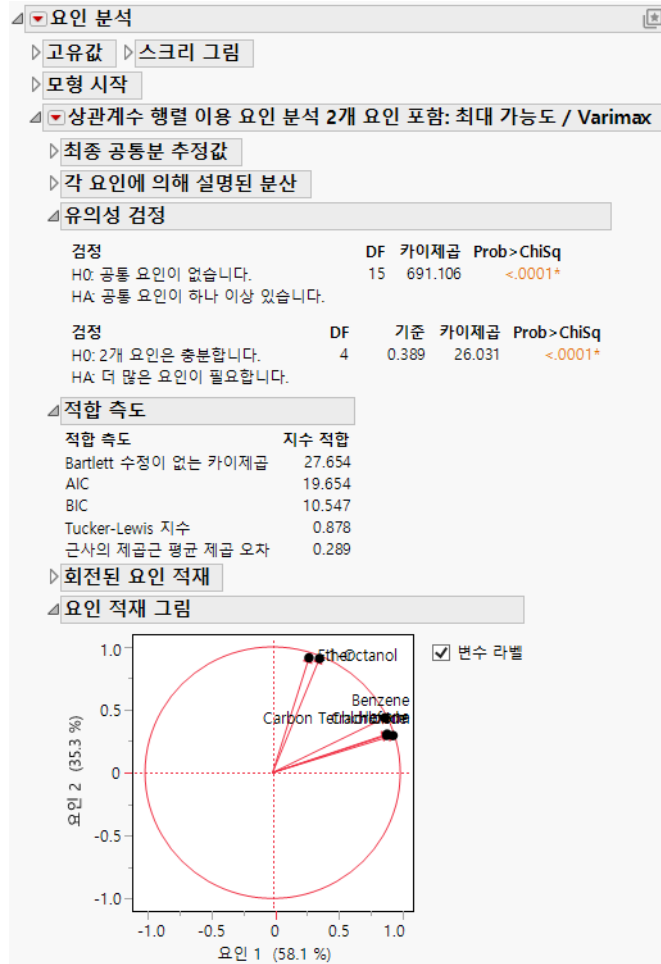
스크리 그림의 팔꿈치 부분을 보면 요인이 두 개인 모형이 적절하다는 것을 알 수 있습니다.

5. "모형 시작" 개요에서 다음과 같이 선택합니다.

- 요인화 방법 = **최대 가능성도**
- 사전 공통분 = **공통 요인 분석**
- 요인 수 = 2
- 회전 방법 = **Varimax**

6. **시작**을 클릭합니다.

그림 9.3 요인 분석 보고서



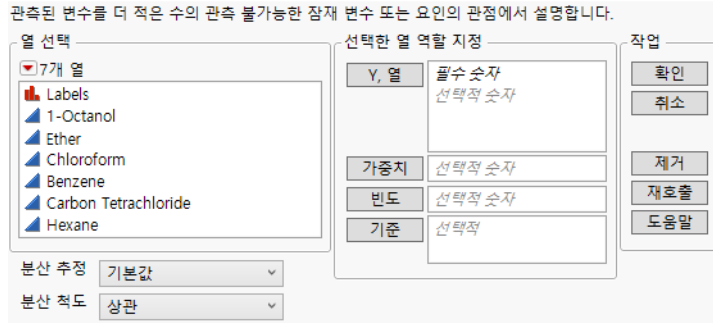
보고서에는 공통분 추정값, 분산 추정값, 유의성 검정, 적합 속도, 회전된 요인 적재 및 요인 적재 그림이 나열됩니다. "회전된 요인 적재" 및 "요인 적재 그림"에서는 요인 1이 Carbon Tetrachloride-Chloroform-Benzene-Hexane 변수 군집과 관련되고, 요인 2가 Ether-1-Octanol 변수 군집과 관련된다는 것을 나타냅니다. 보고서에 표시된 정보에 대한 자세한 내용은 "요인 분석 모형 적합 옵션"에서 확인하십시오.

**팁:** "요인 적재 그림"의 점을 클릭하여 라벨을 선택하고 이동할 수 있습니다. 오른쪽 아래 모서리를 클릭하면 라벨을 더 쉽게 볼 수 있도록 그림 크기가 커집니다.

## 요인 분석 플랫폼 시작

분석 > 다변량 방법 > 요인 분석을 선택하여 요인 분석 플랫폼을 시작합니다.

그림 9.4 요인 분석 시작 창



"열 선택"의 빨간색 삼각형 메뉴에 포함된 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

**Y, 열** 분석할 열입니다. 데이터 유형이 숫자여야 합니다.

**가중치** 데이터 테이블의 각 관측값에 대한 가중치를 포함하는 열입니다. 값이 0보다 큰 행만 분석에 포함됩니다.

**빈도** 분석의 각 행에 빈도를 할당합니다. 데이터를 요약할 경우에 유용합니다.

**기준** 기준 변수의 각 수준에 대해 개별 보고서를 생성합니다. 기준 변수가 둘 이상 할당되면 기준 변수의 가능한 각 수준 조합에 대해 개별 보고서가 생성됩니다.

**분산 추정** 분석을 위한 분산 - 공분산 행렬을 추정하는 방법을 나열합니다. 방법에 대한 자세한 내용은 "다변량 플랫폼 시작"에서 확인하십시오.

**분산 척도** 요인 분석의 분산 척도에 사용되는 방법을 지정합니다. 다음 옵션을 사용할 수 있습니다.

**상관** 상관계수 행렬을 이용한 분석을 수행할 수 있는 기본 방법입니다.

**공분산** 가중치가 변수의 분산인 가중 상관 행렬을 이용한 분석을 수행할 수 있습니다.

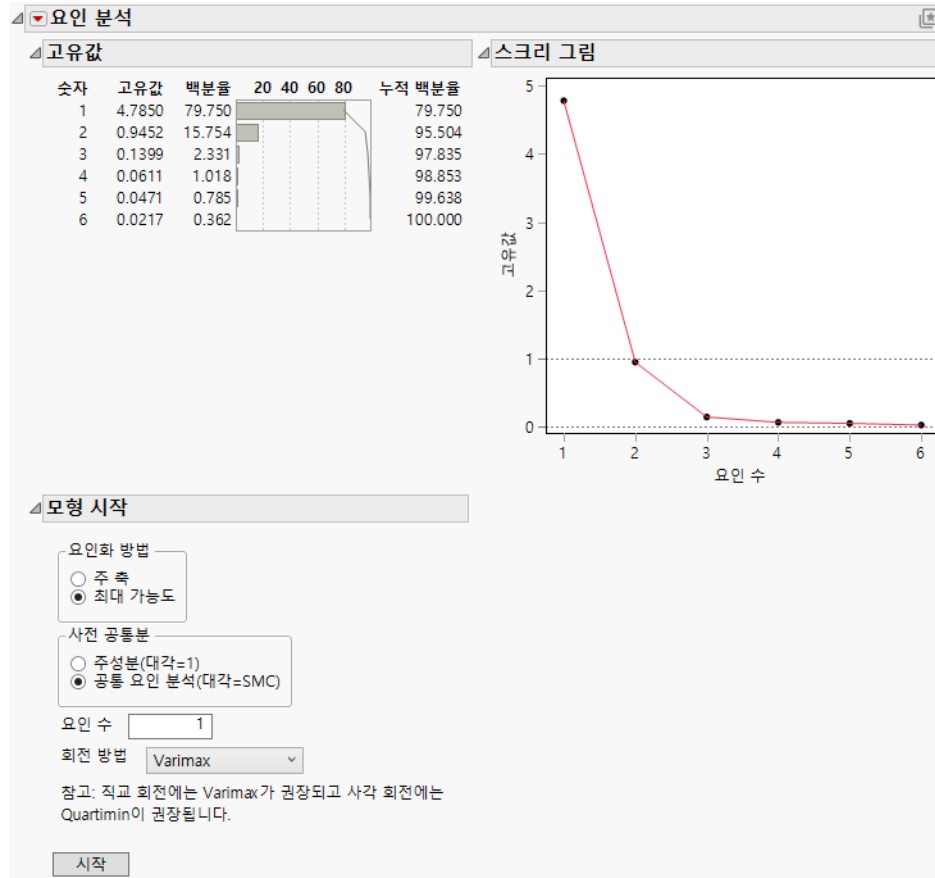
**비척도화 및 비중심화** 이미 중심화 및 척도화된 변수를 분석할 수 있습니다.

## 요인 분석 보고서

초기 요인 분석 보고서에는 고유값과 스크리 그림이 표시됩니다. 고유값은 주성분 분석을 통해 계산됩니다. 스크리 그림에는 이러한 고유값이 그래프로 표시됩니다. "모형 시작"의 초기 요인 수는 1.0을 초과하는 고유값 수와 같습니다. 추출할 요인 수를 변경할 수 있습니다.

요인 수를 결정할 때 스크리 그림을 참고로 사용할 수 있습니다. 스크리 그림이 평평하게 안정되기 전의 고유값 수는 요인 수에 대한 상한을 제공합니다. 이 예에서는 스크리 그림에 의해 두 개의 성분(요인)이 제안됩니다.

그림 9.5 요인 분석 보고서



"고유값" 테이블을 보면 첫 번째 고유값이 변동의 79.75%를 설명하고 두 번째 고유값이 15.75%를 설명한다는 것을 알 수 있습니다. 따라서 처음 두 고유값이 총 변동의 95.50%를 설명합니다. 세 번째 고유값은 변동의 2.33%만 설명하고 나머지 고유값의 기여도는 무시해도 될 정도의 수준입니다. **요인 수** 상자는 처음에 1로 설정되었지만 이 분석을 통해 2개 요인을 추출하는 것이 적절하다는 것을 알 수 있습니다.

## 모형 시작

"모형 시작" 제어판을 사용하여 요인 분석 모형을 구성합니다. **시작**을 클릭하면 요인 분석 결과를 얻을 수 있습니다.

그림 9.6 모형 시작

**요인화 방법** 요인 추출 방법을 정의합니다.

**주축** 행렬의 대각 항목이 변수의 공통분 추정값으로 대체된 축소 상관 행렬 또는 축소 공분산 행렬에 대해 고유값 분해를 수행합니다. 이 방법은 계산이 효율적이지만 가설 검정을 허용하지 않습니다.

**최대 가능성도** 공통 요인 수에 대한 가설을 검정하고 모형 적합 통계량을 구할 수 있습니다.

**참고 :** 최대 가능성도 방법을 사용하려면 양정치 상관 행렬이 필요합니다. 상관 행렬이 양정치가 아닌 경우 주축 방법을 선택하십시오.

**사전 공통분** 각 변수의 공통 요인이 기여하는 분산 비율을 추정하는 방법을 정의합니다.

**주성분 (대각 = 1)** 모든 공통분을 1로 설정하여 모든 요인에 의해 각 변수의 분산이 100% 설명된다는 것을 나타냅니다.

**팁 :** "요인화 방법"을 "주축"으로 설정하고 이 옵션을 사용하면 주성분 분석이 수행됩니다.

**공통 요인 분석 (대각 = SMC)** 공통분을 SMC( 다중상관제곱 ) 계수로 동일하게 설정합니다. 변수가 주어지면 SMC는 다른 모든 변수에 대한 해당 변수의 회귀 분석  $R^2$ 입니다.

**요인 수** 분석에서 추출할 요인 수를 지정합니다. 기본값은 1.0보다 크거나 같은 고유값 수입니다. 요인 수를 1 이상으로 설정할 수 있으며 변수의 수를 초과할 수 없습니다.

**회전 방법** 회전 방법을 정의합니다. 기본값은 "Varimax"입니다. 사용 가능한 회전 방법에 대한 설명은 "회전 방법"에서 확인하십시오.

## 회전 방법

요인 분석 플랫폼에서 회전은 요인을 더 쉽게 해석할 수 있도록 요인의 기준 축을 변경하는 데 사용됩니다. 데이터에서 추출된 요인에 회전이 적용됩니다. 회전 방법은 다양한 복잡성 또는 단순성 함수를 기반으로 합니다. 회전에 대한 자세한 내용은 SAS Institute Inc. (2023c)의 "FACTOR Procedure" 장, Browne(2001) 또는 Frank & Todeschini(1994)에서 확인하십시오.

초기 추출 후에는 요인 간에 상관관계가 없습니다. 직교 변환에 의해 요인이 회전하는 경우 회전된 요인도 상관관계가 없습니다. 사각 변환에 의해 요인이 회전하는 경우 회전된 요인은 상관관계를 가집니다. 사각 회전이 직교 회전보다 해석 가능한 요인을 더 많이 생성합니다. 그러나 요인이 상관되면 변수를 설명할 때 요인 중요도를 결정하는 명확한 측도가 없어집니다.

### 직교 회전 방법

**Varimax** 모든 변수에 대해 요인 적재 제곱의 분산 합계를 최대화합니다. 가장 일반적인 이 방법을 사용하면 각 요인에 대한 각 변수의 적재 값이 작거나 커집니다.  $\gamma=1$  인 Orthomax 와 같습니다.

**Biquartimax** Varimax 및 Quartimax 회전의 균등 가중 해입니다.  $\gamma=0.5$  인 Orthomax 와 같습니다.

**Equamax** Varimax 회전과 Quartimax 회전 사이의 가중 해입니다.  $\gamma=N/2$  인 Orthomax 와 같습니다 ( $N$ =요인 수).

**Factorparsimax** 요인 복잡성을 최소화하는 것이 목표인 해입니다. 변수 복잡성은 알고리즘에서 고려되지 않으므로 이 방법을 사용하면 교차 적재가 발생할 수 있습니다.  $\gamma=N$  인 Orthomax 와 같습니다 ( $N$ =요인 수).

**Orthomax** 가중치가  $\gamma$  로 표시되는 일반적인 가중 회전 방법입니다. 대부분의 특정 직교 회전 방법은 특정  $\gamma$  를 사용하는 Orthomax 회전입니다.

**Parsimax** 변수와 요인 복잡성 간에 균형을 맞춥니다.  $\gamma=(I(N-1))/(I+N-2)$  인 Orthomax 와 같습니다 ( $I$ =항목 수,  $N$ =요인 수).

**Quartimax** 각 변수를 설명하는 데 필요한 요인 수를 최소화합니다.  $\gamma=1$  인 Orthomax 와 같습니다.

### 사각 회전 방법

**Biquartimin** 공분산 비율을 최소화하기 위한 회전입니다.  $\tau=0.5$  인 Oblimin 과 같습니다.

**Covarimin** 사각 Varimax 회전입니다.  $\tau=1$  인 Oblimin 과 같습니다.

**Obbiquartimax** 사각 Biquartimax 회전입니다.

**Obequamax** 사각 Equamax 회전입니다.

**Obfactorparsimax** 사각 요인 Parsimax 회전입니다.

**Oblimin** 가중치가  $\tau$  로 표시되는 일반적인 가중 사각 회전 방법입니다. 대부분의 특정 사각 회전 방법은 특정  $\tau$  를 사용하는 Oblimin 회전입니다.

**Obparsimax** 사각 Parsimax 회전입니다.

**Obquartimax** 사각 Quartimax 회전입니다 (Quartimin 방법과 동등).

**Obvarimax** 사각 Varimax 회전입니다.

**Quartimin** 사각 Quartimin 회전입니다(사각 Quartimax와 동등).  $\tau=0$ 인 Oblimin과 같습니다.

**Promax** Varimax 를 먼저 수행한 후 Procrustes 회전을 사용하여 간단한 구조를 얻는 2 단계 회전입니다. 이 방법은 Oblimin 대신 사용할 수 있으며 계산이 효율적입니다.

## 요인 분석 플랫폼 옵션

"요인 분석" 의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

**고유값** 원래 상관 행렬, 공분산 행렬 또는 비척도화/비중심화 행렬의 고유값 테이블을 표시하거나 숨깁니다. 이 테이블에는 각 고유값이 나타내는 총 분산의 백분율, 기여도 백분율을 보여주는 막대 차트 및 연속된 각 고유값이 기여하는 누적 백분율이 포함됩니다. 1.0 보다 크거나 같은 고유값 수를 분석할 요인 수의 지표로 사용할 수 있습니다.

**스크리 그림** 고유값 대 성분(요인) 수를 나타내는 그림을 표시하거나 숨깁니다. 이 그림을 추가 지표로 사용하여 최대 분산 크기에 기여하는 요인 수를 결정할 수 있습니다. 그림에 표시된 선이 평평하게 안정되는 점을 분석에 필요한 충분 요인 수로 사용할 수 있습니다.

**구형성 Bartlett 검정** 구형성 Bartlett 검정 결과를 표시하거나 숨깁니다. 이 검정은 검정에 대한 카이제곱, DF(자유도) 및  $p$  값(Prob > ChiSq) 을 계산하여 고유값의 분산이 동일한지 여부를 판별합니다. 자세한 내용은 Bartlett(1937, 1954) 연구 자료에서 확인하십시오.

**Kaiser-Meyer-Olkin 검정** KMO(Kaiser-Meyer-Olkin) 검정 결과를 표시하거나 숨깁니다. 이 검정은 기본 요인에 의한 공통 분산일 수 있는 분산 비율의 지표입니다. 검정 통계량은 각 개별 변수와 모든 변수 집합에 대해 계산되며, JMP에서는 MSA(표집 적합성 측도)라고 합니다. 다음 값은 요인 분석에 대한 변수의 적합성을 나타냅니다.

- 0.00 ~ 0.49 허용 불가
- 0.50 ~ 0.59 매우 나쁨
- 0.60 ~ 0.69 조금 나쁨
- 0.70 ~ 0.79 보통
- 0.80 ~ 0.89 조금 좋음
- 0.90 ~ 1.00 매우 좋음

**참고:** 상관 행렬이 특이 행렬인 경우 KMO 검정을 사용할 수 없습니다.

다음 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

**로컬 데이터 필터** 특정 보고서에서 사용되는 데이터를 필터링할 수 있는 로컬 데이터 필터를 표시하거나 숨깁니다.

**다시 실행** 분석을 반복하거나 다시 시작할 수 있는 옵션이 포함되어 있습니다. 이 기능을 지원하는 플랫폼에서 "자동 재계산" 옵션은 해당하는 보고서 창에서 데이터 테이블에 대한 변경 사항을 즉시 반영합니다.

**플랫폼 환경 설정** 현재 플랫폼 환경 설정을 보거나, 현재 JMP 보고서의 설정과 일치하도록 플랫폼 환경 설정을 업데이트할 수 있는 옵션이 포함되어 있습니다.

**스크립트 저장** 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다.

**그룹별 스크립트 저장** 기준 변수의 모든 수준에 대한 플랫폼 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다. 시작 창에서 기준 변수를 지정한 경우에만 사용할 수 있습니다.

---

**참고 :** 이 플랫폼의 추가 옵션은 스크립트를 통해 사용할 수 있습니다. 자세한 내용은 "도움말" 메뉴의 "스크립트 인덱스"에서 여십시오. 또한 "스크립트 인덱스"에서 이 섹션에 설명된 옵션의 스크립트 예제도 찾을 수 있습니다.

---

---

## 요인 분석 모형 적합 옵션

요인 분석 모형 적합의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

**사전 공통분** (공통 요인 분석에만 사용 가능) 각 변수에 대한 공통분의 초기 추정값을 표시하거나 숨깁니다. 변수가 주어지면 이 추정값은 SMC(다중상관제공) 계수이거나, 다른 모든 변수에 대한 해당 변수의 회귀 분석  $R^2$ 입니다.

---

**참고 :** Heywood 사례는 공통분이 1 이상일 때 발생합니다. 이는 고유한 요인 분산(공통 요인으로 설명되지 않는 잔차 분산)이 음수임을 의미합니다. Heywood 사례는 고유 분산에 하한을 지정하거나 공통분에 상한을 지정하여 처리됩니다.

---

**고유값** (공통 요인 분석에만 사용 가능) 축소 상관 행렬의 고유값과 해당 값이 설명하는 공통 분산의 백분율을 표시하거나 숨깁니다. 축소 상관 행렬은 대각 항목이 공통분 추정값으로 대체된 상관 행렬입니다. 고유값은 요인에 의해 설명되는 공통 분산을 나타냅니다. 축소 상관 행렬은 양정치가 아닐 수도 있고 음수 고유값을 가질 수 있으므로 "누적 백분율"이 100%를 초과할 수 있습니다.

이 테이블에는 분석을 위해 유지된 요인 수가 나타납니다.

**정렬되지 않고 회전되지 않은 요인 적재** 정렬 및 회전하기 전의 요인 적재 행렬을 표시하거나 숨깁니다.

**회전되지 않은 요인 적재** 회전하기 전의 요인 적재 행렬을 표시하거나 숨깁니다. 요인 적재는 공통 요인이 변수에 미치는 영향을 측정합니다. 회전되지 않은 요인은 직교하므로 요인 적재

행렬은 변수와 요인 간의 상관 행렬입니다. 적재 절대값이 1에 가까울수록 변수에 대한 요인의 효과가 커집니다.

테이블에서 슬라이더와 **다음보다 작은 절대 적재 값 표시 안 함**의 값을 사용하여 지정된 값 미만인 값을 표시하지 않을 수 있습니다. 표시되지 않는 값은 **텍스트 흐리게**에 지정한 설정에 따라 흐리게 나타납니다.

**텍스트 흐리게** 슬라이더와 값을 사용하여 **다음보다 작은 절대 적재 값 표시 안 함** 값보다 절대 값이 작은 요인 값에 대한 테이블의 글꼴 투명도 그래디언트를 제어할 수 있습니다.

**그림 9.7** 텍스트 흐리게 컨트롤이 있는 회전되지 않은 요인 적재

회전되지 않은 요인 적재		
	요인 1	요인 2
Benzene	0.977736	-0.096606
Carbon Tetrachloride	0.956158	-0.249473
Hexane	0.923401	-0.215765
Chloroform	0.915425	-0.222605
1-Octanol	0.792777	0.575359
Ether	0.725745	0.623805

다음보다 작은 절대 적재 값 표시 안 함

텍스트 흐리게

**참고 :** " 회전되지 않은 요인 적재 " 행렬은 동일한 요인과 관련된 변수가 서로 옆에 나타나도록 정렬됩니다.

**회전 행렬** 요인 적재 그림과 요인 적재 행렬을 회전하는 데 사용된 값을 표시하거나 숨깁니다.

**요인 간 상관** (사각 회전에만 사용 가능) 요인 간 상관 행렬을 표시하거나 숨깁니다.

**목표 행렬** (Promax 회전에만 사용 가능) Varimax 요인 패턴이 회전되는 행렬을 표시하거나 숨깁니다.

**요인 구조** (사각 회전에만 사용 가능) 변수와 공통 요인 간의 상관 행렬을 표시하거나 숨깁니다.

**최종 공통분 추정값** 요인 모형이 적합된 후 공통분 추정값을 표시하거나 숨깁니다. 요인이 적교할 경우 변수의 최종 공통분 추정값은 해당 변수의 적재 제곱합과 같습니다.

**표준 스코어 계수** 회전된 요인을 소스 데이터 테이블에 저장할 때 요인 스코어를 추정하는 데 사용되는 승수 테이블을 표시하거나 숨깁니다.

**각 요인에 의해 설명된 분산** (적교 회전에만 사용 가능) 회전된 각 요인에 의해 설명되는 공통 분산의 분산, 백분율 및 누적 백분율을 표시하거나 숨깁니다.

**다른 요인을 무시하고 각 요인에 의해 설명된 분산** (사각 회전에만 사용 가능) 다른 요인에 관계없이 회전된 각 요인에 의해 설명되는 공통 분산의 분산 및 백분율을 표시하거나 숨깁니다.

**유의성 검정** (최대 가능도 요인화 방법에만 사용 가능) 두 개의 카이제곱 검정 결과를 제공합니다.

첫 번째 검정은 " $H_0$ : 공통 요인이 없습니다."에 대한 것입니다. 이 귀무가설은 변수 간 상호 상관관계를 설명하는 공통 요인이 없음을 나타냅니다. 이 검정은 요인의 상관 행렬이 단위 행렬이라는 귀무가설을 사용하는 구형성 Bartlett 검정입니다 (Bartlett, 1954 참조).

두 번째 검정은 " $H_0$ :  $N$  개 요인은 충분합니다."에 대한 것입니다. 여기서  $N$  은 지정된 요인 수입니다. 이 귀무가설의 기각은 변수 간 상호 상관관계를 설명하기 위해 요인이 더 필요할 수 있음을 나타냅니다 (Bartlett, 1954 참조). "기준" 은 로그 가능도 목적 함수 값입니다.

**적합 척도** ( 최대 가능도 요인화 방법에만 사용 가능 ) "Bartlett 수정이 없는 카이제곱 ", "AIC", "BIC", "Tucker-Lewis 지수" 및 "근사의 제곱근 평균 제곱 오차" 를 포함한 적합 척도를 표시하거나 숨깁니다.

**요인 스코어 척도** "다중 R", "다중 R<sup>2</sup>" 및 "최소 상관" 스코어를 포함한 요인 스코어 결정 척도를 표시하거나 숨깁니다. 이러한 척도는 요인 스코어가 2 차 분석에 유용한지 여부를 평가하는데 사용됩니다.

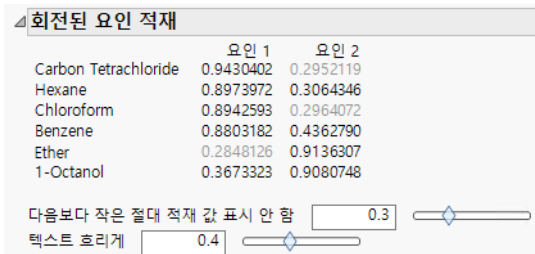
**정렬되지 않고 회전된 요인 적재** 회전 후 정렬되지 않은 요인 적재 행렬을 표시하거나 숨깁니다.

**회전된 요인 적재** 회전 후 요인 적재 행렬을 표시하거나 숨깁니다. 직교 회전을 사용한 경우 이러한 값은 변수와 회전된 요인 간의 상관입니다.

테이블에서 슬라이더와 **다음보다 작은 절대 적재 값 표시 안 함**의 값을 사용하여 지정된 값 미만인 값을 표시하지 않을 수 있습니다. 표시되지 않는 값은 **텍스트 흐리게**에 지정된 설정에 따라 흐리게 나타납니다.

**텍스트 흐리게** 슬라이더와 값을 사용하여 테이블의 글꼴 투명도 그래디언트를 제어할 수 있습니다. 값이 낮을수록 **다음보다 작은 절대 적재 값 표시 안 함**에 지정된 값보다 절대값이 작은 요인 값의 글꼴이 더 투명해집니다.

**그림 9.8** 텍스트 흐리게 컨트롤이 있는 회전된 요인 적재



**참고:** "회전된 요인 적재" 행렬은 동일한 요인과 관련된 변수가 서로 옆에 나타나도록 정렬됩니다.

**요인 적재 그림** 회전된 요인 적재 그림을 표시하거나 숨깁니다. 세 개 이상의 요인을 모델링하는 경우 요인 적재 그림은 그림의 행렬입니다.

**스코어 그림** 추정된 요인 스코어의 산점도를 표시하거나 숨깁니다. 세 개 이상의 요인을 모델링하는 경우 스코어 그림은 그림의 행렬입니다.

**결측값 대치를 사용한 스코어 그림** (결측값이 있는 경우에만 사용 가능) 결측값을 대치 값으로 채우고 추정된 요인 스코어의 산점도를 표시하거나 숨깁니다.

**표시 옵션** 적재 그림에 화살표를 표시하거나 숨길 수 있습니다.

**요인 스코어 저장** 새 계산식 열을 원래 데이터 테이블에 저장합니다. 새 열에는 Thurstone 방법으로 추정된 요인 스코어의 계산식이 포함됩니다.

---

**참고 :** 계산식은 결측값이 있는 행을 계산할 수 없습니다.

---

**SEM 에 대한 모형 규격 복사** 요인 정의를 클립보드에 복사합니다. 그런 다음 정의를 SEM 플랫폼에 붙여 넣을 수 있습니다. SEM JSL 스크립트에는 회전된 최종 적재 행렬을 기반으로 하는 고정 적재가 포함됩니다. 이 스크립트는 **다음보다 작은 절대 적재 값 표시 안 함** 임계보다 큰 적재만 포함합니다.

**결측값 대치를 사용한 요인 스코어 저장** (결측 데이터가 있는 모형에 사용 가능) 새 계산식 열을 원래 데이터 테이블에 저장합니다. 새 열에는 결측값의 대치 값과 함께 요인 스코어의 계산식이 포함됩니다.

**적합 제거** "요인 분석" 보고서에서 모형 적합 결과를 제거합니다. 이 옵션을 사용하면 새 보고서에 대한 "모형 시작" 구성을 변경할 수 있습니다.

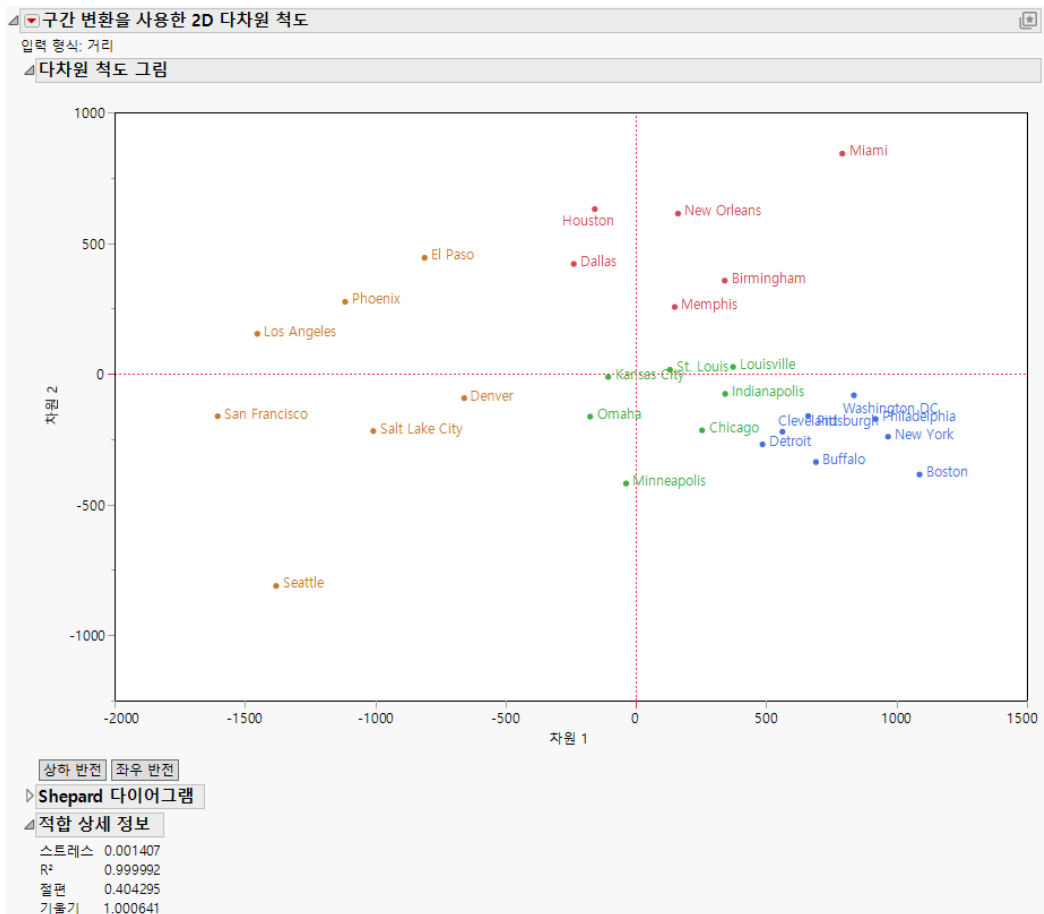
# 10 장

## 다차원 척도법 개체 집합 간의 근접성 시각화

MDS(다차원 척도법)는 개체 집합 간의 근접성 패턴(유사성, 비유사성 또는 거리)을 시각적으로 표현하는 데 사용되는 기법입니다. 예를 들어 도시 간의 거리 행렬이 주어진 경우 MDS를 사용하여 2차원 도시 맵을 생성할 수 있습니다.

다차원 척도법은 연구원이 브랜드, 취향 또는 기타 상품 속성과 관련된 인식 측도를 가지고 있는 소비자 조사에 주로 사용됩니다. MDS는 근접성 또는 속성 집합을 기반으로 개체의 근접성을 시각화하는 데 관심이 있는 다른 많은 분야에 적용할 수 있습니다.

그림 10.1 다차원 척도법의 예



## 목차

다차원 척도법 플랫폼 개요.....	247
다차원 척도법의 예.....	247
다차원 척도법 플랫폼 시작.....	250
다차원 척도 보고서.....	251
다차원 척도 그림.....	251
Shepard 다이어그램.....	252
적합 상세 정보.....	252
다차원 척도법 플랫폼 옵션.....	252
Waern 링크.....	253
다차원 척도법의 추가 예.....	254
다차원 척도법 플랫폼에 대한 통계 상세 정보.....	256
스트레스 함수에 대한 통계 상세 정보.....	256
변환에 대한 통계 상세 정보.....	256
속성 목록 형식에 대한 통계 상세 정보.....	257

---

## 다차원 척도법 플랫폼 개요

MDS(다차원 척도법) 플랫폼에서는 개체 집합 간의 근접성 그림을 생성합니다. 이 그림을 사용하여 데이터 집합의 구조를 시각적으로 탐색할 수 있습니다. MDS는 적은 수의 차원에서 개체 집합 간의 근접성 패턴(거리, 유사성)을 시각화하는 데 사용되는 다변량 기법입니다. MDS는 거리 행렬에 적용됩니다. 스트레스 함수(실제 근접성과 예측된 근접성의 차이)를 최소화하여 MDS 그림의 좌표를 얻습니다.

거리라는 용어가 물리적 거리의 측도를 나타낼 수 있습니다(예: 도시 간). 거리는 정확한 측정보다는 주관적 평가인 경우가 더 많습니다. 근접성은 제품 브랜드 간의 인지된 유사성, 범죄율의 상관관계 또는 표본 국가에 대한 경제적 유사성을 측정할 수 있습니다. 거리를 근접성 또는 유사성(비유사성)이라고 할 수도 있습니다. 속성 목록 데이터가 주어진 경우 먼저 속성 목록에서 거리 행렬이 생성됩니다.

다차원 척도법에 대한 자세한 내용은 Borg와 Groenen(2005) 또는 Jackson(2003) 연구 자료에서 확인하십시오.

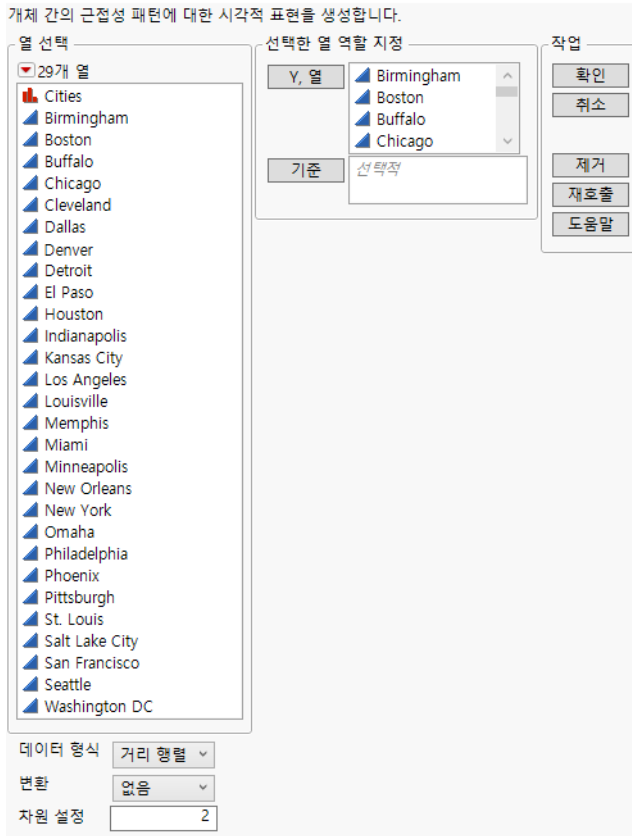
---

## 다차원 척도법의 예

이 예에서는 MDS를 사용하여 쌍별 비행 거리를 기반으로 28개 도시의 2차원 맵을 생성합니다. 비행 거리를 포함하는 데이터 테이블은 거리 행렬입니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Flight Distances.jmp를 엽니다.
2. **분석 > 다변량 방법 > 다차원 척도법**을 선택합니다.
3. Birmingham ~ Washington DC를 선택하고 **Y, 열**을 클릭합니다.
4. "데이터 형식" 메뉴에서 **거리 행렬**을 선택합니다.

그림 10.2 완료된 다차원 척도법 시작 창

5. **확인**을 클릭합니다 .

"다차원 척도 그림"에서 데이터 점을 커서로 가리키면 행 번호 또는 행 라벨이 표시됩니다.  
다음 7개 단계를 수행하여 MDS 그림에 라벨을 지정하고 그림을 회전합니다 .

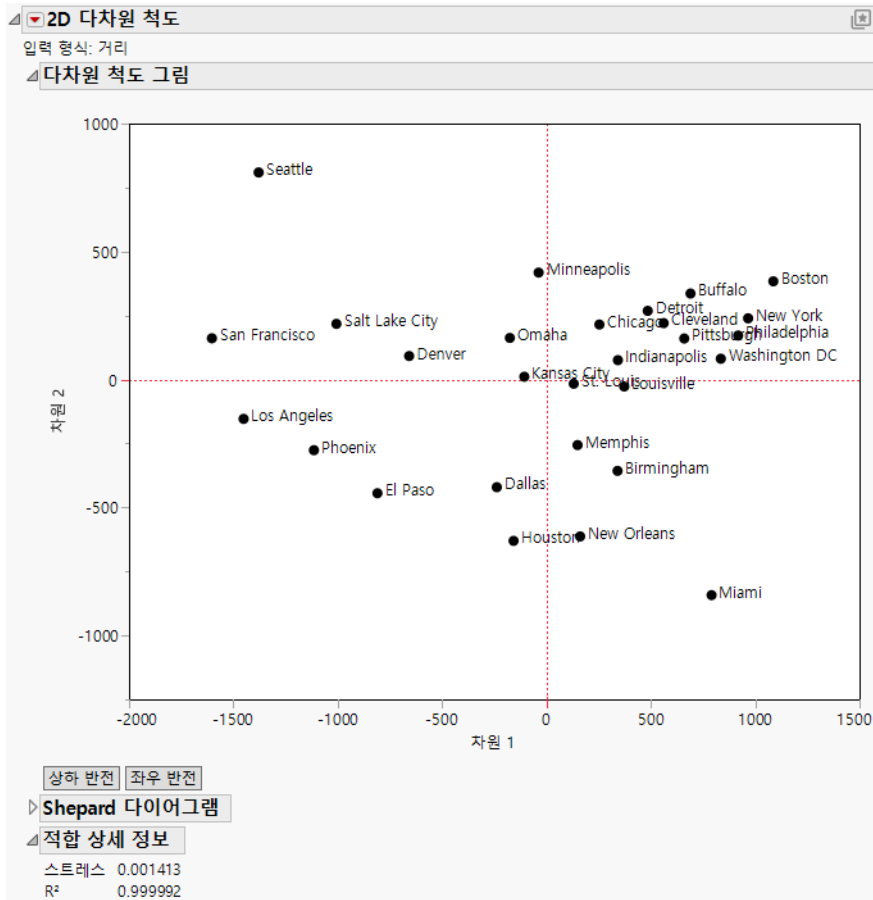
6. **Flight Distances** 데이터 테이블을 선택합니다 .7. **Cities** 열을 마우스 오른쪽 버튼으로 클릭하고 **라벨 / 라벨 해제**를 선택합니다 .8. **행 > 행 선택 > 모든 행 선택**을 선택합니다 .9. **행 > 라벨 / 라벨 해제**를 선택합니다 .

## 10. "다차원 척도 그림"을 선택합니다 .

11. **좌우 반전** 버튼을 클릭합니다 .

"상하 반전" 및 "좌우 반전" 버튼을 사용하여 MDS 그림의 방향을 변경할 수 있습니다. MDS 결과는 방향에 영향을 받지 않습니다. 결과에 지도의 물리적 위치와 같이 알려진 방향이 있는 경우 그림을 회전하거나 뒤집을 수 있습니다 .

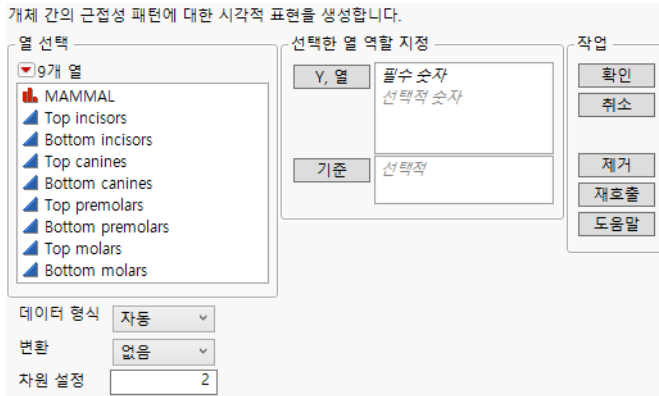
그림 10.3 다차원 척도 그림



## 다차원 척도법 플랫폼 시작

분석 > 다변량 방법 > 다차원 척도법을 선택하여 다차원 척도법 플랫폼을 시작합니다.

그림 10.4 다차원 척도법 시작 창



"열 선택"의 빨간색 삼각형 메뉴에 포함된 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

**Y, 열** 분석할 열입니다. 데이터 유형이 숫자여야 합니다.

**기준** 개별 분석을 정의하는 수준이 포함된 하나 이상의 열입니다. 지정된 열의 각 수준에 대해 해당 행은 사용자가 지정한 다른 변수를 사용하여 분석됩니다. 결과는 개별 보고서에 제공됩니다. 기준 변수가 둘 이상 할당되면 기준 변수의 가능한 각 수준 조합에 대해 개별 보고서가 생성됩니다.

**참고:** 거리 행렬을 사용하는 경우 기준 변수에 해당 기준 변수의 각 수준에 대한 완전 행렬이 필요합니다.

**데이터 형식** MDS는 다음과 같은 두 가지 데이터 형식을 지원합니다. 형식을 지정하거나, JMP에서 데이터 형식을 추론하게 할 수 있습니다.

**자동** 데이터 형식이 추론됩니다. 발견적 방법을 사용하여 거리 행렬 테이블을 식별하고 다른 모든 데이터 테이블은 속성 목록 형식을 갖는 것으로 가정됩니다.

**거리 행렬** 행 수가 열 수와 동일한 완전 대칭, 하삼각 또는 상삼각 행렬입니다. 대각 항목은 0 또는 결측일 수 있습니다.

**속성 목록** 개체의 품질 또는 특성에 대한 측도가 포함된 열 집합입니다. 측도가 동일한 척도를 사용한다고 가정합니다. 일반적으로 열에 개체 이름이 지정됩니다. 개체 열은 분석에 사용되지 않고 MDS 그림에서 데이터 점의 라벨로 사용됩니다.

**변환** 지원되는 변환은 "비", "구간" 및 "순서형"입니다.

**없음** 변환이 사용되지 않습니다.

**비** 데이터에 가장 작은 것부터 큰 것까지 순서가 있고, 값의 차이에 의미가 있으며, 척도 0은 절대 0입니다. MDS 그림을 척도화하는 데 사용됩니다.

**구간** 데이터에 가장 작은 것부터 큰 것까지 순서가 있고, 값의 차이에 의미가 있습니다. MDS 그림을 척도화하고 이동하는 데 사용됩니다.

**순서형** 데이터에 가장 작은 것부터 큰 것까지 순서가 있습니다. 순서형 데이터에 사용됩니다.

**차원 설정** 개체 간의 근접성을 시각적으로 표현하기 위한 차원 수입니다. 일반적으로 2 차원 또는 3 차원이 사용됩니다. 3 차원보다 크면 시각화가 복잡해집니다.

거리 행렬을 사용하는 경우 1에서  $n-1$  사이의 차원을 선택할 수 있습니다 ( $n$  = 개체 수).

속성 목록을 사용하는 경우 다음 사항에 유의하십시오.

-  $n > p$  인 경우 차원 수는 1에서  $p$  사이일 수 있습니다 ( $p$  = 속성 수).

-  $n \leq p$  인 경우 1에서  $n-1$  사이의 차원을 선택할 수 있습니다 ( $n$  = 개체 수).

거리 행렬 처리 및 다차원 척도 최적화는 계산량이 많은 프로세스입니다. 큰 거리 행렬을 처리할 때 진행률 창을 사용하여 프로세스를 모니터링하거나 취소할 수 있습니다. 최적화의 경우 MDS에서는 비볼록 최적화 알고리즘을 사용합니다. 로컬 최적 해를 방지하기 위해 시작 값을 여러 개 사용하는 것이 중요합니다. 중간 규모에서 대규모 데이터 집합의 경우 대화식 진행률 창을 통해 최적화 프로세스를 모니터링하고, 현재 시작에 대한 추정값 또는 현재 해를 수락할 수 있습니다.

---

## 다차원 척도 보고서

초기 MDS 보고서에는 다음 섹션이 포함되어 있습니다.

- "다차원 척도 그림"
- "Shepard 다이어그램"
- "적합 상세 정보"

시작 창에서 적합 차원을 3 이상으로 지정하면 표시할 차원을 선택할 수 있는 컨트롤이 "MDS 그림"에 제공됩니다.

MDS 그림에서 서로 가까이 있는 개체는 유사한 속성을 공유합니다. 그림에 라벨과 색상을 추가하면 유사한 그룹을 식별하는 데 도움이 될 수 있습니다. Shepard 다이어그램과 적합 통계량 요약은 MDS 그림이 개체의 근접성을 얼마나 잘 나타내는지 보여 줍니다.

## 다차원 척도 그림

MDS 그림에는 다차원 척도가 2차원으로 표시됩니다. 그림 아래에는 축을 상하 또는 좌우로 반전할 수 있는 두 개의 버튼이 있습니다. 점 간 근접성을 변경하지 않고 MDS 해를 반사, 회전 또

는 이동할 수 있습니다. 축을 반사 또는 회전하는 것은 맵 방향이 알려진 지리적 개체로 작업할 때 가장 일반적입니다.

시작 창에 세 개 이상의 차원이 지정된 경우 그림 아래의 **차원 선택** 컨트롤을 사용하여 그림에 표시되는 차원을 전환하거나 단계별로 이동할 수 있습니다. 첫 번째 컨트롤은 그림의 가로 축을 정의하고 두 번째 컨트롤은 그림의 세로 축을 정의합니다.

## Shepard 다이어그램

"다차원 척도" 보고서의 "Shepard 다이어그램"은 실제 또는 변환된 근접성 대 예측된 근접성을 보여 주는 그림입니다. 이 그림에는 다차원 척도 그림에 실제 근접성이 얼마나 잘 반영되는지 나타냅니다. Shepard는 실제값 대 예측값 그림과 유사합니다. 빨간색으로 표시된  $Y=X$  선 위에 점이 있는 그림이 이상적입니다.

## 적합 상세 정보

"다차원 척도" 보고서의 "적합 상세 정보" 섹션에는 MDS 근접성이 실제 근접성과 얼마나 일치하는지 요약하는 통계량과 사용된 변환 정보가 제공됩니다.

**스트레스** 적합 절차에서 최소화된 스트레스 함수 (스트레스 1)의 값입니다. 스트레스 값은 0에서 1 사이일 수 있으며 값이 작을수록 더 나은 적합을 나타냅니다.

**$R^2$**  실제 또는 변환된 근접성 대 예측된 근접성의 선형 적합에 대한  $R^2$  값입니다.

**기울기** 비율 또는 구간 변환이 사용된 경우 변환에 대한 기울기가 제공됩니다. 변환된 근접성에 대한 실제 근접성의 선형 회귀 기울기입니다.

**절편** 구간 변환이 사용된 경우 변환에 대한 절편이 제공됩니다. 변환된 근접성에 대한 실제 근접성의 선형 회귀 절편입니다.

---

## 다차원 척도법 플랫폼 옵션

"다차원 척도"의 빨간색 삼각형 메뉴 옵션을 사용하면 필요에 따라 보고서를 사용자 정의할 수 있습니다. 사용할 수 있는 옵션은 분석에 사용하는 데이터 유형 및 차원 수에 따라 결정됩니다.

**MDS 그림** MDS 그림을 표시하거나 숨깁니다.

**진단** MDS에 대한 진단을 제공합니다.

**Shepard 다이어그램** 실제 근접성 (변환이 사용된 경우 변환된 근접성) 대 예측된 근접성 그림을 표시합니다. 이 보고서는 기본적으로 나타냅니다. 자세한 내용은 "**Shepard 다이어그램**"에서 확인하십시오.

**Waern 링크** MDS 그림에 Waern 링크를 표시합니다. 이 옵션을 선택하면 링크를 표시할 부분 (가장 작은 부분 또는 가장 큰 부분)을 결정하는 컨트롤을 사용할 수 있습니다. 자세한 내용은 "**Waern 링크**"에서 확인하십시오.

**좌표 표시** 해 좌표 보고서를 제공합니다. 이는 "다차원 척도 그림"에 표시된 점의 좌표입니다. 보고서에는 최대 3개 차원의 좌표가 표시됩니다. 보고서를 마우스 오른쪽 버튼으로 클릭하고 **열**을 선택하여 보고서에 차원을 추가할 수 있습니다. 최대 차원 수는 시작 창에 설정한 차원 수입니다.

**근접성 표시** 근접성 보고서를 제공합니다. 각 개체 쌍 사이의 원래 근접성과 도출된 근접성 (거리)이 제공됩니다. 쌍은 "시작 개체" 열과 "끝 개체" 열에서 식별됩니다. 변환이 사용된 경우 변환된 근접성도 테이블에 포함됩니다.

**근접성 저장** (데이터 형식이 "속성 목록"인 경우에만 사용 가능) 거리 행렬을 데이터 테이블에 저장합니다.

**3D 그림** (시작 창에서 "차원 설정"에 세 개 이상의 차원을 지정한 경우에만 사용 가능) 처음 세 개 차원의 3D 그림을 표시합니다.

**좌표 저장** 해 좌표를 데이터 테이블에 별도의 열로 저장합니다.

다음 옵션에 대한 자세한 내용은 **JMP 사용**에서 확인하십시오.

**로컬 데이터 필터** 특정 보고서에서 사용되는 데이터를 필터링할 수 있는 로컬 데이터 필터를 표시하거나 숨깁니다.

**다시 실행** 분석을 반복하거나 다시 시작할 수 있는 옵션이 포함되어 있습니다. 이 기능을 지원하는 플랫폼에서 "자동 재계산" 옵션은 해당하는 보고서 창에서 데이터 테이블에 대한 변경 사항을 즉시 반영합니다.

**플랫폼 환경 설정** 현재 플랫폼 환경 설정을 보거나, 현재 JMP 보고서의 설정과 일치하도록 플랫폼 환경 설정을 업데이트할 수 있는 옵션이 포함되어 있습니다.

**스크립트 저장** 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다.

**그룹별 스크립트 저장** 기준 변수의 모든 수준에 대한 플랫폼 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다. 시작 창에서 기준 변수를 지정한 경우에만 사용할 수 있습니다.

---

**참고 :** 이 플랫폼의 추가 옵션은 스크립트를 통해 사용할 수 있습니다. 자세한 내용은 "도움말" 메뉴의 "스크립트 인덱스"에서 여십시오. 또한 "스크립트 인덱스"에서 이 섹션에 설명된 옵션의 스크립트 예제도 찾을 수 있습니다.

---

## Waern 링크

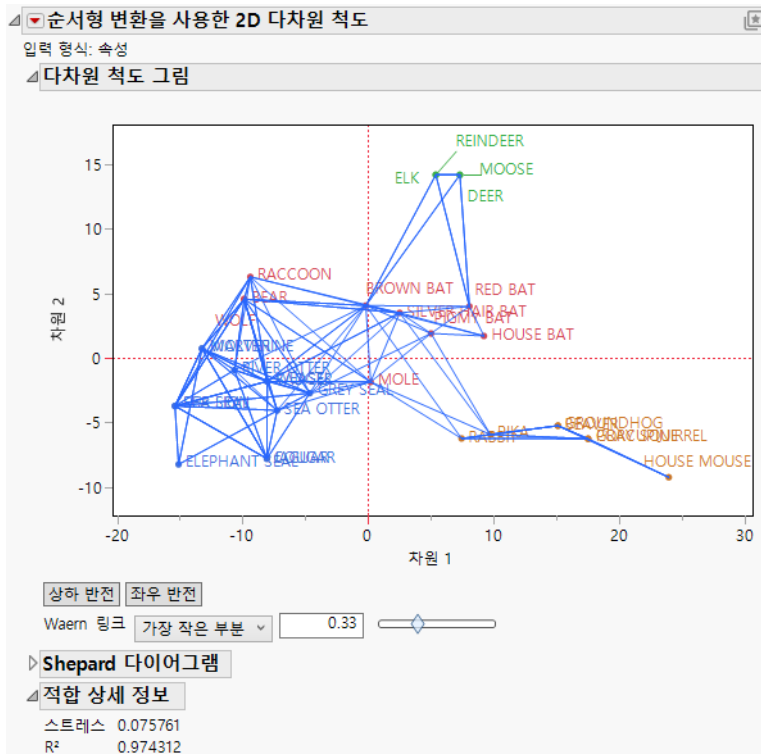
Waern 링크를 표시하면 실제 근접성과 예측된 근접성을 비교하여 다차원 척도법 결과를 시각적으로 확인할 수 있습니다. 이 링크는 실제 근접성을 기반으로 다차원 척도 그림의 점을 연결합니다. 근접성이 가장 작거나 가장 큰 개체가 연결됩니다. 일반적으로 개체 간 근접성의 가장 작은 33%를 고려해야 합니다. MDS 그림이 근접성을 잘 표현하면 가장 작은 실제 근접성의 링크가 그림에서 가장 가까운 개체를 연결해야 합니다. 가장 작은 근접성의 링크가 그림에서 늘어나 멀리 있는 개체를 연결하면 MDS 적합을 의심해 볼 수 있습니다.

### Waern 링크 컨트롤

그림에 링크의 가장 작은 부분을 표시할지 아니면 가장 큰 부분을 표시할지 선택할 수 있는 목록이 있습니다. 표시된 링크 부분은 상자에 값을 입력하거나 슬라이더를 사용하여 제어합니다. **그림 10.5**에서는 **Teeth.jmp** 데이터 테이블에 대한 **Waern** 링크를 보여 줍니다. 이 링크는 가장 작은 33% 부분을 표시합니다.

Waern 링크에 대한 자세한 내용은 [Waern\(1972\)](#) 연구 자료에서 확인하십시오.

**그림 10.5** Waern 링크가 있는 MDS 그림



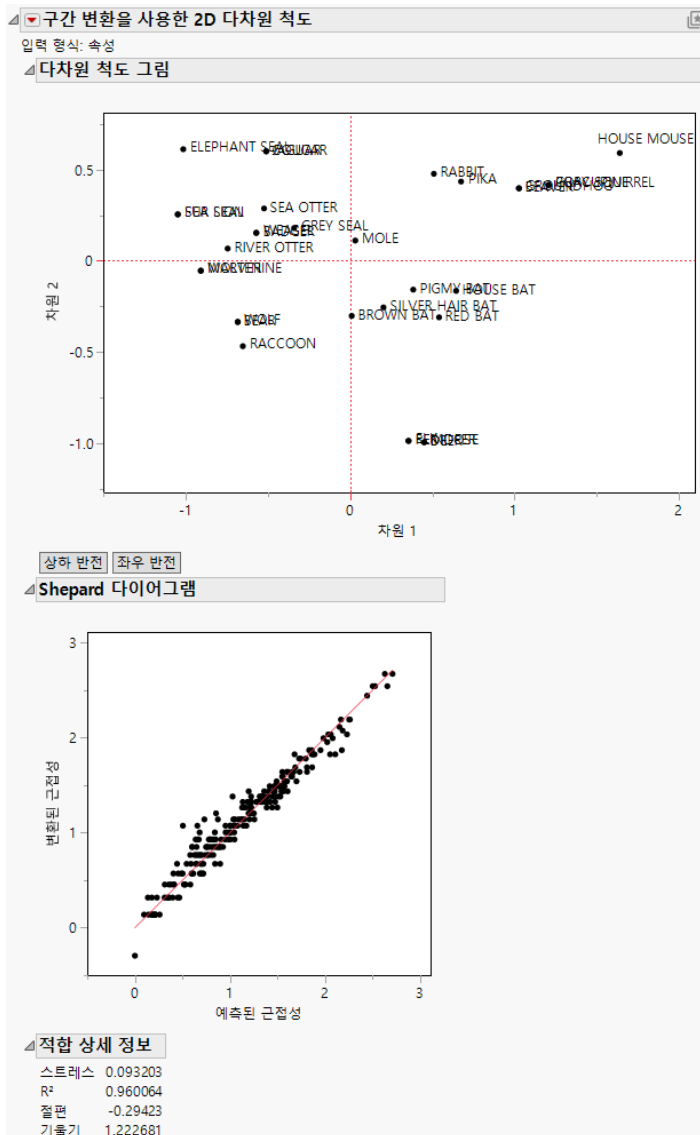
## 다차원 척도법의 추가 예

MDS 를 사용하여 치아를 기반으로 32 마리 포유 동물의 유사성을 탐색합니다. 구간 변환을 사용하여 해당 변환의 결과를 보여 줍니다. 데이터에는 의미를 갖는 순서가 있습니다 (치아 네 개는 치아 두 개보다 2 배 많음). 순서형 변환과 같은 다른 변환을 탐색할 수도 있습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Teeth.jmp** 를 엽니다.
2. **MAMMAL** 열을 마우스 오른쪽 버튼으로 클릭하고 **라벨 / 라벨 해제**를 선택합니다.
3. **행 > 행 선택 > 모든 행 선택**을 선택합니다.

4. 행 > 라벨 / 라벨 해제를 선택합니다 .
5. 분석 > 다변량 방법 > 다차원 척도법을 선택합니다 .
6. Top incisors ~ Bottom molars 를 선택하고 Y, 열을 클릭합니다 .
7. 데이터 형식 > 속성 목록을 선택합니다 .
8. 변환 > 구간을 선택합니다 .
9. 확인을 클릭합니다 .

그림 10.6 다차원 척도 보고서



"Shepard 다이어그램" 과 "적합 상세 정보" 에서는 MDS 그림에 치아 유사성으로 인한 동물의 유사성이 잘 표현되어 있음을 나타냅니다. "스트레스" 통계량 (0.093) 은 낮고, 변환된 근접성 대 예측된 근접성의  $R^2$  적합 (0.96) 은 높습니다. 또한 "적합 상세 정보" 에서는 실제 근접성의 변환을 위한 절편과 기울기를 제공합니다.

## 다차원 척도법 플랫폼에 대한 통계 상세 정보

이 섹션에는 다차원 척도법 플랫폼에 대한 통계 상세 정보가 포함되어 있습니다.

- "스트레스 함수에 대한 통계 상세 정보"
- "변환에 대한 통계 상세 정보"
- "속성 목록 형식에 대한 통계 상세 정보"

### 스트레스 함수에 대한 통계 상세 정보

다차원 척도법 플랫폼에서는 MDS 좌표를 결정하기 위해 quasi-Newton 최적화 방법을 사용하여 스트레스 함수를 최소화합니다. 이 최소화로 인해 차원의 각 쌍별 집합에 대해 도출된 근접성 측도를 최소화하는 좌표 집합이 미리 결정된 수의 차원에 생성됩니다. 데이터가 순서형이면 단조 회귀가 사용됩니다. 그렇지 않으면 표준 최소 제곱 회귀가 사용됩니다.

다음 표기를 사용하여 스트레스를 정의합니다.

- $i, j$  - 개체 수에 대한 인덱스
- $d_{ij}$  - 개체  $i$  와  $j$  사이의 유도 거리
- $\delta_{ij}$  - 개체  $i$  와  $j$  사이의 관측된 상대 거리
- $f(\delta_{rs})$  - 거리에 대한 변환 함수

스트레스 함수는 다음과 같이 정의됩니다.

$$\text{스트레스} = \left[ \frac{\sum_{i < j} [f(\delta_{ij}) - d_{ij}]^2}{\sum_{i < j} d_{ij}^2} \right]^{\frac{1}{2}}$$

이 스트레스 측도는 Kruskal 스트레스, 제1종 또는 간단히 스트레스 1이라고도 합니다.

### 변환에 대한 통계 상세 정보

다차원 척도법 플랫폼에서 변환은 실제 근접성을 척도화하는 데 사용됩니다. 변환은 데이터의 특정 구조를 고려하여 실제 근접성의 MDS 표현을 향상시키는 것으로 간주됩니다. 변환 함수의

모수는 최소화 알고리즘의 추가 모수가 됩니다. 이 섹션에서는 "스트레스 함수에 대한 통계 상세 정보"에 설명된 표기를 사용합니다.

### 비 변환

비율 데이터의 경우 :

$$f(\delta_{rs}) = b\delta_{rs}$$

### 구간 변환

구간 데이터의 경우 :

$$f(\delta_{rs}) = a + b\delta_{rs}$$

### 순서형 변환

순서형 데이터의 경우 데이터가 변환되지 않으며 최소 제곱 회귀 대신 단조 회귀가 알고리즘에 사용됩니다.

## 속성 목록 형식에 대한 통계 상세 정보

데이터 테이블이 속성 목록인 경우 거리 행렬로 변환된 후 다차원 척도법이 적용됩니다. 거리 행렬은 유클리드 거리를 사용하여 계산됩니다. 각 항목 쌍에 대해 다음 계산식을 사용하여 항목 간의 거리를 정의합니다.

$$\delta_{ij} = \sqrt{\sum_k \frac{(x_{ki} - x_{kj})^2}{k}}$$

여기서  $k$ 는 속성 수입니다. 측도가 동일한 척도를 사용한다고 가정합니다.

**참고 :** MDS 플랫폼에 대한 고급 예는 [San Francisco Crime Distances.jmp](#) 샘플 데이터 테이블과 해당 테이블의 소스 스크립트를 참조하십시오. 이 스크립트는 쌍별 상관을 사용하여 거리 행렬을 생성합니다. 그런 다음 결과 거리 행렬을 사용하여 범주 범주 간의 관계를 탐색합니다.



# 11 장

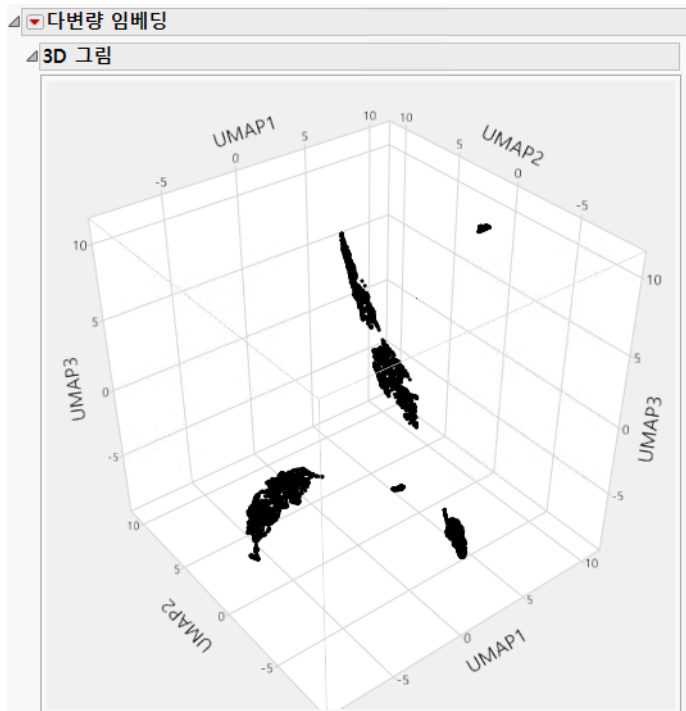
## JMP PRO 다변량 임베딩

### 고차원 데이터를 저차원 공간에 매핑

다변량 임베딩 플랫폼은 JMP Pro 에서만 사용할 수 있습니다.

다변량 임베딩 플랫폼을 사용하면 고차원 공간의 데이터를 저차원 공간에 매핑할 수 있습니다. 대부분의 경우 저차원 공간을 더 쉽게 시각화할 수 있도록 데이터를 2차원 또는 3차원으로 매핑하려고 합니다. 다변량 임베딩 플랫폼에서는 UMAP(Uniform Manifold Approximation and Projection, 균일 매니폴드 근사 및 투영) 방법과 t-SNE(t-Distributed Stochastic Neighbor Embedding, t 분포 확률적 이웃 임베딩) 방법을 사용할 수 있습니다. 두 방법 모두 근접 이웃의 근접을 더 쉽게 식별할 수 있는 방식으로 저차원 공간을 채우려고 합니다.

그림 11.1 다변량 임베딩의 예



**목차**

다변량 임베딩 플랫폼 개요.....	261
다변량 임베딩의 예 .....	262
다변량 임베딩 플랫폼 시작.....	263
다변량 임베딩 보고서 .....	266
다변량 임베딩 플랫폼 옵션.....	266
다변량 임베딩의 추가 예.....	267
다변량 임베딩 플랫폼에 대한 통계 상세 정보 .....	269
t-SNE 방법에 대한 통계 상세 정보 .....	269
경사하강법 알고리즘에 대한 통계 상세 정보.....	271

## JMP PRO 다변량 임베딩 플랫폼 개요

다변량 임베딩 플랫폼에서는 고차원 공간의 점  $\{x_1, x_2, \dots, x_n\}$ 을 저차원 공간의 점  $\{y_1, y_2, \dots, y_n\}$ 에 매핑하는 차원 축소를 수행합니다. 차원 축소의 목표는 고차원 데이터에 있는 중요한 정보를 유지하면서 점을 저차원 공간에 매핑하는 것입니다. 다변량 임베딩 플랫폼에 사용되는 특정 기법은 UMAP(Uniform Manifold Approximation and Projection, 균일 매니폴드 근사 및 투영) 방법과 t-SNE(t-Distributed Stochastic Neighbor Embedding, t 분포 확률적 이웃 임베딩) 방법입니다. UMAP 방법은 매니폴드 학습 기법으로, 비선형 차원 축소라고도 합니다. 이 기법은 Riemannian 기하학 및 대수적 위상수학(May, 1992)을 기반으로 합니다. t-SNE 방법은 Stochastic Neighbor Embedding(Hinton & Roweis, 2002)의 변형입니다.

다변량 임베딩 플랫폼에서 사용할 수 있는 두 가지 차원 축소 방법은 모두  $k$  이웃 기반 학습 알고리즘입니다. 이러한 유형의 알고리즘은 먼저 각 점에 대한 인접 이웃을 찾아 고차원 공간에  $k$  이웃 그래프를 생성합니다. 그런 다음 그래프 구조를 유지하면서 고차원 공간의 점을 저차원 공간에 매핑하는 저차원 매핑이 생성됩니다.

### UMAP 방법 개요

UMAP 방법은 먼저 각 점의 근접 이웃을 찾은 다음  $k$  이웃 그래프를 생성하여 위상 구조를 형성합니다. 기본 설정을 사용하면 각 점이 하나 이상의 다른 이웃, 즉 최근접 이웃에 연결되며 15번째 이후의 이웃에는 연결되지 않습니다. 그 사이의 이웃은 퍼지 영역을 형성합니다. 그런 다음 퍼지 영역의 엣지를 함께 병합하여 고차원 데이터의 위상 표현이 생성됩니다. 엣지가 병합되는 방법에 대한 자세한 내용은 McInnes et al. (2018)에서 확인하십시오.

저차원 매핑을 생성하기 위해 UMAP은 경사하강법을 사용하여 고차원 위상 표현과 저차원 위상 표현 사이의 교차 엔트로피를 최소화합니다(McInnes et al., 2018 참조). UMAP 방법은 계산 시간을 최소화하면서 데이터의 전역 구조를 보존하고 극도로 큰 데이터 집합을 처리할 수 있습니다.

### t-SNE 방법 개요

t-SNE 방법은 점 사이의 쌍별 유사성을 기반으로 합니다. 각 쌍별 유사성은 두 점이 이웃이라는 조건부 확률로 나타냅니다. 고차원 공간에서 거리는 가우시안 분포를 사용하여 조건부 확률로 변환됩니다. 저차원 맵에서 거리는 자유도가 1인 스튜던트  $t$  분포를 사용하여 확률로 변환됩니다. 여기서 t-SNE라는 방법 이름을 얻습니다(van der Maaten & Hinton, 2008 참조).

좋은 저차원 매핑의 경우 고차원 공간에서  $\{x_i, x_j\}$  사이의 쌍별 유사성이 저차원 공간에서  $\{y_i, y_j\}$  사이의 쌍별 유사성과 동일합니다. 이 가정하에 t-SNE 방법은 고차원 유사성과 저차원 유사성 사이의 차이를 최소화하는 저차원 매핑을 찾습니다. Kullback-Leibler 발산 버전을 사용하여 차이를 측정 한 후 경사하강법을 사용하여 최소화합니다. t-SNE 방법에 대한 자세한 내용은 "다변량 임베딩 플랫폼에 대한 통계 상세 정보"에서 확인하십시오.

## JMP PRO 다변량 임베딩의 예

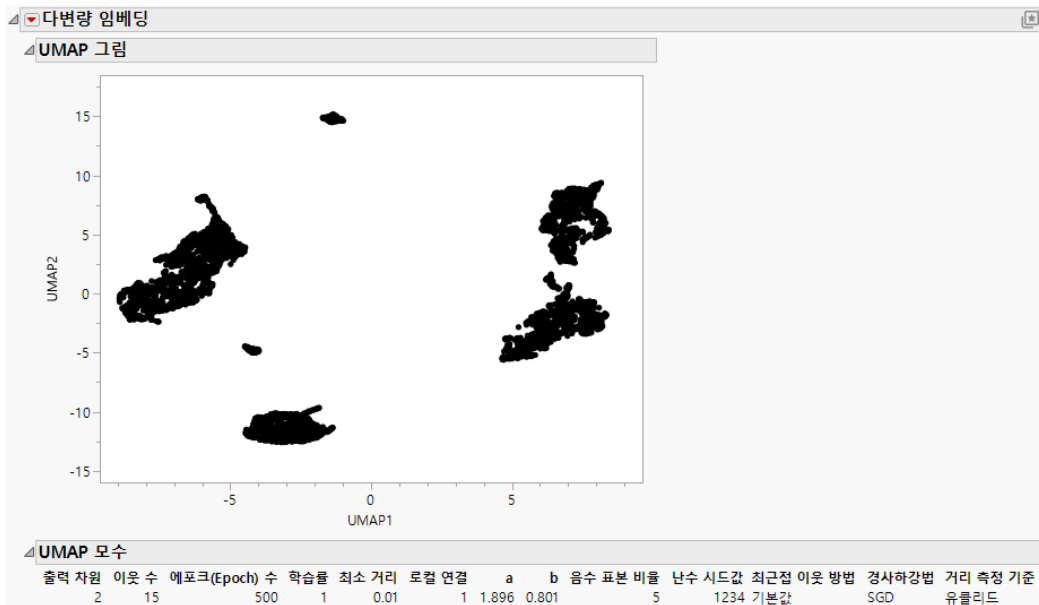
UMAP 방법을 사용하여 고차원 데이터를 2 차원으로 축소합니다 .

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Cytometry.jmp 를 엽니다 .
2. **분석 > 다변량 방법 > 다변량 임베딩**을 선택합니다 .
3. CD3 ~ MCB 를 선택하고 **Y, 열**을 클릭합니다 .
4. " 방법 " 이 "UMAP" 으로 설정되어 있는지 확인합니다 .
5. ( 선택 사항 ) **난수 시드값** 옆에 "1234" 를 입력합니다 .

**참고 :** 위의 시드값을 입력하면 이 예에 표시된 결과를 재현할 수 있습니다 .

6. **확인**을 클릭합니다 .
7. " 반복 기록 " 옆의 회색 삼각형을 클릭합니다 .

그림 11.2 다변량 임베딩 보고서

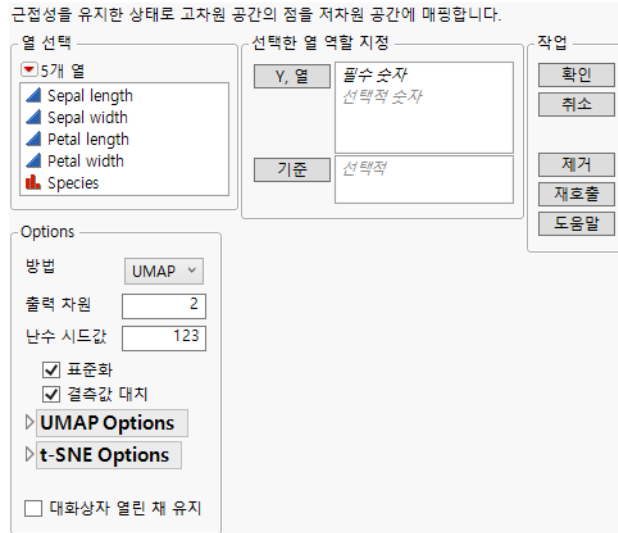


"다변량 임베딩" 보고서 창에는 UMAP 방법을 사용하여 계산된 2차원 그림과 계산 알고리즘에 대해 지정된 모수 설정이 포함되어 있습니다.

## JMP PRO 다변량 임베딩 플랫폼 시작

분석 > 다변량 방법 > 다변량 임베딩을 선택하여 다변량 임베딩 플랫폼을 시작합니다.

그림 11.3 다변량 임베딩 시작 창



"열 선택"의 빨간색 삼각형 메뉴에 포함된 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오. 다변량 임베딩 시작 창에는 다음 옵션이 포함되어 있습니다.

**Y, 열** 저차원 공간에 매핑할 고차원 데이터를 나타내는 열을 지정합니다.

**기준** 각 수준으로 개별 분석을 정의하는 열입니다. 지정된 열의 각 수준에 대해 해당 행은 사용자가 지정된 다른 변수를 사용하여 분석됩니다. 결과는 별도의 테이블과 보고서에 제공됩니다. 기준 변수가 둘 이상 할당되면 기준 변수의 가능한 각 수준 조합에 대해 개별 분석이 생성됩니다.

**방법** 데이터를 저차원 공간에 매핑하는 방법을 지정합니다. UMAP과 t-SNE 중에서 선택합니다.

**출력 차원** 저차원 공간의 성분 또는 차원 수를 지정합니다. 성분 수는 2보다 크거나 같아야 합니다.

**난수 시드값** 이후 플랫폼 실행에 대한 결과를 재현하는 데 사용할 난수 시드값을 지정합니다.

**표준화** 차원 축소에 사용되는 거리를 계산하기 전에 내부적으로 데이터를 표준화합니다.

**결측값 대치** 다변량 SVD(특이값 분해) 기법을 사용하여 데이터의 결측값을 대치하도록 지정합니다.

**참고 :** 데이터에 결측값이 있고 " 결측값 대치 " 옵션을 선택하지 않은 경우 시작 창에서 " 확인 " 을 클릭하면 결측값 대치 창이 표시됩니다. 데이터의 각 행에 하나 이상의 결측값이 포함된 경우 결측값을 대치하거나, Y 열 선택을 변경하거나, 분석을 취소하도록 선택할 수 있습니다. 데이터의 일부 행에 결측값이 없는 경우에는 결측값을 대치하거나, 결측값 대치 없이 계속하거나, 분석을 취소하도록 선택할 수 있습니다.

**UMAP 옵션** UMAP 알고리즘에 사용되는 옵션을 포함합니다. UMAP 알고리즘에서 다음 모두가 사용되는 방식에 대한 자세한 내용은 [McInnes et al. \(2018\)](#) 에서 확인하십시오.

**이웃 수** 각 데이터 점에 대해 발견되는 근접 이웃의 수를 지정합니다. 지정된 근접 이웃 수가 적을수록 UMAP 알고리즘은 데이터의 로컬 구조에 더 집중합니다. 근접 이웃 수가 증가하면 UMAP 알고리즘은 데이터의 전역 구조를 더 많이 포착합니다. "이웃 수" 값의 범위는 2 에서 데이터 관측값 수의 1/4 까지입니다. 기본값은 15 입니다.

**에포크 (Epoch) 수** 저차원 표현을 최적화할 때 사용할 훈련 에포크 (epoch) 수를 지정합니다. 이 값은 전체 훈련 데이터에서 알고리즘이 작동하는 횟수입니다. 기본값은 500 입니다.

**학습률** 계산에서 학습률 값을 지정합니다. 기본값은 1 입니다. 학습률은 모형이 문제에 얼마나 빨리 적응하는지에 영향을 줍니다. 학습률이 너무 크면 알고리즘이 최적 해를 놓칠 수 있습니다. 학습률이 너무 작으면 알고리즘이 수렴하는 데 시간이 오래 걸릴 수 있습니다.

**팁:** 알고리즘이 수렴하지 않거나, 극단값이 있는 임베딩 좌표가 생성되는 경우 학습률 값을 조정해 보십시오.

**최소 거리** 저차원 공간의 점이 서로 떨어져 있을 수 있는 최소 표준화 거리를 지정합니다. 0 에서 0.99 사이의 값을 지정할 수 있습니다. 기본값은 0.01 입니다.

**로컬 연결** 로컬 수준에서 연결된 것으로 간주되는 최근접 이웃 수를 지정합니다. 기본값은 1 이며, 이는 고차원 공간의 모든 점에 하나 이상의 다른 이웃이 연결되어 있다고 가정합니다.

**a** 임베딩 최적화 알고리즘을 제어하는 모수 중 하나를 지정합니다. 이 값이 0 또는 음수로 지정되면  $a$  는 알고리즘에서 비선형 최소 제곱 절차에 의해 계산됩니다.

**b** 임베딩 최적화 알고리즘을 제어하는 모수 중 하나를 지정합니다. 이 값이 0 또는 음수로 지정되면  $b$  는 알고리즘에서 비선형 최소 제곱 절차에 의해 계산됩니다.

**음수 표본 비율** 데이터의 저차원 표현을 찾을 때 양수 1- 심플렉스 표본당 사용할 음수 1- 심플렉스 표본 수를 지정합니다. " 음수 표본 비율 " 값의 범위는 2 에서 20 까지입니다. 기본값은 5 입니다.

**N 이 다음 값보다 큰 경우 배치 모드** 표본 크기가 지정된 값보다 클 때 멀티스레딩을 사용하여 임베딩 좌표를 최적화하도록 지정합니다. 기본값은 4096 입니다.

**최근접 이웃 방법** 최근접 이웃을 찾는 데 사용되는 방법을 지정합니다.

**기본값** 표본 크기와 변수 수에 따라 최근접 이웃 방법을 선택합니다. 관측값 수가 4096 개보다 크고, 변수 수가 1500 개 이하이거나, "거리 측정 기준" 이 "유클리드" 로 설정되지 않은 경우 기본값은 ANNOY 입니다. 그렇지 않은 경우 기본값은 VPTree 입니다.

**VPTree(정확)** VP(Vantage-Point, 기준점) 트리를 사용하여 최근접 이웃 집합을 찾습니다.

**ANNOY(근사)** ANN(Approximate Nearest Neighbors, 근사 최근접 이웃) 방법을 사용하여 최근접 이웃 집합을 찾습니다 (Bernhardsson, 2013 참조). 데이터 집합이 클 경우 둘 중 이 방법이 더 빠르지만 결과가 VPTree 방법보다 정확하지 않을 수 있습니다.

**거리 측정 기준** (ANNOY 를 최근접 이웃 방법으로 지정한 경우에만 적용 가능) 최근접 이웃 간의 거리를 계산하는 데 사용되는 측정 기준을 지정합니다. 거리 측정 기준 옵션은 "유클리드", "각", "Hamming" 및 "맨해튼" 입니다. 기본적으로 유클리드가 거리 측정 기준으로 지정됩니다.

---

**팁 :** 데이터에 이항 또는 범주형 변수가 포함된 경우 비유클리드 거리 측정 기준이 더 적절할 수 있습니다.

---

**경사하강법** 최적화 알고리즘에 사용되는 경사하강법을 지정합니다.

**SGD** Stochastic Gradient Descent 알고리즘을 사용합니다 (Saad, 1998 참조). 이 방법이 기본값입니다.

**ADAM** Adaptive Moment Estimation 방법을 사용합니다 (Kingma, 2014 참조). 이 옵션은 멀티스레딩이 사용되는 경우에만 사용할 수 있습니다.

**t-SNE 옵션** t-SNE 알고리즘에 사용되는 옵션을 포함합니다. 이러한 옵션의 대부분은 "다변량 임베딩 플랫폼에 대한 통계 상세 정보" 에서 설명합니다.

**희소** 고차원 공간에서 조건부 확률 계산에 희소 방법을 사용할지 여부를 지정합니다. 희소 방법을 사용하면 고차원 데이터를 계산할 수 있습니다.

**혼란도** 표본의 유사성 계산과 관련된 혼란도 모수의 값을 지정합니다. 혼란도 모수의 값은 5 에서 50 사이여야 하며, 표본 크기의 1/8 보다 크면 안 됩니다. 기본값은 30 또는 표본 크기의 1/8 중 더 작은 값입니다.

**최대 반복 수** 계산에 사용되는 최대 반복 수를 지정합니다.

**초기 주성분 차원** 초기 랜덤 주성분 분석 단계에서 유지되는 차원 수를 지정합니다. 기본값은 50 입니다.

**수렴 기준** 수렴을 측정하는 데 사용되는 값을 지정합니다. 기본값은 1e-8 입니다.

**초기 척도** 도출된 성분의 초기 척도를 지정합니다. 기본값은 0.0001 입니다.

**Eta** 계산에서 학습률 값을 지정합니다. 기본값은 200 입니다.

**반복 확장** 모멘텀 값이 더 이상 확대되지 않는 반복 수를 지정합니다. 기본값은 250 입니다.

**대화상자 열린 채 유지** 분석을 실행한 후 옵션을 업데이트하고 분석을 다시 실행할 수 있도록 시작 창을 열린 채로 유지합니다.

## JMP PRO 다변량 임베딩 보고서

"다변량 임베딩" 보고서 창에는 그림과 하나 이상의 테이블이 포함되어 있습니다. 이 보고서를 사용하면 저차원 공간을 시각화하고 알고리즘 모수 및 계산 상세 정보를 검토할 수 있습니다.

보고서의 그림 형식은 시작 창의 "출력 차원" 옵션에 지정된 차원 수에 따라 달라집니다. 2차원이 지정된 경우 "t-SNE 그림" 또는 "UMAP 그림" 보고서가 나타납니다. 3차원 이상이 지정된 경우에는 "3D 그림" 보고서가 나타납니다. 이러한 보고서에는 각 관측값에 대해 처음 두 개 또는 세 개의 임베딩 성분 스코어의 2D 또는 3D 산점도가 각각 표시됩니다.

모수 테이블에는 시작 창에서 UMAP 방법 또는 t-SNE 방법에 대해 지정된 모수 값이 포함되어 있습니다. 자세한 내용은 "UMAP 옵션" 및 "t-SNE 옵션"에서 확인하십시오.

시작 창에서 t-SNE 방법을 지정한 경우 "반복 기록" 테이블도 보고서에 표시됩니다. 이 테이블에는 계산 반복을 통한 통계량이 포함되어 있습니다. 각 행은 100 회 연속 반복이 끝난 후의 값을 나타냅니다. "반복 기록" 테이블에는 다음 열이 포함되어 있습니다.

**반복** 반복 번호입니다.

**비용** 해당 반복에서 비용 함수의 값입니다. 이 값은 모든 데이터 점에 대한 Kullback-Leibler 발산의 합입니다.

**최대 기울기** 해당 반복에서 모든 데이터 점에 대한 기울기의 최대값입니다.

"반복 기록" 테이블의 열에 대한 자세한 내용은 "경사하강법 알고리즘에 대한 통계 상세 정보"에서 확인하십시오.

## JMP PRO 다변량 임베딩 플랫폼 옵션

"다변량 임베딩"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

**임베딩 성분 값 저장** 도출된 t-SNE 또는 UMAP 성분을 데이터 테이블의 새 열로 저장합니다.

**PQ 행렬 표시** (t-SNE 방법을 사용하고 시작 창에서 "희소" 옵션을 선택하지 않은 경우에만 사용 가능) t-SNE 계산에 사용된 **P** 행렬과 **Q** 행렬을 포함하는 두 개의 새 데이터 테이블을 생성합니다. 자세한 내용은 "t-SNE 방법에 대한 통계 상세 정보"에서 확인하십시오.

다음 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

**로컬 데이터 필터** 특정 보고서에서 사용되는 데이터를 필터링할 수 있는 로컬 데이터 필터를 표시하거나 숨깁니다.

**다시 실행** 분석을 반복하거나 다시 시작할 수 있는 옵션이 포함되어 있습니다. 이 기능을 지원하는 플랫폼에서 "자동 재계산" 옵션은 해당하는 보고서 창에서 데이터 테이블에 대한 변경 사항을 즉시 반영합니다.

**플랫폼 환경 설정** 현재 플랫폼 환경 설정을 보거나, 현재 JMP 보고서의 설정과 일치하도록 플랫폼 환경 설정을 업데이트할 수 있는 옵션이 포함되어 있습니다.

**스크립트 저장** 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다.

**그룹별 스크립트 저장** 기준 변수의 모든 수준에 대한 플랫폼 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다. 시작 창에서 기준 변수를 지정한 경우에만 사용할 수 있습니다.

---

**참고 :** 이 플랫폼의 추가 옵션은 스크립트를 통해 사용할 수 있습니다. 자세한 내용은 "도움말" 메뉴의 "스크립트 인덱스"에서 여십시오. 또한 "스크립트 인덱스"에서 이 섹션에 설명된 옵션의 스크립트 예제도 찾을 수 있습니다.

---

## 다변량 임베딩의 추가 예

이 예에서는 "다변량 임베딩의 예"와 동일한 데이터 집합을 사용하지만 데이터를 2 차원이 아니라 3 차원으로 축소합니다. 또한 UMAP 방법과 t-SNE 방법 간의 결과를 비교합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Cytometry.jmp 를 엽니다 .
2. **분석 > 다변량 방법 > 다변량 임베딩**을 선택합니다 .
3. CD3 ~ MCB 를 선택하고 **Y, 열**을 클릭합니다 .
4. "방법"에 "t-SNE"를 지정합니다 .
5. "출력 차원" 옆에 "3"을 입력합니다 .
6. (선택 사항)"난수 시드값" 옆에 "1234"를 입력합니다 .

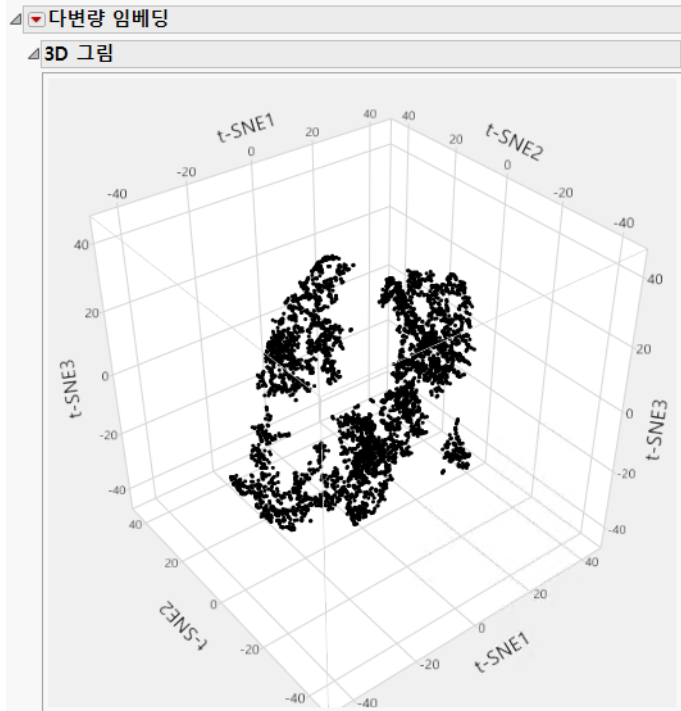
---

**참고 :** 위의 시드값을 입력하면 이 예에 표시된 결과를 재현할 수 있습니다.

---

7. **대화상자 열린 채 유지**를 선택합니다 .
8. **확인**을 클릭합니다 .

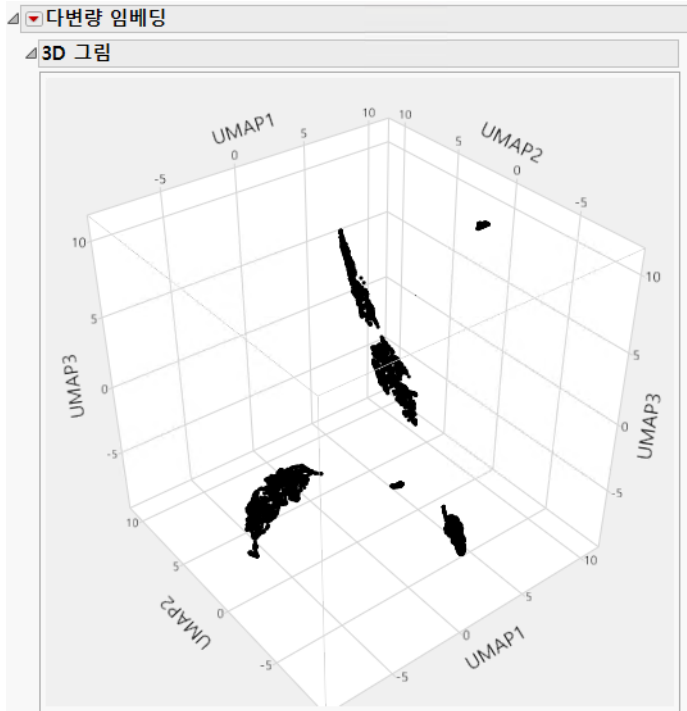
그림 11.4 3D t-SNE 그림의 예



9. 시작 창에서 "방법" 을 "UMAP" 으로 변경합니다.

10. **확인**을 클릭합니다.

그림 11.5 3D UMAP 그림의 예



각 3D 그림에는 해당하는 차원 축소 방법에서 처음 세 개의 임베딩 성분에 대한 좌표가 표시됩니다. UMAP 방법과 t-SNE 방법을 비교하면 UMAP 방법을 사용했을 때 데이터가 더 조밀하게 군집화되면서도 계산 시간이 적게 걸렸습니다. [그림 11.4](#)의 3D t-SNE 그림과 비교할 때 3D UMAP 그림에서 더 뚜렷한 군집을 볼 수 있습니다.

**팁:** 격자를 회전하여 군집을 더 명확하게 볼 수 있습니다.

## JMP PRO 다변량 임베딩 플랫폼에 대한 통계 상세 정보

이 섹션에는 다변량 임베딩 플랫폼에 대한 통계 상세 정보가 포함되어 있습니다.

- "t-SNE 방법에 대한 통계 상세 정보"
- "경사하강법 알고리즘에 대한 통계 상세 정보"

## JMP PRO t-SNE 방법에 대한 통계 상세 정보

t-SNE 방법은  $\{x_i, x_j\}$ 의 고차원 유사성과  $\{y_i, y_j\}$ 의 저차원 유사성 간 차이를 최소화하여 고차원 공간의 점  $\{x_1, x_2, \dots, x_n\}$ 을 저차원 공간의 점  $\{y_1, y_2, \dots, y_n\}$ 에 매핑합니다. 쌍별 유사성은 확률 분포

로 나타냅니다. 고차원 공간에서 조건부 확률  $p_{j|i}$ 는 가우시안 분포를 사용하여 계산됩니다. 다변량 임베딩 플랫폼에서는 조건부 확률을 계산하는 두 가지 방법을 제공합니다.

### 조건부 확률에 대한 희소 근사 계산

시작 창에서 "희소" 옵션을 선택하는 경우 희소 근사를 사용하여  $p_{j|i}$ 가 계산됩니다.  $n$  개의 각 입력에 대해 VP(Vantage-Point, 기준점) 트리를 사용하여 최근접 이웃 집합이 발견됩니다. 그런 다음 최근접 이웃의 부분집합에 대해서만 조건부 확률이 계산됩니다.

$$p_{j|i} = \begin{cases} \frac{\exp(-\|x_i - x_j\|^2 / (2\sigma_i^2))}{\sum_{k \in N_i} \exp(-\|x_i - x_k\|^2 / (2\sigma_i^2))}, & \text{if } j \in N_i \\ 0 & \text{그렇지 않은 경우} \end{cases}$$

이 방정식에서  $N_i$ 는  $x_i$ 의 floor( $3p$ ) 최근접 이웃 집합입니다. 여기서  $p$ 는 시작 창에 정의된 혼란도 모수입니다. 가우시안 분포의 분산  $\sigma_i$ 도 혼란도 모수를 기반으로 합니다. 자세한 내용은 van der Maaten & Hinton(2008) 및 van der Maaten(2014)에서 확인하십시오.

### 조건부 확률에 대한 비희소 계산

시작 창에서 "희소" 옵션을 선택하지 않는 경우 모든 점에 대해  $p_{j|i}$ 가 계산됩니다.

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / (2\sigma_i^2))}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / (2\sigma_i^2))}$$

이 계산에서 가우시안 분포의 분산  $\sigma_i$ 도 혼란도 모수를 기반으로 합니다.

### 결합 확률 분포 계산

t-SNE 방법에서는 조건부 확률이 대칭이라고 가정합니다. 따라서 고차원 공간에서 결합 확률  $p_{ij}$ 는 다음과 같이 대칭 조건부 유사성에 의해 정의됩니다.

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

이때 모든  $i$ 와  $j$ 에 대해  $p_{ij} = p_{ji}$ 입니다. 쌍별 유사성에 관심이 있으므로  $p_{ii} = 0$ 도 가정됩니다.

저차원 매핑의 결합 확률  $q_{ij}$ 는 자유도가 1 인 스튜던트 t 분포를 사용하여 계산됩니다.

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$$

이러한 확률은  $p_{ij}$ 와 동일한 특성을 갖습니다. 즉, 모든  $i$ 와  $j$ 에 대해  $q_{ij} = q_{ji}$ 이고  $q_{ii} = 0$ 입니다.

t-SNE 방법은 결합 확률 분포  $\mathbf{P}$ 와 결합 확률 분포  $\mathbf{Q}$  사이의 단일 Kullback-Leibler 발산을 최소화하여 고차원 공간의 쌍별 유사성과 저차원 공간의 쌍별 유사성 간 차이를 최소화합니다.  $\mathbf{P}$ 와  $\mathbf{Q}$  사이의 Kullback-Leibler 발산은 다음과 같이 계산됩니다.

$$KL(\mathbf{P} \parallel \mathbf{Q}) = \sum_i \sum_j p_{ij} \log(p_{ij}/q_{ij})$$

## JMP PRO 경사하강법 알고리즘에 대한 통계 상세 정보

t-SNE 방법에 사용되는 Kullback-Leibler 발산은 Barnes-Hut 근사에 기반한 경사하강법 알고리즘을 사용하여 최소화됩니다 (van der Maaten, 2014 참조). 이 알고리즘에서는 다음과 같은 표기를 사용합니다.

$\mathbf{X} = \{x_1, x_2, \dots, x_n\}$   $n$  개의 데이터 점이 있는 원래 고차원 데이터

$p$  = 혼란도 모수

$T$  = 반복 수

$\eta$  = 학습률

$t_{inflation}$  = 모멘텀 값이 변경되는 최대 반복 수

$\alpha(t)$  = 반복  $t$ 에서의 모멘텀, 여기서  $t \leq t_{inflation}$ 의 경우  $\alpha(t) = 0.5$ 이고 그렇지 않은 경우  $\alpha(t) = 0.8$

$\mathbf{Y}^{(t)} = \{y_1, y_2, \dots, y_n\}$  반복  $t$ 에서의 저차원 매핑 해

경사하강법 알고리즘의 단계는 다음과 같이 정의됩니다.

1. 고차원 공간의 쌍별 유사성  $p_{ji}$ 를 계산합니다. 지정된 혼란도  $p$ 를 기반으로  $\sigma_i$ 를 선택합니다.
2.  $p_{ij}$ 를 계산합니다.
3. 평균이 0 이고 표준편차가  $10^{-4}$  인 정규 분포에서 생성되는 초기 해  $\mathbf{Y}^{(0)} = \{y_1, y_2, \dots, y_n\}$  을 설정합니다.
4.  $t = 1 \sim T$ 에 대해
  - 저차원 매핑의 쌍별 유사성  $q_{ij}$ 를 계산합니다.

- 비용 함수를 계산합니다.

$$C = \text{KL}(\mathbf{P} \parallel \mathbf{Q}) = \sum_i \sum_j p_{ij} \cdot \log(p_{ij}/q_{ij})$$

- Barnes-Hut 근사를 사용하여 기울기 함수를 계산합니다.

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}$$

- 해를 업데이트합니다.

$$\mathbf{Y}^{(t)} = \mathbf{Y}^{(t-1)} + \eta \left( \frac{\partial C}{\partial \mathbf{Y}} \right) + \alpha(t)(\mathbf{Y}^{(t-1)} - \mathbf{Y}^{(t-2)})$$

다음 조건 중 하나가 충족되면 알고리즘이 중지됩니다.

- $i$ 에 대한 최대 기울기 값이 시작 창에 지정된 수렴 기준보다 작은 경우
- 최대 반복 수  $T$ 에 도달한 경우

# 12 장

## 항목 분석

### 항목 및 대상별 검정 결과 분석

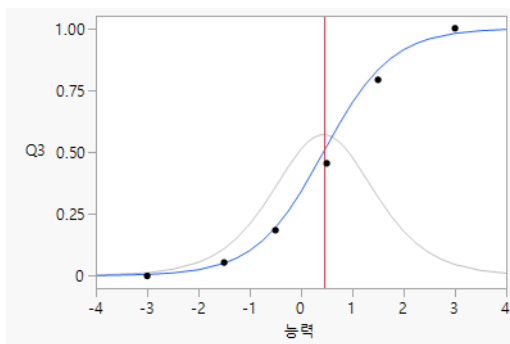
항목 분석 플랫폼을 사용하면 **항목 반응 이론** 모형을 적합시킬 수 있습니다. IRT(항목 반응 이론) 방법은 검사 및 설문지와 같이 측정기의 분석 및 스코어링에 사용됩니다. IRT(항목 반응 이론)는 일련의 방정식을 사용하여 항목을 관측되지 않은(잠재적) 특성 또는 능력과 연관시킵니다. 항목 또는 질문은 직접 관측할 수 없는 기본 잠재 구성의 지표입니다. 데이터를 수집할 때는 대상 능력과 항목 특성 둘 다 알 수 없습니다. 표준화 검사, 인지 발달 및 소비자 선호도를 연구할 때 IRT를 사용할 수 있습니다. IRT는 항목 스코어보다 관측된 총 스코어에 초점을 맞추는 CTT(고전 검사 이론) 대신 사용할 수 있는 방법입니다.

항목 분석 플랫폼에서는 다음과 같은 결과로 IRT 방법을 구현합니다.

- 측정기가 항목 수준에서 스코어링되므로 잠재 반응에 대한 각 항목의 기여도를 파악할 수 있습니다.
- 응답자와 항목에 대한 스코어가 모두 동일한 척도로 산출됩니다.
- 응답자 및 항목 스코어가 단일 그림에 표시됩니다.
- 항목 특성 곡선이 표시됩니다. 이 곡선은 항목과 응답자의 기본 특성 또는 능력 간의 관계를 탐색하는 데 사용할 수 있습니다.

항목 반응 이론에 대한 자세한 내용은 de Ayala(2009) 연구 자료에서 확인하십시오.

**그림 12.1** 항목 분석 특성 그림



## 목차

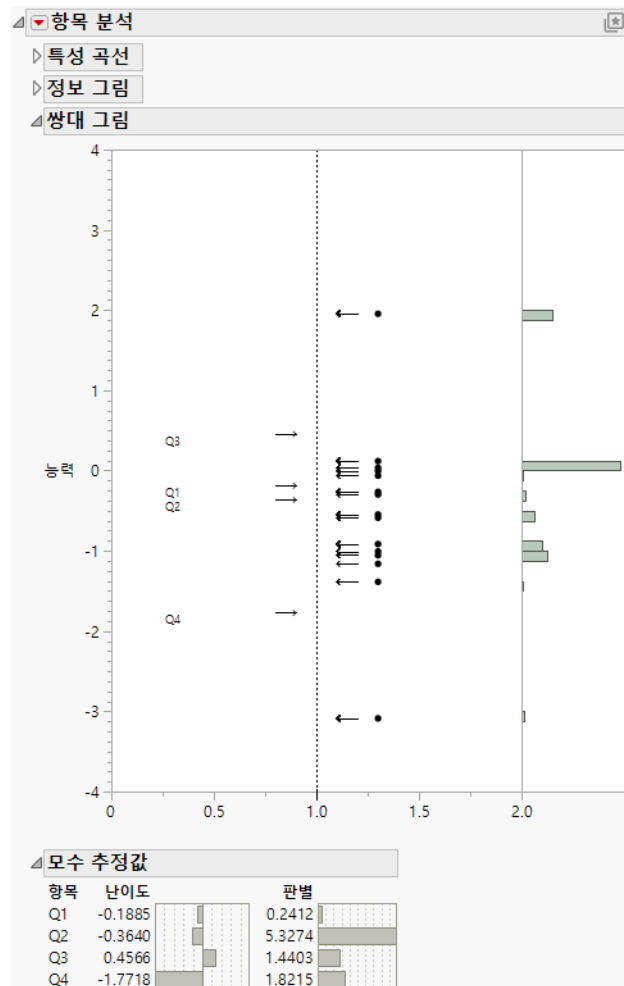
항목 분석의 예 .....	275
항목 분석 플랫폼 시작 .....	277
로지스틱 3PL 모형 상세 정보 .....	278
데이터 형식 .....	278
항목 분석 보고서 .....	278
특성 곡선 .....	278
정보 그림 .....	279
쌍대 그림 .....	280
모수 추정값 .....	280
항목 분석 플랫폼 옵션 .....	281
항목 분석 플랫폼에 대한 통계 상세 정보 .....	281
항목 반응 곡선에 대한 통계 상세 정보 .....	282
항목 반응 곡선 모형에 대한 통계 상세 정보 .....	282
IRT 모형 가정에 대한 통계 상세 정보 .....	284
IRT 모형 적합에 대한 통계 상세 정보 .....	285
능력 계산식에 대한 통계 상세 정보 .....	286

## 항목 분석의 예

이 예에서는 항목 분석 플랫폼을 사용하여 시험 문제와 과목 능력 간의 관계를 이해합니다. 14개 질문에 대한 1263명 학생의 스코어 (1 = 정답, 0 = 오답)가 사용됩니다. 시험 문제는 잠재적 수학 능력을 측정하는 데 사용되는 항목입니다. 문제와 응답자의 수학적 능력 간의 관계를 이해하기 위해 2PL 모형을 사용하여 처음 네 가지 문제를 검토합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **MathScienceTest.jmp**를 엽니다.
2. **분석 > 다변량 방법 > 항목 분석**을 선택합니다.
3. Q1 ~ Q4를 선택하고 **Y, 테스트 항목**을 클릭한 후 **확인**을 클릭합니다.

그림 12.2 항목 반응 보고서

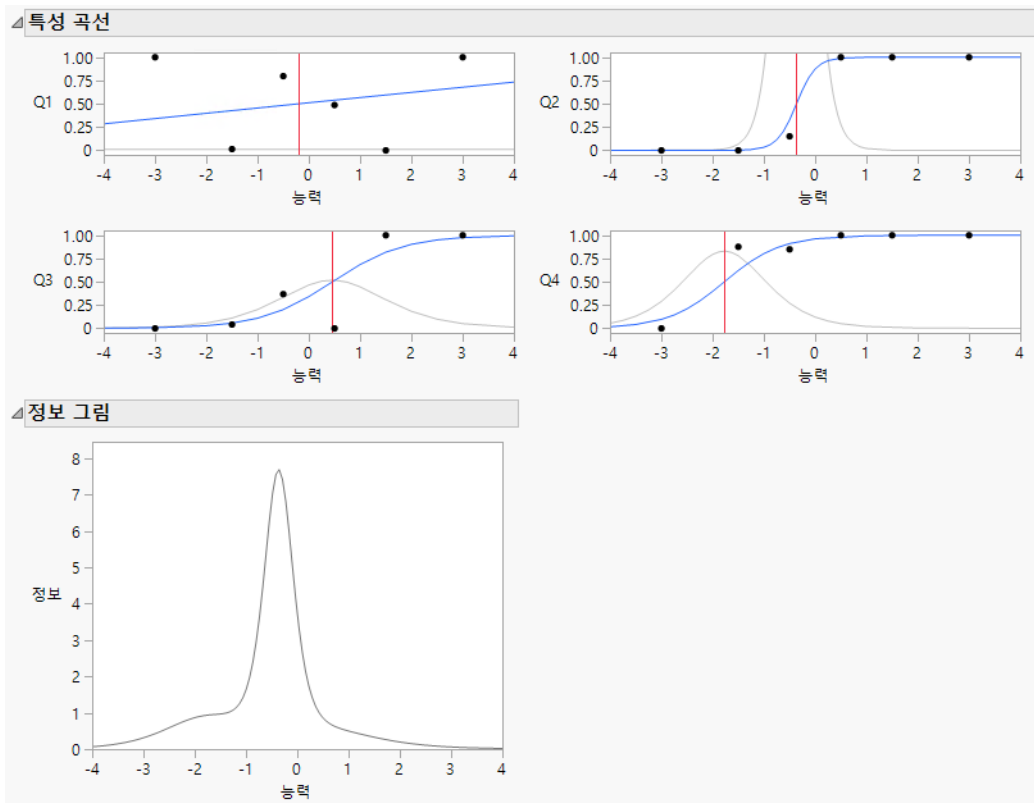


쌍대 그림에서 Q4의 난이도 스코어가 -1.78로 가장 낮으므로 Q4가 가장 쉬운 문제라는 것을 알 수 있습니다. Q3이 난이도 스코어 0.46으로 가장 어렵습니다. 대부분의 응답자는 그래프 가운데 부분의 데이터 점에서 볼 수 있듯이 능력 척도의 중간에서 아래쪽 끝에 속합니다. 히스토그램을 보면 응답자의 약 40%가 능력 척도에서 0보다 약간 위에 있는 것을 알 수 있습니다.

**참고 :** 모두 오답이거나 모두 정답인 개인에 대해서는 능력 스코어가 계산되지 않습니다. 자세한 내용은 "IRT 모형 적합에 대한 통계 상세 정보"에서 확인하십시오.

4. "특성 곡선" 보고서의 회색 표시 아이콘을 클릭하여 엽니다.
5. "항목 분석"의 빨간색 삼각형을 클릭하고 가로 그림 수를 선택합니다.
6. "2"를 입력하고 확인을 클릭합니다.
7. "정보 그림" 보고서의 회색 표시 아이콘을 클릭하여 엽니다.

그림 12.3 항목 반응의 예



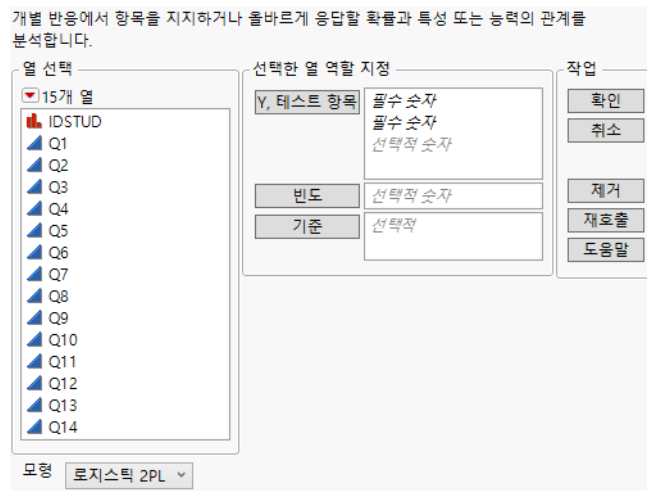
Q1은 특성 곡선과 정보 곡선이 모두 편평합니다. 이는 Q1이 응답자의 수학적 능력을 판별할 수 있는 많은 정보를 제공하지 않는다는 것을 나타냅니다. Q2의 특성 곡선은 가파르므로 Q2가 응답

자 능력을 판별하는 데 유용하다는 것을 알 수 있습니다. 각 그림의 세로선은 특성 곡선의 변곡점에 있습니다. 이 세로선은 응답자가 지정된 문제에 정답을 맞출 확률이 50%인 능력 수준입니다. 정보 그림에서는 네 가지 질문을 함께 분석한 결과 -1과 0 사이의 능력 수준에 대한 정보를 가장 많이 제공한다는 것을 나타냅니다. 난이도가 높은 질문을 모형에 더 많이 포함하면 정보 곡선이 넓어질 수 있습니다.

## 항목 분석 플랫폼 시작

분석 > 다변량 방법 > 항목 분석을 선택하여 항목 분석 플랫폼을 시작합니다.

그림 12.4 항목 분석 시작 창



"열 선택"의 빨간색 삼각형 메뉴에 포함된 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

**Y, 테스트 항목** 분석할 두 개 이상의 열을 할당합니다. 열은 연속형 숫자여야 하고 0 과 1 만 포함할 수 있습니다.

**팁:** 데이터를 0 과 1 로 재코딩해야 하는 경우 열 > 재코딩을 사용합니다. 자세한 내용은 JMP 사용에서 확인하십시오.

**빈도** 이 역할에 빈도 변수를 할당합니다. 이 기능은 데이터를 요약할 경우에 유용합니다.

**기준** 기준 변수의 각 수준에 대해 개별 보고서를 생성합니다. 기준 변수가 둘 이상 할당되면 기준 변수의 가능한 각 수준 조합에 대해 개별 보고서가 생성됩니다.

**모형** 다음 옵션에서 원하는 모형을 지정합니다.

**로지스틱 2PL** 2 모수 로지스틱 모형입니다.

**로지스틱 3PL** 3 모수 로지스틱 모형입니다.

**로지스틱 1PL** Rasch 파라미터화를 사용한 1 모수 로지스틱 모형입니다.

## 로지스틱 3PL 모형 상세 정보

"항목 분석" 시작 창에서 "로지스틱 3PL" 모형을 선택하고 "확인" 을 클릭하면 추측 모수에 대한 별점을 입력하라는 메시지가 나타납니다. 3PL 모형의 경우 별점 기본값은 0 입니다. 그러나  $c$  모수(각 항목에 대한 추측)에 0이 아닌 별점을 입력할 수 있습니다. 이 별점은 능형 회귀에서 사용하는 별점 모수 유형과 유사합니다. 별점은 추정된 추측 모수의 분산에 적용됩니다. 별점을 사용하면 다음과 같은 이점이 있습니다.

- 모형 모수 추정을 안정화합니다.
- 계산 속도가 빠릅니다.
- 약간의 편향은 있지만 항목 전체에서 추측 모수의 변동을 줄입니다.

별점 값이 크면 추측 모수가 0이 되며 값이 작을수록 항목 전체에서 추측 모수의 변동을 줄이는데 도움이 됩니다. 0 값을 사용하면 별점이 없습니다.

## 데이터 형식

항목 분석 플랫폼에는 각 개인에 대한 행과 각 항목에 대한 열이 포함된 데이터 테이블이 필요합니다. 항목 열은 수치형이어야 하고 올바른 반응을 나타내는 0과 올바른 반응을 나타내는 1만 포함할 수 있습니다. MathScienceTest.jmp 샘플 데이터 테이블에서는 14개의 시험 문제에 대한 1,263명의 반응으로 항목 반응 분석을 수행하기 위해 필요한 데이터 형식을 보여 줍니다.

---

## 항목 분석 보고서

"항목 분석" 보고서에는 다음과 같은 섹션이 포함되어 있습니다.

- "특성 곡선"
- "정보 그림"
- "쌍대 그림"
- "모수 추정값"

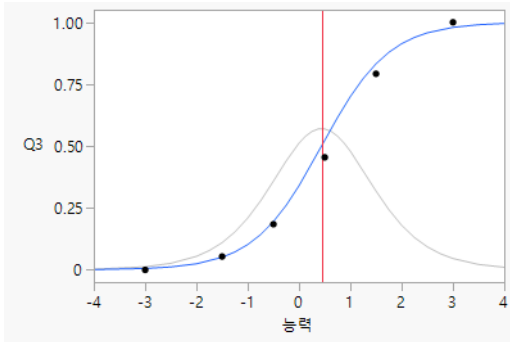
## 특성 곡선

"항목 분석" 보고서의 "특성 곡선" 섹션에는 시작 창에서 지정한 각 항목에 대한 ICC(항목 특성 곡선)가 포함됩니다. "특성 곡선" 섹션은 처음에 닫혀 있습니다.

항목 특성 곡선은 항목에 올바르게 답할 확률 대 능력을 표시합니다. 능력은 표준화된 척도로 측정되므로 능력이 0인 응답자는 평균 능력의 응답자입니다. 고정된 능력 수준에 대해 관측된 정답 확률의 데이터 점이 그림에 표시됩니다. 적합 특성 곡선을 데이터 점과 비교하면 각 개별 항목

에 대한 모형 적합도를 시각적으로 평가할 수 있습니다. 또한 특성 그림에는 배경 정보 곡선이 있으며 특성 곡선 변곡점 위치에 세로선이 있습니다. 배경 정보 곡선은 변곡점에서 최대화되는 항목 특성 곡선의 기울기 그림입니다.

**그림 12.5** 항목 특성 곡선

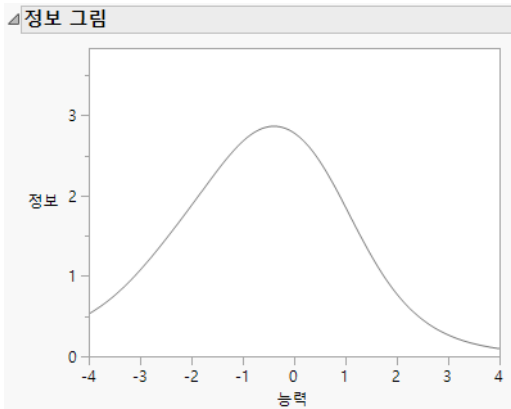


**팁:** "항목 분석"의 빨간색 삼각형 메뉴에 있는 "가로 그림 수" 옵션을 사용하여 보고서의 각 행에 나타나는 특성 곡선의 수를 조정할 수 있습니다.

## 정보 그림

"항목 분석" 보고서의 "정보 그림" 섹션에는 개별 항목 정보 곡선을 합하여 생성되는 전체 정보 곡선 그림이 포함됩니다. 이 정보 그림을 통해 테스트로 측정할 수 있는 적절한 능력 수준을 파악할 수 있습니다. **그림 12.6**에서는 능력 수준이 높은 개인보다 능력이 평균 수준에서 낮은 수준인 개인을 평가하는 데 더 적절한 항목을 사용한 테스트를 보여 줍니다. 이 그림은 처음에 닫혀 있습니다.

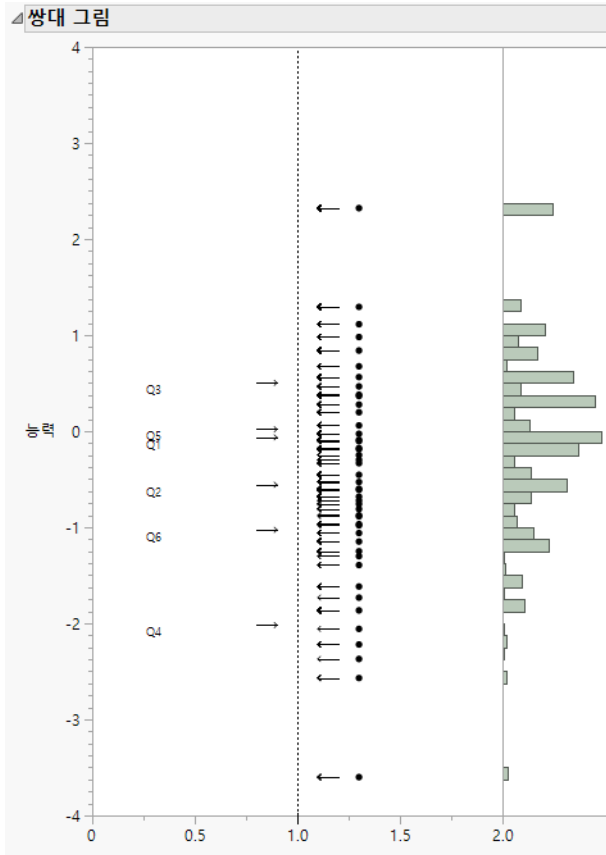
**그림 12.6** 정보 그림



## 쌍대 그림

"항목 분석" 보고서의 "쌍대 그림" 섹션에서는 항목 난이도와 과목 능력을 하나의 그림에 표시합니다. 난이도와 능력은 세로 축에 표시된 공통 표준화 척도를 사용합니다. 항목은 그림 왼쪽에 난이도별로 표시됩니다. 대상은 그림 오른쪽에 데이터 점과 히스토그램으로 표시됩니다. 쌍대 그림을 사용하면 각 항목의 난이도를 각 응답자의 능력과 연관시킬 수 있습니다.

그림 12.7 쌍대 그림



## 모수 추정값

"항목 분석" 보고서의 "모수 추정값" 섹션에는 각 항목에 대한 추정 모수 테이블이 포함됩니다. 제공되는 모수는 분석에 사용된 모형 (1PL, 2PL 또는 3PL) 에 따라 다릅니다.

**항목** 테스트 항목입니다.

**난이도**  $b$  모수 또는 항목 난이도 척도입니다. 난이도 추정값 옆에 난이도 모수 히스토그램이 표시됩니다.

**판별** (2PL 및 3PL 모형에만 사용 가능)  $a$  모수 또는 항목 판별 측도입니다. 판별 추정값 옆에 판별 모수 히스토그램이 표시됩니다.

**하위 접근선** (3PL 모형에만 사용 가능)  $c$  모수 또는 추측 측도입니다.

---

## 항목 분석 플랫폼 옵션

"항목 분석"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

**가로 그림 수** "특성 곡선" 보고서의 각 그림 행에 표시할 ICC 그림 수를 지정할 수 있습니다. 기본값은 해당 하나의 ICC 그림입니다.

**능력 계산식 저장** 능력 계산식을 데이터 테이블의 새 열에 저장합니다. 이 옵션은 IRT Ability() JSL 함수를 사용합니다. 이 함수에 대한 자세한 내용은 도움말 > 스크립트 인덱스에서 확인하십시오.

다음 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

**다시 실행** 분석을 반복하거나 다시 시작할 수 있는 옵션이 포함되어 있습니다. 이 기능을 지원하는 플랫폼에서 "자동 재계산" 옵션은 해당하는 보고서 창에서 데이터 테이블에 대한 변경 사항을 즉시 반영합니다.

**플랫폼 환경 설정** 현재 플랫폼 환경 설정을 보거나, 현재 JMP 보고서의 설정과 일치하도록 플랫폼 환경 설정을 업데이트할 수 있는 옵션이 포함되어 있습니다.

**스크립트 저장** 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다.

**그룹별 스크립트 저장** 기준 변수의 모든 수준에 대한 플랫폼 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다. 시작 창에서 기준 변수를 지정한 경우에만 사용할 수 있습니다.

---

**참고:** 이 플랫폼의 추가 옵션은 스크립트를 통해 사용할 수 있습니다. 자세한 내용은 "도움말" 메뉴의 "스크립트 인덱스"에서 여십시오. 또한 "스크립트 인덱스"에서 이 섹션에 설명된 옵션의 스크립트 예제도 찾을 수 있습니다.

---

---

## 항목 분석 플랫폼에 대한 통계 상세 정보

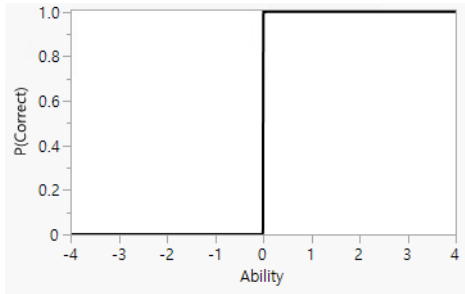
이 섹션에는 항목 분석 플랫폼에 대한 통계 상세 정보가 포함되어 있습니다.

- "항목 반응 곡선에 대한 통계 상세 정보"
- "항목 반응 곡선 모형에 대한 통계 상세 정보"
- "IRT 모형 가정에 대한 통계 상세 정보"
- "IRT 모형 적합에 대한 통계 상세 정보"
- "능력 계산식에 대한 통계 상세 정보"

## 항목 반응 곡선에 대한 통계 상세 정보

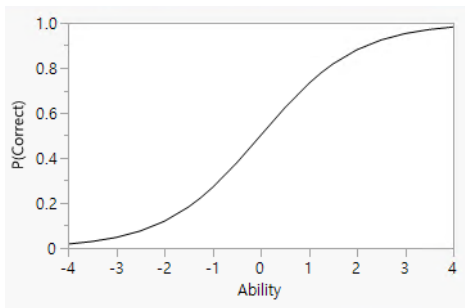
항목 분석 플랫폼에는 세 가지 항목 반응 곡선(항목 특성 곡선)이 표시됩니다. 이러한 곡선은 능력 척도로 정의된 능력과 각 항목 간의 관계를 기술하는 데 사용됩니다. 항목 반응 곡선은 다양한 수준의 능력에 대해 항목에 올바르게 답할 확률을 표시합니다. 완벽한 판별력을 가진 항목의 경우 능력이 임계값 미만인 응답자가 올바르게 답할 확률은 0%이고 능력이 임계값을 초과하는 대상이 올바르게 답할 확률은 100%입니다.

그림 12.8 완벽한 판별력을 가진 항목에 대한 특성 곡선



항목에 올바르게 답할 확률과 능력 간의 일반적인 관계는 하위 점근선과 상위 점근선이 있는 S자 형태의 함수입니다. 응답자의 능력이 향상될수록 항목에 올바르게 답할 확률은 100%까지 증가합니다. 특정 항목에 대한 곡선 모양은 항목의 난이도 및 판별 특성과 관련됩니다.

그림 12.9 일반적 항목 반응 곡선



## 항목 반응 곡선 모형에 대한 통계 상세 정보

항목 분석 플랫폼에는 항목 반응 곡선을 모델링하기 위한 1 모수, 2 모수 및 3 모수 로지스틱 모형이 있습니다. 3PL(3 모수 로지스틱) 모형은 다음과 같이 정의됩니다.

$$P(\theta) = c + \frac{1 - c}{1 + e^{-a(\theta - b)}}$$

- $P(\theta)$  는 능력 수준  $\theta$  에 대해 항목에 올바르게 답할 확률입니다. 항목 반응 이론 모형 적합에 대한 자세한 내용은 "IRT 모형 적합에 대한 통계 상세 정보" 에서 확인하십시오.
- $a$  모수는 변곡점에서 곡선의 경사도를 정의합니다. 이 모수는 항목의 판별력 추정값을 제공합니다.
- $b$  모수는 능력 축에서 변곡점 위치를 정의합니다. 이 모수는 항목의 난이도 추정값을 제공합니다.
- $c$  모수는 하위 점근선입니다. 이 모수는 추측을 통해 항목에 올바르게 답할 확률의 추정값을 제공합니다.
- 2PL 모형의 경우  $c$  모수는 0 으로 설정됩니다.

$$P(\theta) = \frac{I}{1 + e^{-(a)(\theta - b)}}$$

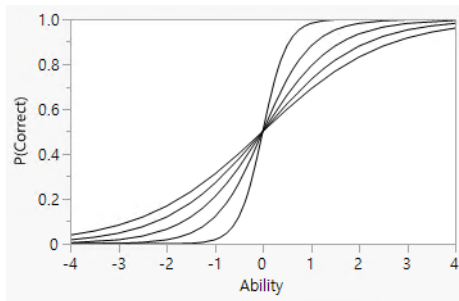
- 1PL 모형의 경우  $c$  모수는 0 으로 설정되고  $a$  모수는 1 로 설정됩니다. 이 파라미터화를 Rasch 모형이라고도 합니다 (Rasch 1980 참조).

$$P(\theta) = \frac{I}{1 + e^{-(\theta - b)}}$$

### a 모수 : 항목 판별

2PL 및 3PL 모형의  $a$  모수, 즉 변곡점에서 곡선의 경사도는 항목의 판별력 측도를 제공합니다. 항목의 판별력 또는 판별은 능력 수준이 낮은 응답자와 능력 수준이 높은 응답자를 항목이 얼마나 잘 구별할 수 있는지를 나타냅니다. 가파른 항목 반응 곡선은 항목의 판별력이 크다는 것을 나타냅니다. 능력 수준이 낮은 응답자는 항목에 올바르게 반응할 확률이 낮은 반면 능력이 높은 응답자는 올바르게 반응할 확률이 높습니다. 비교적 곡선이 편평한 항목은 판별력이 낮습니다. 판별력이 낮은 항목은 측정기에서 탈락할 후보입니다.

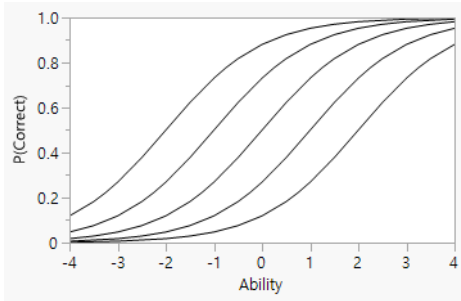
그림 12.10 여러  $a$  값에 대한 로지스틱 모형



## b 모수 : 항목 난이도

$b$  모수, 즉 능력에 대한 변곡점 위치는 항목 난이도 측도를 제공합니다. 능력 척도에서 변곡점 위치가 오른쪽으로 더 멀리 있는 항목 반응 곡선은 변곡점이 왼쪽에 있는 항목보다 답하기 어려운 항목을 나타냅니다. 1PL 및 2PL 모형에서  $b$  모수는 항목에 올바르게 답할 확률이 50%가 되기 위해 필요한 능력 수준의 추정값을 제공합니다.

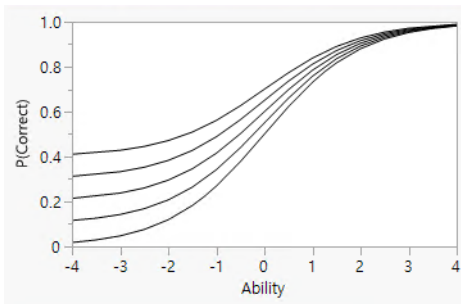
그림 12.11 여러  $b$  값에 대한 로지스틱 곡선



## c 모수 : 추측

3PL 모형에서  $c$  모수, 즉 항목 반응 곡선의 하위 점근선은 추측 모수 측도를 제공합니다. 0이 아닌 하위 점근선은 능력 수준이 매우 낮은 사람이 항목에 올바르게 답할 확률이 0이 아님을 나타냅니다.

그림 12.12 여러  $c$  값에 대한 로지스틱 모형



## IRT 모형 가정에 대한 통계 상세 정보

항목 분석 플랫폼에서는 2PL 모형이 기본 모형입니다. 1PL 모형은 모든 항목의 판별력이 동일하다고 가정할 수 있을 때 적절합니다. 이 가정이 적절하지 않을 때는 2PL 또는 3PL 모형을 사용해야 합니다. 2PL 모형은 특히 작은 데이터 집합의 경우 3PL 모형보다 수치 안정성이 큼니다. 또한 2PL 모형에서  $b$ 는 응답자가 항목에 올바르게 답할 확률이 50%가 되기 위해 필요한 능력 수준으로 해석할 수 있습니다.

IRT 모형은 기본 특성이 1차원이라고 가정합니다. 즉, 기본 잠재 구성이 1개입니다. 서로 복잡하게 상호 작용하는 몇 가지 특성을 측정하는 경우 1차원 모형이 적절하지 않습니다. IRT 모형은 연속형 잠재 변수에 적절합니다. 범주형 잠재 변수의 경우 잠재 계층 모형을 고려해야 합니다. 자세한 내용은 "잠재 계층 분석"에서 확인하십시오. IRT 모형은 항목 불변으로 간주됩니다. 항목 불변성은  $P(\theta)$ 가 능력 수준이  $\theta$ 인 개인 집합에 대한 올바른 반응 확률로 해석된다는 의미입니다. 능력 수준이 동일한 많은 개인이 항목에 답한 경우  $P(\theta)$ 는 항목에 올바르게 답할 사람의 비율을 예측합니다. 이는 IRT 모형이 테스트 대상 그룹에 관계없이 동일한 모수를 갖는다는 의미입니다. 또한 IRT 모형은 국소적 독립을 가정합니다. 즉, 잠재적 구성이 고려된 후에는 항목이 서로 독립적입니다.

## IRT 모형 적합에 대한 통계 상세 정보

항목 분석 플랫폼에서 항목 반응 이론 모형은 MMLE(주변 최대 가능도 추정)를 사용하여 적합됩니다. MMLE는 JLE(결합 최대 가능도 추정) 대신 사용할 수 있는 방법입니다. MMLE는 대상을 임의 효과로 처리합니다. 항목과 능력은 조건부 확률로 관련됩니다. 계산식은 다음과 같이 정의됩니다.

$$p(\mathbf{x}|\theta, \boldsymbol{\eta}) = \prod_{j=1}^L p_j(\theta)^{x_j} (1-p_j(\theta))^{1-x_j}$$

여기서  $p(\mathbf{x}|\theta, \boldsymbol{\eta})$ 는 대상 능력  $\theta$ 와 항목 모수 벡터  $\boldsymbol{\eta}$ 가 주어졌을 때 반응 벡터  $\mathbf{x}$ 의 확률입니다. 항목 모수의 수는 사용된 모형(1PL, 2PL 또는 3PL)에 따라 다릅니다.

MMLE는 항목 모수 추정값을 얻기 위해 가우시안 구적을 사용하여 대상 효과를 통합합니다. 반응 벡터  $\mathbf{x}$ 의 확률은 다음과 같이 계산됩니다.

$$p(\mathbf{x}) = \int_{-\infty}^{\infty} p(\mathbf{x}|\theta, \boldsymbol{\eta})g(\theta|\nu)d\theta$$

여기서  $g(\theta|\nu)$ 는 대상의 분포이고  $\nu$ 는 모집단 위치와 척도 모수의 벡터입니다. JMP에서는 평균이 0이고 표준편차가 1인 정규 분포가  $g(\theta|\nu)$ 에 사용됩니다.

**참고 :** 테스트 질문에 대한 결측값은 올바르게 않은 반응으로 처리됩니다. 모두 올바르게 않게 답한 개인 또는 모두 올바르게 답한 개인에 대해서는 능력 스코어가 계산되지 않습니다. 이러한 대상에 대한 반응 패턴이 모형 추정에 포함됩니다.

IRT 모형 적합을 위한 MMLE 절차는 두 단계로 임의 효과 모형을 적합시키는 것과 비교할 수 있습니다. 능력 모수는 분산이 1인 임의 효과로 처리됩니다. 첫 번째 단계에서는 가우시안 구적을 사용하여 이러한 임의 효과가 통합됩니다. 항목 모수는 능력 모수가 통합된 상태에서 주변 가능도의 ML을 사용하여 추정되는 고정 효과로 처리됩니다. 능력 모수는 본질적으로 완전 비통합(결합) 가능도를 사용하여 추정되는 최량 선형 비편향 예측입니다. 이때 항목 모수를 알려진 것으로 처리하고 첫 번째 단계에서 얻은 값으로 고정합니다.

$L$  개 항목에 대한  $2^L$  반응 패턴이 있습니다. 각 패턴의 능력 수준은  $\theta$  가 수렴할 때까지 다음을 적용하여 반응 패턴에 대한 확률이 가장 높은 능력 수준을 찾는 방식으로 계산할 수 있습니다.

$$\theta_i^{t+1} = \theta_i^t - \frac{X_i - \sum_{j=1}^L p_{ij}(t)}{L - \sum_{j=1}^L p_{ij}(t)(1 - p_{ij}(t))}$$

다음은 각 요소에 대한 설명입니다.

$\theta$  = 반응 패턴을 얻을 가능성을 최대화합니다.

$t$  = 반복 수

$L$  = 항목 수

$X_i$  = 관측된 스코어

$p_{ij}$  = 항목 모수를 기반으로  $i$  번째 사람이  $j$  번째 항목에 대해 올바르게 반응할 확률

## 능력 계산식에 대한 통계 상세 정보

"항목 분석"의 빨간색 삼각형 메뉴에 있는 "능력 계산식 저장" 옵션은 능력 계산식을 데이터 테이블의 새 열에 저장합니다. 이 계산식을 사용하여 데이터 테이블에 추가된 다른 대상을 스코어링하거나, 새 테이블에 복사하여 새 대상 집합을 스코어링할 수 있습니다.

테이블에 저장된 함수를 **IRT Ability** 함수라고 합니다. 항목 모수 추정값은 이 함수의 행렬에 저장됩니다. 이 함수에 대한 자세한 내용은 도움말 > 스크립트 인덱스에서 확인하십시오.

# 13 장

## 계층적 군집화 군집 트리를 사용하여 관측값 그룹화

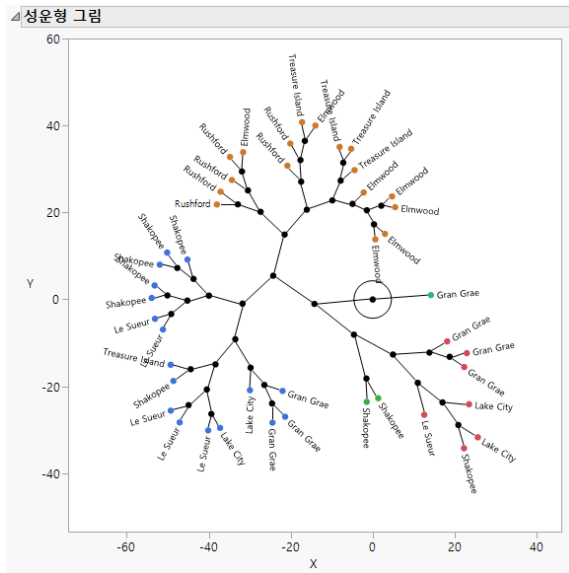
군집화는 여러 변수에 대해 유사한 값을 공유하는 관측값을 함께 그룹화하는 다변량 기법입니다. 이 방법을 사용하여 데이터의 군집 구조를 이해할 수 있습니다.

계층적 군집화는 군집을 연속해서 결합합니다. 이 방법은 각 관측값을 자체 군집화하는 것으로 시작합니다. 그런 다음 각 단계에서 거리가 가장 가까운 두 군집을 단일 군집으로 결합합니다. 결과는 덴드로그램이라고 하는 트리로 표시됩니다.

일반적으로 계층적 군집화는 행 수가 수만 개 이하인 작은 데이터 테이블에 유용합니다. 이 알고리즘은 시간이 많이 소요되며 큰 데이터 테이블의 경우 느리게 실행될 수 있습니다. 그러나 계층적 군집화 플랫폼은 계산 시간을 줄이고 더 큰 데이터 테이블을 군집화하는 데 유용한 "빠른 Ward"와 "하이브리드 Ward"라는 두 가지 방법도 제공합니다.

**참고:** 계층적 군집화는 문자 열을 지원하고 K 평균 군집화 또는 정규 혼합에는 숫자 열이 필요합니다.

그림 13.1 성운형 그림의 예



## 목차

계층적 군집화 플랫폼 개요.....	289
관측값 군집화 플랫폼 개요.....	289
계층적 군집화의 예.....	291
계층적 군집화 플랫폼 시작.....	293
계층적 군집화 보고서.....	298
덴드로그램.....	298
군집화 기록.....	299
계층적 군집화 플랫폼 옵션.....	299
계층적 군집화의 추가 예.....	303
거리 행렬의 예.....	304
공간 측도를 사용한 웨이퍼 결함 분류의 예.....	306
계층적 군집화 플랫폼에 대한 통계 상세 정보.....	308
공간 측도에 대한 통계 상세 정보.....	309
거리 방법에 대한 통계 상세 정보.....	310
근접 이웃 결함 주기에 대한 통계 상세 정보.....	311

## 계층적 군집화 플랫폼 개요

계층적 군집화 방법은 각 관측값이 자체 군집을 형성하는 것으로 시작합니다. 군집화 과정의 각 단계에서는 모든 군집 쌍 간의 거리를 계산하고 가장 가까운 두 군집을 결합합니다. 모든 점이 하나의 군집에 포함될 때까지 이 과정이 계속됩니다. 계층적 군집화는 사용하는 결합 방식 때문에 병합 군집화라고도 합니다.

병합 과정은 덴드로그램이라는 트리로 표현됩니다. JMP에서는 군집 수를 결정하는 데 도움이 되도록 거리 그래프를 제공합니다. 군집 간의 거리가 실제적으로 더 이상 중요하지 않은 시점을 확인하여 군집 수를 선택할 수 있습니다.

계층적 군집화는 문자 열도 지원합니다. 두 가지 방법을 사용하여 거리를 정의할 수 있습니다.

- 열이 순서형인 경우 군집화에 사용되는 값은 순서가 지정된 범주의 인덱스로, 연속형 데이터 처럼 처리됩니다. 이러한 값은 연속형 데이터처럼 표준화됩니다.
- 열이 명목형인 경우 범주가 일치하는 두 관측값 간의 거리는 0입니다. 범주가 다르면 거리가 1입니다.

계층적 군집화를 사용하면 "평균", "중심", "Ward", "단일" 및 "완전"이라는 다섯 가지 규칙 중에서 선택하여 군집 간 거리를 정의할 수 있습니다. 각 규칙은 서로 다른 군집 시퀀스를 생성할 수 있습니다. 또한 Ward 방법을 기반으로 군집 간 거리를 정의하는 "빠른 Ward"와 "하이브리드 Ward"라는 두 가지 추가 방법이 있습니다.

**팁:** 계층적 군집화 과정은 "빠른 Ward" 방법을 사용하는 경우를 제외하고  $n$  개의 관측값에 대해  $n(n+1)/2$  개의 거리로 시작합니다. 따라서 이 방법은  $n$ 이 큰 경우 실행 시간이 오래 걸릴 수 있습니다. 숫자 관측값 수가 많은 경우에는 K 평균 군집화 또는 정규 혼합 방법을 고려해 보십시오.

계층적 군집화는 JMP에서 관측값을 군집화하기 위해 제공하는 네 가지 플랫폼 중 하나입니다. 네 가지 방법을 모두 비교하려면 "[관측값 군집화 플랫폼 개요](#)"의 내용을 참조하십시오.

## 관측값 군집화 플랫폼 개요

군집화는 여러 변수에 대해 유사한 값을 공유하는 관측값을 함께 그룹화하는 다변량 기법입니다. 일반적으로 관측값은  $p$  차원 공간에 고르게 분산되지 않습니다. 여기서  $p$ 는 변수 수입니다. 대신 관측값은 군집을 형성합니다. 이러한 군집을 식별하면 데이터를 더 세밀하게 파악할 수 있습니다.

**참고:** JMP에서는 변수를 군집화할 수 있는 플랫폼도 제공합니다. 자세한 내용은 "[변수 군집화](#)"에서 확인하십시오.

JMP에서 제공하는 다음 네 가지 플랫폼을 사용하여 관측값을 군집화할 수 있습니다.

- 계층적 군집화는 작은 데이터 테이블과 큰 데이터 테이블에 모두 유용하며 문자 데이터를 허용합니다. 계층적 군집화는 행을 계층적 순서로 결합하여 트리로 표현합니다. 트리가 생성

되면 데이터에 가장 적절한 군집 수를 선택할 수 있습니다. 자세한 내용은 "[계층적 군집화](#)"에서 확인하십시오.

- K 평균 군집화는 최대 수백만 개의 행이 있는 큰 테이블에 적절하며 숫자 데이터만 허용합니다. 군집 수  $k$ 를 미리 지정해야 합니다. 이 알고리즘은 군집 시드 점을 추측합니다. 그런 다음 군집에 점을 번갈아 할당하고 군집 중심을 다시 계산하는 반복 과정을 수행합니다. 자세한 내용은 "[K 평균 군집화](#)"에서 확인하십시오.
- 정규 혼합은 중첩될 수 있는 다변량 정규 분포의 혼합에서 데이터를 수집한 경우에 적절하며 숫자 데이터만 허용합니다. 다변량 이상치가 있는 경우 균등 분포를 가정하고 이상치 군집을 사용할 수 있습니다. 자세한 내용은 "[정규 혼합](#)"에서 확인하십시오.

군집 수를 미리 지정해야 합니다. 최대 가능도는 혼합 비율과 평균, 표준편차 및 상관을 공동으로 추정하는 데 사용됩니다. 각 그룹에 속할 확률이 각 점에 할당됩니다. EM 알고리즘은 추정값을 구하는 데 사용됩니다.

- 잠재 계층 분석은 대부분의 변수가 범주형일 때 적절합니다. 군집 수를 미리 지정해야 합니다. 이 알고리즘은 다항 혼합 분포를 가정하는 모형을 적합시킵니다. 각 관측값에 대해 소속 군집의 최대 가능도 추정값이 계산됩니다. 관측값은 소속 확률이 가장 큰 군집으로 분류됩니다. 자세한 내용은 "[잠재 계층 분석](#)"에서 확인하십시오.

**표 13.1** 군집화 방법 요약

방법	데이터 유형 또는 모델링 유형	데이터 테이블 크기	군집 수 지정
계층적 군집화	임의	하이브리드 Ward를 사용하는 경우 최대 수십만 개의 행  빠른 Ward를 사용하는 경우 최대 200,000개 행  다른 방법을 사용하는 경우 최대 5,000개 행	아니요
K 평균 군집화	수치형	최대 수백만 개 행	예
정규 혼합	수치형	모든 크기	예
잠재 계층 분석	명목형 또는 순서형	모든 크기	예

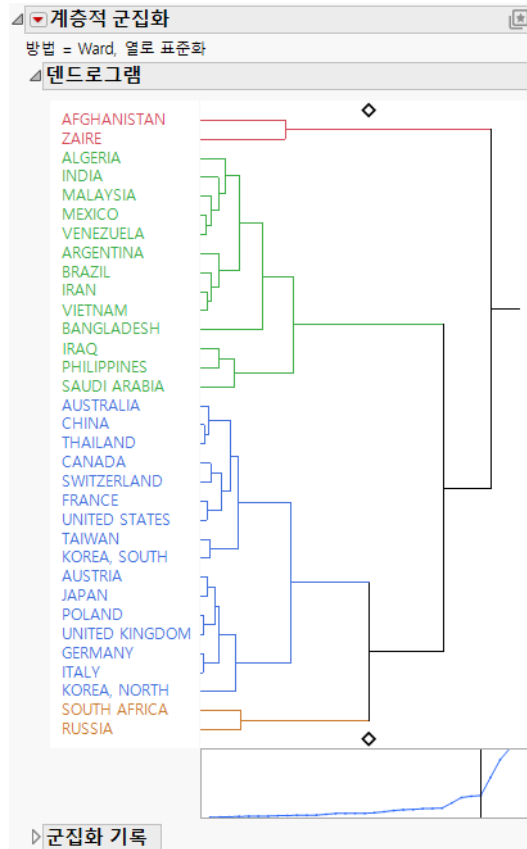
일부 군집화 플랫폼에는 데이터의 이상치를 처리하는 옵션이 있습니다. 그러나 데이터에 이상치가 있으면 분석하기 전에 먼저 이상치를 탐색하는 것이 가장 좋습니다. 이 작업은 이상치 탐색 유틸리티를 사용하여 수행할 수 있습니다. 자세한 내용은 [예측 및 전문 모델링](#)에서 확인하십시오.

## 계층적 군집화의 예

이 예에서는 데이터의 군집을 조사하기 위해 2009 년 인구 1,000 명당 조출생률 및 사망률을 기준으로 여러 국가를 그룹화합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Birth Death Subset.jmp 를 엽니다.
2. **분석 > 군집화 > 계층적 군집화**를 선택합니다.
3. birth 와 death 를 선택하고 **Y, 열**을 클릭합니다.
4. country 를 선택하고 **라벨**을 클릭합니다.  
이렇게 선택하고 " 확인 " 을 클릭하면 표시된 덴드로그램에 행 번호 대신 country 열을 사용하여 라벨을 지정할 수 있습니다.
5. **확인**을 클릭합니다.
6. " 계층적 군집화 " 의 빨간색 삼각형을 클릭하고 **군집 색 표시**를 선택합니다.

그림 13.2 계층적 군집화 보고서



덴드로그램에는 군집화 수행 방법이 표시됩니다. 덴드로그램을 왼쪽에서 오른쪽으로 읽으면 군집화 과정을 볼 수 있습니다. 각 단계는 가장 가까운 두 군집을 단일 군집으로 결합하는 것으로 구성됩니다.

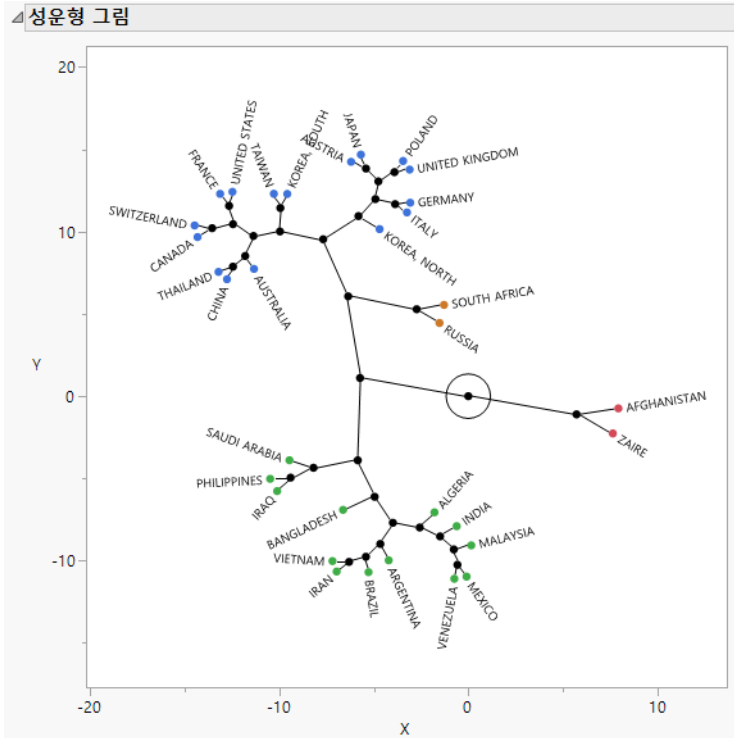
덴드로그램에서 군집 간의 상대 거리는 군집을 연결하는 세로선 사이의 가로 거리로 지정됩니다. 예를 들어 예를 들어 Mexico 와 Venezuela 로 구성된 군집과의 차이는 Malaysia 보다 Afghanistan 과 Zaire 가 더 큼니다.

다이아몬드는 네 개의 군집 위치에서 설정됩니다. 네 개의 군집 모형을 형성하기 위해 가장 최근에 결합된 두 군집은 Algeria ~ Bangladesh 로 구성된 군집과 Iraq ~ Saudi Arabia 로 구성된 군집입니다. 이러한 두 군집 간의 거리는 다이아몬드 가 4 로 설정될 때 거리 그림에서 세로선으로 표시된 점입니다. 거리는 "군집화 기록" 보고서에서 "군집 수" = 4 옆에 제공됩니다. 여기서 거리는 1.618708760 이고, Algeria 와 Iraq 로 시작하는 군집이 결합되어 네 개의 군집을 생성한다는 것을 알 수 있습니다.

거리 그래프를 보면 네 개의 군집에서 기울기가 눈에 띄게 변했습니다. 기울기 변화는 네 개의 군집이 남아 있는 점까지 결합된 군집의 차이가 비교적 작다는 것을 나타냅니다. 따라서 군집 수를 4 로 선택하는 것이 좋습니다. 이 값은 기본적으로 표시된 군집 수입니다.

7. "계층적 군집화"의 빨간색 삼각형을 클릭하고 **성운형 그림**을 선택합니다.

그림 13.3 성운형 그림



이 성운형 그림은 국가를 끝점으로 배열하고 각 군집 결합을 새 점으로 배열합니다. 선은 소속 군집을 나타냅니다. 군집 결합 사이의 선 길이는 결합된 군집 사이의 거리와 거의 같습니다.

다. 성운형 그림을 보면 Zaire 와 Afghanistan 을 포함하는 군집과 두 개의 각 주요 군집 간의 거리가 거의 비슷합니다.

## 계층적 군집화 플랫폼 시작

분석 > 군집화 > 계층적 군집화를 선택하여 계층적 군집화 플랫폼을 시작합니다.

그림 13.4 계층적 군집화 시작 창

유사한 관측값을 연쇄적으로 군집 계층 구조로 군집화합니다.

**열 선택**

3개 열

- country
- birth
- death

**선택한 열 역할 지정**

Y, 열    선택적

정렬    선택적 숫자

라벨    선택적

기준    선택적

**작업**

확인

취소

제거

재호출

도움말

**옵션**

**방법**

- Ward
- 평균
- 중심
- 단일
- 완전
- 빠른 Ward
- 하이브리드 Ward

데이터 형식    일반적인 데이터

표준화 기준    열

로버스트 표준화

길쭉값 대치

이월 군집화

**고급 옵션**

Hybrid Goal    400

Hybrid Cycles    30

Hybrid Initial K    10

Hybrid RandomPCA Dim    0

Hybrid Log Details

저장된 군집 테이블 사용

"열 선택"의 빨간색 삼각형 메뉴에 포함된 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

**Y, 열** 관측값을 군집화하는 데 사용되는 변수입니다.

**정렬** 지정된 열에 기반한 평균 값을 기준으로 군집을 정렬합니다.

**팁:** 주성분 분석을 수행하여 얻은 첫 번째 주성분을 정렬 열로 사용합니다. 군집은 이러한 값을 기준으로 정렬됩니다.

**속성 ID** (쌍인 데이터를 데이터 구조로 선택한 경우에만 사용 가능) 쌍이는 변수를 지정합니다.

**개체 ID** ( 요약 데이터 또는 쌓인 데이터를 데이터 구조로 선택한 경우에만 사용 가능 ) 측정값이 쌓이는 각 유닛에 대한 고유 식별자를 제공하는 하나 이상의 열입니다.

**라벨** 보고서의 덴드로그램에 라벨을 지정하는 데 사용되는 열입니다.

---

**참고 :** 선택한 데이터 구조가 **거리 행렬 데이터**인 경우 라벨 열의 데이터 유형은 문자여야 합니다.

---

**기준** 각 수준으로 개별 분석을 정의하는 열입니다. 지정된 열의 각 수준에 대해 해당하는 행이 분석됩니다. 결과는 개별 보고서에 제공됩니다. 기준 변수가 둘 이상 할당되면 기준 변수의 가능한 각 수준 조합에 대해 개별 분석이 생성됩니다.

**방법** 군집을 정의하기 위한 거리를 계산하는 데 사용되는 방법을 지정합니다. 각 방법에 의해 정의된 거리가 최소화되도록 군집이 결합됩니다. 거리 계산식은 " **거리 방법에 대한 통계 상세 정보** " 에서 확인하십시오.

**Ward** 두 군집 간의 거리를 모든 변수에 대해 합산된 두 군집 간의 ANOVA 제곱합으로 정의합니다. 각 생성 단계마다 이전 생성 단계에서 얻은 두 군집을 병합하여 얻을 수 있는 모든 파티션에 대해 군집 내 제곱합이 최소화됩니다. 제곱합을 총 제곱합으로 나누어 분산 비율 (준부분 상관 제곱) 을 제공하면 제곱합을 더 쉽게 해석할 수 있습니다.

Ward 방법은 다변량 정규 혼합, 구형성 공분산 행렬 및 등확률 표집을 가정하고 계층의 각 수준에서 가능도를 최대화하도록 군집을 결합합니다.

Ward 방법은 관측값 수가 적은 군집을 결합하는 경향이 있으며 관측값 수가 거의 같은 군집을 생성하는 쪽으로 크게 편향되어 있습니다. 또한 이상치에도 매우 민감합니다. 자세한 내용은 Milligan(1980) 연구 자료에서 확인하십시오.

**평균** 두 군집 간의 거리를 관측값 쌍 간의 평균 거리로 정의합니다. 평균 연결법은 분산이 작은 군집을 결합하는 경향이 있으며 분산이 같은 군집을 생성하는 쪽으로 약간 편향되어 있습니다. 자세한 내용은 Sokal 과 Michener(1958) 연구 자료에서 확인하십시오.

**중심** 두 군집 간의 거리를 군집 평균 간의 유클리드 거리 제곱으로 정의합니다. 중심 방법은 대부분의 다른 계층적 방법보다 이상치에 더 로버스트하지만 다른 측면에서는 Ward 방법 또는 평균 연결법만큼 잘 수행되지 않을 수 있습니다. 자세한 내용은 Milligan(1980) 연구 자료에서 확인하십시오.

**단일** 두 군집 간의 거리를 한 군집의 관측값과 다른 군집의 관측값 사이의 최소 거리로 정의합니다. 단일 연결법은 바람직한 이론적 특성을 많이 가지고 있지만 몬테카를로 연구에서는 잘 수행되지 않았습니다. 자세한 내용은 Jardine 과 Sibson(1971), Fisher 와 Van Ness(1971), Hartigan(1981) 및 Milligan(1980) 연구 자료에서 확인하십시오. 단일 연결법은 Florek et al. (1951a, 1951b) 에 의해 시작되었으며 이후에 McQuitty(1957) 와 Sneath(1957) 에 의해 재창조되었습니다.

단일 연결법은 군집 모양에 제약 조건을 두지 않으므로 길고 불규칙한 군집을 감지하는 대신 조밀한 군집 복구 성능이 저하됩니다. 단일 연결법은 주 군집을 분리하기 전에 분포의 꼬리를 잘라내는 경향이 있습니다. 자세한 내용은 Hartigan(1981) 연구 자료에서 확인하십시오.

**완전** 두 군집 간의 거리를 한 군집의 관측값과 다른 군집의 관측값 사이의 최대 거리로 정의합니다. 완전 연결법은 지름이 거의 같은 군집을 생성하는 쪽으로 크게 편향되어 있으며 중간 이상치에 의해 심하게 왜곡될 수 있습니다. 자세한 내용은 Milligan(1980) 연구 자료에서 확인하십시오.

**빠른 Ward** Ward 방법을 사용하여 두 군집 간의 거리를 정의합니다. "빠른 Ward"는 근접 이웃 체인 알고리즘을 사용하여 Ward 거리를 계산합니다. 이 알고리즘을 사용하면 거리 행렬을 계산할 필요가 없으므로 계산 시간이 단축됩니다. "빠른 Ward"는 행 수가 2,000 개를 초과할 때마다 자동으로 사용됩니다.

**하이브리드 Ward** 군집화를 두 단계로 나누는 알고리즘을 적용합니다. 첫 번째 단계는 근접 이웃 결합 주기를 사용하여 예비 군집을 생성하는 전처리 단계입니다. 자세한 내용은 "[군집 이웃 결합 주기에 대한 통계 상세 정보](#)"에서 확인하십시오. 이 작업은 계층적 군집화 루틴에 전달되는 테이블의 크기를 줄이기 위해 수행됩니다. 특정 횟수의 주기가 수행되거나 특정 수의 군집이 생성된 후 Ward 방법을 사용하여 나머지 군집이 형성됩니다. 이 방법은 군집화할 항목이 수만 또는 수십만 개일 때 유용합니다.

---

**참고 :** 빠른 Ward 방법과 달리 이 방법은 전체 Ward 방법과 동일한 계층을 생성하지 않습니다. 그러나 항목 수가 많은 경우, 특히 컴퓨팅 코어가 여러 개 있고 근접 이웃 검색에 멀티스레딩을 사용할 수 있는 경우 계산 시간이 더 적게 걸립니다.

---

**데이터 형식** 다변량 거리를 계산하는 데 사용되는 데이터의 형식을 지정합니다.

**일반 데이터** 각 관측값마다 하나의 행과 각 변수마다 하나의 열이 있는 직사각형 데이터입니다.

**요약 데이터** 하나 이상의 식별 열의 수준으로 요약된 데이터입니다. 이 옵션을 선택하면 시작 창에 "개체 ID" 텍스트 상자가 나타납니다. 식별 열을 개체 ID로 지정합니다. "요약 데이터" 옵션은 수준 평균을 계산하고 이러한 평균을 입력 데이터로 처리합니다.

**거리 행렬 데이터** 관측값 간의 거리로 구성된 데이터입니다. 관측값이  $n$ 개인 경우 거리 테이블에는  $n$ 개의 행과  $n+1$ 개의 열이 있어야 합니다. 하나의 열(대개 첫 번째 열)에  $n$ 개의 각 관측값에 대한 고유 식별자가 포함되어야 합니다. 나머지 열에는 해당 관측값과  $n$ 개 관측값 사이의 거리가 포함됩니다. 다음 사항에 유의하십시오.

- 점은 자신과의 거리가 0 이므로 테이블의 대각 요소는 0 이거나 결측이어야 합니다. 0이 아닌 값 또는 비결측 값은 0으로 처리되며 보고서에 이에 대한 설명이 나타납니다.
- 거리 열은 대칭 정방 행렬이거나, 아래쪽 또는 위쪽 부분에 결측 항목이 있는 상삼각 또는 하삼각 행렬일 수 있습니다. 거리가 정방 행렬로 제공된 경우 테이블이 대칭이 아니면 보고서에 경고가 나타납니다.
- 다른 데이터 구조로 시작한 후 거리 행렬을 저장할 수 있습니다. 자세한 내용은 "[거리 행렬 저장](#)"에서 확인하십시오.

"거리 행렬 데이터" 옵션을 선택하는 경우 거리 열을 "Y, 열"로 입력하고 식별자 열을 "라벨"로 입력합니다. 라벨 열의 데이터 유형은 문자여야 합니다. 예는 "[거리 행렬의 예](#)"에서 확인하십시오.

**쌓인 데이터** 관심 반응이 하나이고 각 개체에 대한 행이 여러 개 있는 데이터입니다.

"쌓인 데이터" 옵션을 선택하면 시작 창에 "속성 ID" 및 "개체 ID" 텍스트 상자가 나타납니다.

- 단일 열을 "Y, 열" 로 입력합니다.
- "Y, 열" 변수의 그룹화를 설명하는 열을 "속성 ID" 로 입력합니다. 두 개의 열만 입력하고 "공간 척도 추가" 를 선택하면 군집 분석에 사용할 공간 성분을 추가할 수 있습니다. 자세한 내용은 "공간 척도 추가" 에서 확인하십시오.
- 개체에 대한 식별 열을 "개체 ID" 로 입력합니다.

수행되는 분석은 "속성 ID" 열을 기준으로 "Y, 열" 변수를 분할한 후 반응 열을 표준화하지 않고 계층적 군집화를 수행하는 것과 같습니다.

**팁:** 이 옵션을 "공간 척도 추가" 옵션과 함께 사용하여 2 차원 공간 군집화를 수행할 수 있습니다. 예를 들어 웨이퍼 데이터는 주로 각 다이당 하나의 행을 사용하여 기록됩니다. 주요 관심은 웨이퍼를 군집화하는 것입니다. 자세한 내용은 "공간 척도를 사용한 웨이퍼 결합 분류의 예" 에서 확인하십시오.

**주의:** 측정값 열이 하나이므로 쌓인 데이터에는 데이터 표준화가 적절하지 않습니다.

**표준화 기준** 군집화하기 전에 값을 표준화하는 방법을 지정합니다. 연속형 열과 순서형 열에 대해 서로 다른 측정 척도 문제를 해결하는 데 유용합니다.

**표준화되지 않음** 원래 데이터를 사용합니다.

**열** 각 열의 값에서 열 평균을 빼고 열 표준편차로 나누어 표준화합니다.

**행** 각 행의 값에서 행 평균을 빼고 행 표준편차로 나누어 표준화합니다.

**열 및 행** 먼저 열 평균과 행 평균을 모두 빼 다음, 다시 총 평균을 더하여 값을 표준화합니다. 그러면 이중 중심화된 데이터의 표준편차를 사용하여 값이 척도화됩니다.

**로버스트 표준화** 연속형 열과 순서형 열의 평균 및 표준편차 추정값에 대한 이상치의 영향을 줄입니다. 이 옵션은 평균 및 표준편차의 Huber M- 추정값을 사용합니다 (Huber 1964, Huber 1973, Huber 와 Ronchetti 2009 참조). 이상치가 있는 열의 경우 이 옵션을 사용하면 다변량 거리를 결정할 때 표준화된 값을 더 크게 표현할 수 있습니다.

**참고:** "표준화 기준" 옵션을 사용하고 "로버스트 표준화" 를 선택하는 경우 지정한 표준화 방법에 대해 로버스트 평균 및 표준편차가 사용됩니다.

**결측값 대체** 결측값을 대체합니다. 변수 수가 50 개 이하이거나 행 수의 절반 미만이면 다변량 정규 대체법이 사용됩니다. 그렇지 않으면 다변량 SVD 대체법이 사용됩니다.

다변량 정규 대체법은 쌍별 공분산을 계산하여 반응 열에 대한 공분산 행렬을 생성합니다. 그런 다음 주어진 관측값에 대해 결측값이 없는 모든 예측 변수를 사용한 회귀 예측과 동일

한 방법으로 각 결측값이 대치됩니다. 생성된 공분산 행렬이 양정치가 아니면 열 평균을 사용하여 결측값이 대치됩니다.

다변량 SVD 대치법은 특이값 분해를 사용하여 공분산 행렬 생성을 방지합니다. 자세한 내용은 예측 및 전문 모델링에서 확인하십시오.

---

**주의:** 결측값 대치에서는 군집이 없고, 다변량 정규 분포의 데이터를 가져왔으며, 값이 랜덤으로 완전히 결측되었다고 가정합니다. 대개 이러한 가정은 실제로는 적절하지 않으므로 이 기능을 사용할 때 주의해야 합니다. 그러나 이 기능을 사용하면 데이터의 대부분을 삭제하는 것보다 더 유의한 결과를 얻을 수 있습니다.

---

**공간 축소 추가** (" 쌓인 데이터 " 옵션을 데이터 형식으로 선택한 경우에만 사용 가능) 데이터가 쌓여 있고 공간 좌표 (예: 가로 및 세로 좌표) 에 해당하는 두 속성 열을 포함하는 경우 이 옵션을 선택합니다. 이 옵션을 선택하면 결합 패턴을 군집화하는 데 도움이 되도록 공간 성분을 선택하고 가중치를 지정할 수 있는 창이 열립니다. 이 방법은 특별하며 매우 특정한 설정에만 적용할 수 있습니다. 자세한 내용은 "공간 축소에 대한 통계 상세 정보" 및 "공간 축도를 사용한 웨이퍼 결합 분류의 예" 에서 확인하십시오.

**이원 군집화** (" 일반 데이터 " 또는 " 요약 데이터 " 옵션을 데이터 형식으로 선택한 경우에만 사용 가능) 지정된 열과 행 둘 다를 기준으로 군집화합니다. 맨 아래의 Y 변수 덴드로그램과 함께 색상 맵이 덴드로그램에 추가됩니다. 일반적으로 이원 군집화의 경우 변수가 동일한 척도로 측정되므로 데이터를 표준화하지 않습니다.

**고급 옵션** 하이브리드 Ward 방법에 대한 고급 옵션을 지정합니다.

**하이브리드 목표** 계층적 군집화 루틴으로 전환하기 전에 허용되는 최대 군집 수를 지정합니다. 계층적 군집화 루틴이 시작될 때 군집 수는 "하이브리드 목표" 보다 작거나 같아야 합니다. "하이브리드 목표" 의 기본값은 400 입니다.

**하이브리드 주기** 계층적 군집화 루틴으로 전환하기 전에 수행되는 근접 이웃 결합 주기의 최소 수를 지정합니다. "하이브리드 주기" 의 기본값은 30 입니다.

**하이브리드 초기 K** 근접 이웃 결합 주기에 사용되는 초기 이웃 수를 지정합니다. 이웃 수는 이전 주기에서 발견된 고유한 인접 이웃 수에 따라 증가하거나 감소할 수 있습니다. "하이브리드 초기 K" 의 기본값은 10 입니다.

**하이브리드 랜덤 PCA 차원** 랜덤화 PCA 차원 축소 기법에 사용할 차원 수를 지정합니다. 이 기법은 "하이브리드 랜덤 PCA 차원" 값이 0 보다 클 때 사용되며 속도가 훨씬 향상됩니다. 랜덤화 PCA 기법은 근사 주성분을 계산하여 문제의 차원을 줄이고 점 사이의 거리를 근사화합니다. 자세한 내용은 Halko, Martinsson & Tropp(2011) 에서 확인하십시오.

**하이브리드 로그 상세 정보** 하이브리드 Ward 방법의 각 상태와 타이밍을 로그에 표시할지 여부를 지정합니다.

**저장된 군집 테이블 사용** 별도의 군집 기록 테이블을 사용하여 군집화를 지정합니다.

**경고 - 비결측 데이터가 충분하지 않습니다 .**

" 요약 데이터 " 또는 " 쌓인 데이터 " 형식을 사용하는 경우 " 비결측 데이터가 충분하지 않습니다 ." 라는 JMP 경고를 이해하기 어려울 수 있습니다 . 다음과 같은 경우에 경고가 발생합니다 .

- 데이터 형식을 " 일반 데이터 " 로 선택한 경우 모든 행 또는 하나를 제외한 모든 행에서 " Y, 열 " 변수의 값이 하나 이상 결측일 때 경고가 발생합니다 .
- 데이터 형식을 " 요약 데이터 " 로 선택한 경우 데이터가 " 개체 ID " 열에 대해 요약되고 모든 행 또는 하나를 제외한 모든 행에서 요약된 " Y, 열 " 변수의 값이 하나 이상 결측일 때 경고가 발생합니다 . 군집화 플랫폼에서 분석하는 데이터 구조를 확인하려면 **테이블 > 요약**을 선택한 후 " 개체 ID " 열을 " 그룹 " 으로 입력하고 " Y, 열 " 변수를 통계량 > 평균으로 입력합니다 .
- 데이터 형식을 " 쌓인 데이터 " 로 선택한 경우 " 속성 ID " 열에 대해 데이터가 분할되고 모든 행 또는 하나를 제외한 모든 행에서 분할 " Y, 열 " 의 값이 하나 이상 결측일 때 경고가 발생합니다 . 군집화 플랫폼에서 분석하는 데이터 구조를 확인하려면 **테이블 > 분할**을 선택한 후 " 속성 ID " 열을 " 분할 기준 " 으로, " Y, 열 " 변수를 " 분할 열 " 로, " 개체 ID " 열을 " 그룹 " 으로 각각 입력합니다 .

**팁 :** 결측값이 있는 개체를 식별하는 메시지도 로그에 출력됩니다 .

**계층적 군집화 보고서**

" 계층적 군집화 " 보고서에는 사용된 방법 , 덴드로그램 및 " 군집화 기록 " 테이블이 표시됩니다 . 시작 창에서 라벨 열을 할당한 경우 해당 열의 값으로 덴드로그램의 각 관측값을 식별합니다 .

- " 덴드로그램 "
- " 군집화 기록 "

**덴드로그램**

" 계층적 군집화 " 보고서의 " 덴드로그램 " 섹션에는 덴드로그램과 거리 그래프가 포함되어 있습니다 . 덴드로그램은 관측값이 군집으로 병합되는 것을 나타내는 트리 다이어그램입니다 . 덴드로그램은 군집의 비유사성 정도에 대한 정보도 제공합니다 .

덴드로그램을 왼쪽에서 오른쪽으로 읽으면 군집화 과정을 볼 수 있습니다 . 각 단계는 가장 가까운 두 군집을 단일 군집으로 결합하는 것으로 구성됩니다 .

- 군집 결합은 세로선에 의해 연결되는 가로선으로 표시됩니다 .
- 세로선의 가로 위치는 지정된 수의 군집을 형성하기 위해 가장 최근에 결합된 두 군집 사이의 거리를 나타냅니다 .

**참고 :** 관측값 수가 256 개 미만인 경우 거리는 거리 그래프에 표시된 거리에 비례합니다 . 그렇지 않은 경우 기하 간격이 사용됩니다 . 자세한 내용은 " 덴드로그램 척도 " 에서 확인하십시오 .

다음과 같은 작업을 수행할 수 있습니다.

- 덴드로그램의 위쪽 또는 아래쪽에 있는 다이아몬드 모양 핸들을 클릭한 채로 드래그하여 주어진 군집 수를 식별합니다.
- 군집 줄기를 클릭하여 덴드로그램과 데이터 테이블에서 군집의 모든 멤버를 선택합니다.

## 거리 그래프

거리 그래프는 덴드로그램 아래에 나타나는 그림입니다. 이 그래프에는 두 군집이 단일 군집으로 결합되는 각 단계에 대한 점이 있습니다. 가로 좌표는 군집 수를 나타내며 왼쪽에서 오른쪽으로 감소합니다. 점의 세로 좌표는 주어진 단계에서 결합된 군집 간의 거리입니다.

덴드로그램에서 다이아몬드 모양 핸들 중 하나를 클릭한 채로 드래그하면 선택된 군집 수를 제어할 수 있습니다. 다이아몬드를 클릭한 채로 드래그하면 군집 수에 따라 움직이는 세로선이 그림에 나타납니다. 거리 그래프의 기울기가 평평해지는 점이 종종 있습니다. 이러한 점은 자연적인 단락점을 나타내며 군집 수를 결정하는 데 도움이 됩니다.

덴드로그램과 거리 그래프의 관계에 대한 예는 "[계층적 군집화의 예](#)"에서 확인하십시오.

## 군집화 기록

"계층적 군집화" 보고서의 "군집화 기록" 테이블에는 군집화 기록을 설명하는 다음과 같은 열이 포함되어 있습니다.

**군집 수** 피 군집 개체와 군집 개체로 표시된 결합이 수행된 후 생성되는 군집 수를 나열합니다. 군집 수는 첫 번째 결합으로 시작합니다 ( $n - 1$  개의 군집이 있는 경우). 여기서  $n$  은 개체 수입니다. 보고서에는 모든 개체가 하나의 군집에 포함될 때까지 군집 수가 내림차순으로 나열됩니다. 군집화 기록은 이러한 방식으로 덴드로그램의 순서 (왼쪽에서 오른쪽으로)를 따릅니다.

**거리** 시작 창에서 선택한 거리 방법에 따라 계산된 군집 간의 거리입니다. 자세한 내용은 "[방법](#)"에서 확인하십시오.

**피 군집 개체** 덴드로그램에서 결합되는 첫 번째 군집의 대표 개체입니다. "피 군집 개체" 열에 표시된 군집 순서와 대표 개체는 데이터 정렬 방식에 따른 것이며 내재된 의미는 없습니다.

**군집 개체** 덴드로그램에서 결합되는 두 번째 군집의 대표 개체입니다. "군집 개체" 열에 표시된 군집 순서와 대표 개체는 데이터 정렬 방식에 따른 것이며 내재된 의미는 없습니다.

---

## 계층적 군집화 플랫폼 옵션

"계층적 군집화"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

**군집 색 표시** 소속 군집에 따라 덴드로그램 및 연결된 결합 막대의 라벨 색상을 지정합니다. 데이터 테이블의 행에도 해당 색상을 할당합니다. 군집 수를 변경하면 색상이 업데이트됩니다. 이 옵션을 선택 취소하면 군집 수에 따라 색상이 더 이상 업데이트되지 않습니다.

**군집 표시** 행이 속한 군집에 해당하는 데이터 테이블의 행에 표식을 할당합니다. 군집 수를 변경하면 표식이 업데이트됩니다. 이 옵션을 선택 취소하면 군집 수에 따라 표식이 더 이상 업데이트되지 않습니다.

**군집 수** 행 군집 수를 지정하고 덴드로그램 슬라이더를 해당 숫자에 배치합니다.

**군집 기준** ("거리 행렬 데이터"를 데이터 형식으로 선택한 경우에는 사용 불가능) CCC(Cubic Clustering Criterion) 테이블을 표시하거나 숨깁니다. CCC는 군집 1~데이터 테이블에 있는 관측값 수의 약 1/10에 해당하는 군집에 대해 표시됩니다. CCC는 특정 시나리오에서 군집 수를 추정하는 데 사용할 수 있습니다. 일반적으로 CCC 값이 클수록 군집 수 측면에서 더 나은 적합을 나타냅니다. CCC를 해석하는 방법에 대한 보다 구체적인 지침은 SAS Institute Inc. (1983)에서 확인하십시오. 이 기준은 거리 기반 군집화 알고리즘과 함께 사용할 수 있습니다.

---

**주의:** 군집이 길거나 불규칙한 모양인 경우 Cubic Clustering Criterion을 군집화 기준으로 사용하면 안 됩니다.

---

**덴드로그램 표시** "덴드로그램" 보고서를 표시하거나 숨깁니다.

**덴드로그램 척도** 덴드로그램 척도를 지정하기 위한 다음 옵션을 포함합니다.

**거리 간격** 시작 창에 지정된 거리 방법을 기반으로 두 결합점 사이의 가로 거리를 해당 점에서 결합된 두 군집 간의 거리로 표시합니다. 거리 척도는 거리 그래프에 사용되는 척도와 동일하며 덴드로그램의 기본 척도입니다.

**균등 간격** 두 결합점 사이의 가로 거리를 동일하게 표시합니다.

**기하 간격** 군집 수가 증가함에 따라 결합점 사이의 가로 거리를 증가합니다. 이 옵션은 개체가 많고, 큰 군집보다 작은 군집을 더 잘 보이게 표시하려는 경우에 유용합니다.

**거리 그래프** 덴드로그램 아래에 거리 그림을 표시하거나 숨깁니다.

**군집 수 핸들 표시** 덴드로그램에서 군집 수를 수동으로 변경하는 데 사용되는 핸들을 표시하거나 숨깁니다.

**선택 행 확대 / 축소** 덴드로그램에서 군집을 선택한 후 특정 군집을 선택하고 확대합니다. 또는 군집을 두 번 클릭하여 확대할 수 있습니다. 원래 보기로 돌아가려면 "확대 / 축소 해제"를 사용합니다.

**확대 / 축소 해제** 확대 / 축소 후 덴드로그램을 원래 보기로 되돌립니다.

**선택한 군집에서 피벗** 현재 선택된 군집의 두 하위 군집 순서를 역순으로 바꿉니다.

**위치 지정** 라벨 및 덴드로그램의 다른 부분에 대한 위치를 변경하는 옵션을 제공합니다.

**색상 맵** 값별로 색상이 지정된 각 "Y, 열" 변수를 표시하는 색상 맵 또는 히트맵을 추가할 수 있습니다. 하위 메뉴에 몇 가지 색상 테마 선택 옵션이 제공됩니다. 색상 맵을 제거하려면 **색상 맵 > 없음**을 선택합니다.

**추가 색상 맵 열** ("일반 데이터"를 데이터 형식으로 선택한 경우에만 사용 가능) 지정된 열에 대한 색상 맵을 추가합니다.

**범례** 색상 맵에 사용된 색상의 범례를 표시하거나 숨깁니다. 지정된 각 열에 대한 별도의 범례가 있습니다. 이 옵션은 색상 맵이 활성화된 경우에만 사용할 수 있습니다.

**참고:** 열 수가 400 개를 초과하면 색상 맵에 사용된 색상에 대한 표준화된 스코어와 함께 단일 범례가 표시됩니다.

**이원 군집화** ("일반 데이터" 또는 "요약 데이터"를 데이터 형식으로 선택한 경우에만 사용 가능) 지정된 열과 행 둘 다를 기준으로 군집화합니다. 맨 아래의 Y 변수 덴드로그램과 함께 색상 맵이 덴드로그램에 추가됩니다. 일반적으로 이원 군집화의 경우 변수가 동일한 척도로 측정되므로 데이터를 표준화하지 않습니다.

**열 군집화** (이원 군집화를 사용하는 경우에만 사용 가능) 이원 군집화에서 열을 군집화하기 위한 옵션을 제공합니다.

**열 군집 수** 열 군집 수를 지정합니다.

**열 군집 기준** 전체 열 군집 수 범위에 대한 CCC(Cubic Clustering Criterion) 테이블을 표시하거나 숨깁니다. CCC는 군집 수를 추정하는 데 사용됩니다. 이 통계량은 거리 기반 군집화 알고리즘과 함께 사용할 수 있습니다. CCC 값이 클수록 군집 수 측면에서 더 적합하다는 것을 나타냅니다. 자세한 내용은 SAS Institute Inc. (1983)에서 확인하십시오.

**열 군집 저장** 열의 소속 군집 정보를 포함하는 새 데이터 테이블을 저장합니다.

**군집 저장** 소속 군집 정보를 포함하는 새 데이터 테이블을 저장합니다. 시작 창에서 "공간 축소 추가"를 선택한 경우 군집 번호가 Hough 데이터 테이블에도 저장됩니다.

**군집 평균 저장** 각 군집에서 각 열의 평균과 행 수를 포함하는 새 데이터 테이블을 생성합니다.

**기타 저장** 추가 저장 옵션이 포함된 하위 메뉴를 표시합니다.

**가장 근접한 군집에 대한 계산식 저장** 가장 가까운 군집에 대한 계산식을 포함하는 데이터 테이블 열을 생성합니다. 이 옵션은 각 군집의 중심에 대한 유클리드 거리 제곱을 계산하고 가장 가까운 군집을 선택합니다. 군집이 다르게 결정되므로 이 계산식은 계층적 군집화에서 제공된 군집 할당을 재현하지 못할 수도 있습니다. 그러나 군집 할당은 매우 유사합니다. (요약 데이터, 거리 행렬 데이터 또는 쌓인 데이터를 선택한 경우에는 사용 불가능)

**군집 기록 저장** "군집화 기록" 보고서의 정보를 포함하는 새 데이터 테이블을 생성합니다.

**표시 순서 저장** 행이 덴드로그램에 나타나는 순서를 포함하는 데이터 테이블 열을 생성합니다.

**거리 행렬 저장** 관측값 사이의 거리를 포함하는 새 데이터 테이블을 생성합니다.

**성운형 좌표 저장** 성운형 그림의 좌표를 데이터 테이블에 저장합니다. ("요약 데이터", "거리 행렬 데이터" 또는 "쌓인 데이터"를 선택한 경우에는 사용 불가능)

**군집 계층 저장** 사용자 덴드로그램에 대한 스크립트를 작성하는 데 필요한 정보를 포함하는 데이터 테이블을 생성합니다. 각 군집 결합에 대해 세 개의 행이 있습니다. 첫 번째 행은 군집 개체, 두 번째 행은 피 군집 개체, 세 번째 행은 군집 중심, 크기 및 기타 정보를 제공하는 결과에 대한 것입니다.

**군집 트리 저장** JMP 와 SAS 간의 군집 트리를 비교하는 데 필요한 정보를 포함하는 새 데이터 테이블을 생성합니다. 각 군집 결합에 대해 새 군집마다 군집 크기와 기타 정보가 포함된 행이 각각 하나씩 있습니다.

**군집화 기록** "군집화 기록" 보고서를 표시하거나 숨깁니다. 자세한 내용은 "**군집화 기록**"에서 확인하십시오.

**군집 요약** ("거리 행렬 데이터"를 선택한 경우에는 사용 불가능) 다음 정보를 포함하는 보고서를 표시하거나 숨깁니다.

**군집 평균** 각 군집에 대해 관측값 수(쌓인 데이터의 경우 개체 ID) 및 각 변수의 평균을 제공하는 테이블입니다.

**군집 표준편차** 각 군집에 대해 관측값 수(쌓인 데이터의 경우 개체 ID) 및 각 변수의 표준편차를 제공하는 테이블입니다.

**군집 평균 그림** 군집 평균의 평행 그림 또는 2 차원 히트맵입니다.

**쌓인 데이터**를 선택하고 두 개의 "속성 ID" 변수가 있는 경우 외에는 평행 그림이 표시됩니다. 평행 그림의 경우 각 변수의 축이 축소됩니다.

- "표준화 기준" 옵션에 "열"을 선택한 경우 축 범위는 평균의 위아래로 2 표준편차이며, 여기서 표준편차와 평균은 원시 데이터에 대해 계산됩니다. 군집 평균이 이 범위를 벗어나면 이 값을 포함하도록 축이 확장됩니다.
- "표준화 기준" 옵션에 "열" 이외의 항목을 선택한 경우 척도가 표시되는 공통 세로 축이 있습니다. 척도는 "그래프 빌더"의 "균등 척도" 옵션과 동일합니다.

**쌓인 데이터**를 선택하고 두 개의 "속성 ID" 변수가 있는 경우 각 군집에 대해 각 위치에서 Y 변수의 평균을 보여 주는 2 차원 그림이 표시됩니다. 이러한 그림은 "파랑 - 회색 - 빨강" 색상 그래디언트를 사용하여 색상이 적용됩니다.

**열 요약** 각 변수에 대해 군집에 의해 설명되는 변동 비율을 나타내는  $R^2$  값을 제공합니다. 이 값은 군집의 변수 회귀에 대한  $R^2$  값입니다. 이 옵션은  $R^2$  값의 막대 차트도 제공합니다.

**늦은 결합 이상치** 알고리즘에서 가장 늦게 군집화된 관측값을 포함하는 테이블을 표시하거나 숨깁니다. 이 테이블의 관측값은 알고리즘이 80% 완료될 때까지도 하나의 군집을 구성합니다. 이러한 각 관측값은 알고리즘 후반까지 자체의 군집으로 남아 있기 때문에 데이터 집합에서 잠재적 이상치입니다.

**성운형 그림** 계층적 군집화 덴드로그램의 정보를 표시하는 대체 방법을 표시하거나 숨깁니다. 각 관측값(행)을 끝점으로 나타내고 각 군집 결합을 새 점으로 나타냅니다. 표시된 선은 소속 군집을 나타냅니다. 선 길이는 군집 간 거리를 나타냅니다. 선이 길수록 군집 간의 거리가 더 멀다는 것을 나타냅니다.

성운형 그림의 선을 커서로 가리키면 길이를 확인할 수 있습니다. 그러나 길이 값은 서로에 대해서만 의미가 있습니다. 축 척도, 점 방향 및 선 각도는 임의적입니다. 단지, 노드 끝 사

이에 간격을 두고 그림이 복잡해 보이지 않도록 값이 결정되므로 데이터 집합이 더 큰 경우 중요합니다.

끝점의 라벨을 해제하려면 성운형 그림 안쪽을 마우스 오른쪽 버튼으로 클릭하고 **라벨 표시**를 선택 취소합니다.

**산점도 행렬** ("일반 데이터"를 데이터 형식으로 선택한 경우에만 사용 가능) 모든 변수를 사용하여 산점도 행렬을 생성합니다.

**평행 좌표 그림** ("일반 데이터"를 데이터 형식으로 선택한 경우에만 사용 가능) 각 군집에 대한 평행 좌표 그림을 생성합니다. 축은 "군집 평균 그림"에 설명된 대로 척도화됩니다. 자세한 내용은 "군집 평균 그림"에서 확인하십시오.

**군집 처리 비교** (Shift 키를 누른 채 "계층적 군집화"의 빨간색 삼각형을 클릭한 경우에만 사용 가능) 반응 열과 2 수준 처리 열을 선택합니다. "계층적으로 군집화된 차이" 보고서를 생성합니다.

다음 옵션에 대한 자세한 내용은 **JMP 사용**에서 확인하십시오.

**로컬 데이터 필터** 특정 보고서에서 사용되는 데이터를 필터링할 수 있는 로컬 데이터 필터를 표시하거나 숨깁니다.

**다시 실행** 분석을 반복하거나 다시 시작할 수 있는 옵션이 포함되어 있습니다. 이 기능을 지원하는 플랫폼에서 "자동 재계산" 옵션은 해당하는 보고서 창에서 데이터 테이블에 대한 변경 사항을 즉시 반영합니다.

**플랫폼 환경 설정** 현재 플랫폼 환경 설정을 보거나, 현재 JMP 보고서의 설정과 일치하도록 플랫폼 환경 설정을 업데이트할 수 있는 옵션이 포함되어 있습니다.

**스크립트 저장** 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다.

**그룹별 스크립트 저장** 기준 변수의 모든 수준에 대한 플랫폼 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다. 시작 창에서 기준 변수를 지정한 경우에만 사용할 수 있습니다.

---

**참고:** 이 플랫폼의 추가 옵션은 스크립트를 통해 사용할 수 있습니다. 자세한 내용은 "도움말" 메뉴의 "스크립트 인덱스"에서 여십시오. 또한 "스크립트 인덱스"에서 이 섹션에 설명된 옵션의 스크립트 예제도 찾을 수 있습니다.

---

## 계층적 군집화의 추가 예

이 섹션에는 계층적 군집화 플랫폼을 사용하는 예가 포함되어 있습니다.

- "거리 행렬의 예"
- "공간 측도를 사용한 웨이퍼 결함 분류의 예"

## 거리 행렬의 예

이 예에서는 데이터에 관측값 간의 거리가 있으므로 "계층적 군집화" 시작 창에서 "거리 행렬 데이터" 옵션을 사용합니다. 거리 행렬에 대한 올바른 데이터 테이블 구조는 다음으로 구성됩니다.

- 문자 데이터 유형을 사용하는 식별자 열 (대개 첫 번째 열)
  - $n$  개 열의 집합. 여기서  $n$  은 행 수이기도 합니다. 이러한  $n$  개의 열은 대각선에 0 또는 결측값이 있는 대칭 행렬을 정의합니다.
1. **도움말 > 샘플 데이터 폴더**를 선택하고 `Flight Distances.jmp` 를 엽니다.
  2. **분석 > 군집화 > 계층적 군집화**를 선택합니다.
  3. "데이터 형식" 옆의 목록에서 **거리 행렬 데이터**를 선택합니다.
  4. `Cities` 를 선택하고 **라벨**을 클릭합니다.
  5. 나머지 열을 모두 선택하고 **Y, 열**을 클릭합니다.

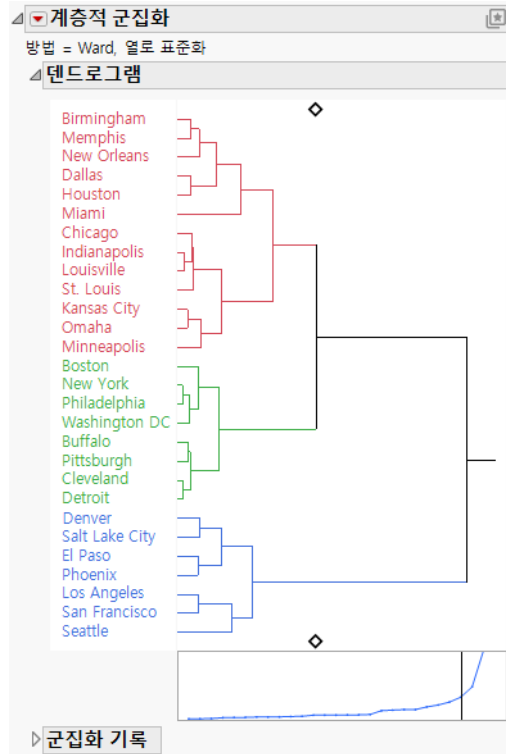
그림 13.5 완료된 거리 행렬 시작 창

유사한 관측값을 연속적으로 군집 계층 구조로 군집화합니다.

6. **확인**을 클릭합니다.

7. "계층적 군집화"의 빨간색 삼각형을 클릭하고 **군집 색 표시**를 선택합니다.

그림 13.6 Flight Distances 에 대한 덴드로그램 보고서



비행 거리 "덴드로그램" 보고서의 다이아몬드 위치는 모형이 도시를 세 개의 군집으로 그룹화했음을 나타냅니다. 군집은 덴드로그램에서 색상으로 구분됩니다. 보고서 해석 방법에 대한 자세한 내용은 "덴드로그램"에서 확인하십시오.

## 공간 측도를 사용한 웨이퍼 결함 분류의 예

이 예에서는 계층적 군집화 플랫폼의 "공간 측도" 라는 특별한 군집화 옵션을 사용합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Wafer Stacked.jmp 를 엽니다.
2. **분석 > 군집화 > 계층적 군집화**를 선택합니다.
3. " 데이터 형식 " 옆의 목록에서 **쌍인 데이터**를 선택합니다.  
쌍인 데이터에 대한 추가 옵션이 시작 창에 나타납니다.
4. Defects 를 선택하고 **Y, 열**을 클릭합니다.
5. X\_Die 와 Y\_Die 를 선택하고 **속성 ID** 를 클릭합니다.
6. Lot 와 Wafer 를 선택하고 **개체 ID** 를 클릭합니다.
7. **공간 측도 추가**를 선택합니다.

그림 13.7 완료된 군집화 시작 창

유사한 관측값을 연속적으로 군집 계층 구조로 군집화합니다.

**열 선택**

6개 열

- Lot
- Wafer
- Lot\_Wafer Label
- X\_Die
- Y\_Die
- Defects

**선택한 열 역할 지정**

Y, 열  Defects

정렬  선택적 숫자

속성 ID  X\_Die  
 Y\_Die

개체 ID  Lot  
 Wafer

라벨  선택적

기준  선택적

**작업**

확인

취소

제거

재로출

다음달

**옵션**

**방법**

- Ward
- 평균
- 중심심
- 단일
- 완전
- 빠른 Ward
- 하이브리드 Ward

데이터 형식

표준화 기준

로버스트 표준화

결측값 대치

이원 군집화

공간 축도 추가

8. **확인**을 클릭합니다.

그림 13.8 공간 성분 창

공간 성분 선택

변수	숫자	가중치
<input checked="" type="checkbox"/> 속성	1423	1
<input checked="" type="checkbox"/> 각도, 파이	18	<input type="text" value="1"/>
<input checked="" type="checkbox"/> 반지름, 원	21	<input type="text" value="1"/>
<input checked="" type="checkbox"/> 스트릭 각도	18	<input type="text" value="1"/>
<input checked="" type="checkbox"/> 스트릭 위치	10	<input type="text" value="1"/>
<input type="checkbox"/> 샷(shot) 내의 위치		<input type="text" value="1"/>
<input type="checkbox"/> 샷(shot)		<input type="text" value="1"/>
샷(shot) 가로 크기	<input type="text" value="0"/>	
샷(shot) 세로 크기	<input type="text" value="0"/>	

1423 개 위치에서 **Defects** 가 측정되었으므로 1423 개의 "속성" 변수가 있습니다.

9. **확인**을 클릭하여 "공간" 창의 선택 내용을 적용합니다.

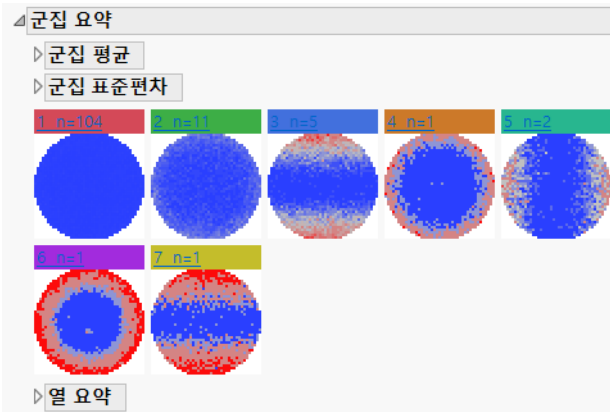
" 계층적 군집화 " 보고서와 "Wafer Stacked Defects 공간 " 데이터 테이블이라는 두 개의 창이 열립니다.

10. " 덴드로그램 " 그림에서 맨 위에 있는 다이아몬드 모양 핸들을 클릭한 채로 드래그하여 다양한 군집 수를 탐색합니다 .

핸들을 드래그하면 덴드로그램 아래의 거리 그래프에서 세로선이 해당 군집 수로 움직입니다 . 세로 좌표는 주어진 단계에서 결합된 군집 간의 거리를 제공합니다 . 군집 수가 7일 때 그래프가 수평이 되는 것 같습니다 .

11. " 계층적 군집화 " 의 빨간색 삼각형을 클릭하고 **군집 수**를 선택합니다 .
12. "7" 을 입력하고 **확인**을 클릭합니다 .
13. " 계층적 군집화 " 의 빨간색 삼각형을 클릭하고 **군집 요약**을 선택합니다 .

**그림 13.9** 군집 요약 보고서



웨이퍼 맵은 각 군집에 대한 결합의 공간적 특성을 나타냅니다 . 군집 1에는 104 개의 웨이퍼가 포함되어 있으며 비교적 적은 수의 결함이 웨이퍼 전체에 퍼져 있습니다 . 군집 3에는 5 개의 웨이퍼가 있으며 상단 및 하단 반구체의 극단에 결함이 집중되어 있습니다 . 군집 분석으로 생성된 데이터 테이블에서 개별 웨이퍼에 대한 맵과 Hough 공간 맵을 볼 수 있습니다 . 자세한 내용은 " 공간 측도에 대한 통계 상세 정보 " 에서 확인하십시오 .

## 계층적 군집화 플랫폼에 대한 통계 상세 정보

이 섹션에는 계층적 군집화 플랫폼에 대한 통계 상세 정보가 포함되어 있습니다 .

- " 공간 측도에 대한 통계 상세 정보 "
- " 거리 방법에 대한 통계 상세 정보 "
- " 근접 이웃 결합 주기에 대한 통계 상세 정보 "

## 공간 측도에 대한 통계 상세 정보

계층적 군집화 플랫폼에서 "공간 측도 추가" 옵션을 사용하려면 데이터가 쌓여 있고, 공간 좌표에 해당하는 두 개의 속성 열이 포함되어 있어야 합니다. 일부 공간 측도는 Hough 변환을 사용하여 생성됩니다. 자세한 내용은 White et al. (2008) 및 Ballard(1981) 연구 자료에서 확인하십시오. 자세한 내용은 "공간 측도를 사용한 웨이퍼 결함 분류의 예"에서 확인하십시오.

### 공간 성분 선택 창

시작 창에서 다음을 수행하면 "공간 성분 선택" 창이 나타납니다.

- 데이터 구조를 "쌓인 데이터" 로 선택
- 공간 좌표에 해당하는 두 개의 열을 "속성 ID" 로 지정
- "개체 ID" 지정
- "공간 측도 추가" 선택

"공간 성분 선택" 창에서 군집 분석에 사용할 공간 성분을 선택하고 가중치를 지정합니다. 이러한 성분은 군집 분석에 사용되는 변수를 생성하는 데 사용됩니다. 각 개체에 대한 행이 있는 새 테이블이 열립니다. 이 테이블에는 각 개체에 대해 계산된 공간 성분이 포함되어 있습니다.

**변수** 군집 분석에서 생성 및 사용되는 변수 유형입니다. 변수는 공간 성분과 반응 Y 를 사용하여 생성됩니다.

**속성** 각 개체에 대해 두 개의 "속성 ID" 변수로 정의된 각 위치에서 계산된 Y 변수의 값입니다.

**각도, 파이** 파이 모양 또는 반구체 모양을 반영하는 변수입니다.

**반지름, 원** 원형 모양을 반영하는 변수입니다.

**스트릭 각도** 동일한 각도의 스트릭을 반영하는 변수입니다.

**스트릭 위치** 동일한 공간 위치의 스트릭을 반영하는 변수입니다.

**샷 (shot) 내의 위치** 샷 (shot) 에서 다이 위치를 기반으로 하는 변수입니다. "샷 (shot) 내의 위치" 변수는 "ShotPos[vShotSize, hShotSize]" 로 표시되며, 여기서 vShotSize 와 hShotSize 는 정의된 세로 및 가로 샷 (shot) 크기입니다.

**샷 (shot)** 개체가 있는 직사각형을 식별하는 변수이며, 직사각형에 있는 개체의 가로 및 세로 위치 수를 지정합니다. 샷 (shot) 이라는 용어는 반도체 웨이퍼 데이터에서 웨이퍼 전체에 대해 함께 이미징되는 다이를 식별하기 위해 사용됩니다.

"샷 (shot) 가로 크기" 및 "샷 (shot) 세로 크기" 에 값을 입력합니다. 가로 샷 (shot) 크기를 4 로 지정하고 세로 샷 (shot) 크기를 5 로 지정하면 한 샷 (shot) 에 최대 20 개의 다이가 있음을 나타냅니다. 생성되는 총 식별자 수는 다음과 같이 계산됩니다.

$$\text{floor}[(\text{hSize}+\text{hShotSize}-1)/\text{hShotSize}] * \text{floor}[(\text{vSize}+\text{vShotSize}-1)/\text{vShotSize}]$$

여기서 hSize 와 vSize 는 각각 최대 가로 위치 수와 최대 세로 위치 수이며, hShotSize = 샷 (shot) 가로 크기이고 vShotSize = 샷 (shot) 세로 크기입니다.

---

**참고 :** 샷 (shot) 변수는 "Shot[vert, horiz]" 로 표시되며 , 여기서 vert 와 horiz 는 각각 세로 및 가로 다이 위치를 나타냅니다 .

---

**개수** 주어진 유형으로 생성되는 총 변수 수입니다 .

**가중치** 군집을 결정하는 데 사용되는 주어진 유형의 변수에 대한 중요도 척도입니다 .

## 공간 척도 보고서

"공간 성분 선택" 창에서 "확인"을 클릭하면 두 개의 창이 나타납니다 .

### 계층적 군집화 보고서

쌍인 데이터와 두 개의 속성 ID를 사용하여 분석을 수행하면 "군집 요약" 보고서에 Y 변수의 공간 맵이 표시됩니다 . 각 그림은 "속성 ID" 변수로 정의된 각 위치에 대한 군집 평균을 표시하는 2 차원 그림입니다 . 이 그림에는 파랑-회색-빨강 색상 그래디언트와 분위수 척도가 사용됩니다 . 분위수 척도를 사용하면 이상치로 인한 영향이 줄어듭니다 .

### 공간 데이터 테이블

공간 척도 데이터 테이블에는 고유한 각 개체 ID 에 대한 행이 있습니다 . 열은 Y 변수를 나타내기 위해 파랑-회색-빨강 기본 색상 그래디언트를 사용하여 표시됩니다 . 이 테이블에는 다음 열이 포함됩니다 .

**Object** 두 개의 "속성 ID" 변수로 정의된 각 공간 위치에서 Y 변수의 히트맵을 표시하는 표현식 열입니다 .

**Hough** 각 개체에 대한 Hough 공간의 히트맵을 표시하는 표현식 열입니다 . 자세한 내용은 White et al. (2008) 연구 자료에서 확인하십시오 .

**공간 척도** 각 개체에 대해 계산된 값을 표시하는 각 공간 척도의 열입니다 . 셀은 값에 따라 색상이 지정됩니다 .

## 거리 방법에 대한 통계 상세 정보

이 섹션에서는 "계층적 군집화" 시작 창에서 선택한 방법을 기반으로 거리를 계산하는 데 사용되는 계산식을 제공합니다 . 방법에 대한 설명은 "방법"에서 확인하십시오 .

계산식에는 다음 표기가 사용됩니다 . 여기서 소문자 기호는 일반적으로 관측값과 관련되고 대문자 기호는 군집과 관련됩니다 .

$n$  = 관측값 수

$v$  = 변수 수

$x_i$  =  $i$  번째 관측값

$C_K$  =  $K$  번째 군집 ( $\{1, 2, \dots, n\}$  의 부분집합)

$N_K = C_K$ 의 관측값 수

$\bar{x}$  = 표본 평균 벡터

$\bar{x}_K = C_K$  군집의 평균 벡터

$\|x\| = \mathbf{x}$  요소의 제곱합 제곱근 ( $\mathbf{x}$  벡터의 유클리드 길이)

$$d(x_i, x_j) = \|x_i - x_j\|^2$$

**평균 연결법** 평균 연결 군집 방법의 거리는 다음과 같습니다.

$$D_{KL} = \sum_{i \in C_K} \sum_{j \in C_L} \frac{d(x_i, x_j)}{N_K N_L}$$

**중심 방법** 중심 군집화 방법의 거리는 다음과 같습니다.

$$D_{KL} = \|\bar{x}_K - \bar{x}_L\|^2$$

**Ward** Ward 방법의 거리는 다음과 같습니다.

$$D_{KL} = \frac{\|\bar{x}_K - \bar{x}_L\|^2}{\frac{1}{N_K} + \frac{1}{N_L}}$$

**단일 연결법** 단일 연결 군집 방법의 거리는 다음과 같습니다.

$$D_{KL} = \min_{i \in C_K} \min_{j \in C_L} d(x_i, x_j)$$

**완전 연결법** 완전 연결 군집 방법의 거리는 다음과 같습니다.

$$D_{KL} = \max_{i \in C_K} \max_{j \in C_L} d(x_i, x_j)$$

## 근접 이웃 결합 주기에 대한 통계 상세 정보

계층적 군집화 플랫폼에서 근접 이웃 결합 주기는 하이브리드 Ward 방법의 첫 번째 단계에 사용됩니다. 이 작업은 계층적 군집화 루틴에 전달되는 테이블의 크기를 줄이기 위해 수행됩니다. 근접 이웃 결합 주기 알고리즘의 규격은 다음과 같습니다.

**하이브리드 목표** 알고리즘이 중지되기 전에 허용되는 최대 군집 수를 지정합니다. 기본값은 400입니다.

**하이브리드 주기** 알고리즘이 중지되기 전에 수행되는 근접 이웃 결합 주기의 최소 수를 지정합니다. 기본값은 30입니다.

**하이브리드 초기 K** 근접 이웃 결합 주기에 사용되는 초기 이웃 수를 지정합니다. 기본값은 10입니다.

근접 이웃 결합 주기 알고리즘은 다음 단계를 반복합니다.

1. 최근접 이웃을 효율적으로 찾기 위해 VP(Vantage-Point, 기준점) 트리가 생성됩니다.
2. 각 항목마다  $k$  개의 최근접 이웃이 결정됩니다.
3. 근접 이웃 쌍이 거리별로 정렬됩니다.
4. 거리가 가장 짧은 쌍의 절반에 대해 각 쌍의 항목이 이 주기에서 다른 항목과 아직 결합되지 않은 경우 해당 항목을 결합합니다. 결합된 항목은 다음 주기의 항목이 됩니다.
5. 최소 주기 수 (하이브리드 주기) 에 도달할 때까지 **1 단계 ~ 4 단계**를 반복합니다.
  - 항목 수가 하이브리드 목표보다 작거나 같으면 중지합니다.
  - 항목 수가 하이브리드 목표보다 큰 경우 항목 수가 하이브리드 목표 미만일 때까지 **1 단계 ~ 4 단계**를 계속 반복합니다.

각 주기에서 결합된 쌍의 수가 적으면 다음 주기에 대해 최근접 이웃 수, 즉  $k$ 가 증가합니다.  $k$ 의 값은 이전 주기에서 충분한 수의 쌍이 결합되면 이후 주기에서 감소할 수 있습니다.  $k$ 의 값은 다음 규칙에 따라 증가하고 감소합니다.

- **4 단계**에서 20% 미만의 쌍이 결합되면  $k$  값이 10 씩 증가합니다.
- **4 단계**에서 10% 미만의 쌍이 결합되면  $k$  값이 20 씩 증가합니다.
- **4 단계**에서 5% 미만의 쌍이 결합되면  $k$  값이 30 씩 증가합니다.
- **4 단계**에서 30% 를 초과하는 쌍이 결합되면  $k$  값이 10 씩 감소합니다.

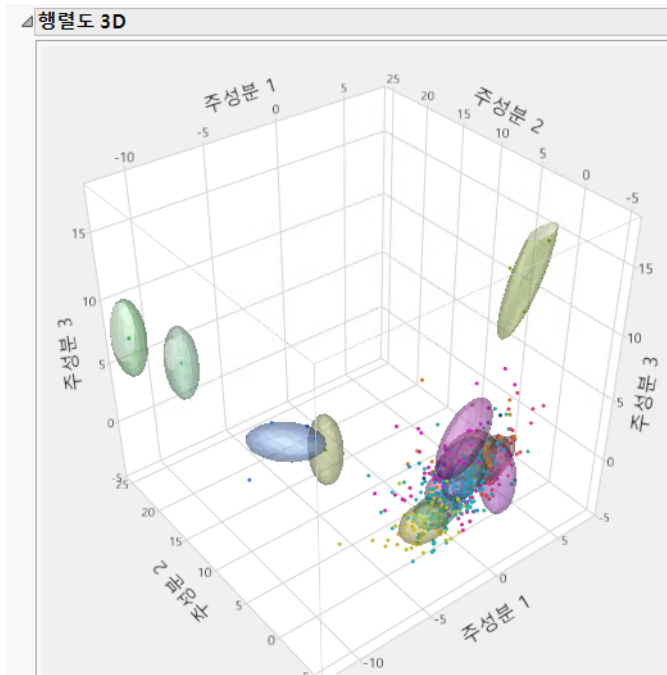
# 14 장

## K 평균 군집화 거리를 사용하여 관측값 그룹화

K 평균 군집화 플랫폼을 사용하여 여러 변수에 대해 유사한 값을 공유하는 관측값을 그룹화할 수 있습니다. 약 200 개 ~ 100,000 개의 관측값을 포함하는 큰 데이터 테이블에  $k$  평균 방법을 사용합니다.

K 평균 군집화 플랫폼에서는 관측값을 분할하는 반복 알고리즘을 사용하여 지정된 수의 군집을 생성합니다.  $k$  평균이라고 하는 이 방법은 군집 중심까지의 거리를 최소화하기 위해 관측값을 군집으로 분할합니다. 군집 수  $k$ 를 미리 지정해야 합니다. 그러나 여러  $k$  값의 결과를 비교하여 데이터에 대한 최적 군집 수를 선택할 수 있습니다.

그림 14.1 3D 행렬도



## 목차

K 평균 군집화 플랫폼 개요.....	315
관측값 군집화 플랫폼 개요.....	315
K 평균 군집화의 예.....	316
K 평균 군집화 플랫폼 시작.....	320
반복적 군집화 보고서.....	320
반복적 군집화 옵션.....	322
K 평균 보고서.....	322
군집 비교 보고서.....	322
K 평균 군집 수 보고서.....	323
자기 조직화 지도.....	325
자기 조직화 지도 제어판.....	325
자기 조직화 지도 보고서.....	326
SOM 알고리즘 설명.....	326
K 평균 군집화의 추가 예.....	327

## K 평균 군집화 플랫폼 개요

K 평균 군집화 플랫폼에서는 반복 적합 프로세스를 사용하여 지정된 수의 군집을 형성합니다.  $k$  평균 알고리즘은 먼저 군집 시드라고 하는  $k$ 개의 점 집합을 군집 평균의 초기 추측 값으로 선택합니다. 각 관측값은 가장 가까운 군집 시드에 할당되어 임시 군집 집합을 형성합니다. 그런 다음 시드가 군집 평균으로 대체되고 점이 재할당되며 군집에서 더 이상 변화가 없을 때까지 프로세스가 계속됩니다.

$k$  평균 알고리즘은 EM 알고리즘의 특수한 경우로, 여기서  $E$ 는 기대 (Expectation),  $M$ 은 최대화 (Maximization)를 나타냅니다.  $k$  평균 알고리즘의 경우 임시 군집 평균의 계산은 기대 단계를 나타내고 가장 가까운 군집에 점을 할당하는 것은 최대화 단계를 나타냅니다.

K 평균 군집화는 숫자 열만 지원합니다. K 평균 군집화는 모델링 유형(명목형 및 순서형)을 무시하고 모든 숫자 열을 연속형으로 처리합니다.

군집 수  $k$  또는  $k$  값 범위를 미리 지정해야 합니다. 그러나 여러  $k$  값의 결과를 비교하여 데이터에 대한 최적 군집 수를 선택할 수 있습니다.

K 평균 군집화에 대한 배경 정보는 SAS Institute Inc. (2023d)의 "FASTCLUS Procedure" 장 및 Hastie et al. (2009)에서 확인하십시오.

K 평균 군집화는 JMP에서 관측값을 군집화하기 위해 제공하는 네 가지 플랫폼 중 하나입니다. 네 가지 방법을 모두 비교하려면 "[관측값 군집화 플랫폼 개요](#)"의 내용을 참조하십시오.

## 관측값 군집화 플랫폼 개요

군집화는 여러 변수에 대해 유사한 값을 공유하는 관측값을 함께 그룹화하는 다변량 기법입니다. 일반적으로 관측값은  $p$  차원 공간에 고르게 분산되지 않습니다. 여기서  $p$ 는 변수 수입니다. 대신 관측값은 군집을 형성합니다. 이러한 군집을 식별하면 데이터를 더 세밀하게 파악할 수 있습니다.

**참고:** JMP에서는 변수를 군집화할 수 있는 플랫폼도 제공합니다. 자세한 내용은 "[변수 군집화](#)"에서 확인하십시오.

JMP에서 제공하는 다음 네 가지 플랫폼을 사용하여 관측값을 군집화할 수 있습니다.

- 계층적 군집화는 작은 데이터 테이블과 큰 데이터 테이블에 모두 유용하며 문자 데이터를 허용합니다. 계층적 군집화는 행을 계층적 순서로 결합하여 트리로 표현합니다. 트리가 생성되면 데이터에 가장 적절한 군집 수를 선택할 수 있습니다. 자세한 내용은 "[계층적 군집화](#)"에서 확인하십시오.
- K 평균 군집화는 최대 수백만 개의 행이 있는 큰 테이블에 적절하며 숫자 데이터만 허용합니다. 군집 수  $k$ 를 미리 지정해야 합니다. 이 알고리즘은 군집 시드 점을 추측합니다. 그런 다음 군집에 점을 번갈아 할당하고 군집 중심을 다시 계산하는 반복 과정을 수행합니다. 자세한 내용은 "[K 평균 군집화](#)"에서 확인하십시오.

- 정규 혼합은 중첩될 수 있는 다변량 정규 분포의 혼합에서 데이터를 수집한 경우에 적절하며 숫자 데이터만 허용합니다. 다변량 이상치가 있는 경우 균등 분포를 가정하고 이상치 군집을 사용할 수 있습니다. 자세한 내용은 "정규 혼합"에서 확인하십시오.  
군집 수를 미리 지정해야 합니다. 최대 가능도는 혼합 비율과 평균, 표준편차 및 상관을 공동으로 추정하는 데 사용됩니다. 각 그룹에 속할 확률이 각 점에 할당됩니다. EM 알고리즘은 추정값을 구하는 데 사용됩니다.
- 잠재 계층 분석은 대부분의 변수가 범주형일 때 적절합니다. 군집 수를 미리 지정해야 합니다. 이 알고리즘은 다항 혼합 분포를 가정하는 모형을 적합시킵니다. 각 관측값에 대해 소속 군집의 최대 가능도 추정값이 계산됩니다. 관측값은 소속 확률이 가장 큰 군집으로 분류됩니다. 자세한 내용은 "잠재 계층 분석"에서 확인하십시오.

표 14.1 군집화 방법 요약

방법	데이터 유형 또는 모델링 유형	데이터 테이블 크기	군집 수 지정
계층적 군집화	임의	하이브리드 Ward 를 사용하는 경우 최대 수십만 개의 행  빠른 Ward 를 사용하는 경우 최대 200,000개 행  다른 방법을 사용하는 경우 최대 5,000개 행	아니요
K 평균 군집화	수치형	최대 수백만 개 행	예
정규 혼합	수치형	모든 크기	예
잠재 계층 분석	명목형 또는 순서형	모든 크기	예

일부 군집화 플랫폼에는 데이터의 이상치를 처리하는 옵션이 있습니다. 그러나 데이터에 이상치가 있으면 분석하기 전에 먼저 이상치를 탐색하는 것이 가장 좋습니다. 이 작업은 이상치 탐색 유틸리티를 사용하여 수행할 수 있습니다. 자세한 내용은 예측 및 전문 모델링에서 확인하십시오.

## K 평균 군집화의 예

이 예에서는 K 평균 군집화 플랫폼을 사용하여 세포분석의 관측값을 군집화합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Cytometry.jmp 를 엽니다.
2. **분석 > 군집화 > K 평균 군집화**를 선택합니다.
3. CD3, CD8, CD4 및 MCB 를 선택하고 **Y, 열**을 클릭합니다.

4. **확인**을 클릭합니다.
5. **군집 수** 옆에 "3" 을 입력합니다.
6. **군집 범위 (선택 사항)** 옆에 "15" 를 입력합니다.  
군집 범위가 15 로 설정되었으므로 플랫폼에서 3 개 ~ 15 개 군집에 대한 적합을 제공합니다.  
그런 다음 원하는 군집 수를 결정할 수 있습니다.
7. **시작**을 클릭합니다.

**그림 14.2** 군집 비교 보고서

군집 비교			
방법	군집 수	CCC 최선	
K 평균 군집화	3	23.1784	
K 평균 군집화	4	8.80709	
K 평균 군집화	5	29.5123	
K 평균 군집화	6	52.5517	
K 평균 군집화	7	49.5876	
K 평균 군집화	8	56.5308	
K 평균 군집화	9	54.053	
K 평균 군집화	10	69.8707	
K 평균 군집화	11	70.5239	최적 CCC
K 평균 군집화	12	61.5326	
K 평균 군집화	13	68.1277	
K 평균 군집화	14	66.4044	
K 평균 군집화	15	69.9928	

"군집 비교" 보고서가 보고서 창 맨 위에 나타납니다. 최량 적합은 가장 큰 CCC 값에 의해 결정됩니다. 이 경우에는 11 개 군집을 적합시킬 때 최량 적합이 발생합니다.

8. "K 평균 군집 수 =11" 보고서가 나올 때까지 스크롤합니다.

**그림 14.3** K 평균 군집 수 =11 보고서

K 평균 군집 수 =11				
개별적으로 척도화된 열				
군집 요약				
군집	개수	단계	기준	
1	816	24	0	
2	482			
3	157			
4	577			
5	856			
6	498			
7	447			
8	549			
9	377			
10	123			
11	118			

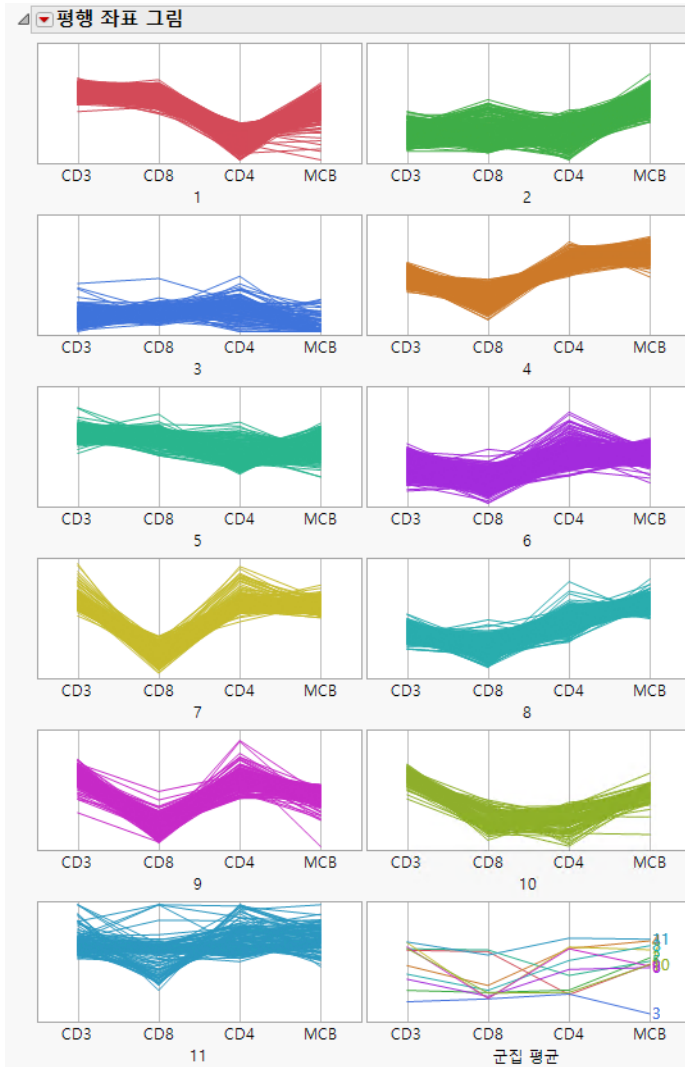
  

군집 평균				
군집	CD3	CD8	CD4	MCB
1	314.073529	300.270833	106.615196	183.4375
2	140.091286	116.365145	127.024896	205.014523
3	90.7898089	87.5477707	109.356688	15.0191083
4	247.426343	148.064125	314.074523	261.831889
5	320.287383	307.372664	193.117991	193.174065
6	188.686747	99.0863454	220.062249	171.682731
7	347.496644	89.9574944	320.782998	231.557047
8	210.258652	126.125683	260.327869	245.588342
9	328.206897	89.7824934	312.909814	175.965517
10	322.105691	113.715447	116.666667	181.731707
11	349.864407	284.813559	360.932203	267.474576

"군집 요약" 보고서에는 11 개 군집 각각의 관측값 수가 표시됩니다. "군집 평균" 보고서에는 각 군집에 대한 표지 관측값의 평균이 표시됩니다.

9. "K 평균 군집 수 =11" 의 빨간색 삼각형을 클릭하고 **평행 좌표 그림** 을 선택합니다.

그림 14.4 Cytometry 데이터에 대한 평행 좌표 그림

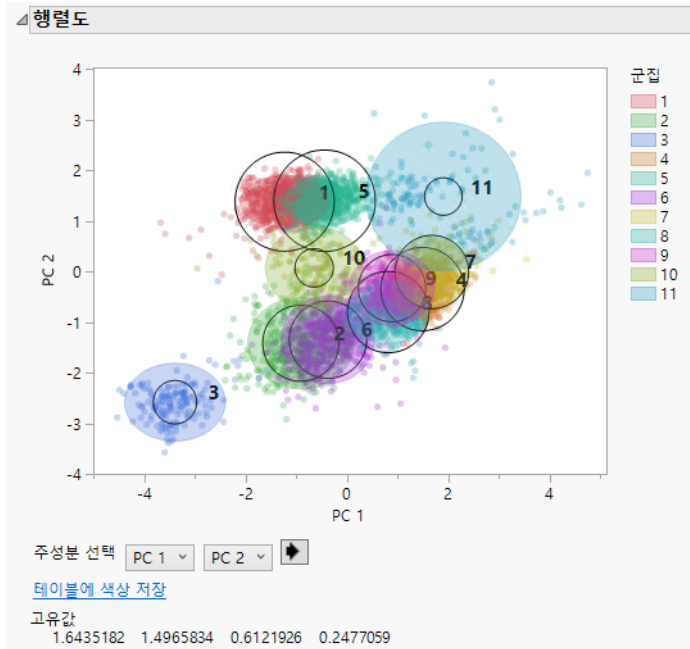


"평행 좌표 그림"에는 각 군집의 관측값 구조가 표시됩니다. 이러한 그림을 사용하여 군집이 어떻게 다른지 확인할 수 있습니다. 군집 4, 6, 7, 8, 9 는 비교적 CD8 값이 낮고 CD4 값이 높은 경향이 있습니다. 반면에 군집 1 은 CD8 값이 더 높고 CD4 값이 더 낮습니다.

10. "K 평균 군집 수 =11" 의 빨간색 삼각형을 클릭하고 **행렬도** 를 선택합니다.

- 11. (선택 사항) "K 평균 군집 수 =11" 의 빨간색 삼각형을 클릭하고 **행렬도 옵션 > 행렬도 선 표시** 를 선택 취소합니다.

그림 14.5 Cytometry 데이터에 대한 행렬도



군집 색상을 식별하는 범례가 그림 오른쪽에 표시됩니다 . 처음 두 주성분을 기반으로 다른 군집과 가장 많이 떨어져 있는 군집은 3, 10 및 11 군집입니다 . 이는 다른 군집에 대한 그림과 다른 **그림 14.4** 의 평행 좌표 그림에서도 입증됩니다 . 그림 아래의 목록을 사용하여 다른 주 성분 조합에 대한 행렬도를 확인할 수 있습니다 .

## K 평균 군집화 플랫폼 시작

분석 > 군집화 > K 평균 군집화를 선택하여 K 평균 군집화 플랫폼을 시작합니다.

그림 14.6 K 평균 군집화 시작 창

수치형 변수 기반의 행을 지정된 수의 군집으로 군집화합니다.

열 선택

8개 열

- ForSc
- SideSc
- CD3
- CD4
- MCB
- Prin1
- Prin2

선택한 열 역할 지정

Y, 열

필수 숫자  
필수 숫자  
선택적 숫자

가중치

선택적 숫자

빈도

선택적 숫자

기준

선택적

작업

확인

취소

제거

재호출

도움말

개별적으로 척도화된 열

"열 선택"의 빨간색 삼각형 메뉴에 포함된 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

**Y, 열** 관측값을 군집화하는 데 사용되는 변수입니다.

**참고 :** K 평균 군집화는 숫자 열만 지원합니다.

**가중치** 분석의 각 행에 가중치를 할당하는 숫자 값이 포함된 열입니다.

**빈도** 분석의 각 행에 빈도를 할당하는 숫자 값이 포함된 열입니다.

**기준** 각 수준으로 개별 분석을 정의하는 열입니다. 지정된 열의 각 수준에 대해 해당하는 행이 분석됩니다. 결과는 개별 보고서에 제공됩니다. 기준 변수가 둘 이상 할당되면 기준 변수의 가능한 각 수준 조합에 대해 개별 분석이 생성됩니다.

### 시작 창 옵션

**개별적으로 척도화된 열** 각 열을 다른 열과 독립적으로 척도화합니다. 변수가 공통 측정 척도를 공유하지 않고 한 변수가 군집화 과정을 지배하지 않게 하려는 경우 이 옵션을 사용합니다. 예를 들어 한 변수의 값이 0에서 1000 사이이고 다른 변수의 값이 0에서 10 사이일 수 있습니다. 이 경우 군집화 과정이 첫 번째 변수에 의해 지배되지 않도록 이 옵션을 사용할 수 있습니다.

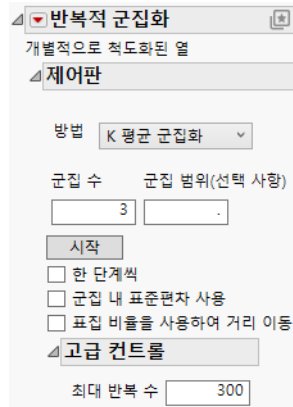
"확인"을 클릭하면 "제어판"이 나타납니다. 자세한 내용은 "반복적 군집화 보고서"에서 확인하십시오.

## 반복적 군집화 보고서

"K 평균 군집화" 플랫폼의 "반복적 군집화" 보고서에는 모형 적합을 위한 제어판이 표시됩니다. 다른 수의 군집을 반복적으로 적합시키거나, "군집 범위" 옵션을 사용하여 범위를 지정할 수 있습

니다. 모형을 적합시키면 해당 보고서가 창에 추가됩니다. 자세한 내용은 "K 평균 보고서"에서 확인하십시오.

그림 14.7 반복적 군집화 제어판



" 제어판 "에는 다음 옵션이 있습니다.

**방법** 다음과 같은 군집화 방법을 사용할 수 있습니다.

**K 평균 군집화** 이 장에서 설명합니다.

**자기 조직화 지도** " 자기 조직화 지도 제어판 "에서 설명합니다.

**군집 수** 형성할 군집 수를 지정합니다.

**군집 범위 (선택 사항)** 형성할 군집 수에 대한 상한을 제공합니다. 여기에 숫자를 입력하면 플랫폼에서 "군집 수"와 "군집 범위 (선택 사항)"로 입력한 값 사이의 모든 정수에 대해 개별 분석을 생성합니다.

**시작** 군집을 자동으로 적합시킵니다 (" 한 단계씩 "을 선택한 경우 제외).

**한 단계씩** 군집화 과정을 한 번에 한 단계씩 반복할 수 있습니다. " 한 단계씩 "을 선택하고 " 시작 "을 클릭하면 군집 할당이 없지만 " 시작 " 및 " 단계 "버튼이 포함된 "K 평균 군집화" 보고서가 나타납니다.

- 한 번에 한 단계씩 반복하려면 " 단계 "버튼을 클릭합니다.

- 군집을 자동으로 적합시키려면 " 시작 "버튼을 클릭합니다.

**군집 내 표준편차 사용** 각 군집 내 관측값에 대해 각 변수의 표준편차 추정값을 사용하여 거리를 척도화합니다. 이 옵션을 선택하지 않으면 각 변수의 전체 표준편차 추정값을 사용하여 거리를 척도화합니다.

**표집 비율을 사용하여 거리 이동** 군집 크기에 따라 거리를 조정합니다. 군집 크기가 같지 않은 경우 관측값은 더 큰 군집에서 나올 사전 확률이 더 높으므로 더 큰 군집에 관측값이 할당될 확률이 더 높아야 합니다.

**고급 컨트롤** 다음과 같은 고급 컨트롤을 사용할 수 있습니다.

**최대 반복 수** EM 알고리즘의 수렴 단계에 대한 최대 반복 수입니다.

## 반복적 군집화 옵션

이 섹션에서는 "반복적 군집화"의 빨간색 삼각형 메뉴에 포함된 옵션을 다룹니다.

다음 옵션에 대한 자세한 내용은 **JMP 사용**에서 확인하십시오.

**로컬 데이터 필터** 특정 보고서에서 사용되는 데이터를 필터링할 수 있는 로컬 데이터 필터를 표시하거나 숨깁니다.

**다시 실행** 분석을 반복하거나 다시 시작할 수 있는 옵션이 포함되어 있습니다. 이 기능을 지원하는 플랫폼에서 "자동 재계산" 옵션은 해당하는 보고서 창에서 데이터 테이블에 대한 변경 사항을 즉시 반영합니다.

**플랫폼 환경 설정** 현재 플랫폼 환경 설정을 보거나, 현재 JMP 보고서의 설정과 일치하도록 플랫폼 환경 설정을 업데이트할 수 있는 옵션이 포함되어 있습니다.

**스크립트 저장** 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다.

**그룹별 스크립트 저장** 기준 변수의 모든 수준에 대한 플랫폼 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다. 시작 창에서 기준 변수를 지정한 경우에만 사용할 수 있습니다.

---

**참고:** 이 플랫폼의 추가 옵션은 스크립트를 통해 사용할 수 있습니다. 자세한 내용은 "도움말" 메뉴의 "스크립트 인덱스"에서 여십시오. 또한 "스크립트 인덱스"에서 이 섹션에 설명된 옵션의 스크립트 예제도 찾을 수 있습니다.

---

## K 평균 보고서

"제어판"에서 "시작"을 클릭하면 "군집 비교" 보고서와 하나 이상의 "K 평균" 보고서가 나타납니다. K 평균 보고서 이름은  $k$  (적합된 군집 수)에 따라 동적으로 "K 평균 군집 수 =< $k$ >"로 지정됩니다. 수행한 각 적합에 대한 "K 평균 군집 수 =< $k$ >" 보고서가 나타납니다.

- "군집 비교 보고서"
- "K 평균 군집 수 보고서"

## 군집 비교 보고서

K 평균 군집화 플랫폼의 "군집 비교" 보고서에서는 다양한 모형을 비교하기 위한 적합 통계량을 제공합니다. 적합 통계량은 CCC(Cubic Clustering Criterion)입니다. CCC 값이 클수록 더 나은 적합을 나타냅니다. 최량 적합은 "최선" 열에 "최적 CCC"를 지정하여 표시됩니다. 자세한 내용은 SAS Institute Inc. (1983)에서 확인하십시오. 상수 열은 CCC 계산에 포함되지 않습니다.

## K 평균 군집 수 보고서

K 평균 군집화 플랫폼의 각 "K 평균 군집 수" 보고서에는 각 군집에 대한 다음과 같은 요약 통계량이 제공됩니다.

- "군집 요약" 보고서에는 군집 수와 각 군집의 관측값 수 및 필요한 반복 수가 제공됩니다.
- "군집 평균" 보고서에는 각 군집에서 각 변수에 대한 관측값의 평균이 제공됩니다.
- "군집 표준편차" 보고서에는 각 군집에서 각 변수에 대한 관측값의 표준편차가 제공됩니다.

## K 평균 군집 수 보고서 옵션

각 "K 평균 군집 수" 보고서에는 다음과 같은 빨간색 삼각형 메뉴 옵션이 있습니다.

**행렬도** 데이터의 처음 두 주성분으로 표현된 점 및 군집의 그림을 군집 색상을 식별하는 범례와 함께 표시합니다. 군집 중심을 기준으로 원이 그려지고 원 크기는 군집 내의 개수에 비례합니다. 음영 영역은 평균 주위의 밀도 등고선입니다. 기본적으로 이 영역은 해당 군집의 관측값 중 90%가 속하는 위치를 나타냅니다 (Mardia et al. 1980 참조). 그림 아래의 목록을 사용하여 그림 축을 다른 주성분으로 변경할 수 있습니다. 또는 화살표 버튼을 사용하여 가능한 모든 축 조합을 순환할 수 있습니다. 군집 색상을 데이터 테이블에 저장하는 옵션도 그림 아래에 있습니다. 자세한 내용은 "테이블에 색상 저장"에서 확인하십시오. 고유값은 내림차순으로 표시됩니다.

---

**참고:** 시작 창에서 "개별적으로 척도화된 열"을 선택한 경우 행렬도에 상관 행렬이 사용됩니다. "개별적으로 척도화된 열"을 선택하지 않은 경우에는 행렬도에 공분산 행렬이 사용됩니다.

---

**행렬도 옵션** 행렬도 모양을 제어하기 위한 다음 옵션을 포함합니다.

**행렬도 선 표시** 행렬도 선을 표시합니다. 라벨이 지정된 선은 주성분으로 정의된 부분 공간에서 공변량의 방향을 표시합니다. 이러한 선은 각 변수와 각 주성분의 연관성 정도를 나타냅니다.

**행렬도 선 위치** 행렬도 선의 위치와 반지름 척도를 지정할 수 있습니다. 기본적으로 선은 점 (0,0)에서 시작됩니다. 그림에서 선을 드래그하거나 이 옵션을 사용하여 좌표를 지정할 수 있습니다. 반지름 척도 옵션을 사용하여 선이 더 잘 보이게 선 척도를 조정할 수도 있습니다.

**행렬도 등고선 밀도** 밀도 등고선에 대한 신뢰 수준을 지정할 수 있습니다. 기본 신뢰 수준은 90%입니다.

**군집 표시** 데이터 테이블의 행에 군집을 식별하는 표식을 할당합니다.

**행렬도 3D** 데이터의 3 차원 행렬도를 표시합니다. 변수가 세 개 이상인 경우에만 사용할 수 있습니다.

**평행 좌표 그림** 각 군집에 대한 평행 좌표 그림을 생성합니다. 이 그림 보고서에는 데이터와 평균을 표시하거나 숨기는 옵션이 있습니다. 자세한 내용은 그래프 분석에서 확인하십시오.

**산점도 행렬** 모든 Y 변수를 사용하여 산점도 행렬을 표시하거나 숨깁니다. 각 산점도에는 현재 군집 수를 기반으로 하는 밀도 타원이 포함되어 있습니다.

**SOM 히트맵** (자기 조직화 지도에만 사용 가능) 군집화에 사용된 Y 변수 중 하나로 색상이 지정된 SOM 군집 평균의 히트맵을 표시하거나 숨깁니다. "히트맵에 적용할 색상 열 선택" 옆의 메뉴를 사용하여 Y 변수를 변경합니다.

**참고 :** 히트맵의 군집은 위에서 아래로, 오른쪽에서 왼쪽으로 배치되어 구성됩니다. 즉, 첫 번째 군집은 오른쪽 상단 모서리에 있고 마지막 군집은 왼쪽 하단 모서리에 있습니다.

**테이블에 색상 저장** 데이터 테이블의 행에 군집을 식별하는 색상을 할당합니다. 보고서 창에 "행렬도"가 있는 경우 데이터 테이블에 저장된 색상은 행렬도의 군집 색상과 일치합니다. "행렬도"에서 색상을 변경하고 "테이블에 색상 저장" 옵션을 다시 선택하면 테이블의 색상이 행렬도의 색상과 일치하도록 업데이트됩니다.

**참고 :** 저장 옵션을 선택하면 저장된 각 열에 해당 특정 열의 데이터에 대한 군집 수를 지정하는 "노트" 열 특성이 포함됩니다. 이렇게 하면 둘 이상의 군집 적합에서 열을 저장하고 열 특성을 사용하여 저장된 열을 가져온 군집화 적합을 식별할 수 있습니다.

**군집 저장** 다음 두 열을 데이터 테이블에 저장합니다.

- 군집 열은 주어진 행이 할당된 군집의 번호를 포함합니다.
- (자기 조직화 지도에는 사용 불가능) 거리 열은 주어진 관측값과 군집 평균 사이의 유클리드 거리 제곱을 포함합니다. 각 변수에 대해 관측값과 해당 변수의 군집 평균 간의 차이를 변수의 전체 표준편차로 나눕니다. 변수 전체에 대해 이러한 척도화된 차이의 제곱합이 계산됩니다.

**군집 거리 저장** (자기 조직화 지도에는 사용 불가능) 거리 열을 데이터 테이블에 저장합니다. 이 열은 **군집 저장** 옵션에서 얻은 거리 열과 동일합니다.

**SOM 격자 저장** (자기 조직화 지도에만 사용 가능) 새 열을 데이터 테이블에 저장합니다. 새 열에는 각 관측값에 대해 최대 확률 분류 군집의 SOM 격자 행 및 열 번호가 포함됩니다.

**군집 계산식 저장** "군집 계산식"이라는 계산식 열을 데이터 테이블에 저장합니다. 이 계산식은 각 소속 군집을 식별합니다.

**거리 계산식 저장** (자기 조직화 지도에는 사용 불가능) "거리 계산식"이라는 계산식 열을 데이터 테이블에 저장합니다. 이 계산식은 할당된 군집까지의 거리를 계산합니다.

**K 군집 거리 저장** (자기 조직화 지도에는 사용 불가능) 각 군집 중심까지의 유클리드 거리 제곱을 포함하는 k 개 열을 저장합니다.

**K 거리 계산식 저장** (자기 조직화 지도에는 사용 불가능) 각 군집 중심까지의 유클리드 거리 제곱 계산식을 포함하는 k 개 열을 저장합니다.

**군집 계산식 게시** "군집 계산식 저장" 옵션에 사용된 것과 동일한 스코어링 코드를 "계산식 저장소"에 게시합니다.

**군집 시뮬레이션** 군집 평균 및 표준편차를 사용하여 Y 변수에 대한 시뮬레이션된 군집 관측값을 포함하는 새 데이터 테이블을 생성합니다.

제거 군집화 보고서를 제거합니다.

## 자기 조직화 지도

SOM(자기 조직화 지도) 기법은 Teuvo Kohonen(1989, 1990)에 의해 개발되었으며 다른 신경망 지지자 및 통계학자에 의해 확장되었습니다. K 평균 군집화 플랫폼에서 SOM 기법을 구현할 수 있습니다. 예는 "[K 평균 군집화의 추가 예](#)"에서 확인하십시오.

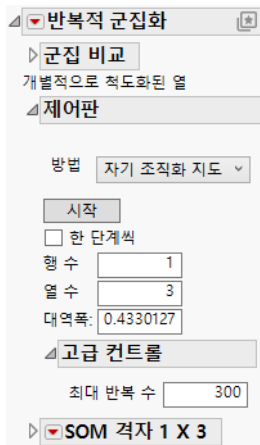
원래 SOM은 원래의 신경망 알고리즘처럼 학습 프로세스였지만 여기서 구현된 버전은  $k$  평균 군집화의 변형입니다. SOM 문헌에서는 이 변형을 국소 가중 선형 평활기를 사용한 배치 알고리즘이라고 합니다.

SOM의 목표는 군집 격자의 특정 레이아웃에 군집을 형성하는 것뿐만 아니라 SOM 격자에서 서로 가까운 군집의 점이 다변량 공간에서도 서로 가까이 있도록 하는 것입니다. 전통적  $k$  평균 군집화에서는 군집 구조가 임의적이지만 SOM의 군집은 격자 구조를 가집니다. 격자 구조는 군집을 2차원으로 해석하는 데 도움이 됩니다. 즉, 서로 가까운 군집이 멀리 있는 군집보다 더 유사합니다. 자세한 내용은 "[SOM 알고리즘 설명](#)"에서 확인하십시오.

## 자기 조직화 지도 제어판

K 평균 군집화 플랫폼의 반복적 군집화 제어판에 있는 "방법" 목록에서 "자기 조직화 지도" 옵션을 선택합니다.

그림 14.8 제어판의 자기 조직화 지도 옵션



제어판의 일부 옵션은 "[반복적 군집화 보고서](#)"에 설명되어 있습니다. 나머지 옵션은 아래에서 설명합니다.

**행 수** 군집 격자의 행 수입니다.

**열 수** 군집 격자의 열 수입니다.

**대역폭** 중심 예측을 위한 인접 군집의 효과를 지정합니다. 대역폭이 작을수록 더 가까운 군집에 가중치가 더 많이 지정됩니다.

**고급 컨트롤** 다음과 같은 고급 컨트롤을 사용할 수 있습니다.

**최대 반복 수** EM 알고리즘의 수렴 단계에 대한 최대 반복 수입니다.

## 자기 조직화 지도 보고서

K 평균 군집화 플랫폼의 SOM 보고서는 요청된 격자 크기에 따라 이름이 지정됩니다. "대역폭"은 "SOM 격자" 보고서 맨 위에 제공됩니다. 보고서 자체는 "K 평균 군집 수" 보고서와 유사합니다. 자세한 내용은 "[K 평균 군집 수 보고서](#)"에서 확인하십시오. "군집 비교" 보고서에는 총 군집 수와 요청된 행 수가 표시됩니다.

"자기 조직화 지도"의 빨간색 삼각형 옵션에 대한 자세한 내용은 "[K 평균 군집 수 보고서 옵션](#)"에서 확인하십시오.

그림 14.9 자기 조직화 지도 보고서

**군집 비교**

방법	군집 수	행 수	CCC	최선
자기 조직화 지도	6	2	2.87879	최적 CCC

개별적으로 적도화된 열

**SOM 격자 2 X 3**

개별적으로 적도화된 열

대역폭: 0.4330127

**군집 요약**

군집	개수	단계	기준
1	28	12	0
2	20		
3	31		
4	21		
5	28		
6	22		

**군집 평균**

군집	Sepal length	Sepal width	Petal length	Petal width
1	6.98187076	3.15339949	5.85011658	2.16765916
2	6.14554483	2.62528195	5.11988902	1.79042135
3	6.20930867	3.05499872	4.51264582	1.46300228
4	5.46721399	2.56978376	3.79572186	1.15921467
5	5.22582725	3.67826605	1.47738883	0.27226749
6	4.7258207	3.10958329	1.44133246	0.21218254

**군집 표준편차**

## SOM 알고리즘 설명

이 섹션에는 K 평균 군집화 플랫폼의 SOM 구현 단계가 포함되어 있습니다.

- 다차원 공간에 대한 적절한 범위를 제공하는 방식으로 초기 군집 시드가 선택됩니다. JMP에서 주성분을 사용하여 데이터의 가장 큰 변동을 포착하는 두 방향을 결정합니다.

- 그런 다음 JMP 가 각 방향의 중간에서 2.5 표준편차를 갖는 이 주성분 공간에 격자를 배치합니다. 이 격자를 변수의 원래 공간으로 다시 변환하여 군집 시드가 결정됩니다.
- k 평균과 마찬가지로 군집 할당이 진행됩니다. 각 점이 가장 가까운 위치의 군집에 할당됩니다.
- k 평균과 마찬가지로 각 군집에 대해 평균이 추정됩니다. 그런 다음 JMP 에서 이러한 평균을 사용하여 가중 회귀를 설정합니다. 이때 각 변수를 회귀의 반응으로 지정하고 SOM 격자 좌표를 회귀변수로 지정합니다. 가중치 함수는 중심이 추정되는 군집에 큰 가중치를 제공하는 커널 함수를 사용합니다. SOM 격자의 군집에서 더 멀리 있는 군집에 더 작은 가중치가 제공됩니다. 새 군집 평균은 이 회귀 분석의 예측값입니다.
- 과정이 수렴될 때까지 이러한 반복이 계속됩니다.

## K 평균 군집화의 추가 예

이 예에서는 SOM( 자기 조직화 지도 ) 을 사용하여 붓꽃 관측값을 올바른 종으로 군집화합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Iris.jmp 를 엽니다.
2. **분석 > 군집화 > K 평균 군집화**를 선택합니다.
3. Sepal length, Sepal width, Petal length 및 Petal width 를 선택하고 **Y, 열**을 클릭합니다.
4. **확인**을 클릭합니다.
5. " 제어판 " 의 " 방법 " 메뉴에서 **자기 조직화 지도**를 선택합니다.
6. **행 수**를 "1" 로 설정하고 **열 수**를 "2" 로 설정합니다.
7. **시작**을 클릭합니다.
8. " 제어판 " 보고서를 엽니다.
9. **행 수**를 "1" 로 설정하고 **열 수**를 "3" 으로 설정합니다.
10. **시작**을 클릭합니다.
11. " 제어판 " 보고서를 엽니다.
12. **행 수**를 "2" 로 설정하고 **열 수**를 "2" 로 설정합니다.
13. **시작**을 클릭합니다.

그림 14.10 SOM 군집 비교

4 군집 비교			
방법	군집 수	행 수	CCC 최선
자기 조직화 지도	2	1	3.35952
자기 조직화 지도	3	1	4.95837 최적 CCC
자기 조직화 지도	4	2	3.64938

" 군집 비교 " 보고서가 보고서 창 맨 위에 나타납니다. 최량 적합한 가장 큰 CCC 값에 의해 결정됩니다. 가장 큰 CCC 를 제공하는 군집 수는 종의 수와 같은 3 입니다.

14. "SOM 격자 1 X 3" 보고서가 나올 때까지 스크롤합니다. 분류가 완벽하지 않습니다. 각 군집은 각각 50 개의 행이 있는 각 종을 나타내야 합니다.

그림 14.11 Iris.jmp 에 대한 SOM 보고서

SOM 격자 1 대 3				
개별적으로 적도화된 열				
대역폭: 0.4330127				
군집 요약				
군집	개수	단계	기준	
1	45	10	0	
2	55			
3	50			
군집 평균				
군집	Sepal length	Sepal width	Petal length	Petal width
1	6.81555029	3.09552593	5.54004585	1.98001626
2	5.81996034	2.7503061	4.29253797	1.39367974
3	5.00599574	3.427976	1.46203714	0.24601317

15. 데이터 테이블에서 Species 열을 선택하고 **행 > 열 값에 따른 색상 또는 표식**을 선택합니다.
16. "표식"에서 **클래식** 옵션을 선택합니다.
17. **확인**을 클릭합니다.
18. "SOM 격자 1 X 3" 옆의 빨간색 삼각형 메뉴를 클릭하고 **행렬도**를 선택합니다.
19. "SOM 격자 1 X 3" 옆의 빨간색 삼각형 메뉴를 클릭하고 **행렬도 옵션 > 행렬도 선 표시**를 선택합니다.

그림 14.12 SOM 행렬도

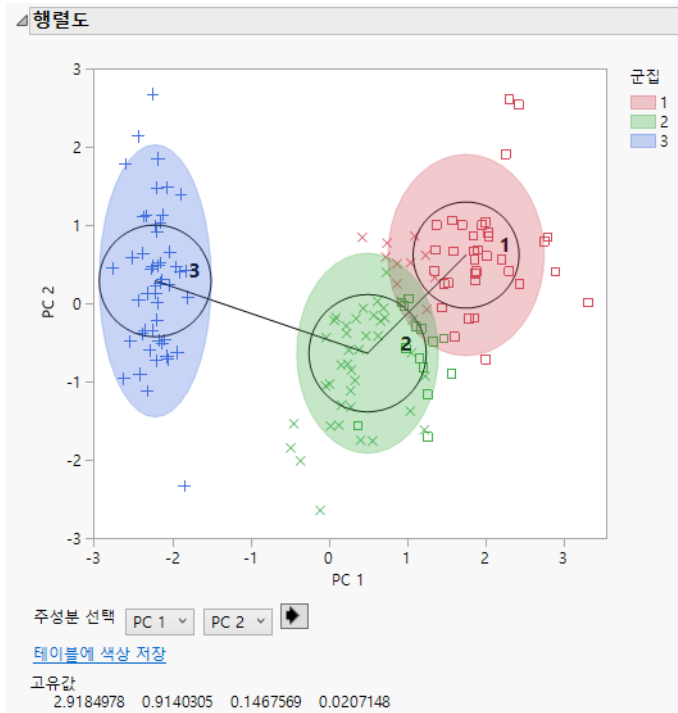


그림 14.12 에서는 군집 3 의 모든 행이 setosa 종으로 올바르게 식별되었음을 보여 줍니다 . 다른 두 종 , virginica 와 versicolor 는 약간 중첩되어 서로 잘못 볼 수 있습니다 .



# 15 장

## 정규 혼합

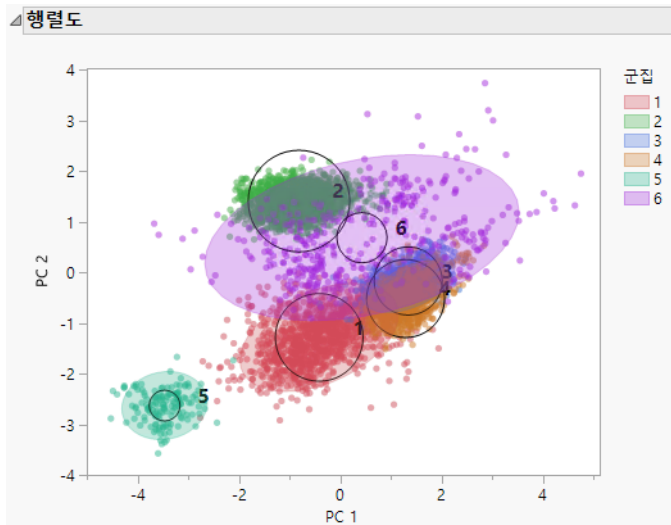
### 확률을 사용하여 관측값 그룹화

정규 혼합은 다변량 정규 분포의 혼합을 사용하여 관측값의 결합 확률 분포가 근사화된다는 가정을 기반으로 하는 반복 기법입니다. 이러한 혼합은 서로 다른 군집을 나타냅니다. 개별 군집은 다변량 정규 분포를 따릅니다.

군집이 잘 분리되면 계층적 군집화 및  $k$  평균 군집화가 제대로 작동합니다. 그러나 군집이 중첩된 경우에는 정규 혼합이 더 나은 방법입니다. 이 방법은 경계에 기반한 임의의 군집 할당 대신 소속 확률을 기반으로 하기 때문입니다.

중첩된 정규 분포의 데이터를 가져오는 경우 정규 혼합을 사용하여 군집화합니다. 군집 수를 미리 지정해야 합니다.

그림 15.1 정규 혼합 행렬도



**목차**

정규 혼합 플랫폼 개요.....	333
관측값 군집화 플랫폼 개요.....	333
정규 혼합 군집화의 예.....	334
정규 혼합 플랫폼 시작.....	336
정규 혼합 보고서.....	337
정규 혼합 옵션.....	338
개별 정규 혼합 보고서.....	338
군집 비교 보고서.....	339
정규 혼합 군집 수 보고서.....	339
정규 혼합 플랫폼에 대한 통계 상세 정보.....	341

## 정규 혼합 플랫폼 개요

정규 혼합은 수치 변수를 위한 반복 군집화 기법입니다. 그러나 각 군집 내에서 기대 반응의 비율도 예측합니다. 정규 혼합은 서로 다른 군집을 나타내는 다변량 정규 분포의 혼합을 사용하여 측정값 열의 결합 확률 분포를 근사화할 수 있다고 가정합니다. 각 군집에 대해 평균 벡터와 공분산 행렬이 추정됩니다. 자세한 내용은 McLachlan과 Krishnan(1997) 연구 자료 및 Hand et al. (2001)의 Section 9.6에서 확인하십시오.

**참고:** 정규 혼합 알고리즘에는 군집 중심에 대한 랜덤 추측을 반복하는 작업이 포함됩니다. 이 때문에 분석을 실행할 때마다 결과가 약간 다를 수 있습니다.

다변량 이상치가 있다고 의심되는 경우 두 가지 옵션이 있습니다. 이상치 군집 또는 이상치 탐색 유틸리티를 사용할 수 있습니다. 이상치 군집 옵션은 균등 분포를 가정하고 표준 정규 혼합 방법보다 이상치에 덜 민감합니다. 이상치 탐색 유틸리티를 사용하면 분석 전에 이상치를 탐색하고 처리할 수 있습니다. 자세한 내용은 "[이상치 군집](#)" 및 [예측 및 전문 모델링](#)에서 확인하십시오.

정규 혼합은 JMP에서 관측값을 군집화하기 위해 제공하는 네 가지 플랫폼 중 하나입니다. 네 가지 방법을 모두 비교하려면 "[관측값 군집화 플랫폼 개요](#)"의 내용을 참조하십시오.

## 관측값 군집화 플랫폼 개요

군집화는 여러 변수에 대해 유사한 값을 공유하는 관측값을 함께 그룹화하는 다변량 기법입니다. 일반적으로 관측값은  $p$  차원 공간에 고르게 분산되지 않습니다. 여기서  $p$ 는 변수 수입니다. 대신 관측값은 군집을 형성합니다. 이러한 군집을 식별하면 데이터를 더 세밀하게 파악할 수 있습니다.

**참고:** JMP에서는 변수를 군집화할 수 있는 플랫폼도 제공합니다. 자세한 내용은 "[변수 군집화](#)"에서 확인하십시오.

JMP에서 제공하는 다음 네 가지 플랫폼을 사용하여 관측값을 군집화할 수 있습니다.

- 계층적 군집화는 작은 데이터 테이블과 큰 데이터 테이블에 모두 유용하며 문자 데이터를 허용합니다. 계층적 군집화는 행을 계층적 순서로 결합하여 트리로 표현합니다. 트리가 생성되면 데이터에 가장 적절한 군집 수를 선택할 수 있습니다. 자세한 내용은 "[계층적 군집화](#)"에서 확인하십시오.
- K 평균 군집화는 최대 수백만 개의 행이 있는 큰 테이블에 적합하며 숫자 데이터만 허용합니다. 군집 수  $k$ 를 미리 지정해야 합니다. 이 알고리즘은 군집 시드 점을 추측합니다. 그런 다음 군집에 점을 번갈아 할당하고 군집 중심을 다시 계산하는 반복 과정을 수행합니다. 자세한 내용은 "[K 평균 군집화](#)"에서 확인하십시오.
- 정규 혼합은 중첩될 수 있는 다변량 정규 분포의 혼합에서 데이터를 수집한 경우에 적합하며 숫자 데이터만 허용합니다. 다변량 이상치가 있는 경우 균등 분포를 가정하고 이상치 군집을 사용할 수 있습니다. 자세한 내용은 "[정규 혼합](#)"에서 확인하십시오.

군집 수를 미리 지정해야 합니다. 최대 가능도는 혼합 비율과 평균, 표준편차 및 상관을 공동으로 추정하는 데 사용됩니다. 각 그룹에 속할 확률이 각 점에 할당됩니다. EM 알고리즘은 추정값을 구하는 데 사용됩니다.

- 잠재 계층 분석은 대부분의 변수가 범주형일 때 적절합니다. 군집 수를 미리 지정해야 합니다. 이 알고리즘은 다항 혼합 분포를 가정하는 모형을 적합시킵니다. 각 관측값에 대해 소속 군집의 최대 가능도 추정값이 계산됩니다. 관측값은 소속 확률이 가장 큰 군집으로 분류됩니다. 자세한 내용은 "잠재 계층 분석"에서 확인하십시오.

표 15.1 군집화 방법 요약

방법	데이터 유형 또는 모델링 유형	데이터 테이블 크기	군집 수 지정
계층적 군집화	임의	하이브리드 Ward 를 사용하는 경우 최대 수십만 개의 행  빠른 Ward 를 사용하는 경우 최대 200,000 개 행  다른 방법을 사용하는 경우 최대 5,000 개 행	아니요
K 평균 군집화	수치형	최대 수백만 개 행	예
정규 혼합	수치형	모든 크기	예
잠재 계층 분석	명목형 또는 순서형	모든 크기	예

일부 군집화 플랫폼에는 데이터의 이상치를 처리하는 옵션이 있습니다. 그러나 데이터에 이상치가 있으면 분석하기 전에 먼저 이상치를 탐색하는 것이 가장 좋습니다. 이 작업은 이상치 탐색 유틸리티를 사용하여 수행할 수 있습니다. 자세한 내용은 예측 및 전문 모델링에서 확인하십시오.

## 정규 혼합 군집화의 예

이 예에서는 정규 혼합 플랫폼을 사용하여 세포분석에서 감지된 네 개의 표지 판독값을 기반으로 관측값을 군집화합니다. 세포분석법은 세포의 다양한 특성을 측정하는 데 사용됩니다. 세포 표지를 측정하면 특정 질병을 진단하는 데 도움이 됩니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Cytometry.jmp 를 엽니다.
2. **분석 > 군집화 > 정규 혼합**을 선택합니다.
3. CD3, CD8, CD4 및 MCB 를 선택하고 **Y, 열**을 클릭합니다.
4. **확인**을 클릭합니다.

5. 군집 수 옆에 "6" 을 입력합니다 .
6. 시작을 클릭합니다 .

---

**참고 :** 알고리즘에 랜덤 시작 값이 있으므로 결과가 다를 수 있습니다 .

---

**그림 15.2** 정규 혼합 군집 수 =6 보고서

정규 혼합

군집 비교

방법	군집 수	BIC	AICc	최선
정규 혼합	6	208033	207456	가장 작은 BIC 가장 작은 AICc

제어판

정규 혼합 군집 수 =6

군집 요약

군집	개수	비율
1	393	0.08550
2	944	0.18467
3	1194	0.24341
4	147	0.02932
5	720	0.14049
6	1602	0.31661

군집 평균

군집	CD3	CD8	CD4	MCB
1	336.456533	193.385204	238.384846	206.503738
2	233.766979	140.184238	298.679748	256.78221
3	173.951704	109.760064	187.942955	195.539918
4	87.8647731	86.2326743	107.791654	10.1473172
5	338.61353	86.6669975	315.560339	208.345296
6	317.251206	306.00611	150.866465	189.617978

군집 표준편차

-1*로그 가능도	BIC	AICc
103637.52	208033.06	207456.3

정규 혼합에 대한 상관

" 군집 요약 " 보고서에는 6 개 군집 각각의 관측값 수가 표시됩니다 . " 군집 평균 " 보고서에는 각 군집에 대한 표지 관측값의 평균이 표시됩니다 .

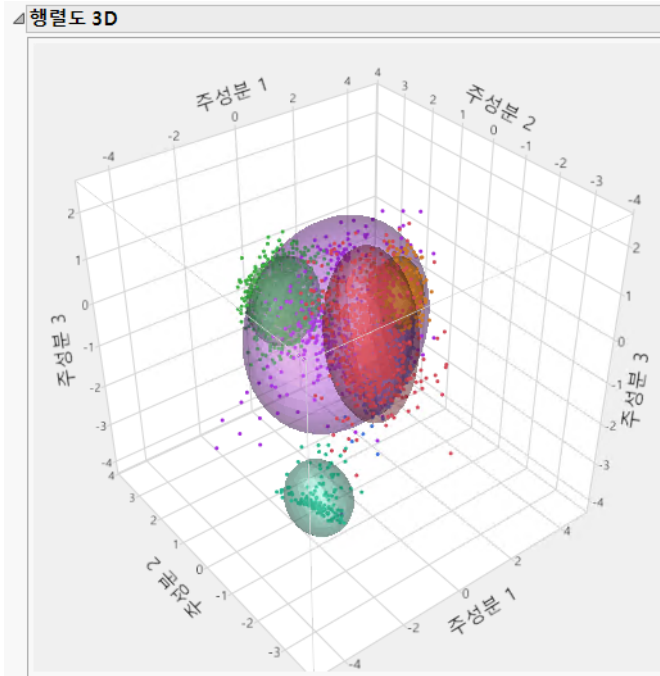
7. " 정규 혼합 군집 수 =6 " 옆의 빨간색 삼각형을 클릭하고 **행렬도 3D** 를 선택합니다 .

---

**참고 :** 알고리즘에 랜덤 시작 값이 있으므로 3D 행렬도가 다르게 나타날 수 있습니다 .

---

그림 15.3 Cytometry 데이터에 대한 3D 행렬도



이 그림에는 군집에 적합한 정규 밀도의 등고선이 표시됩니다. 한 군집이 처음 세 개의 주성분을 기반으로 다른 군집과 뚜렷하게 분리된 것 같습니다.

## 정규 혼합 플랫폼 시작

분석 > 군집화 > 정규 혼합을 선택하여 정규 혼합 플랫폼을 시작합니다.

그림 15.4 정규 혼합 시작 창

정규 혼합을 사용하여 수치형 변수 기반의 행을 지정된 수의 군집으로 군집화합니다.

열 선택	선택한 열 역할 지정	작업													
<ul style="list-style-type: none"> <li>▼ 8개 열</li> <li>▲ ForSc</li> <li>▲ SideSc</li> <li>▲ CD3</li> <li>▲ CD8</li> <li>▲ CD4</li> <li>▲ MCB</li> <li>▲ Prin1</li> <li>▲ Prin2</li> </ul>	<table border="1"> <tr> <td>Y, 열</td> <td>필수 숫자 선택적 숫자</td> </tr> <tr> <td>가중치</td> <td>선택적 숫자</td> </tr> <tr> <td>빈도</td> <td>선택적 숫자</td> </tr> <tr> <td>기준</td> <td>선택적</td> </tr> </table>	Y, 열	필수 숫자 선택적 숫자	가중치	선택적 숫자	빈도	선택적 숫자	기준	선택적	<table border="1"> <tr> <td>확인</td> </tr> <tr> <td>취소</td> </tr> <tr> <td>제거</td> </tr> <tr> <td>재호출</td> </tr> <tr> <td>도움말</td> </tr> </table>	확인	취소	제거	재호출	도움말
Y, 열	필수 숫자 선택적 숫자														
가중치	선택적 숫자														
빈도	선택적 숫자														
기준	선택적														
확인															
취소															
제거															
재호출															
도움말															

"열 선택"의 빨간색 삼각형 메뉴에 포함된 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

**Y, 열** 관측값을 군집화하는 데 사용되는 변수입니다.

**참고 :** 정규 혼합 군집화는 숫자 열만 지원합니다.

**가중치** 분석의 각 행에 가중치를 할당하는 숫자 값이 포함된 열입니다.

**빈도** 분석의 각 행에 빈도를 할당하는 숫자 값이 포함된 열입니다.

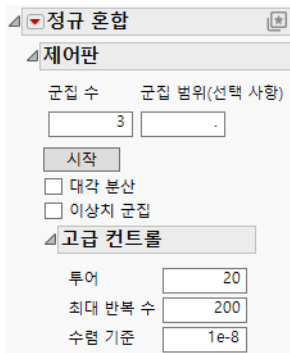
**기준** 각 수준으로 개별 분석을 정의하는 열입니다. 지정된 열의 각 수준에 대해 해당하는 행이 분석됩니다. 결과는 개별 보고서에 제공됩니다. 기준 변수가 둘 이상 할당되면 기준 변수의 가능한 각 수준 조합에 대해 개별 분석이 생성됩니다.

"확인"을 클릭하면 "제어판"이 나타납니다. 자세한 내용은 "[개별 정규 혼합 보고서](#)"에서 확인하십시오.

## 정규 혼합 보고서

"정규 혼합" 시작 창에서 "확인"을 클릭하면 "정규 혼합" 보고서 창에 모형 적합을 위한 "제어판"이 표시됩니다. "제어판"을 반복적으로 사용하여 다양한 수의 군집을 적합시키거나, "군집 범위" 옵션을 사용하여 범위를 지정할 수 있습니다. 모형을 적합시키면 해당 보고서가 창에 추가됩니다. 자세한 내용은 "[개별 정규 혼합 보고서](#)"에서 확인하십시오.

**그림 15.5** 정규 혼합에 대한 제어판



정규 혼합 제어판에는 다음과 같은 옵션이 있습니다.

**군집 수** 형성할 군집 수를 지정합니다.

**군집 범위 (선택 사항)** 형성할 군집 수에 대한 상한을 제공합니다. 여기에 숫자를 입력하면 플랫폼에서 **군집 수**와 **군집 범위 (선택 사항)**로 입력한 값 사이의 모든 정수에 대해 개별 분석을 생성합니다.

**시작** 군집을 적합시킵니다.

**대각 분산** 공분산 행렬의 비대각 요소를 0으로 제한합니다. 플랫폼에서는 변수 간에 상관관계가 없는 다변량 정규 분포를 적합시킵니다.

**참고 :** 변수보다 관측값 수가 적은 경우 특이 공분산 행렬이 생성되는 것을 방지하기 위해 "대각 분산" 옵션이 필요할 때가 있습니다. 또한 많은 수의 변수에 대해 매우 큰 공분산 행렬을 추정하는 것을 방지하는 데 사용할 수도 있습니다.

**이상치 군집** 정규 군집에 속하지 않는 이상치를 포착하기 위해 군집을 적합시킵니다. 이 군집이 생성되면 군집 0으로 지정되고 관측값 수가 "군집 요약" 보고서에 나타납니다. 이상치 군집에 속하는 관측값의 분포는 관측값을 포함하는 초입방체에 대해 균등하다고 가정합니다.

**고급 컨트롤** 다음과 같은 고급 컨트롤을 사용할 수 있습니다.

**투어** 추정 프로세스의 독립적 재시작 횟수입니다. 다시 시작할 때마다 시작 값이 다릅니다. 독립적으로 시작하면 국소 해를 찾지 않도록 할 수 있습니다.

**최대 반복 수** EM 알고리즘의 수렴 단계에 대한 최대 반복 수입니다.

**수렴 기준** EM 반복이 종료되는 가능성도 차이입니다.

## 정규 혼합 옵션

이 섹션에서는 "정규 혼합"의 빨간색 삼각형 메뉴에 포함된 옵션을 다룹니다.

다음 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

**로컬 데이터 필터** 특정 보고서에서 사용되는 데이터를 필터링할 수 있는 로컬 데이터 필터를 표시하거나 숨깁니다.

**다시 실행** 분석을 반복하거나 다시 시작할 수 있는 옵션이 포함되어 있습니다. 이 기능을 지원하는 플랫폼에서 "자동 재계산" 옵션은 해당하는 보고서 창에서 데이터 테이블에 대한 변경 사항을 즉시 반영합니다.

**플랫폼 환경 설정** 현재 플랫폼 환경 설정을 보거나, 현재 JMP 보고서의 설정과 일치하도록 플랫폼 환경 설정을 업데이트할 수 있는 옵션이 포함되어 있습니다.

**스크립트 저장** 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다.

**참고 :** 이 플랫폼의 추가 옵션은 스크립트를 통해 사용할 수 있습니다. 자세한 내용은 "도움말" 메뉴의 "스크립트 인덱스"에서 여십시오. 또한 "스크립트 인덱스"에서 이 섹션에 설명된 옵션의 스크립트 예제도 찾을 수 있습니다.

## 개별 정규 혼합 보고서

정규 혼합의 "제어판"에서 "시작"을 클릭하면 "군집 비교" 보고서와 하나 이상의 개별 "정규 혼합" 보고서가 나타납니다. 보고서 이름은  $k$ (적합된 군집 수)에 따라 동적으로 "정규 혼합 군집 수= $k$ "로 지정됩니다. 수행한 각 적합에 대한 "정규 혼합 군집 수= $k$ " 보고서가 나타납니다.

- "군집 비교 보고서"

- "정규 혼합 군집 수 보고서"

## 군집 비교 보고서

정규 혼합 플랫폼의 "군집 비교" 보고서에서는 다양한 모형을 비교하기 위한 적합 통계량을 제공합니다. 적합 통계량은 BIC와 AICc입니다. 각 값이 작을수록 더 나은 적합을 나타냅니다. 최량 적합은 "최선" 열에 표시됩니다.

## 정규 혼합 군집 수 보고서

각 "정규 혼합 군집 수" 보고서에서는 각 군집에 대해 다음과 같은 요약 통계량을 제공합니다.

- "군집 요약" 보고서에는 각 군집에 대한 관측값 수와 비율이 제공됩니다.
- "군집 평균" 보고서에는 각 군집에서 각 변수에 대한 관측값의 평균이 제공됩니다.
- "군집 표준편차" 보고서에는 각 군집에서 각 변수에 대한 관측값의 표준편차가 제공됩니다.
- "-1\* 로그 가능도" 테이블에는 음의 로그 가능도, BIC 및 AICc가 제공됩니다. 자세한 내용은 선형 모형 적합에서 확인하십시오.
- "정규 혼합에 대한 상관" 보고서에는 각 군집에 대한 추정 상관 행렬이 제공됩니다.

## 정규 혼합 군집 수 보고서 옵션

각 "정규 혼합 군집 수" 보고서에는 다음과 같은 빨간색 삼각형 메뉴 항목이 포함되어 있습니다.

**행렬도** 데이터의 처음 두 주성분으로 표현된 점 및 군집의 그림을 군집 색상을 식별하는 범례와 함께 표시합니다. 군집 중심을 기준으로 원이 그려지고 원 크기는 군집 내의 개수에 비례합니다. 음영 영역은 평균 주위의 밀도 등고선입니다. 기본적으로 이 영역은 해당 군집의 관측값 중 90%가 속하는 위치를 나타냅니다 (Mardia et al. 1980 참조). 그림 아래의 목록을 사용하여 그림 축을 다른 주성분으로 변경할 수 있습니다. 또는 화살표 버튼을 사용하여 가능한 모든 축 조합을 순환할 수 있습니다. 군집 색상을 데이터 테이블에 저장하는 옵션도 그림 아래에 있습니다. 자세한 내용은 "테이블에 색상 저장"에서 확인하십시오. 고유값은 내림차순으로 표시됩니다.

---

**참고:** 행렬도에서는 항상 상관 행렬을 사용하여 주성분을 계산합니다.

---

**행렬도 옵션** 행렬도 모양을 제어하기 위한 옵션을 포함합니다.

**행렬도 선 표시** 행렬도 선을 표시합니다. 라벨이 지정된 선은 주성분으로 정의된 부분 공간에서 공변량의 방향을 표시합니다. 이러한 선은 각 변수와 각 주성분의 연관성 정도를 나타냅니다.

**행렬도 선 위치** 행렬도 선의 위치와 반지름 척도를 지정할 수 있습니다. 기본적으로 선은 점 (0,0)에서 시작됩니다. 그림에서 선을 드래그하거나 이 옵션을 사용하여 좌표를 지정할 수 있습니다. 반지름 척도 옵션을 사용하여 선이 더 잘 보이게 선 척도를 조정할 수도 있습니다.

**행렬도 등고선 밀도** 밀도 등고선에 대한 신뢰 수준을 지정할 수 있습니다. 기본 신뢰 수준은 90% 입니다.

**군집 표시** 데이터 테이블의 행에 군집을 식별하는 표식을 할당합니다.

**행렬도 3D** 데이터의 3 차원 행렬도를 표시합니다. 변수가 세 개 이상인 경우에만 사용할 수 있습니다.

**평행 좌표 그림** 각 군집에 대한 평행 좌표 그림을 생성합니다. 이 그림 보고서에는 데이터와 평균을 표시하거나 숨기는 옵션이 있습니다. 자세한 내용은 **그래프 분석**에서 확인하십시오.

**산점도 행렬** 모든 Y 변수를 사용하여 산점도 행렬을 생성합니다.

**테이블에 색상 저장** 데이터 테이블의 행에 군집을 식별하는 색상을 할당합니다. 보고서 창에 "행렬도"가 있는 경우 데이터 테이블에 저장된 색상은 행렬도의 군집 색상과 일치합니다. "행렬도"에서 색상을 변경하고 "테이블에 색상 저장" 옵션을 다시 선택하면 테이블의 색상이 행렬도의 색상과 일치하도록 업데이트됩니다.

---

**참고:** 저장 옵션을 선택하면 저장된 각 열에 해당 특정 열의 데이터에 대한 군집 수를 지정하는 "노트" 열 특성이 포함됩니다. 이렇게 하면 둘 이상의 군집 적합에서 열을 저장하고 열 특성을 사용하여 저장된 열을 가져온 군집화 적합을 식별할 수 있습니다.

---

**군집 저장** 주어진 행이 할당된 군집의 번호를 포함하는 군집 열을 데이터 테이블에 저장합니다. 정규 혼합의 경우 이 군집은 최대 확률 분류 군집입니다.

**군집 계산식 저장** 군집 계산식이라는 계산식 열을 데이터 테이블에 저장합니다. 이 계산식은 행이 속한 군집을 식별합니다.

**군집 계산식 게시** "군집 계산식 저장" 옵션에 사용된 것과 동일한 스코어링 코드를 "계산식 저장소"에 게시합니다. "군집 계산식 게시"를 선택하고 "계산식 저장소" 내의 모형에서 "스크립트 실행"을 선택하는 경우, 데이터 테이블에 저장된 열은 "군집 계산식 저장"을 선택할 때 저장된 열과 일치해야 합니다.

**혼합 확률 저장** 각 군집에 대해 관측값이 해당 군집에 속할 확률을 포함하는 군집 확률 <k> 열을 추가합니다.

**혼합 계산식 저장** 혼합 확률을 계산하는 데 사용된 계산식을 포함하는 열을 데이터 테이블에 추가합니다. 이러한 계산식 열은 제외된 데이터 또는 테이블에 추가한 데이터의 확률을 스코어링하는 데 사용됩니다.

**거리 계산식 <k>** 관측값에서 계산된 군집 <k>에 대한 추정된 다변량 정규 밀도 함수입니다.

**총 거리** 거리 계산식 열의 합계입니다. 이 열의 계산식은 "밀도 계산식 저장" 옵션으로 생성된 혼합 밀도 열의 계산식과 동일합니다.

**확률 계산식 <k>** 관측값이 군집 <k>에 속할 확률입니다. 이러한 열에는 "혼합 확률 저장" 옵션으로 생성된 군집 확률 <k> 열의 값을 제공하는 계산식이 포함됩니다. 혼합 확률을 계산하기 위한 열 계산식은 다음과 같습니다.

$$\text{확률 계산식 } \langle k \rangle = \frac{\text{거리 계산식 } \langle k \rangle}{\text{총 거리}}$$

**밀도 계산식 저장** 정규 혼합에 대한 추정 밀도 함수를 포함하는 혼합 밀도 열을 데이터 테이블에 추가합니다.

**군집 시뮬레이션** 혼합 밀도를 사용하여 예측 변수 값을 시뮬레이션합니다. 이러한 값과 해당 값이 분류되는 군집을 새 데이터 테이블에 저장합니다.

**제거** 군집화 보고서를 제거합니다.

---

## 정규 혼합 플랫폼에 대한 통계 상세 정보

정규 혼합에서는 Newton-Raphson 알고리즘보다 안정적인 EM 알고리즘을 사용하여 적합을 수행합니다. 또한 JMP에서는 공분산 행렬이 특이 행렬인 경우 이를 원활하게 처리할 수 있는 베이지안 정규화 버전의 EM 알고리즘을 사용합니다. 추정값은 초기 추측에 크게 의존하므로 플랫폼에서 여러 번 투어를 반복합니다. 각 투어마다 초기 중심에 대해 랜덤 선택된 점이 있습니다.

투어를 여러 번 반복하면 추정 프로세스에 시간이 많이 소요되므로 큰 문제의 경우 상당히 오래 기다려야 합니다. 컨트롤을 사용하여 투어 및 반복 제한을 지정할 수 있습니다.



# 16 장

## 잠재 계층 분석 범주형 변수의 관측값 그룹화

잠재 계층 분석을 사용하면 범주형 반응 변수에 대한 관측값 군집을 찾을 수 있습니다. 잠재 변수는 관측 불가능한 그룹화 변수입니다. 잠재 변수의 각 수준을 잠재 계층이라고 합니다. 잠재 계층 분석 플랫폼에서는 잠재 계층 모형을 적합시키고 각 관측값에 대한 잠재 계층 또는 최대 확률 분류 군집을 결정합니다. 대부분의 상황에서 주제 관련 전문가가 잠재 계층 분석 결과를 사용하여 계층의 특성을 기반으로 각 잠재 계층에 대한 정의를 생성합니다.

그림 16.1 잠재 계층 분석의 예

모수 추정값														
군집	전체	sex		marital status		country			size			type		
		Female	Male	Married	Single	American	European	Japanese	Large	Medium	Small	Family	Sporty	Work
군집 1	0.28553	0.3763	0.6237	0.4160	0.5840	0.1555	0.2306	0.6139	0.0009	0.2839	0.7152	0.0190	0.9635	0.0175
군집 2	0.25943	0.4066	0.5934	0.6740	0.3260	0.0271	0.0562	0.9168	0.0029	0.2942	0.7029	0.6419	0.0368	0.3213
군집 3	0.20712	0.6793	0.3207	0.7526	0.2474	0.5618	0.1376	0.3005	0.0025	0.9914	0.0062	0.8013	0.1950	0.0036
군집 4	0.19697	0.4707	0.5293	0.9922	0.0078	0.8532	0.0826	0.0642	0.4965	0.2047	0.2989	0.7651	0.0020	0.2329
군집 5	0.05095	0.1837	0.8163	0.0324	0.9676	0.8477	0.1468	0.0054	0.7766	0.1180	0.1054	0.4366	0.0957	0.4677
군집	전체	sex		marital status		country			size			type		
군집 1	0.28553													
군집 2	0.25943													
군집 3	0.20712													
군집 4	0.19697													
군집 5	0.05095													

## 목차

잠재 계층 분석 플랫폼 개요 .....	345
관측값 군집화 플랫폼 개요 .....	345
잠재 계층 분석의 예 .....	346
잠재 계층 분석 플랫폼 시작 .....	349
잠재 계층 분석 보고서 .....	350
군집 비교 보고서 .....	350
잠재 계층 모형 보고서 .....	350
잠재 계층 분석 플랫폼 옵션 .....	352
잠재 계층 모형 옵션 .....	353
잠재 계층 분석 플랫폼의 추가 예 .....	354
잠재 계층 분석 플랫폼에 대한 통계 상세 정보 .....	355
잠재 계층 모형 적합에 대한 통계 상세 정보 .....	355
최대 군집 수에 대한 통계 상세 정보 .....	356

## 잠재 계층 분석 플랫폼 개요

잠재 계층 분석 플랫폼에서는 범주형 반응 변수에 잠재 계층 모형을 적합시키고 각 관측값에 대한 잠재 계층 또는 최대 확률 분류 군집을 결정합니다. 잠재 변수는 관측 불가능한 그룹화 변수입니다. 잠재 변수의 각 수준을 잠재 계층이라고 합니다. 예를 들어 위험 우선 순위에 따라 그룹화된 설문 조사 응답자 군집이 잠재 계층일 수 있습니다.

모형은 다항 혼합 모형의 형태를 취합니다. 모형에는  $\gamma$  모수와  $\rho$  모수의 두 가지 모수 집합이 있습니다.  $\gamma$  모수는 소속 군집의 전체 확률을 나타냅니다.  $\rho$  모수는 소속 군집을 조건으로 주어진 반응을 관측할 확률을 나타냅니다. 잠재 계층은 이러한 조건부 확률의 패턴을 특징으로 합니다.

분석 결과가 유의하려면 주제 관련 전문가가 플랫폼에서 생성되는 군집을 해석해야 합니다. 이 전문가는 잠재 계층의 특성을 조사하고 해당 특성을 기반으로 각 계층에 대한 정의를 생성합니다.

**참고:** 반응 열에 결측값이 있는 행은 분석에서 제외됩니다.

잠재 계층 모형에 대한 자세한 내용은 Collins와 Lanza(2010) 및 Goodman(1974) 연구 자료에서 확인하십시오.

잠재 계층 분석은 JMP에서 관측값을 군집화하기 위해 제공하는 네 가지 플랫폼 중 하나입니다. 네 가지 방법을 모두 비교하려면 "[관측값 군집화 플랫폼 개요](#)"의 내용을 참조하십시오.

## 관측값 군집화 플랫폼 개요

군집화는 여러 변수에 대해 유사한 값을 공유하는 관측값을 함께 그룹화하는 다변량 기법입니다. 일반적으로 관측값은  $p$  차원 공간에 고르게 분산되지 않습니다. 여기서  $p$ 는 변수 수입니다. 대신 관측값은 군집을 형성합니다. 이러한 군집을 식별하면 데이터를 더 세밀하게 파악할 수 있습니다.

**참고:** JMP에서는 변수를 군집화할 수 있는 플랫폼도 제공합니다. 자세한 내용은 "[변수 군집화](#)"에서 확인하십시오.

JMP에서 제공하는 다음 네 가지 플랫폼을 사용하여 관측값을 군집화할 수 있습니다.

- 계층적 군집화는 작은 데이터 테이블과 큰 데이터 테이블에 모두 유용하며 문자 데이터를 허용합니다. 계층적 군집화는 행을 계층적 순서로 결합하여 트리로 표현합니다. 트리가 생성되면 데이터에 가장 적절한 군집 수를 선택할 수 있습니다. 자세한 내용은 "[계층적 군집화](#)"에서 확인하십시오.
- K 평균 군집화는 최대 수백만 개의 행이 있는 큰 테이블에 적절하며 숫자 데이터만 허용합니다. 군집 수  $k$ 를 미리 지정해야 합니다. 이 알고리즘은 군집 시드 점을 추측합니다. 그런 다음 군집에 점을 번갈아 할당하고 군집 중심을 다시 계산하는 반복 과정을 수행합니다. 자세한 내용은 "[K 평균 군집화](#)"에서 확인하십시오.

- 정규 혼합은 중첩될 수 있는 다변량 정규 분포의 혼합에서 데이터를 수집한 경우에 적절하며 숫자 데이터만 허용합니다. 다변량 이상치가 있는 경우 균등 분포를 가정하고 이상치 군집을 사용할 수 있습니다. 자세한 내용은 "정규 혼합"에서 확인하십시오.  
군집 수를 미리 지정해야 합니다. 최대 가능도는 혼합 비율과 평균, 표준편차 및 상관을 공동으로 추정하는 데 사용됩니다. 각 그룹에 속할 확률이 각 점에 할당됩니다. EM 알고리즘은 추정값을 구하는 데 사용됩니다.
- 잠재 계층 분석은 대부분의 변수가 범주형일 때 적절합니다. 군집 수를 미리 지정해야 합니다. 이 알고리즘은 다항 혼합 분포를 가정하는 모형을 적합시킵니다. 각 관측값에 대해 소속 군집의 최대 가능도 추정값이 계산됩니다. 관측값은 소속 확률이 가장 큰 군집으로 분류됩니다. 자세한 내용은 "잠재 계층 분석"에서 확인하십시오.

표 16.1 군집화 방법 요약

방법	데이터 유형 또는 모델 링 유형	데이터 테이블 크기	군집 수 지정
계층적 군집화	임의	하이브리드 Ward 를 사용하는 경우 최대 수십만 개의 행  빠른 Ward 를 사용하는 경우 최대 200,000개 행  다른 방법을 사용하는 경우 최대 5,000개 행	아니요
K 평균 군집화	수치형	최대 수백만 개 행	예
정규 혼합	수치형	모든 크기	예
잠재 계층 분석	명목형 또는 순서형	모든 크기	예

일부 군집화 플랫폼에는 데이터의 이상치를 처리하는 옵션이 있습니다. 그러나 데이터에 이상치가 있으면 분석하기 전에 먼저 이상치를 탐색하는 것이 가장 좋습니다. 이 작업은 이상치 탐색 유틸리티를 사용하여 수행할 수 있습니다. 자세한 내용은 예측 및 전문 모델링에서 확인하십시오.

## 잠재 계층 분석의 예

이 예에서는 잠재 계층 모형을 적합시켜 12개 설문 조사 질문에 대한 학생들의 응답을 기반으로 학생 군집을 식별합니다. 응답은 2005년 미국 고등학생을 대상으로 한 설문 조사에서 나왔습니다. 설문 조사에서 학생들에게 건강 위험 행동과 관련된 다양한 선다형 질문을 했습니다. 분석 대상 열은 선다형 설문 조사 질문에서 응답을 두 개의 계층(Yes/No)으로 분류하여 얻었습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Health Risk Survey.jmp 를 엽니다.

- Health Risk Survey 데이터 테이블에서 Launch LCA Platform 스크립트 옆의 녹색 삼각형을 클릭합니다.

이 스크립트는 12 개의 관심 대상 열을 선택하고 "잠재 계층 분석" 시작 창을 연 후 12 개 열을 Y 로 입력합니다.

---

**참고 :** LCA 플랫폼을 직접 시작하려면 분석 > 군집화 > 잠재 계층 분석을 선택합니다.

---

- 최대** 옆의 상자에 "5" 를 입력합니다.  
이 옵션은 3 개의 군집 및 최대 5 개의 군집에 대한 잠재 계층 모형을 적합시킵니다.
- 확인**을 클릭합니다.

**그림 16.2** 군집 비교 보고서

군집 수	-1*로그 가능성도	BIC	AIC 최선
3	38713	777762	775019
4	38207.1	768844	765163
5	379648	765196	760576

가장 작은 BIC 가장 작은 AIC

"잠재 계층 분석" 개요에는 "군집 비교" 보고서와 세 개의 개별 잠재 계층 모형 보고서가 포함됩니다. 잠재 계층 모형 보고서에는 3 개, 4 개, 5 개 군집에 대한 모형이 각각 표시됩니다. "군집 비교" 보고서를 보면 5 개 군집 모형의 BIC 및 AIC 가 가장 작으며 이는 세 모형 중 이 모형이 최량 적합 모형이라는 것을 나타냅니다. 즉, 이 모형이 분석 대상입니다.

- "5 개 군집의 잠재 계층 모형" 보고서에서 "모수 추정값" 아래의 막대 차트를 검토합니다. 다음 사항에 유의하십시오.
  - 군집 1 은 모든 위험 행동에 대한 대답이 대부분 "No" 입니다.
  - 군집 2 는 13 세 이전의 네 가지 위험 행동에 대해 "Yes" 대답이 많습니다.
  - 군집 3 은 음주 운전 및 지난 30 일 동안 5 회 이상 음주에 대해 "Yes" 대답이 많습니다.
  - 군집 4 는 13 세 이전의 행동을 제외한 대부분의 위험 행동에 대해 "Yes" 대답이 많습니다.
  - 군집 5 는 대부분의 위험 행동에 대해 "Yes" 대답이 가장 많습니다.
 이 정보를 사용하여 군집에 의미 있는 이름을 지정합니다.
- "5 개 군집의 잠재 계층 모형" 옆의 빨간색 삼각형을 클릭하고 **군집 이름 바꾸기**를 선택합니다.
  - 군집 1: Low Risk 입력
  - 군집 2: Early Risk-Takers 입력
  - 군집 3: Drinkers 입력
  - 군집 4: Late High Risk 입력
  - 군집 5: High Risk 입력
- 확인**을 클릭합니다.

---

**참고 :** 새 군집 이름은 스크립트에 저장되지 않습니다.

---

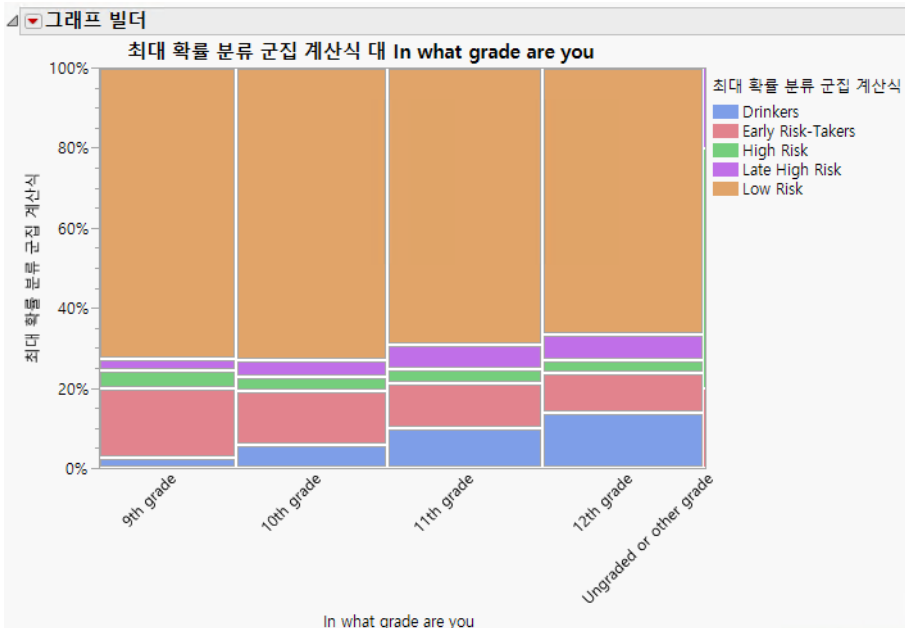
그림 16.3 모수 추정값 보고서 ( 일부 )

군집		Drove 1- times when drinking		Smoked cigarette before 13		Smoked daily for 30 days		Had first drink before 13		Five+ drinks 1- past 30 days		Tried marijuana before 13		Used cocaine 1- times in life		Sniffed glue 1- times in life	
	정제	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Low Risk	0.67525	0.9872	0.0128	0.9624	0.0376	0.9741	0.0259	0.8648	0.1352	0.8975	0.1025	0.9943	0.0057	0.9930	0.0070	0.9440	0.0
Early Risk-Takers	0.13852	0.8963	0.1037	0.4457	0.5543	0.7408	0.2592	0.3418	0.6582	0.5965	0.4035	0.6539	0.3461	0.9517	0.0483	0.8209	0.1
Drinkers	0.09626	0.4307	0.5693	0.8662	0.1338	0.7619	0.2381	0.7610	0.2390	0.0169	0.9832	0.9665	0.0335	0.9092	0.0908	0.8689	0.1
Late High Risk	0.05146	0.6848	0.3152	0.8223	0.1777	0.4921	0.5079	0.7703	0.2297	0.3494	0.6506	0.9114	0.0886	0.2685	0.7315	0.5432	0.4
High Risk	0.03851	0.4943	0.5057	0.1333	0.8667	0.2882	0.7118	0.1188	0.8812	0.1270	0.8730	0.2145	0.7855	0.1650	0.8350	0.4711	0.5

그림 16.3에서는 분석의 처음 8개 변수에 대한 모수 추정값을 보여 줍니다. 보고서 창에 새 군집 이름이 나타납니다.

- 이제 "In what grade are you" 라는 인구 통계 질문에 대해 소속 군집을 비교합니다.
- "5개 군집의 잠재 계층 모형" 옆의 빨간색 삼각형을 클릭하고 혼합 및 군집 계산식 저장을 선택합니다.
  - 그래프 > 그래프 빌더를 선택합니다.
  - In what grade are you 를 X로 입력합니다.
  - 최대 확률 분류 군집 계산식을 Y로 입력합니다.
  - 모자이크 요소를 선택합니다.
  - 완료를 클릭합니다.

그림 16.4 소속 군집 대 학년의 모자이크 그림

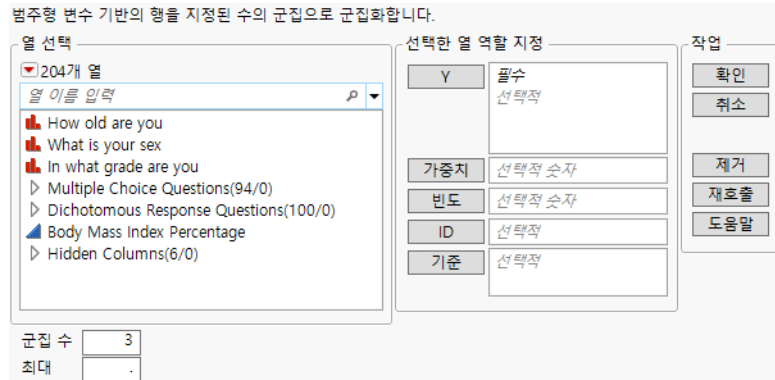


대부분의 응답자가 "Low Risk" 군집에 속한다는 것을 알 수 있습니다. "Drinkers" 라벨이 지정된 계층은 학년이 올라갈수록 더 많은 응답자를 포함합니다.

## 잠재 계층 분석 플랫폼 시작

분석 > 군집화 > 잠재 계층 분석을 선택하여 잠재 계층 분석 플랫폼을 시작합니다.

그림 16.5 잠재 계층 분석 시작 창



"별 선택"의 빨간색 삼각형 메뉴에 포함된 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

잠재 계층 분석 플랫폼 시작 창에는 다음 옵션이 포함되어 있습니다.

**Y** 분석할 하나 이상의 열입니다. 명목형, 순서형 또는 다중 반응 모델링 유형의 열을 분석할 수 있습니다. 명목형 또는 순서형 반응을 분석하려면 두 개 이상의 열이 필요합니다. 반응이 여러 개 포함되고 모델링 유형이 다중 반응인 경우에는 열이 하나만 필요합니다.

**가중치** 분석의 각 행에 가중치를 할당하는 숫자 값이 포함된 열입니다.

**빈도** 분석의 각 행에 빈도를 할당하는 숫자 값이 포함된 열입니다.

**ID** 개별 응답자를 식별하는 데 사용되는 열입니다. 이 ID는 일부 출력 테이블에 사용됩니다.

**기준** 변수의 각 수준에 대한 개별 분석으로 구성된 보고서를 생성하는 열입니다. 기준 변수가 둘 이상 할당되면 기준 변수의 가능한 각 수준 조합에 대해 개별 분석이 생성됩니다.

**군집 수** 분석에서 계산할 군집 수입니다.

**최대** 최대 군집 수를 지정합니다. 이 수가 "군집 수"에 지정된 값을 초과하면 "군집 수"와 "최대" 사이의 범위에 있는 각 정수 값과 동일한 군집 수로 모형 보고서가 생성됩니다. 이러한 보고서는 "잠재 계층 분석" 보고서 개요의 일부로 나타납니다.

**주의:** LCA 모형이 적절하게 적합시킬 수 있는 최대 군집 수가 있습니다. 최대값보다 많은 군집을 요청하면 보고서 창에 경고 메시지가 나타납니다. LCA 플랫폼에서는 열에 지원되는 최대 군집 수까지 적합시킵니다. 최대 군집 수를 결정하는 방법에 대한 자세한 내용은 "[최대 군집 수에 대한 통계 상세 정보](#)"에서 확인하십시오.

## 잠재 계층 분석 보고서

초기 "잠재 계층 분석" 보고서에는 "군집 비교" 보고서와 지정된 각 군집 수에 대한 잠재 계층 모형 보고서가 포함됩니다.

### 군집 비교 보고서

잠재 계층 분석 플랫폼의 "군집 비교" 보고서에는 다양한 모형을 비교하기 위한 적합 통계량이 표시됩니다. 적합 통계량은 음의 로그 가능도(-1\*로그 가능도), BIC 및 AIC입니다. 각 값이 작을수록 더 나은 적합을 나타냅니다. 최량 적합은 최선 열에 표시됩니다. 자세한 내용은 선형 모형 적합에서 확인하십시오.

### 잠재 계층 모형 보고서

각 잠재 계층 모형 보고서 이름은  $k$ (적합된 군집 수)에 따라 동적으로 "< $k$ > 군집의 잠재 계층 모형"으로 지정됩니다. 보고서에는 다음 결과와 개요가 포함됩니다.

- "모형 요약"
- "모수 추정값"
- "전치된 모수 추정값"
- "효과 크기"
- "MDS 그림"
- "혼합 확률"

#### 모형 요약

기본적으로 지정된 수의 군집에 대한 모형 요약이 각 잠재 계층 모형 보고서 맨 위에 나타납니다. 모형 요약에는 -1\*로그 가능도, 모수 수, BIC 및 AIC가 포함됩니다. 이러한 요약 값을 사용하여 모형이 데이터에 얼마나 잘 적합되는지 확인할 수 있습니다. -1\*로그 가능도, BIC 및 AIC 값이 작을수록 더 나은 적합을 나타냅니다. 자세한 내용은 선형 모형 적합에서 확인하십시오. "모수 수" 값은 잠재 계층 모형의 고유 모수 수를 제공합니다. 자세한 내용은 "[잠재 계층 분석 플랫폼에 대한 통계 상세 정보](#)"에서 확인하십시오.

### 모수 추정값

"모수 추정값" 보고서는 기본적으로 표시되며 테이블 형식과 그래픽 형식의 모수 추정값 요약어를 포함합니다. 각 요약에는 모형 군집에 해당하는 행이 포함됩니다.

"전체" 열에는 관측값이 각 군집에 속할 확률이 표시됩니다. 이는  $\gamma$  모수입니다. 자세한 내용은 "[잠재 계층 분석 플랫폼에 대한 통계 상세 정보](#)"에서 확인하십시오.

표시된 나머지 열은 "잠재 계층 분석" 시작 창에 지정된 Y 열에 따라 세로 구분자로 그룹화됩니다.

- 범주형 반응 열의 각 그룹에는 각 반응 내의 각 수준에 대한 열이 있습니다. 각 그룹에서 주어진 행과 열의 값은 행으로 식별된 군집에 관측값이 속할 경우 열이 나타내는 반응의 조건부 확률입니다. 이는  $\rho$  모수입니다.
- 다중 반응 열의 각 그룹에는 다중 반응 내의 각 범주에 대한 열이 있습니다. 각 그룹에서 주어진 행과 열의 값은 행으로 식별된 군집에 관측값이 속할 경우 지정된 범주의 하위 수준에서 반응의 조건부 확률입니다. 이는  $\rho$  모수입니다.

그래픽 보기에는 조건부 확률 값이 **점유율 차트**로 표시됩니다. 각 군집 및 각 Y에 대해 주어진 소속 군집의 조건부 확률이 가로 누적 막대 차트로 표시됩니다. 이항 또는 명목형 반응 열의 경우 이러한 차트의 백분율 합계는 각 반응에 대해 1입니다. 다중 반응 열의 경우 각 범주의 하위 수준에 대한 백분율이며 합계가 1이 아닙니다. 막대 누적은 값 테이블에 표시되는 변수의 순서를 따릅니다. 막대를 커서로 가리켜 변수의 수준 또는 범주를 볼 수도 있습니다.

---

**팁:** "모수 추정값" 보고서의 한 테이블에서 하나 이상의 행을 선택하여 해당 군집에 할당된 관측값을 선택할 수 있습니다.

---

### 전치된 모수 추정값

"전치된 모수 추정값" 보고서에는 "모수 추정값" 보고서 테이블의 전치 테이블이 포함됩니다. 여기서는 군집이 열로 표시됩니다. 분석에 포함된 각 Y 열의 각 반응 범주에 대해 각 군집의 조건부 확률이 표시됩니다.

---

**참고:** "전체" 열의 추정값은 전치된 테이블에 포함되지 않습니다.

---

### 효과 크기

"효과 크기" 테이블은 기본적으로 표시되며 군집에 대해 Y 열을 비교합니다. 이 테이블의 각 행에 있는 통계량은 Y 열의 수준 또는 범주별로 소속 군집에 대한 기대 개수의 분할표 분석에서 얻습니다. 기대 개수는 각 군집의 관측값 수에 Y 열의 각 수준 또는 범주에 대한 조건부 확률을 곱한 값입니다.

각 반응에서 군집별 수준의 기대 개수 분할표에 대해 Pearson 카이제곱 통계량  $\chi^2$  이 계산됩니다.  $n$  이 관측값 수를 나타낸다고 하겠습니다. "효과 크기" 열의 값은 다음과 같이 정의됩니다.

$$\text{효과 크기} = \sqrt{\frac{\chi^2}{n}}$$

"LR Logworth" 열의 각 값은  $-\log_{10}(p_{LR})$  을 나타내며, 여기서  $p_{LR}$  은 기대 개수의 분할표에 대한 가능도비 검정  $p$  값입니다. 2를 초과하는 Logworth 값은 0.01 유의 수준의 유의성에 해당합니다.

**팁 :** "효과 크기" 테이블에서 하나 이상의 행을 선택하여 데이터 테이블에서 해당 열을 선택할 수 있습니다.

### MDS 그림

"MDS 그림"은 기본적으로 표시되며 각 군집마다 하나의 점을 포함합니다. 이 그림은 군집 근접성을 2차원으로 표현한 것입니다. 서로 더 가까운 군집이 더 유사합니다. 이 그림은  $\rho$  모수의 비 유사성 행렬에서 생성됩니다. MDS 그림에 대한 자세한 내용은 "[다차원 척도법](#)"에서 확인하십시오.

### 혼합 확률

"혼합 확률" 테이블에는 각 행의 소속 군집 확률이 표시됩니다. "최대 확률 분류 군집" 열은 각 행의 소속 확률이 가장 높은 군집을 나타냅니다.

**참고 :** 하나 이상의 Y 열에 대한 결측값이 포함된 행은 분석에서 제외되고 "혼합 확률" 테이블에 나타나지 않습니다.

## 잠재 계층 분석 플랫폼 옵션

"잠재 계층 분석"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

**새 군집 수** 다른 개수의 군집을 사용하여 다른 분석을 실행할 수 있습니다. 새 분석 보고서가 현재 보고서에 추가됩니다.

다음 옵션에 대한 자세한 내용은 [JMP 사용](#)에서 확인하십시오.

**로컬 데이터 필터** 특정 보고서에서 사용되는 데이터를 필터링할 수 있는 로컬 데이터 필터를 표시하거나 숨깁니다.

**다시 실행** 분석을 반복하거나 다시 시작할 수 있는 옵션이 포함되어 있습니다. 이 기능을 지원하는 플랫폼에서 "자동 재계산" 옵션은 해당하는 보고서 창에서 데이터 테이블에 대한 변경 사항을 즉시 반영합니다.

**플랫폼 환경 설정** 현재 플랫폼 환경 설정을 보거나, 현재 JMP 보고서의 설정과 일치하도록 플랫폼 환경 설정을 업데이트할 수 있는 옵션이 포함되어 있습니다.

**스크립트 저장** 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다.

**그룹별 스크립트 저장** 기준 변수의 모든 수준에 대한 플랫폼 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다. 시작 창에서 기준 변수를 지정한 경우에만 사용할 수 있습니다.

---

**참고 :** 이 플랫폼의 추가 옵션은 스크립트를 통해 사용할 수 있습니다. 자세한 내용은 "도움말" 메뉴의 "스크립트 인덱스"에서 여십시오. 또한 "스크립트 인덱스"에서 이 섹션에 설명된 옵션의 스크립트 예제도 찾을 수 있습니다.

---

## 잠재 계층 모형 옵션

"<k> 군집의 잠재 계층 모형"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

**모형 보고서** 사용 가능한 모형 보고서를 표시하거나 숨길 수 있습니다. 모형 보고서에 대한 자세한 내용은 "잠재 계층 분석 보고서"에서 확인하십시오.

**군집별 색상** 데이터 테이블의 각 행을 최대 확률 분류 군집에 따라 색상을 적용합니다. 예는 "잠재 계층 분석 플랫폼의 추가 예"에서 확인하십시오.

**혼합 및 군집 계산식 저장** 각 군집에 대한 계산식 열과 최대 확률 분류 군집에 대한 계산식 열을 데이터 테이블에 저장합니다.

**군집 계산식만 저장** 최대 확률 분류 군집을 결정하는 계산식이 포함된 열을 데이터 테이블에 저장합니다.

**JMP PRO 확률 계산식 게시** 확률 계산식을 생성하여 계산식 저장소 플랫폼에 계산식 열 스크립트로 저장합니다. "계산식 저장소" 보고서가 열려 있지 않으면 "계산식 저장소" 보고서가 생성됩니다. 자세한 내용은 예측 및 전문 모델링에서 확인하십시오.

**혼합 확률 저장** "혼합 확률" 테이블의 값을 데이터 테이블의 해당 행에 저장합니다.

**군집만 저장** 각 행에 대해 최대 확률 분류 군집이 포함된 새 열을 데이터 테이블에 저장합니다. 이 열에는 계산식이 포함되지 않습니다.

**군집 이름 바꾸기** 보고서의 군집에 의미 있는 이름을 지정할 수 있습니다.

---

**참고 :** 보고서에 난수 시드값을 지정한 경우 외에는 새 군집 이름이 스크립트에 저장되지 않습니다. 스크립트를 통해 보고서를 시작할 때만 난수 시드값을 설정할 수 있습니다.

---

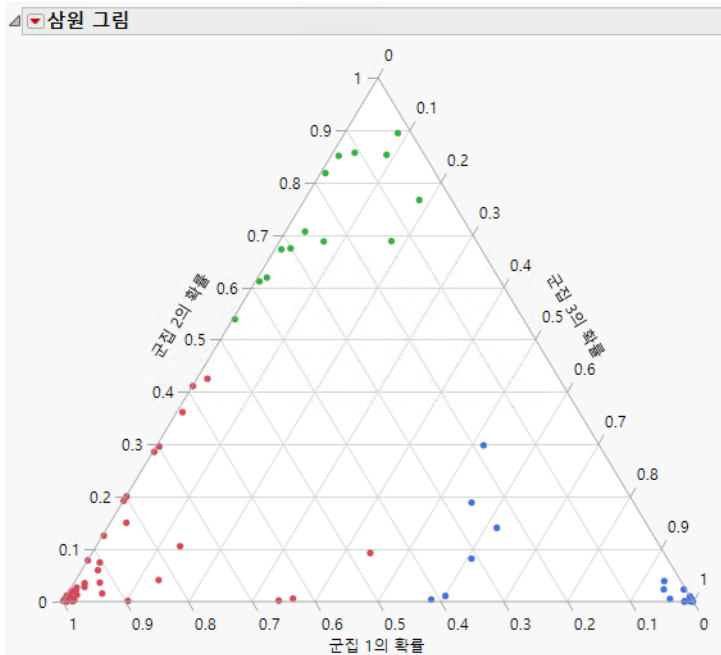
**적합 제거** 보고서 창에서 지정된 군집화 보고서를 제거합니다.

## 잠재 계층 분석 플랫폼의 추가 예

이 예에서는 소속 군집 확률을 시각화하는 그림을 생성합니다. 자동차 소유자 및 제조사에 대한 설문 조사 데이터가 있고 자동차 소유자를 세 군집으로 분류하려고 합니다. 군집이 세 개인 경우 삼원 그림을 사용하면 시각화하는 데 유용합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Car Poll.jmp 를 엽니다 .
2. **분석 > 군집화 > 잠재 계층 분석**을 선택합니다 .
3. age 를 제외한 모든 열을 선택하고 **Y** 를 클릭합니다 .
4. **확인**을 클릭합니다 .
5. "3 개 군집의 잠재 계층 모형 " 옆의 빨간색 삼각형을 클릭하고 **군집별 색상**을 선택합니다 .
6. "3 개 군집의 잠재 계층 모형 " 옆의 빨간색 삼각형을 클릭하고 **혼합 확률 저장**을 선택합니다 .
7. Car Poll 데이터 테이블 창의 열 목록에서 **LCA 군집 확률** 열 그림을 선택합니다 .
8. **그래프 > 삼원 그림**을 선택합니다 .
9. **X, 그림에 그리기**를 클릭합니다 .
10. **확인**을 클릭합니다 .

그림 16.6 소속 군집 확률의 삼원 그림



각 관측값에 대한 군집 확률 삼원 그림에서 대부분의 소속 군집 확률이 꼭지점 근처에 있습니다 . 이는 한 군집에 대해 높은 값이 있고 다른 두 군집에 대해 낮은 값이 있음을 나타냅니다 .

그러나 그림 중간에 몇 개의 점이 있으며, 이는 해당 관측값이 어느 군집에도 소속될 확률이 높지 않음을 나타냅니다. 이러한 관측값은 더 철저한 조사가 필요하거나, 데이터를 더 잘 표현하기 위해 군집이 더 필요하다는 것을 나타낼 수 있습니다.

**참고:** 난수 시드값이 지정되지 않았으므로 결과가 다를 수 있습니다.

## 잠재 계층 분석 플랫폼에 대한 통계 상세 정보

이 섹션에는 잠재 계층 분석에 대한 통계 상세 정보가 포함되어 있습니다.

- "잠재 계층 모형 적합에 대한 통계 상세 정보"
- "최대 군집 수에 대한 통계 상세 정보"

## 잠재 계층 모형 적합에 대한 통계 상세 정보

이 섹션에서는 잠재 계층 분석 플랫폼에서 적합되는 잠재 계층 모형에 대해 설명합니다. 잠재 계층 모형에 대한 자세한 내용은 Collins와 Lanza(2010) 및 Agresti(2013) 연구 자료에서 확인하십시오.

**참고:** 텍스트 탐색기 플랫폼에서 사용되는 LCA 알고리즘은 문서 용어 행렬의 희소성을 활용합니다. 이러한 이유로 텍스트 탐색기 플랫폼의 LCA 결과는 잠재 계층 분석 플랫폼의 결과와 정확하게 매칭되지 않습니다.

$j = 1, \dots, J$ 가 관측된 반응 열을 나타낸다고 하겠습니다. 이는 잠재 계층 분석 플랫폼 시작 창의  $Y$  열입니다.  $j$  열의 수준 수를  $R_j$ 로 나타냅니다.

$J$  변수의 다차원 분할표에는  $W = R_1 * \dots * R_J$  셀이 포함됩니다. 이러한 각 셀은  $J$  변수에 대한 반응 패턴으로 정의됩니다. 따라서 각 반응 패턴은  $\mathbf{y} = y_1, \dots, y_j$  형식의  $J$  길이 벡터입니다.  $\mathbf{Y}$ 를 행 벡터로 간주되는 모든 반응 패턴의  $W \times J$  배열로 정의합니다.  $\mathbf{Y}$ 에 있는 각 요소  $y_w$ 의 확률은  $\Pr(\mathbf{y}_w)$ 입니다. 이러한 확률의 합은 1입니다.

$$\sum_{w=1}^W Pr(\mathbf{y}_w) = 1$$

다음 표기를 고려하십시오.

- $C$  = 잠재 계층 모형의 군집 수
- $\gamma_c$  = 군집  $c$ 에 소속될 확률.  $\gamma_c$ 는 잠재 계층 출현율입니다. 이러한 모수의 합은 1입니다.
- $r_{jk}$  =  $j$  번째 반응의  $k$  번째 수준
- $\rho_{jklc}$  =  $c$  계층의 소속을 조건으로  $j$  열에서  $r_{jk}$  반응을 관측할 확률.  $\rho_{jklc}$ 는 항목-반응 확률입니다. 지정된 군집 및 반응 변수  $j$ 에 대해  $\rho_{jklc}$ 의 합은 1입니다.

- $I(y_j = r_{j,k}) = y_j$  반응이  $j$  번째 반응의  $k$  번째 수준이면 1, 그렇지 않으면 0 인 표시 함수

반응  $\mathbf{y}_w = y_{1w}, \dots, y_{jw}$  의 특정 벡터를 관측할 확률은  $C$  개의 각 잠재 계층에 대해 해당 반응 벡터를 관측할 조건부 확률의 합계입니다.

$$\Pr(\mathbf{y}) = \sum_{c=1}^C \gamma_c \prod_{j=1}^J \prod_{k=1}^{R_j} \rho_{j,k|c}^{I(y_j = r_{j,k})}$$

이 방정식은 "잠재 계층 모형"의 빨간색 삼각형 메뉴에서 "혼합 및 군집 계산식 저장" 옵션을 선택하여 데이터 테이블에 저장할 수 있는 **확률 계산식** 군집 계산식의 분모입니다. **확률 계산식** 군집 열의 계산식은  $\Pr(\text{군집} = c \mid \mathbf{y}_w)$  을 제공하며, 이는  $\Pr(\mathbf{y}_w \mid \text{군집} = c) / \Pr(\mathbf{y}_w)$  과 같습니다.

잠재 계층 모형의  $\gamma$  및  $\rho$  모수는 반복 EM(Expectation-Maximization) 알고리즘을 사용하여 추정됩니다. 잠재 계층 모형의 고유 모수 수는 다음과 같이 정의됩니다.

$$(C-1) + C \sum_{j=1}^J (R_j - 1)$$

## 최대 군집 수에 대한 통계 상세 정보

잠재 계층 분석 모형에 적합시킬 수 있는 최대 군집 수는 모형 자유도에 따라 다릅니다. 잠재 계층 분석 모형의 자유도는 열에 의해 생성된 분할표 크기를 기반으로 합니다. 분할표 크기는 테이블에서 하나 이상의 관측값이 포함된 셀 개수이며  $K$  로 나타냅니다. 모든 셀에 하나 이상의 관측값이 포함된 경우  $K$  는 반응 열의 수준 수를 곱한 값입니다. 자유도 계산식은 다음과 같이 정의됩니다.

$$DF = K - \{nCluster - 1 + nCluster(nTotalLevels - nCols)\} - 1$$

다음은 각 요소에 대한 설명입니다.

nCluster = 군집 수

nTotalLevels = 반응 열 수준의 합계

nCols = 반응 열 수

LCA 모형을 적절하게 적합시키려면 자유도가 양수여야 합니다. 따라서  $DF > 0$  을 위해 최대 군집 수는 다음과 같이 정의됩니다.

$$\max(nCluster) < \text{floor}[K / (1 + nTotalLevels - nCols)]$$

# 17 장

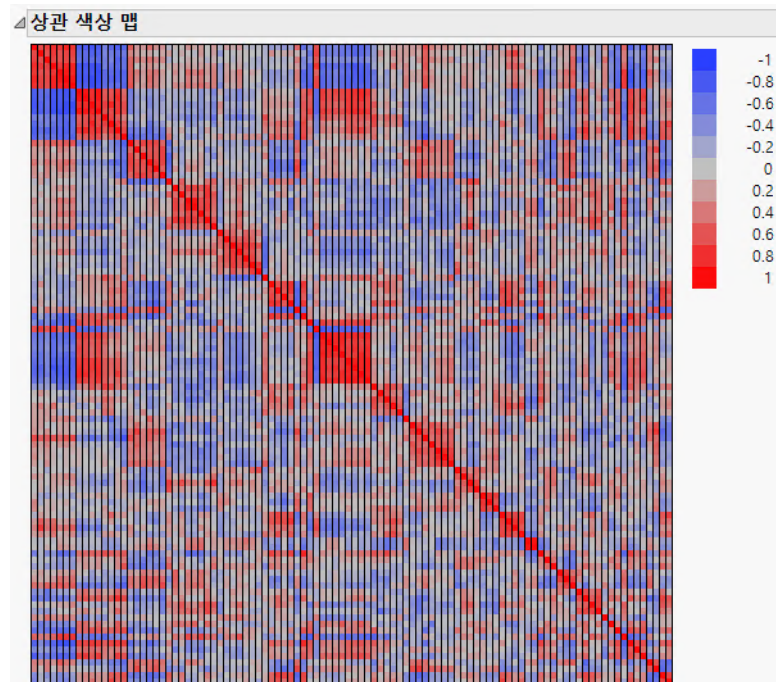
## 변수 군집화

### 유사한 변수를 대표 그룹으로 그룹화

변수 군집화는 유사한 변수를 대표 그룹으로 그룹화하는 방법을 제공합니다. 각 군집은 단일 성분 또는 변수로 나타낼 수 있습니다. 성분은 군집에 있는 모든 변수의 선형 결합입니다. 아니면 군집을 해당 군집에서 가장 대표적인 멤버로 식별되는 변수로 나타낼 수도 있습니다.

변수 군집화를 차원 축소 방법으로 사용할 수 있습니다. 큰 변수 집합을 사용하여 모델링하는 대신 군집 성분 또는 군집에서 가장 대표적인 변수를 사용하여 데이터 변동의 대부분을 설명할 수 있습니다. 또한 변수 군집화를 사용한 차원 축소는 주성분을 사용한 차원 축소보다 의미 있는 해석이 가능한 경우가 많습니다.

그림 17.1 변수 상관 맵의 예



## 목차

변수 군집화 플랫폼 개요.....	359
변수 군집화 플랫폼의 예.....	359
변수 군집화 플랫폼 시작.....	361
변수 군집화 보고서.....	361
상관 색상 맵.....	362
군집 요약.....	362
군집 멤버.....	362
표준화 성분.....	363
변수 군집화 플랫폼 옵션.....	363
변수 군집화 플랫폼의 추가 예.....	364
상관 색상 맵의 예.....	364
차원 축소를 위한 변수 군집화 플랫폼의 예.....	365
변수 군집화 플랫폼에 대한 통계 상세 정보.....	368

---

## 변수 군집화 플랫폼 개요

변수 군집화 플랫폼에서는 변수의 선형 결합인 성분을 유사한 변수로 구성된 군집으로 생성합니다. 이는 분석에 포함된 모든 변수의 선형 결합인 성분을 생성하는 주성분 분석과 다릅니다. 변수 군집화 플랫폼에서는 전체 변수 집합이 군집으로 분할됩니다. 각 군집에 대해 해당 군집 변수의 첫 번째 주성분을 사용하여 **군집 성분**이 생성됩니다. 이 군집 성분은 해당 군집의 변수 중에서 가능한 한 많은 변동을 설명하는 선형 결합입니다.

변수 군집화 옵션을 차원 축소 방법으로 사용할 수 있습니다. 흔히 큰 변수 집합에서 변동의 상당 부분을 군집 성분 또는 군집에서 가장 대표적인 변수로 나타낼 수 있습니다. 그런 다음 이러한 새 변수를 예측 또는 기타 모델링 기법에 사용할 수 있습니다. 일반적으로 새 군집 기반 변수는 모든 변수에 기반한 주성분보다 해석하기가 쉽습니다.

공통 변수 집합에서 생성된 주성분은 직교합니다. 그러나 군집 성분은 구분되는 변수 집합에서 생성되므로 직교하지 않습니다.

큰 변수 집합을 사용하는 경우 변수 군집화 플랫폼에서는 특이값 분해를 기반으로 하는 알고리즘을 사용하여 계산 시간을 단축합니다. 추가 배경 정보는 "[와이드 선형 방법 및 특이값 분해](#)"에서 확인하십시오.

---

## 변수 군집화 플랫폼의 예

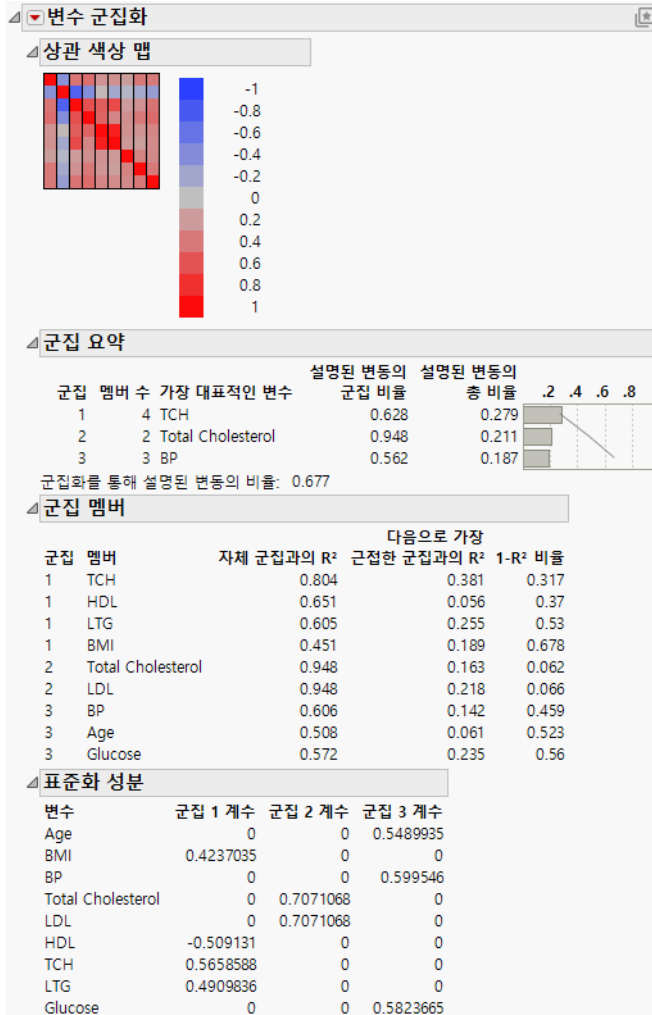
이 예에서는 당뇨병 진행 경과를 모델링하는 데 사용되는 연속형 기준 변수를 군집화합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Diabetes.jmp 를 엽니다.
2. **분석 > 군집화 > 변수 군집화**를 선택합니다.
3. Gender 를 제외한 Age ~ Glucose 열 (Age, BMI, BP, Total Cholesterol, LDL, HDL, TCH, LTG 및 Glucose) 을 선택하고 **Y, 열**을 클릭합니다.

변수 군집화에는 연속형 수치 변수가 필요하므로 Gender 열을 포함할 수 없습니다.

4. **확인**을 클릭합니다.

그림 17.2 Diabetes 데이터에 대한 변수 군집화 보고서



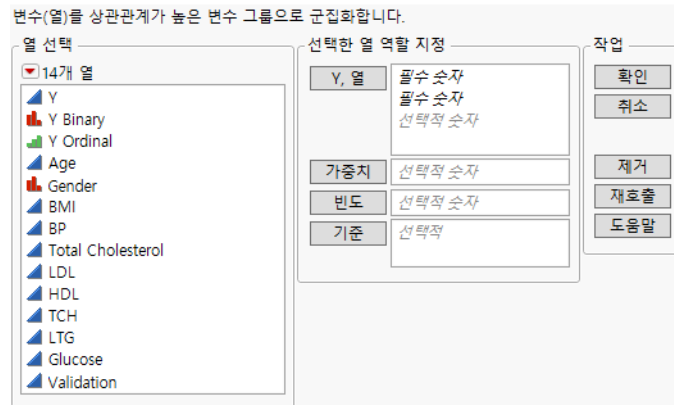
"군집 요약" 보고서에는 다음과 같이 변수가 세 개의 군집으로 그룹화되어 있습니다.

- "군집 멤버" 보고서에 표시된 것처럼 군집 1은 TCH, HDL, LTG 및 BMI로 구성됩니다. "군집 요약" 보고서에 따르면 TCH가 군집 1의 가장 대표적인 변수이고, 군집 1 변수의 경우 변동의 62.8%가 첫 번째 주성분에 의해 설명됩니다.
- 군집 2는 Total Cholesterol과 LDL로 구성됩니다. "군집 요약" 보고서에 따르면 Total Cholesterol이 군집 2의 가장 대표적인 변수이고, 군집 2 변수의 경우 변동의 94.8%가 첫 번째 주성분에 의해 설명됩니다.
- 군집 3은 BP, Age 및 Glucose로 구성됩니다. "군집 요약" 보고서에 따르면 BP가 가장 대표적인 변수이고, 군집 3 변수의 경우 변동의 56.2%가 첫 번째 주성분에 의해 설명됩니다.

## 변수 군집화 플랫폼 시작

분석 > 군집화 > 변수 군집화를 선택하여 변수 군집화 플랫폼을 시작합니다.

그림 17.3 변수 군집화 시작 창



"열 선택"의 빨간색 삼각형 메뉴에 포함된 옵션에 대한 자세한 내용은 JMP 사용에서 확인하십시오.

**Y, 열** 군집화할 변수입니다. 변수는 수치형이고 연속형이어야 합니다.

**가중치** 분석의 각 행에 가중치를 할당하는 숫자 값이 포함된 열입니다.

**빈도** 분석의 각 행에 빈도를 할당하는 숫자 값이 포함된 열입니다.

**기준** 각 수준으로 개별 분석을 정의하는 열입니다. 지정된 열의 각 수준에 대해 해당하는 행이 분석됩니다. 결과는 개별 보고서에 제공됩니다. 기준 변수가 둘 이상 할당되면 기준 변수의 가능한 각 수준 조합에 대해 개별 분석이 생성됩니다.

## 변수 군집화 보고서

기본적으로 변수 군집화 플랫폼의 "변수 군집화" 보고서에는 다음 섹션이 포함되어 있습니다.

- "상관 색상 맵"
- "군집 요약"
- "군집 멤버"
- "표준화 성분"

**팁:** "변수 군집화" 테이블에서 행을 선택하면 데이터 테이블의 해당 열도 선택됩니다. Ctrl 키를 누른 채 행을 클릭하면 데이터 테이블에서 열이 선택 취소됩니다.

## 상관 색상 맵

"변수 군집화" 보고서의 "상관 색상 맵" 섹션에는 변수 간 상관관계에 대한 색상 맵이 표시됩니다. 변수는 "군집 멤버" 테이블에 나열된 순서대로 배열됩니다. 이렇게 배열하면 같은 군집의 멤버가 상관 그림에서 인접하게 배치됩니다. 자세한 내용은 "[상관 색상 맵의 예](#)"에서 확인하십시오.

**팁 :** 색상 맵에서 정사각형을 커서로 가리키면 해당 정사각형에 포함된 변수와 상관 값을 볼 수 있습니다.

같은 군집의 변수는 다른 군집의 변수보다 절대 상관이 더 높은 경향이 있습니다(더 진한 빨간색 또는 파란색). 따라서 상관 맵에서 주어진 성분의 변수에 해당하는 셀로 형성된 정사각형이 대각선을 따라 두드러지는 경우가 많습니다.

상관은 행렬 방법을 사용하여 계산됩니다. 이 방법은 변수에 대한 결측 데이터가 있는 관측값을 상관 계산에서 제외합니다. 행렬 추정 방법에 대한 자세한 내용은 "[분산 추정 방법에 대한 통계 상세 정보](#)"에서 확인하십시오.

## 군집 요약

"변수 군집화" 보고서의 "군집 요약" 테이블에는 다음 열이 포함되어 있습니다.

**군집** 군집 식별자입니다.

**멤버 수** 군집의 변수 수입니다.

**가장 대표적인 변수** 군집 성분과의 상관 제곱이 가장 큰 군집 변수입니다.

**설명된 변동의 군집 비율** 군집의 변수 중 첫 번째 주성분에 의해 설명되는 군집 변동 비율입니다. 군집의 변수가 하나뿐이면 이 값이 1입니다. 이 통계량은 모든 변수가 아니라 군집 내의 변수만 기반으로 합니다.

**설명된 변동의 총 비율** 군집 성분에 의해 설명되는 전체 변동 비율입니다. 이것은 각 군집 내의 변수만 사용하여 첫 번째 주성분을 계산하는 것과 동일합니다.

테이블 아래에는 모든 군집 성분에 의해 설명되는 총 변동 비율에 대한 내용이 나와 있습니다.

## 군집 멤버

"변수 군집화" 보고서의 "군집 멤버" 테이블에는 다음 열이 포함되어 있습니다.

**군집** 군집 식별자입니다.

**멤버** 군집에 포함된 변수입니다.

**자체 군집과의 R<sup>2</sup>** 변수와 해당 군집 성분의 상관 제곱입니다.

**다음으로 가장 근접한 군집과의 R<sup>2</sup>** 다음으로 가장 근접한 군집에 대한 변수와 군집 성분의 상관 제곱입니다. 다음으로 가장 근접한 군집은 변수와 군집 성분의 상관 제곱이 두 번째로 높은 군집입니다.

**1-R<sup>2</sup> 비율** 변수가 속한 군집과 다음으로 가장 근접한 군집 사이의 상대적 근접성에 대한 측도입니다. 이 값은 다음과 같이 정의됩니다.

$$(1 - \text{자체 군집과의 } R^2) / (1 - \text{다음으로 가장 근접한 군집과의 } R^2)$$

## 표준화 성분

"변수 군집화" 보고서의 "표준화 성분" 테이블에는 군집 성분을 정의하는 계수가 나열됩니다. 이러한 계수는 각 군집 내에서 첫 번째 주성분의 고유 벡터입니다.

---

## 변수 군집화 플랫폼 옵션

"변수 군집화"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

**상관 색상 맵** "상관 색상 맵" 그림을 표시하거나 숨깁니다. 자세한 내용은 "상관 색상 맵"에서 확인하십시오.

**군집 요약** "군집 요약" 보고서를 표시하거나 숨깁니다. 자세한 내용은 "군집 요약"에서 확인하십시오.

**군집 멤버** "군집 멤버" 보고서를 표시하거나 숨깁니다. 자세한 내용은 "군집 멤버"에서 확인하십시오.

**군집 성분** "표준화 성분" 보고서를 표시하거나 숨깁니다. 자세한 내용은 "표준화 성분"에서 확인하십시오.

**군집 성분 저장** 군집 성분이라는 열 그룹을 데이터 테이블에 저장합니다. 각 열은 군집 <i>i</i> 성분으로 이름이 지정되고 군집 성분을 비중심화 및 비척도화 변수로 표현하는 계산식을 포함합니다.

**모형 적합 시작** "모형 효과 생성" 목록에 각 군집의 가장 대표적인 변수가 입력된 "모형 규격" 창을 엽니다. 이 옵션을 사용하여 가장 대표적인 변수를 기반으로 모형을 생성할 수 있습니다.

---

**팁:** 성분을 사용하여 모형을 적합시키려면 먼저 **군집 성분 저장** 옵션을 선택합니다. 그런 다음 "모형 효과 생성" 목록에서 각 군집에 대한 가장 대표적인 변수를 원하는 군집 성분 열로 바꿉니다.

---

다음 옵션에 대한 자세한 내용은 **JMP 사용**에서 확인하십시오.

**로컬 데이터 필터** 특정 보고서에서 사용되는 데이터를 필터링할 수 있는 로컬 데이터 필터를 표시하거나 숨깁니다.

**다시 실행** 분석을 반복하거나 다시 시작할 수 있는 옵션이 포함되어 있습니다. 이 기능을 지원하는 플랫폼에서 "자동 재계산" 옵션은 해당하는 보고서 창에서 데이터 테이블에 대한 변경 사항을 즉시 반영합니다.

**플랫폼 환경 설정** 현재 플랫폼 환경 설정을 보거나, 현재 JMP 보고서의 설정과 일치하도록 플랫폼 환경 설정을 업데이트할 수 있는 옵션이 포함되어 있습니다.

**스크립트 저장** 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다.

**그룹별 스크립트 저장** 기준 변수의 모든 수준에 대한 플랫폼 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다. 시작 창에서 기준 변수를 지정한 경우에만 사용할 수 있습니다.

**참고:** 이 플랫폼의 추가 옵션은 스크립트를 통해 사용할 수 있습니다. 자세한 내용은 "도움말" 메뉴의 "스크립트 인덱스"에서 여십시오. 또한 "스크립트 인덱스"에서 이 섹션에 설명된 옵션의 스크립트 예제도 찾을 수 있습니다.

## 변수 군집화 플랫폼의 추가 예

이 섹션에는 변수 군집화 플랫폼을 사용하는 예가 포함되어 있습니다.

- "상관 색상 맵의 예"
- "차원 축소를 위한 변수 군집화 플랫폼의 예"

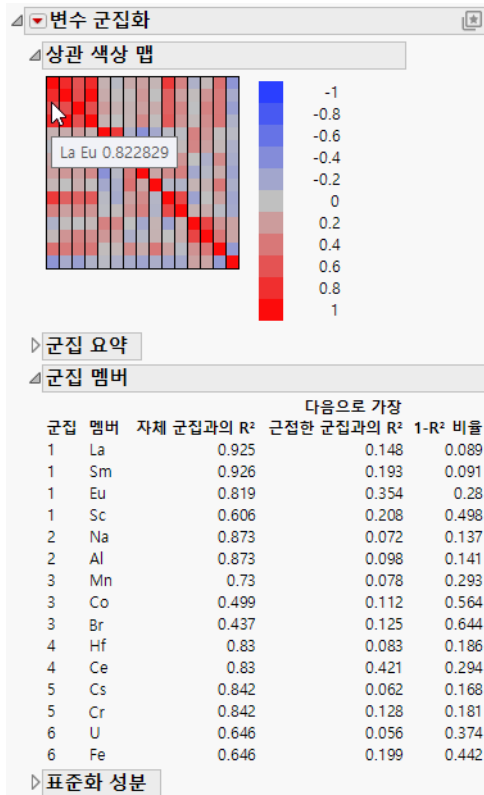
### 상관 색상 맵의 예

이 예에서는 변수 군집화 플랫폼을 사용하여 상관 색상 맵을 생성하고 검토합니다. 색상 맵은 변수 간 상관관계 패턴을 찾는 데 유용합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Cherts JMP 를 엽니다.
2. **분석 > 군집화 > 변수 군집화**를 선택합니다.
3. 모든 연속형 변수를 선택하고 **Y, 열**을 클릭합니다.
4. **확인**을 클릭합니다.
5. "군집 요약" 및 "표준화 성분" 보고서를 닫습니다.
6. 색상 맵에서 두 번째 행과 첫 번째 열의 셀을 커서로 가리킵니다.

이 셀에 해당하는 변수가 La 와 Eu 이고 변수 상관이 0.822829 임을 보여 주는 툴팁이 나타납니다.

그림 17.4 Cherts.jmp 에 대한 상관 색상 맵



"군집 멤버" 보고서에서는 군집 1에 네 개의 변수가 있음을 보여 줍니다. "상관 색상 맵"에서 이 네 개 변수에 해당하는 왼쪽 위의 4x4 정사각형 셀은 뚜렷한 양의 상관관계 패턴을 보여 줍니다. 색상 맵에는 두 개 군집(군집 4 및 5)의 변수에 대한 양의 상관관계 패턴도 표시됩니다. 색상 맵에서 군집 6의 두 변수에 해당하는 오른쪽 아래의 2x2 정사각형 셀은 변수에 음의 상관관계가 있음을 나타냅니다. 자세한 내용은 "상관 색상 맵"에서 확인하십시오.

## 차원 축소를 위한 변수 군집화 플랫폼의 예

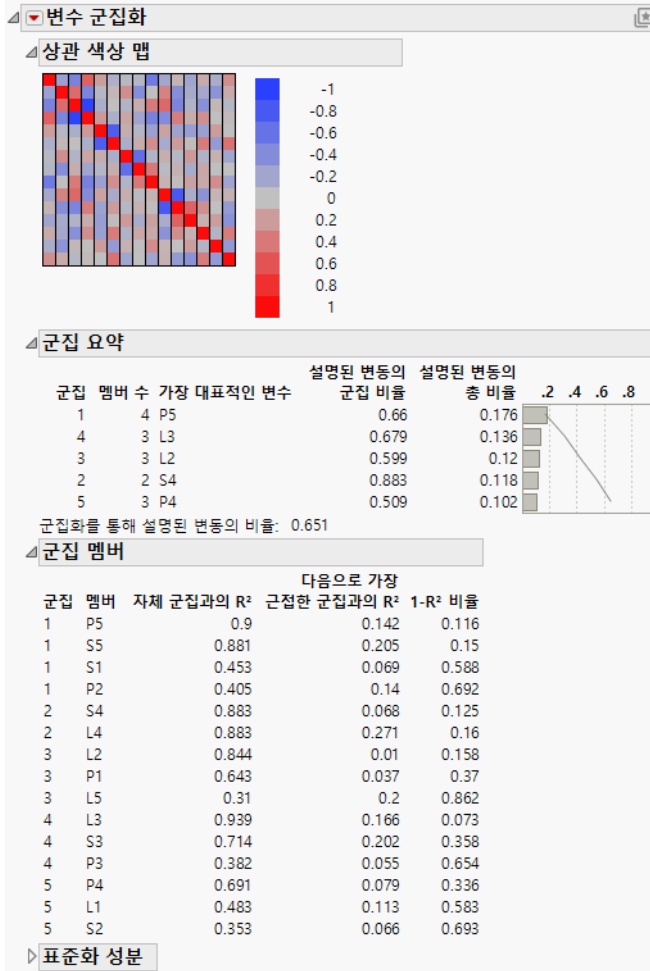
이 예에서는 변수 군집화 플랫폼을 모델링을 위한 차원 축소 도구로 사용합니다. 데이터 테이블에는 반응 변수를 예측하는 데 사용되는 15개의 변수가 포함되어 있으며 이 수를 줄이려고 합니다.

### 변수 군집화

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Penta.jmp 를 엽니다.
2. **분석 > 군집화 > 변수 군집화**를 선택합니다.
3. logRAI 를 제외한 모든 연속형 변수를 선택하고 **Y, 열**을 클릭합니다.

4. **확인**을 클릭합니다.
5. "변수 군집화"의 빨간색 삼각형을 클릭하고 **군집 성분 저장**을 선택합니다.  
그룹화된 다섯 개의 계산식 열이 데이터 테이블에 추가됩니다.

그림 17.5 Penta.jmp 에 대한 변수 군집화 보고서



"군집 요약" 및 "군집 멤버" 보고서에서는 변수가 다섯 개 그룹으로 군집화되어 다섯 개의 군집 성분 변수가 있음을 보여 줍니다.

## 모형 적합

이제 다음 두 모형을 적합시키고 비교하여 logRAI 를 예측합니다.

- 모든 연속형 변수를 예측 변수로 사용하는 모형

- 군집 성분을 예측 변수로 사용하는 모형
- 1. " 변수 군집화 " 의 빨간색 삼각형을 클릭하고 **모형 적합 시작**을 선택합니다 .
- 2. logRAI 를 선택하고 **Y** 를 클릭합니다 .  
가장 대표적인 변수를 나타내는 다섯 개 군집이 " 모형 효과 생성 " 목록에 입력되어 있습니다 .  
그러나 모든 예측 변수를 입력하려고 합니다 .
- 3. 모든 연속형 변수 (S1 ~ P5) 를 선택하고 **추가**를 클릭합니다 .  
Obs Name 을 포함하지 않도록 주의하십시오 .
- 4. **대화상자 열린 채 유지** 옆의 상자를 선택합니다 .
- 5. **실행**을 클릭합니다 .

**그림 17.6** 모든 연속형 예측 변수를 사용한 모형의 최소 제곱 적합 보고서

반응 log RAI

▷ 효과 요약

▲ 적합 요약

R <sup>2</sup>	0.929316
Adj-R <sup>2</sup>	0.853582
제곱근 평균 제곱 오차	0.331225
반응 평균	0.734333
관측값 수(또는 가중치 합)	30

▲ 분산 분석

소스	DF	제곱합	평균 제곱	F 비
모형	15	20.193596	1.34624	12.2709
오차	14	1.535941	0.10971	Prob > F
수정 합계	29	21.729537		<.0001*

▲ 모수 추정값

항	추정값	표준 오차	t 비	Prob> t
절편	-0.802632	0.924946	-0.87	0.4002
P5	2.0563803	1.651272	1.25	0.2335
S4	-0.062354	0.134935	-0.46	0.6511
L2	0.0860287	0.061206	1.41	0.1817
L3	0.3185383	0.080091	3.98	0.0014*
P4	0.4136598	0.394449	1.05	0.3121
S1	-0.09783	0.038948	-2.51	0.0249*
L1	0.032362	0.049732	0.65	0.5258
P1	-0.107951	0.085209	-1.27	0.2259
S2	0.086703	0.044276	1.96	0.0704
P2	0.0847235	0.086297	0.98	0.3429
S3	-0.037728	0.055602	-0.68	0.5085
P3	-0.027313	0.233655	-0.12	0.9086
L4	-0.029756	0.152012	-0.20	0.8476
S5	2.7123146	2.222039	1.22	0.2424
L5	-0.209128	0.270401	-0.77	0.4521

▷ 효과 검정

▷ 효과 상세 정보

- 6. " 모형 적합 " 창의 " 모형 효과 생성 " 목록에서 모든 변수를 선택하고 **제거**를 클릭합니다 .
- 7. 군집 성분 그룹을 선택하고 **추가**를 클릭합니다 .
- 8. **실행**을 클릭합니다 .

그림 17.7 군집 성분을 예측 변수로 사용한 모형의 최소 제곱 적합 보고서

반응 log RAI				
▷ 효과 요약				
▷ 적합 요약				
R <sup>2</sup>		0.8214		
Adj-R <sup>2</sup>		0.784191		
제곱근 평균 제곱 오차		0.402125		
반응 평균		0.734333		
관측값 수(또는 가중치 합)		30		
▷ 분산 분석				
소스	DF	제곱합	평균 제곱	F 비
모형	5	17.848635	3.56973	22.0757
오차	24	3.880902	0.16170	Prob > F
수정 합계	29	21.729537		<.0001*
▷ 모수 추정값				
항	추정값	표준 오차	t 비	Prob> t
절편	0.6651552	0.074349	8.95	<.0001*
군집 1 성분	-0.018483	0.056013	-0.33	0.7443
군집 2 성분	0.0035891	0.069032	0.05	0.9590
군집 3 성분	0.2043072	0.066714	3.06	0.0053*
군집 4 성분	0.5754553	0.065423	8.80	<.0001*
군집 5 성분	-0.046594	0.069786	-0.67	0.5107
▷ 효과 검정				
▷ 효과 상세 정보				

다섯 개의 군집 성분만 예측 변수로 포함하는 모형은 상당한 양의 반응 변동을 설명합니다(수정 R<sup>2</sup>=0.784). 15개의 예측 변수를 모두 사용하는 모형의 수정 R<sup>2</sup>은 약간 더 높은 0.853입니다(그림 17.6 참조).

## 변수 군집화 플랫폼에 대한 통계 상세 정보

변수 군집화 알고리즘은 더 이상 분할할 수 없을 때까지 반복적으로 변수 군집을 분할하고 변수를 군집에 재할당합니다. 초기 군집은 모든 변수로 구성됩니다. 이 알고리즘은 SAS에서 개발했으며 PROC VARCLUS(SAS Institute Inc. 2023g)에서 구현되었습니다.

**참고:** 이 알고리즘은 "Y, 열" 목록의 변수에 대해 결측값이 없는 관측값만 사용합니다.

알고리즘의 반복 단계는 다음과 같습니다.

- 모든 군집에 대해 다음을 수행합니다.
  - 각 군집의 변수에 대한 주성분을 계산합니다.
  - 모든 군집의 두 번째 고유값이 1 보다 작으면 알고리즘을 종료합니다.
- 다음 단계에 따라 두 번째 고유값이 가장 큰 (1 보다 큼) 군집을 두 개의 새 군집으로 분할합니다.
  - 직교 사각 회전을 사용하여 현재 군집의 변수에 대한 주성분을 회전합니다.

- b. 현재 군집에서 첫 번째 회전된 주성분에 대한 상관 계수가 두 번째 주성분에 대한 상관 계수보다 높은 변수로 구성된 군집 하나를 정의합니다.
  - c. 원래 군집의 나머지 변수로 구성된 다른 군집을 정의합니다. 이러한 변수는 두 번째 주성분과 더 높은 상관관계가 있습니다.
  - d. 두 개의 새 군집에 대한 주성분을 계산합니다.
3. 데이터 집합의 변수를 다른 군집에 할당해야 하는지 여부를 결정합니다. 각 변수에 대해 다음을 수행합니다.
- a. 각 군집의 첫 번째 주성분과 변수의 상관을 계산합니다.
  - b. 상관이 가장 큰 변수를 군집에 포함합니다.

---

**참고 :** 직교 사각 회전은 원시 **Quartimax** 회전이라고도 합니다. 자세한 내용은 **Harris** 와 **Kaiser(1964)** 연구 자료에서 확인하십시오.

---



# 부록 A

## 통계 상세 정보 다변량 방법

---

이 부록에서는 와이드 선형 방법에 대한 내용과 특이값 분해 사용에 대해 설명합니다.

## 목차

와이드 선형 방법 및 특이값 분해.....	373
특이값 분해.....	373

## 와이드 선형 방법 및 특이값 분해

군집화, 주성분 및 판별 플랫폼의 와이드 선형 방법을 사용하면 수천 개 또는 수백만 개의 변수가 있는 데이터 집합을 분석할 수 있습니다. 대부분의 다변량 기법에서는 공분산 행렬 또는 공분산 행렬의 역행렬을 계산해야 합니다. 다변량 분석에 많은 수의 변수가 포함될 경우 공분산 행렬이 엄청나게 커질 수 있으므로 공분산 행렬 또는 역행렬을 계산하는 데 문제가 있고 계산 시간이 많이 소요됩니다.

데이터가  $n$ 개의 행과  $p$ 개의 열로 구성되어 있다고 가정해 보겠습니다. 공분산 행렬의 계수는  $n$ 과  $p$  중 더 작은 값입니다. 가로형 데이터 집합에서는 보통  $p$ 가  $n$ 보다 훨씬 큼니다. 이러한 경우 공분산 행렬의 역행렬은 최대  $n$ 개의 0이 아닌 고유값을 가집니다. 와이드 선형 방법에서는 이 사실을 특이값 분해와 함께 사용하여 효율적인 계산을 제공합니다. 자세한 내용은 "[특이값 분해 계산](#)"에서 확인하십시오.

### 특이값 분해

SVD(특이값 분해)를 사용하면 선형 변환을 회전, 척도화, 다른 회전으로 표현할 수 있습니다. SVD에 따르면  $n \times p$  행렬  $\mathbf{X}$ 를 다음과 같이 쓸 수 있습니다.

$$\mathbf{X} = \mathbf{U} \text{Diag}(\Lambda) \mathbf{V}'$$

$r$ 을  $\mathbf{X}$ 의 계수라고 하겠습니다.  $r \times r$  단위 행렬을  $\mathbf{I}_r$ 로 나타냅니다.

$\mathbf{U}$ ,  $\text{Diag}(\Lambda)$  및  $\mathbf{V}$  행렬은 다음과 같은 특성을 가집니다.

$\mathbf{U}$ 는  $\mathbf{U}'\mathbf{U} = \mathbf{I}_r$ 인  $n \times r$  반직교 행렬입니다.

$\mathbf{V}$ 는  $\mathbf{V}'\mathbf{V} = \mathbf{I}_r$ 인  $p \times r$  반직교 행렬입니다.

$\text{Diag}(\Lambda)$ 는 열 벡터  $\Lambda = (\lambda_1, \lambda_2, \dots, \lambda_r)$ 로 지정되는 양수 대각 요소가 포함된  $r \times r$  대각 행렬입니다. 여기서  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0$ 입니다.

$\lambda_i$ 는  $\mathbf{X}$ 의 0이 아닌 특이값입니다.

다음은 SVD를 정방 행렬의 스펙트럼 분해와 관련시킵니다.

- $\lambda_i$ 의 제곱은  $\mathbf{X}'\mathbf{X}$ 의 0이 아닌 고유값입니다.
- $\mathbf{V}$ 의  $r$  열은  $\mathbf{X}'\mathbf{X}$ 의 고유 벡터입니다.

**참고:** 문헌에는  $\mathbf{U}$ ,  $\mathbf{V}$  행렬 및 특이값을 포함하는 행렬의 차원과 관련된 다양한 규칙이 있습니다. 그러나 차이가  $\mathbf{X}$ 의 계수까지의 분해에 실제적인 영향을 미치지 않습니다.

특이값 분해에 대한 자세한 내용은 Press et al. (1998, Section 2.6) 연구 자료에서 확인하십시오.

## 공분산 행렬

이 섹션에서는 SVD(특이값 분해)를 사용하여 공분산 행렬의 고유 벡터와 고유값을 구하는 방법을 설명합니다. 관심 행렬이 적어도 하나의 큰 차원을 갖는 경우 SVD를 계산하는 것이 공분산 행렬 및 고유값 분해를 계산하는 것보다 훨씬 더 효율적입니다.

$n$ 을 관측값 수,  $p$ 를 관심 있는 다변량 분석에 포함된 변수의 수라고 하겠습니다. 데이터 값의  $n \times p$  행렬을  $\mathbf{X}$ 로 나타냅니다.

SVD는 대개 표준화된 데이터에 적용됩니다. 값을 표준화하려면 평균을 뺀 후 표준편차로 나눕니다. 표준화된 데이터 값의  $n \times p$  행렬을  $\mathbf{X}_s$ 로 나타냅니다. 그러면 표준화된 데이터의 공분산 행렬은  $\mathbf{X}$ 의 상관 행렬이며 다음과 같이 정의됩니다.

$$\text{Cov} = \mathbf{X}_s' \mathbf{X}_s / (n - 1)$$

SVD를  $\mathbf{X}_s$ 에 적용하여  $\mathbf{X}_s' \mathbf{X}_s$ 의 고유 벡터와 고유값을 얻을 수 있습니다. 이렇게 하면  $\mathbf{X}$  행렬이 심한 가로형(열이 많음) 또는 세로형(행이 많음)일 때 고유 벡터와 고유값을 효율적으로 계산할 수 있습니다. 이 기법은 와이드 PCA의 기본입니다. 자세한 내용은 "[주성분 보고서](#)"에서 확인하십시오.

## 공분산 역행렬

일부 다변량 기법을 사용할 때 공분산 행렬의 역행렬을 계산해야 합니다. 이 섹션에서는 SVD를 사용하여 공분산 행렬의 역행렬을 계산하는 방법을 설명합니다.

표준화된 데이터 행렬을  $\mathbf{X}_s$ 로 나타내고  $\mathbf{S} = \mathbf{X}_s' \mathbf{X}_s$ 를 정의합니다. 특이값 분해를 사용하면  $\mathbf{S}$ 를 다음과 같이 쓸 수 있습니다.

$$\mathbf{S} = (\mathbf{U} \text{Diag}(\Lambda) \mathbf{V}')' (\mathbf{U} \text{Diag}(\Lambda) \mathbf{V}') = \mathbf{V} \text{Diag}(\Lambda)^2 \mathbf{V}'$$

$\mathbf{S}$ 가 완전 계수이면  $\mathbf{V}$ 는  $p \times p$  직교정규 행렬이고  $\mathbf{S}^{-1}$ 을 다음과 같이 쓸 수 있습니다.

$$\mathbf{S}^{-1} = (\mathbf{V} \text{Diag}(\Lambda)^2 \mathbf{V}')^{-1} = \mathbf{V} \text{Diag}(\Lambda)^{-2} \mathbf{V}'$$

$\mathbf{S}$ 가 완전 계수가 아니면  $\text{Diag}(\Lambda)^{-1}$ 을 일반화 역행렬  $\text{Diag}(\Lambda)^+$ 로 대체할 수 있습니다. 여기서  $\text{Diag}(\Lambda)$ 의 대각 요소는 역수로 대체됩니다. 즉,  $\mathbf{S}$ 의 일반화 역행렬이 다음과 같이 정의됩니다.

$$\mathbf{S}^- = \mathbf{V} (\text{Diag}(\Lambda)^+)^2 \mathbf{V}'$$

이 일반화 역행렬은 SVD만 사용하여 계산할 수 있습니다.

와이드 선형 판별 분석에 SVD를 적용하는 데 대한 자세한 내용은 "[와이드 선형 판별 방법](#)"에서 확인하십시오.

## 특이값 분해 계산

다변량 방법 플랫폼에서 JMP는 Golub와 Kahan(1965) 연구 자료에 제안된 방법에 따라 행렬의 SVD를 계산합니다. Golub와 Kahan 방법은 2단계 절차를 포함합니다. 첫 번째 단계에서는  $\mathbf{M}$  행렬을 이중 대각 행렬  $\mathbf{J}$ 로 축소합니다. 두 번째 단계에서는  $\mathbf{J}$ 의 특이값을 계산합니다. 이 값은 원래 행렬  $\mathbf{M}$ 의 특이값과 동일합니다. 일반적으로  $\mathbf{M}$  행렬의 열은 변수가 계산에 미치는 효과를 균등화하기 위해 표준화됩니다. Golub와 Kahan 방법은 계산이 효율적입니다.



다음 자료가 다변량 방법에서 참조됩니다.

- Agresti, A. (2013). *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: John Wiley & Sons.
- Baglama, J., and Reichel, L. (2005). "Augmented implicitly restarted Lanczos bidiagonalization methods." *SIAM Journal on Scientific Computing* 27:19–42.
- Ballard, D. H. (1981). "Generalizing the Hough Transform to Detect Arbitrary Shapes." *Pattern Recognition* 13:111–122.
- Bartlett, M. S. (1937). "Properties of sufficiency and statistical tests." *Proceedings of the Royal Society of London, Series A* 160:268–282.
- Bartlett, M. S. (1954). "A Note on the Multiplying Factors for Various Chi Square Approximations." *Journal of the Royal Statistical Society, Series B* 16:296–298.
- Benzécri, J. P. (1979). "Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, addendum et erratum à [BIN. MULT]." *Cahiers de l'Analyse des Données* 4:377–378.
- Bentler, P. M. (1990). "Comparative Fit Indexes in Structural Models." *Psychological Bulletin* 107:238.
- Bentler, P. M., and Freeman, E. H. (1983). "Tests for Stability in Linear Structural Equation Systems." *Psychometrika* 48:143–145.
- Bernhardsson, E. (2013). "Approximate Nearest Neighbors in C++/Python optimized for memory usage and loading/saving to disk." <https://github.com/spotify/annoy>
- Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.
- Bollen, K. A., Harden, J. J., Ray, S., and Zavisca, J. (2014). "BIC and Alternative Bayesian Information Criteria in the Selection of Structural Equation Models." *Structural Equation Modeling* 21:1–19.
- Borg, I., and Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications*. 2nd ed. New York: Springer.
- Boulesteix, A.-L., and Strimmer, K. (2007). "Partial Least Squares: A Versatile Tool for the Analysis of High-Dimensional Genomic Data." *Briefings in Bioinformatics* 8:32–44.
- Browne, M. (2001). "An Overview of Analytic Rotation in Exploratory Factor Analysis." *Multivariate Behavioral Research* 36:111–150.
- Browne, M. W., and Cudeck, R. (1993). "Alternative Ways of Assessing Model Fit." In *Testing Structural Equation Models*, edited by K. A. Bollen, and J. S. Long, 136–162. Newbury Park, CA: Sage Publications.

- Chen, F. F. (2007). "Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance." *Structural Equation Modeling* 14:464–504.
- Chen, F., and Rohe, K. (2023). "A new basis for sparse principal component analysis." *Journal of Computational and Graphical Statistics* 1–14.
- Collins, L., and Lanza, S. (2010). *Latent Class and Latent Transition Analysis*. Hoboken NJ: John Wiley & Sons.
- Cox, I., and Gaudard, M. (2013). *Discovering Partial Least Squares with JMP*. Cary, NC: SAS Institute Inc.
- Cronbach, L. J. (1951). "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16:297–334.
- Cudeck, R., and MacCallum, R. C., eds. (2007). *Factor Analysis at 100, Historical Developments and Future Directions*. Mahwah, NJ: Lawrence Erlbaum Associates.
- de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: Guilford Press.
- De Jong, S. (1993). "SIMPLS: An Alternative Approach to Partial Least Squares Regression." *Chemometrics and Intelligent Laboratory Systems* 18:251–263.
- Denham, M. C. (1997). "Prediction Intervals in Partial Least Squares." *Journal of Chemometrics* 11:39–52.
- Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikstrom, C., and Wold, S. (2006). *Multi- and Megavariate Data Analysis Basic Principles and Applications (Part I)*. Chapter 4. Umetrics.
- Finkbeiner, C. (1979). "Estimation for the Multiple Factor Model when Data are Missing." *Psychometrika* 44:409–420.
- Fisher, L., and Van Ness, J. W. (1971). "Admissible Clustering Procedures." *Biometrika* 58:91–104.
- Florek, K., Lukaszewicz, J., Perkal, J., and Zubrzycki, S. (1951a). "Sur la liaison et la division des points d'un ensemble fini." *Colloquium Mathematicae* 2:282–285.
- Florek, K., Lukaszewicz, J., Perkal, J., and Zubrzycki, S. (1951b). "Taksonomia Wroclawska." *Przegląd Antropologiczny* 17:193–211.
- Frank, I. E., and Todeschini, T. (1994). *The Data Analysis Handbook*. New York: Elsevier.
- Friedman, J. H. (1989). "Regularized Discriminant Analysis." *Journal of the American Statistical Association* 84:165–175.
- Garthwaite, P. (1994). "An Interpretation of Partial Least Squares." *Journal of the American Statistical Association* 89:122–127.
- Golub, G. H., and Kahan, W. (1965). "Calculating the singular values and pseudo-inverse of a matrix." *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis* 2:205–224.
- Goodman, L. A. (1974). "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models." *Biometrika* 61:215–231.
- Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

- Halko, N., Martinsson, P. G., & Tropp, J. A. (2011). "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions." *SIAM Review* 53:217–288.
- Hancock, G. R., and Mueller, R. O. (2001). "Rethinking Construct Reliability within Latent Variable Systems." In *Structural Equation Modeling: Present and Future – A Festschrift in Honor of Karl Jöreskog*, edited by R. Cudeck, S. du Toit, and D. Sörbom, 195–216. Lincolnwood, IL: Scientific Software International.
- Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. Cambridge, MA: MIT Press.
- Harris, C. W., and Kaiser, H. F. (1964). "Oblique Factor Analytic Solutions by Orthogonal Transformation." *Psychometrika* 32:363–379.
- Hartigan, J. A. (1981). "Consistency of Single Linkage for High-Density Clusters." *Journal of the American Statistical Association* 76:388–394.
- Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer Verlag.
- Hinton, G. E., and Roweis, S. T. (2002). "Stochastic Neighbor Embedding." *Advances in Neural Information Processing Systems* 15:833–840.
- Hoskuldsson, A. (1988). "PLS Regression Methods." *Journal of Chemometrics* 2:211–228.
- Hoefding, W. (1948). "A Non-Parametric Test of Independence." *Annals of Mathematical Statistics* 19:546–557.
- Hu, L.-T., and Bentler, P. M. (1999). "Cutoff Criteria for Fit Indices in Covariance Structure Analysis: Conventional Criteria versus New Alternatives." *Structural Equation Modeling* 6:1–55.
- Huber, P. J. (1964). "Robust Estimation of a Location Parameter." *Annals of Mathematical Statistics* 35:73–101.
- Huber, P. J. (1973). "Robust Regression: Asymptotics, Conjecture, and Monte Carlo." *Annals of Statistics* 1:799–821.
- Huber, P. J., and Ronchetti, E. M. (2009). *Robust Statistics*. 2nd ed. Hoboken, NJ: John Wiley & Sons.
- Jackson, J. E. (2003). *A User's Guide to Principal Components*. Hoboken, NJ: John Wiley & Sons.
- Jardine, N., and Sibson, R. (1971). *Mathematical Taxonomy*. New York: John Wiley & Sons.
- Jöreskog, K. G. (1977). "Factor Analysis by Least-Squares and Maximum Likelihood Methods." In *Statistical Methods for Digital Computers*, edited by K. Enslein, A. Ralston, and H. Wilf, 125 - 165. New York: John Wiley & Sons.
- Kingma, D. P., and Ba, J. (2014). "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980*.
- Kline, R. B. (2016). *Principles and Practice of Structural Equation Methodology*. 4th ed. New York: The Guilford Press.
- Kohonen, T. (1989). *Self-Organization and Associative Memory*. 3rd ed. Vol. 8 of Springer Series in Information. Berlin: Springer-Verlag.

- Kohonen, T. (1990). "The Self-Organizing Map." *Proceedings of the IEEE* 78:1464–1480.
- Lindberg, W., Persson, J.-A., and Wold, S. (1983). "Partial Least-Squares Method for Spectrofluorimetric Analysis of Mixtures of Humic Acid and Ligninsulfonate." *Analytical Chemistry* 55:643–648.
- LeRoux, B., and Rouanet, H. (2010). Multiple Correspondence Analysis. Vol.07–163 of Sage University Paper Series on Quantitative Applications in the Social Sciences. Thousand Oaks, CA: Sage Publications.
- Maiti, S. S., and Mukherjee, B. N. (1991). "Two New Goodness-of-Fit Indices for Covariance Matrices with Linear Structures." *British Journal of Mathematical and Statistical Psychology* 44:153–180.
- Mardia, K., Kent, J., and Bibby, J. (1980). *Multivariate Analysis*. New York: Academic Press.
- Mason, R. L., and Young, J. C. (2002). *Multivariate Statistical Process Control with Industrial Applications*. Philadelphia: SIAM.
- May, J. P. (1992). *Simplicial Objects in Algebraic Topology*. Vol. 11. Chicago: University of Chicago Press.
- Maydeu-Olivares, A., Shi, D., and Rosseel, Y. (2017). "Assessing Fit in Structural Equation Models: A Monte-Carlo Evaluation of RMSEA Versus SRMR Confidence Intervals and Tests of Close Fit." *Structural Equation Model: A Multidisciplinary Journal* 25:389–402.
- McDonald, R. P. (1999). *Test Theory: A Unified Approach*. Mahwah, NJ: Erlbaum.
- McDonald, R. P., and Marsh, H. W. (1990). "Choosing a Multivariate Model: Noncentrality and Goodness of Fit." *Psychological Bulletin* 107:247–255.
- McInnes, L., Healy, J., and Melville, J. (2018). "UMAP: Uniform manifold approximation and projection for dimension reduction." *arXiv preprint arXiv:1802.03426*.
- McLachlan, G. J., and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: John Wiley & Sons.
- McQuitty, L. L. (1957). "Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies." *Educational and Psychological Measurement* 17:207–229.
- Milligan, G. W. (1980). "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms." *Psychometrika* 45:325–342.
- Nelson, P. R. C., Taylor, P. A., and MacGregor, J. F. (1996). "Missing Data Methods in PCA and PLS: Score calculations with incomplete observations." *Chemometrics and Intelligent Laboratory Systems* 35:45–65.
- Nunnally, J. C. (1978). *Psychometric theory*. 2nd ed. New York: McGraw-Hill.
- Penny, K. I. (1996). "Appropriate Critical Values When Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance." *Journal of the Royal Statistical Society, Series C* 45:73–81.
- Press, W. H, Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1998). *Numerical Recipes in C: The Art of Scientific Computing*. 2nd ed. Cambridge, England: Cambridge University Press.
- Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.

- Rodriguez, A., Reise, S. P., and Haviland, M. G. (2016). "Evaluating Bifactor Models: Calculating and Interpreting Statistical Indices." *Psychological Methods* 21:137.
- Saad, D. (1998). "Online algorithms and stochastic approximations." *Online Learning* 5(3):6.
- SAS Institute Inc. (1983). *SAS Technical Report A-108: Cubic Clustering Criterion*. Cary, NC: SAS Institute Inc. [https://support.sas.com/kb/22/add1/fusion\\_22540\\_1\\_a108\\_5903.pdf](https://support.sas.com/kb/22/add1/fusion_22540_1_a108_5903.pdf).
- SAS Institute Inc. (2023a). "The CALIS Procedure." *SAS/STAT<sup>®</sup> User's Guide*. Cary, NC: SAS Institute Inc. [https://go.documentation.sas.com/api/collections/pgmsascdc/9.4\\_3.5/docsets/statug/content/calis.pdf](https://go.documentation.sas.com/api/collections/pgmsascdc/9.4_3.5/docsets/statug/content/calis.pdf).
- SAS Institute Inc. (2023b). "The CANDISC Procedure." *SAS/STAT<sup>®</sup> User's Guide*. Cary, NC: SAS Institute Inc. [https://go.documentation.sas.com/api/collections/pgmsascdc/9.4\\_3.5/docsets/statug/content/candisc.pdf](https://go.documentation.sas.com/api/collections/pgmsascdc/9.4_3.5/docsets/statug/content/candisc.pdf).
- SAS Institute Inc. (2023c). "The FACTOR Procedure." *SAS/STAT<sup>®</sup> User's Guide*. Cary, NC: SAS Institute Inc. [https://go.documentation.sas.com/api/collections/pgmsascdc/9.4\\_3.5/docsets/statug/content/factor.pdf](https://go.documentation.sas.com/api/collections/pgmsascdc/9.4_3.5/docsets/statug/content/factor.pdf).
- SAS Institute Inc. (2023d). "The FASTCLUS Procedure." *SAS/STAT<sup>®</sup> User's Guide*. Cary, NC: SAS Institute Inc. [https://go.documentation.sas.com/api/collections/pgmsascdc/9.4\\_3.5/docsets/statug/content/fastclus.pdf](https://go.documentation.sas.com/api/collections/pgmsascdc/9.4_3.5/docsets/statug/content/fastclus.pdf).
- SAS Institute Inc. (2023e). "The MIXED Procedure." *SAS/STAT<sup>®</sup> User's Guide*. Cary, NC: SAS Institute Inc. [https://go.documentation.sas.com/api/collections/pgmsascdc/9.4\\_3.5/docsets/statug/content/mixed.pdf](https://go.documentation.sas.com/api/collections/pgmsascdc/9.4_3.5/docsets/statug/content/mixed.pdf).
- SAS Institute Inc. (2023f). "The PLS Procedure." *SAS/STAT<sup>®</sup> User's Guide*. Cary, NC: SAS Institute Inc. [https://go.documentation.sas.com/api/collections/pgmsascdc/9.4\\_3.5/docsets/statug/content/pls.pdf](https://go.documentation.sas.com/api/collections/pgmsascdc/9.4_3.5/docsets/statug/content/pls.pdf).
- SAS Institute Inc. (2023g). "The VARCLUS Procedure." *SAS/STAT<sup>®</sup> User's Guide*. Cary, NC: SAS Institute Inc. [https://go.documentation.sas.com/api/collections/pgmsascdc/9.4\\_3.5/docsets/statug/content/varclus.pdf](https://go.documentation.sas.com/api/collections/pgmsascdc/9.4_3.5/docsets/statug/content/varclus.pdf).
- Schafer, J., and Strimmer, K. (2005). "A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics." *Statistical Applications in Genetics and Molecular Biology* 4 Article 32.
- Sneath, P. H. A. (1957). "The Application of Computers to Taxonomy." *Journal of General Microbiology* 17:201–226.
- Sokal, R. R., and Michener, C. D. (1958). "A Statistical Method for Evaluating Systematic Relationships." *University of Kansas Science Bulletin* 38:1409–1438.
- Steiger, J. H. (1989). *EzPATH: A Supplementary Module for SYSTAT and SYGRAPH*. Evanston, IL: Systat.

- Steiger, J. H. (1990). "Structural Model Evaluation and Modification: An Interval Estimation Approach." *Multivariate Behavioral Research* 25:173–180.
- Tobias, R. D. (1995). "An Introduction to Partial Least Squares Regression." In *Proceedings of the Twentieth Annual SAS Users Group International Conference*, 1250–1257. Cary, NC: SAS Institute Inc.  
<https://support.sas.com/resources/papers/proceedings-archive/SUGI95/Sugi-95-210%20Tobias.pdf>.
- Tracy, N. D., Young, J. C., and Mason, R. R. (1992). "Multivariate Control Charts for Individual Observations." *Journal of Quality Technology* 24:88–95.
- Umetrics. (1995). *Multivariate Analysis (3-day course)*. Winchester, MA.
- Van der Maaten, L. (2014). "Accelerating t-SNE using tree-based algorithms." *The Journal of Machine Learning Research* 15:3221–3245.
- Van der Maaten, L., and Hinton, G. E. (2008). "Visualizing Data using t-SNE." *Journal of Machine Learning Research* 9(11).
- Waern, Y. (1972). "Structure in Similarity Matrices: A Graphic Approach." *Scandinavian Journal of Psychology* 13:5–16.
- West, S. G., Taylor, A. B., and Wu, W. (2012). "Model Fit and Model Selection in Structural Equation Modeling." In *Handbook of Structural Equation Modeling*, edited by R. H. Hoyle, 209–231. New York: The Guilford Press.
- White, K. P., Jr., Kundu, B., and Mastrangelo, C. M. (2008). "Classification of Defect Clusters on Semiconductor Wafers Via the Hough Transform." *IEEE Transactions on Semiconductor Manufacturing* 21:272–278.
- Wold, S. (1994). "PLS for Multivariate Linear Modeling." In *QSAR: Chemometric Methods in Molecular Design. Methods and Principles in Medicinal Chemistry*, edited by H. van de Waterbeemd, pp. 195–218. Weinheim, Germany: Verlag-Chemie.
- Wold, S., Sjostrom, M., and Eriksson, L. (2001). "PLS-Regression: A Basic Tool of Chemometrics." *Chemometrics and Intelligent Laboratory Systems* 58:109–130.