



버전 17

예측 및 전문 모델링

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

Marcel Proust

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

17.0

The correct bibliographic citation for this manual is as follows: JMP Statistical Discovery LLC 2022. *JMP® 17 Predictive and Specialized Modeling*. Cary, NC: JMP Statistical Discovery LLC

JMP® 17 Predictive and Specialized Modeling

Copyright © 2022, JMP Statistical Discovery LLC, Cary, NC, USA

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

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

October 2022

JMP® and all other JMP Statistical Discovery LLC product or service names are registered trademarks or trademarks of 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 license information about third-party software distributed with JMP software, refer to <http://support.sas.com/thirdpartylicenses>.

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

jmp.com/getstarted

목차

예측 및 전문 모델링

1	JMP 알아보기	15
	설명서 및 추가 리소스	
	JMP 설명서에 사용되는 서식 규칙	17
	JMP 도움말	18
	JMP 설명서 라이브러리	18
	JMP 학습을 위한 추가 리소스	24
	JMP 검색	25
	JMP 자습서	25
	샘플 데이터 테이블	25
	통계 및 JSL 용어에 대해 알아보기	26
	JMP 팁 및 힌트 알아보기	26
	JMP 툴팁	26
	JMP 사용자 커뮤니티	26
	무료 온라인 Statistical Thinking 교육 과정	27
	JMP 새로운 사용자를 위한 입문 키트	27
	Statistics Knowledge 포털	27
	JMP 교육 과정	27
	사용자가 작성한 JMP 설명서	27
	JMP 시작하기 창	27
	JMP 기술 지원	28
2	예측 및 전문 모델링 소개	29
	모델링 기법 개요	
3	신경망	31
	노드와 층을 사용하여 비선형 모형 적합	
	신경망 플랫폼 개요	33
	신경망의 예	33
	신경망 플랫폼 시작	36
	신경망 시작 창	36
	모형 시작 제어판	38
	신경망의 검증 방법	39

은닉층 구조	40
부스트	40
신경망 적합 옵션	41
모형 보고서	43
훈련 및 검증 적합 측도	43
혼동 통계량	44
신경망 플랫폼 옵션	44
신경망 모형 옵션	45

4 파티션 모형

결정 트리를 사용하여 데이터 탐색 및 모델링

파티션 플랫폼 개요	49
파티션 플랫폼의 예	49
파티션 플랫폼 시작	52
결측값 정보화	53
분할 보고서	54
분할 그림	54
분할 버튼	56
적합 통계량 요약	56
노드 보고서	57
파티션 플랫폼 옵션	58
적합 상세 정보 표시	61
수익 행렬 지정	63
ROC 곡선	64
향상도 곡선	65
파티션 검증	67
파티션 플랫폼의 추가 예	68
연속형 반응의 예	68
결측값 정보화의 예	71
수익 행렬 및 결정 행렬 보고서의 예	74
파티션 플랫폼에 대한 통계 상세 정보	77
반응 및 요인에 대한 통계 상세 정보	77
분할 기준에 대한 통계 상세 정보	77

5 붓스트랩 포레스트

여러 트리를 평균화하여 모형 적합

붓스트랩 포레스트 플랫폼 개요	81
범주형 반응이 있는 붓스트랩 포레스트의 예	81
연속형 반응이 있는 붓스트랩 포레스트의 예	84
붓스트랩 포레스트 플랫폼 시작	86

규격 창	88
붓스트랩 포레스트 보고서	90
모형 검증 집합 요약	91
규격	91
전체 통계량	91
누적 검증	93
트리별 요약	93
붓스트랩 포레스트 플랫폼 옵션	94
6 부스티드 트리	97
각각 이전 계층에 기반한 여러 트리 계층 적합	
부스티드 트리 플랫폼 개요	99
범주형 반응이 있는 부스티드 트리의 예	99
연속형 반응이 있는 부스티드 트리의 예	101
부스티드 트리 플랫폼 시작	103
규격 창	105
부스티드 트리 보고서	107
모형 검증 집합 요약	109
규격	109
전체 통계량	109
누적 검증	110
부스티드 트리 플랫폼 옵션	110
부스티드 트리 플랫폼에 대한 통계 상세 정보	112
과대적합 벌점에 대한 통계 상세 정보	112
7 K 최근접 이웃	113
인접한 관측값을 사용하여 반응 값 예측	
K 최근접 이웃 플랫폼 개요	115
범주형 반응이 있는 K 최근접 이웃의 예	115
연속형 반응이 있는 K 최근접 이웃의 예	117
K 최근접 이웃 플랫폼 시작	119
K 최근접 이웃 보고서	120
연속형 반응	121
범주형 반응	121
K 최근접 이웃 플랫폼 옵션	122
모형 적합 옵션	123
8 Naive Bayes	125
확률적 가정을 사용하여 관측값 분류	
Naive Bayes 플랫폼 개요	127
Naive Bayes 의 예	127

Naive Bayes 플랫폼 시작	130
Naive Bayes 보고서	131
적합 상세 정보 보고서	131
반응 열 보고서	132
혼동 행렬 보고서	132
ROC 곡선	132
Naive Bayes 플랫폼 옵션	132
Naive Bayes 의 추가 예	134
Naive Bayes 플랫폼에 대한 통계 상세 정보	135
Naive Bayes 알고리즘에 대한 통계 상세 정보	135
저장된 확률 계산식에 대한 통계 상세 정보	136
9 서포트 벡터 머신	139
서포트 벡터를 사용하여 반응 범주 예측	
서포트 벡터 머신 개요	141
서포트 벡터 머신의 예	141
서포트 벡터 머신 플랫폼 시작	144
모형 시작 제어판	144
서포트 벡터 머신 보고서	147
모형 비교 보고서	147
서포트 벡터 머신 모형 보고서	148
서포트 벡터 머신 플랫폼 옵션	150
서포트 벡터 머신 모형 보고서 옵션	151
서포트 벡터 머신의 추가 예	152
10 모형 선별	155
다중 예측 모형 적합 및 비교	
모형 선별의 예	157
모형 선별 플랫폼 시작	159
모형 선별 보고서	162
폴드 간 요약 보고서	162
훈련, 검증 및 테스트 적합 측도	163
모형 선별 플랫폼 옵션	164
결정 임계값 보고서	167
11 모형 비교	169
적합 모형의 예측 능력 비교	
모형 비교의 예	171
모형 비교 플랫폼 시작	174
모형 비교 보고서	175
모형 비교 플랫폼 옵션	176

연속형 반응	176
범주형 반응	176
모형 비교의 추가 예	177
12 검증 열 생성	181
데이터를 훈련, 검증 및 테스트 데이터 집합으로 나누기	
검증 열 생성 플랫폼 개요	183
검증 열 생성 플랫폼의 예	183
검증 열 생성 플랫폼 시작	185
검증 열 생성 보고서	188
비율 또는 상대 비율 지정	188
절단점 설정	188
폴드 수 설정	189
옵션	189
시작	189
검증 열 생성 플랫폼 옵션	189
검증 열 생성 플랫폼의 추가 예	190
13 계산식 저장소	193
모형 관리 및 스코어링 코드 생성	
계산식 저장소 플랫폼 개요	195
계산식 저장소의 예	195
계산식 저장소 플랫폼 시작	196
예측 계산식을 계산식 저장소에 게시하는 플랫폼	196
계산식 저장소 플랫폼 옵션	197
계산식 저장소 모형 옵션	199
계산식 저장소 플랫폼에서 스코어링 코드 생성	199
14 곡선 적합	201
기본 제공 비선형 모형을 데이터에 적합	
곡선 적합 플랫폼 개요	203
곡선 적합 플랫폼의 예	203
곡선 적합 플랫폼 시작	206
곡선 적합 보고서	208
모형 비교 보고서	208
모형 적합 보고서	209
곡선 적합 옵션	210
모형 없는 비교 보고서	212
모형 적합 옵션	213
병렬 검정	215
모수 추정값 비교	217

동등성 검정	218
곡선 적합 플랫폼의 추가 예	219
곡선 적합 플랫폼에 대한 통계 상세 정보	221
곡선 적합 모형에 대한 통계 상세 정보	222
용해 곡선 분석에 대한 통계 상세 정보	232
15 비선형 회귀	235
데이터에 사용자 비선형 모형 적합	
비선형 플랫폼의 예	237
비선형 플랫폼 시작	239
비선형 적합 보고서	240
비선형 플랫폼 옵션	244
모형 라이브러리를 사용하여 계산식 생성	248
비선형 모형 라이브러리 사용자 정의	250
비선형 플랫폼의 추가 예	251
최대 가능도의 예	252
이항 오차가 있는 프로빗 모형의 예	255
Poisson 손실 함수의 예	255
모수 한계 설정의 예	257
왼쪽 중도절단 데이터 분석의 예	260
Weibull 손실 함수 적합의 예	261
단순 생존 분포 적합의 예	264
비선형 플랫폼에 대한 통계 상세 정보	266
프로파일 가능도 신뢰 한계에 대한 통계 상세 정보	266
사용자 손실 함수에 대한 통계 상세 정보	267
도함수에 대한 통계 상세 정보	268
유효 비선형 모델링에 대한 통계 상세 정보	270
16 함수 데이터 탐색기	273
함수 데이터 분석 수행	
함수 데이터 탐색기 플랫폼 개요	275
함수 데이터 탐색기의 예	275
함수 데이터 탐색기 플랫폼 시작	279
함수 데이터 탐색기 보고서	282
모형 보고서	283
함수 데이터 탐색기 플랫폼 옵션	288
함수 데이터 탐색기 그룹 옵션	290
데이터 처리 옵션	291
모형 보고서 옵션	295
함수 데이터 탐색기 플랫폼의 추가 예	297

여러 함수 공정의 예	297
함수 DOE 분석의 예	301
함수 데이터 탐색기 플랫폼에 대한 통계 상세 정보	303
함수 모형 적합에 대한 통계 상세 정보	304
함수 요약 보고서에 대한 통계 상세 정보	304
17 가우스 과정	307
평활 모형을 사용하여 데이터 적합	
가우스 과정 플랫폼의 예	309
가우스 과정 플랫폼 시작	311
가우스 과정 보고서	312
실제값 대 예측값 그림	312
모형 보고서	312
주변 모형 그림	313
가우스 과정 플랫폼 옵션	313
가우스 과정 플랫폼의 추가 예	314
가우스 과정 모형의 예	314
범주형 예측 변수가 있는 가우스 과정 모형의 예	316
가우스 과정 플랫폼에 대한 통계 상세 정보	318
연속형 예측 변수가 있는 모형에 대한 통계 상세 정보	318
범주형 예측 변수가 있는 모형에 대한 통계 상세 정보	319
분산 계산식 파라미터화에 대한 통계 상세 정보	320
모형 적합에 대한 통계 상세 정보	320
18 시계열 분석	321
시계열 모형 및 전이 함수 적합	
시계열 플랫폼 개요	323
시계열 플랫폼의 예	323
시계열 플랫폼 시작	326
시계열 보고서	327
시계열 그래프	328
시계열 기본 진단 차트	328
시계열 플랫폼 옵션	330
시계열 진단	330
차분 및 분해	331
시계열 모형	333
추가 옵션	340
추가 시계열 보고서	341
차분 보고서	341
분해 보고서	342

모형 비교 보고서	344
모형 보고서	346
상태 공간 평활 보고서	350
전이 함수 모형 보고서	352
스펙트럼 밀도 보고서	353
시계열 플랫폼의 추가 예	354
시간 ID 열 생성의 예	354
Box-Cox 변환의 예	359
홀드백 집합을 사용한 예	361
시계열 플랫폼에 대한 통계 상세 정보	363
스펙트럼 밀도에 대한 통계 상세 정보	363
X-11 분해에 대한 통계 상세 정보	364
지수 평활 모형에 대한 통계 상세 정보	364
ARIMA 모형에 대한 통계 상세 정보	368
전이 함수에 대한 통계 상세 정보	369
19 시계열 예측	371
다중 시계열 예측	
시계열 예측의 예	373
시계열 예측 플랫폼 시작	375
데이터 형식	375
시계열 예측 보고서	376
시간 패턴 분석	376
시간에 대한 그룹 요약	377
모델링 규격	378
시계열 예측 플랫폼 옵션	381
시계열 예측 모형 보고서	382
20 매칭 쌍 분석	383
동일한 개체에 대한 측정값 비교	
매칭 쌍 플랫폼 개요	385
매칭 쌍 비교의 예	385
매칭 쌍 플랫폼 시작	386
여러 Y 열	387
매칭 쌍 보고서	388
차이 그림 및 보고서	388
그룹 간	389
매칭 쌍 플랫폼 옵션	389
그룹 간 매칭 쌍 비교의 예	390
매칭 쌍 플랫폼에 대한 통계 상세 정보	391

매칭 쌍의 그래픽에 대한 통계 상세 정보	391
반응 상관에 대한 통계 상세 정보	392
21 이상치 탐색	395
단변량 및 다변량 데이터에서 이상치 찾기	
이상치 탐색 개요	397
분위수 범위 이상치	397
로버스트 적합 이상치	397
로버스트 PCA 이상치	397
K 최근접 이웃 이상치	398
이상치 탐색의 예	399
이상치 탐색 플랫폼 시작	402
이상치 탐색 보고서	405
분위수 범위 이상치 보고서	406
로버스트 적합 이상치 보고서	410
로버스트 PCA 이상치 보고서	413
K 최근접 이웃 보고서	414
이상치 탐색 플랫폼 옵션	415
이상치 탐색 플랫폼의 추가 예	416
22 결측값 탐색	419
결측값 탐색 및 대치	
결측값 탐색 플랫폼 개요	421
다변량 정규 대치법	421
다변량 SVD 대치법	421
다변량 RPCA 대치법	422
자동 데이터 결측값 대치	422
결측값 탐색 플랫폼의 예	423
결측값 탐색 플랫폼 시작	425
결측값 탐색 보고서	425
명령	426
결측값 대치 보고서	428
결측값 탐색 플랫폼 옵션	428
23 패턴 탐색	431
데이터에서 일반적이지 않은 특징이나 패턴 탐색	
패턴 탐색 플랫폼 개요	433
패턴 탐색 플랫폼의 예	433
패턴 탐색 플랫폼 시작	435
패턴 탐색 보고서	436
제어판	436

단변량 요약	436
단변량 패턴	436
패턴 탐색 옵션	437
단변량 패턴 보고서 옵션	440
패턴 탐색 플랫폼의 추가 예	441
패턴 탐색 플랫폼에 대한 통계 상세 정보	442
24 반응 변수 선별	445
대규모 데이터의 여러 반응 변수 검정	
반응 변수 선별 플랫폼 개요	447
반응 변수 선별의 예	448
반응 변수 선별 플랫폼 시작	452
반응 변수 선별 보고서	455
반응 변수 선별 그림	455
반응 변수 선별 보고서 테이블	457
반응 변수 선별 플랫폼 옵션	460
평균 데이터 테이블	463
평균 차이 데이터 테이블	464
p 값 데이터 테이블	466
모형 적합의 반응 변수 선별 분석법	466
모형 적합의 반응 변수 선별 시작	467
반응 변수 선별 적합 보고서	468
반응 변수 선별 적합 옵션	470
반응 변수 선별의 추가 예	472
실제적 유의성 및 동등성 검정의 예	472
최대 LogWorth 옵션의 예	474
로버스트 적합의 예	476
반응 변수 선별 분석법의 예	479
화산 (Volcano) 그림의 예	480
반응 변수 선별 플랫폼에 대한 통계 상세 정보	482
25 예측 변수 선별	483
유의한 효과를 위한 여러 예측 변수 선별	
예측 변수 선별 플랫폼 개요	485
예측 변수 선별의 예	485
예측 변수 선별 플랫폼 시작	487
예측 변수 선별 보고서	488
예측 변수 선별 플랫폼 옵션	488

26	연관성 분석	489
	장바구니 분석 수행	
	연관성 분석 플랫폼 개요	491
	연관성 분석 플랫폼의 예	492
	연관성 분석 플랫폼 시작	494
	데이터 형식	495
	연관성 분석 보고서	495
	빈발 항목 집합	496
	규칙	496
	연관성 분석 플랫폼 옵션	497
	특이값 분해 개요	497
	SVD 보고서	498
	SVD 보고서 옵션	499
	주제 분석 개요	500
	주제 분석 보고서	500
	주제 분석 보고서 옵션	501
	연관성 분석 플랫폼의 추가 예	501
	연관성 분석 플랫폼에 대한 통계 상세 정보	504
	빈발 항목 집합 생성에 대한 통계 상세 정보	504
	성능 측도에 대한 통계 상세 정보	504
27	공정 기록 탐색기	507
	복잡한 공정 기록에서 문제 성분 식별	
	공정 기록 탐색기의 예	509
	공정 기록 탐색기 플랫폼 시작	511
	데이터 형식	512
	공정 기록 탐색기 보고서	512
	공정 기록 탐색기 플랫폼 옵션	512
A	통계 상세 정보	517
	예측 및 전문 모델링	
	엔트로피 R^2	519
	JMP 모델링의 검증	519
	검증 열 역할	520
	K 폴드 및 홀드백 검증	522
	트리 방법의 예측 확률	523
B	참조 자료	525
C	기술 라이선스 고지 사항	529

1 장

JMP 알아보기 설명서 및 추가 리소스

설명서 서식 규칙, 각 JMP 문서에 대한 설명, 도움말 시스템, 기타 지원 제공 위치 등 JMP 설명서에 대해 알아봅니다.

목차

JMP 설명서에 사용되는 서식 규칙	17
JMP 도움말	18
JMP 설명서 라이브러리	18
JMP 학습을 위한 추가 리소스	24
JMP 검색	25
JMP 자습서	25
샘플 데이터 테이블	25
통계 및 JSL 용어에 대해 알아보기	26
JMP 팁 및 힌트 알아보기	26
JMP 톨팁	26
JMP 사용자 커뮤니티	26
무료 온라인 Statistical Thinking 교육 과정	27
JMP 새로운 사용자를 위한 입문 키트	27
Statistics Knowledge 포털	27
JMP 교육 과정	27
사용자가 작성한 JMP 설명서	27
JMP 시작하기 창	27
JMP 기술 지원	28

JMP 설명서에 사용되는 서식 규칙

설명서 자료가 가리키는 화면 정보를 쉽게 알아볼 수 있도록 다음과 같은 서식 규칙이 사용됩니다.

- 샘플 데이터 테이블 이름, 열 이름, 경로 이름, 파일 이름, 파일 확장자 및 폴더는 **Helvetica** (또는 [sans-serif online](#)) 글꼴로 표시됩니다.
- 코드는 **Lucida Sans Typewriter**(또는 [monospace online](#)) 글꼴로 표시됩니다.
- 코드 출력은 **Lucida Sans Typewriter** 기울임꼴 (또는 [monospace italic online](#)) 글꼴로 표시되고 앞의 코드보다 더 많은 들여쓰기가 적용됩니다.
- **Helvetica bold**(또는 [bold sans-serif online](#)) 서식은 작업을 수행하기 위해 사용자가 선택하는 다음과 같은 항목을 나타냅니다.
 - 버튼
 - 체크박스
 - 명령
 - 선택 가능한 목록 이름
 - 메뉴
 - 옵션
 - 탭 이름
 - 텍스트 상자
- 다음 항목은 기울임꼴로 표시됩니다.
 - 중요하거나 JMP 와 관련된 정의가 있는 단어 또는 구
 - 설명서 제목
 - 변수
- JMP Pro 에만 해당되는 기능에는 JMP Pro 아이콘  이 표시됩니다. JMP Pro 기능의 개요는 jmp.com/software/pro 에서 확인할 수 있습니다.

참고 : 참고 섹션에는 특수 정보와 제한 사항이 표시됩니다.

팁 : 팁 섹션에는 유용한 정보가 표시됩니다.

JMP 도움말

"도움말" 메뉴의 "JMP 도움말"에서는 JMP 기능, 통계적 방법 및 JSL(JMP Scripting Language)에 대한 정보를 검색할 수 있습니다. 다음과 같은 몇 가지 방법으로 JMP 도움말을 열 수 있습니다.

- Windows 에서 **도움말 > JMP 도움말**을 선택하여 JMP 도움말을 검색하고 봅니다.
- Windows 에서 F1 키를 눌러 기본 브라우저로 도움말 시스템을 엽니다.
- 데이터 테이블 또는 보고서 창의 특정 부분에 대한 도움말을 확인합니다. **도구** 메뉴에서 도움말 도구 를 선택한 후 데이터 테이블 또는 보고서 창의 아무 곳이나 클릭하면 해당 영역에 대한 도움말이 표시됩니다.
- JMP 창 내에서 **도움말** 버튼을 클릭합니다.

참고: JMP 도움말은 인터넷이 연결되어 있어야 사용할 수 있습니다. 인터넷이 연결되어 있지 않으면 **도움말 > JMP 설명서 라이브러리**를 선택하여 단일 PDF 파일에서 모든 설명서를 검색할 수 있습니다. 자세한 내용은 "[JMP 설명서 라이브러리](#)"에서 확인하십시오.

JMP 설명서 라이브러리

도움말 시스템 콘텐츠는 JMP 설명서 라이브러리라는 단일 PDF 파일로도 제공됩니다. 이 파일을 열려면 **도움말 > JMP 설명서 라이브러리**를 선택합니다. Documentation PDF Files 추가기능을 다운로드하여 JMP 라이브러리에 있는 각 문서의 개별 PDF 파일을 검색할 수도 있습니다. <https://community.jmp.com>에서 사용 가능한 추가기능을 다운로드하십시오.

다음 표에서는 JMP 라이브러리에 포함된 각 문서의 용도와 내용을 설명합니다.

문서 제목	문서 용도	문서 내용
JMP 살펴보기	JMP에 익숙하지 않다면 이 설명서부터 시작하십시오.	JMP에 대해 소개하고 데이터 생성 및 분석을 시작하기 위한 정보를 제공합니다. 또한 결과를 공유하는 방법도 알아볼 수 있습니다.
JMP 사용	JMP 데이터 테이블과 기본적인 작업을 수행하는 방법에 대해 알아봅니다.	데이터 가져오기, 열 특성 수정, 데이터 정렬, SAS 연결 등을 비롯하여 JMP의 모든 영역에 걸친 일반적인 JMP 개념 및 기능을 다룹니다.

문서 제목	문서 용도	문서 내용
기본 분석	이 문서를 사용하여 기본적인 분석을 수행합니다.	"분석" 메뉴의 다음 플랫폼에 대해 설명합니다. • 분포 • X로 Y 적합 • 테이블 생성 • 텍스트 탐색기 분석 > X로 Y 적합을 통해 이변량 분석, 일원 ANOVA 및 분할 분석을 수행하는 방법을 다룹니다. 붓스트랩을 사용하여 표집 분포에 근사한 값을 산출하는 방법과 시뮬레이션 플랫폼을 사용하여 모수 재표집을 수행하는 방법도 포함되어 있습니다.
Essential Graphing	데이터에 이상적인 그래프를 찾습니다.	"그래프" 메뉴의 다음 플랫폼에 대해 설명합니다. • 그래프 빌더 • 3D 산점도 • 등고선 그림 • 버블 그림 • 평행 그림 • 셀 그림 • 산점도 행렬 • 삼원 그림 • 트리맵 • 차트 • 중첩 그림 이 설명서에서는 배경 맵 및 사용자 맵을 생성하는 방법도 다룹니다.
Profilers	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.

문서 제목	문서 용도	문서 내용
실험 설계 가이드	실험 설계 방법을 알아보고 적절한 표본 크기를 결정합니다.	"DOE" 메뉴의 모든 항목을 다룹니다.
선형 모형 적합	모형 적합 플랫폼과 이 플랫폼의 다양한 분석법에 대해 알아봅니다.	"분석" 메뉴의 모형 적합 플랫폼에서 사용할 수 있는 다음 분석법에 대해 설명합니다. • 표준 최소 제곱 • 단계별 • 일반화 회귀 • 혼합 모형 • 일반화 선형 혼합 모형 • MANOVA • 로그 선형 분산 • 명목형 로지스틱 • 순서형 로지스틱 • 일반화 선형 모형

문서 제목	문서 용도	문서 내용
예측 및 전문 모델링	추가 모델링 기법에 대해 알아봅니다.	분석 > 예측 모델링 메뉴의 다음 플랫폼에 대해 설명합니다. • 신경망 • 파티션 • 붓스트랩 포레스트 • 부스티드 트리 • K 최근접 이웃 • Naive Bayes • 서포트 벡터 머신 • 모형 비교 • 모형 선별 • 검증 열 생성 • 계산식 저장소 분석 > 전문 모델링 메뉴의 다음 플랫폼에 대해 설명합니다. • 곡선 적합 • 비선형 • 함수 데이터 탐색기 • 가우스 과정 • 시계열 • 매칭 쌍 분석 > 선별 메뉴의 다음 플랫폼에 대해 설명합니다. • 이상치 탐색 • 결측값 탐색 • 패턴 탐색 • 반응 변수 선별 • 예측 변수 선별 • 연관성 분석 • 공정 기록 탐색기

문서 제목	문서 용도	문서 내용
다변량 방법	몇 개의 변수를 동시에 분석하기 위한 기법을 알아봅니다.	분석 > 다변량 방법 메뉴의 다음 플랫폼에 대해 설명합니다. • 다변량 • 주성분 • 판별 • 부분 최소 제곱 • 다중 대응 분석 • 구조 방정식 모형 • 요인 분석 • 다차원 척도법 • 다변량 임베딩 • 항목 분석 분석 > 군집화 메뉴의 다음 플랫폼에 대해 설명합니다. • 계층적 군집화 • K 평균 군집화 • 정규 혼합 • 잠재 계층 분석 • 변수 군집화

문서 제목	문서 용도	문서 내용
품질 및 공정 방법	공정 평가 및 개선을 위한 도구에 대해 알아봅니다.	분석 > 품질 및 공정 메뉴의 다음 플랫폼에 대해 설명합니다. • 관리도 빌더 및 개별 관리도 • 측정 시스템 분석 (EMP 및 유형 1 게이지) • 계량형 / 계수형 게이지 차트 • 공정 변수 선별 • 공정 능력 • 모형 기반 다변량 관리도 • 레거시 관리도 • 파레토도 • 다이어그램 • 규격 한계 관리 • OC 곡선
Reliability and Survival Methods	제품 또는 시스템의 신뢰도 평가 및 향상 방법과 사람 및 제품의 생존 데이터 분석 방법을 알아봅니다.	분석 > 신뢰성 및 생존 메뉴의 다음 플랫폼에 대해 설명합니다. • 수명 분포 • 수명 분포 적합 • 누적 손상 • 재발 분석 • 열화 • 반복 측정 열화 • 파괴 열화 • 신뢰도 예측 • 신뢰도 성장 • 신뢰도 블록 다이어그램 • 수리 가능 시스템 시뮬레이션 • 생존 • 모수 생존 모형 적합 • 비례 위험 모형 적합

문서 제목	문서 용도	문서 내용
Consumer Research	소비자 선호도를 연구하고 해당 정보를 사용하여 보다 나은 제품 및 서비스를 개발하기 위한 방법을 알아봅니다.	분석 > 소비자 조사 메뉴의 다음 플랫폼에 대해 설명합니다. • 범주형 • 선택 • 최대차이 • Uplift • 다중 요인 분석
유전학	JMP에서 유전자 데이터를 분석하고, 해당 데이터를 사용하여 최적의 유전자 교배를 예측하기 위한 육종 프로그램을 시뮬레이션하는 데 사용할 수 있는 방법을 알아봅니다.	분석 > 유전학 메뉴의 다음 플랫폼에 대해 설명합니다. • 표지자 통계량 • 표지자 시뮬레이션
Scripting Guide	강력한 JSL(JMP 스크립트 언어)의 활용 방법을 알아봅니다.	스크립트 작성/디버깅, 데이터 테이블 조작, 표시 상자 생성 및 JMP 응용 프로그램 생성 등의 다양한 주제를 다룹니다.
JSL Syntax Reference	다양한 JSL 함수 및 인수, 그리고 개체 및 표시 상자로 보내는 메시지에 대해 알아봅니다.	JSL 명령의 구문, 예제 및 참고 사항이 포함되어 있습니다.

JMP 학습을 위한 추가 리소스

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

- ["JMP 검색"](#)
- ["JMP 자습서"](#)
- ["샘플 데이터 테이블"](#)
- ["통계 및 JSL 용어에 대해 알아보기"](#)
- ["JMP 팁 및 힌트 알아보기"](#)
- ["JMP 툴팁"](#)
- ["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 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.

JMP 자습서

도움말 > 자습서를 선택하여 JMP 자습서에 액세스할 수 있습니다. **자습서** 메뉴의 첫 번째 항목은 **자습서 디렉터리**입니다. 이 항목을 클릭하면 모든 자습서가 범주별로 그룹화되어 있는 새 창이 열립니다.

JMP에 익숙하지 않다면 **초보자 자습서**부터 시작하십시오. 이 자습서에서는 JMP 인터페이스를 단계별로 안내하며 JMP를 사용하는 데 필요한 기본 사항을 설명합니다.

나머지 자습서는 실험 설계, 표본 평균과 상수의 비교 같은 JMP의 특정 측면을 이해하는 데 유용합니다.

샘플 데이터 테이블

JMP 설명서 모음에 포함된 모든 예에서는 샘플 데이터를 사용합니다. 샘플 데이터 디렉터리를 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 디렉터리에 설치되어 있습니다.

Windows: C:\Program Files\SAS\JMP\17\Samples\Data

macOS: \Library\Application Support\JMP\17\Samples\Data

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

샘플 데이터를 사용한 예를 보려면 **도움말 > 샘플 인덱스**를 선택하고 "교육 자료" 섹션으로 이동하십시오. 교육 자료에 대한 자세한 내용은 jmp.com/tools에서 확인하십시오.

통계 및 JSL 용어에 대해 알아보기

통계 용어에 대한 도움말을 보려면 "도움말 > 통계 분석 인덱스"를 선택합니다. JSL 스크립트 및 예제에 대한 도움말을 보려면 **도움말 > 스크립트 인덱스**를 선택합니다.

통계 분석 인덱스 통계 용어에 대한 정의를 제공합니다.

스크립트 인덱스 JSL 함수, 개체 및 표시 상자에 대한 정보를 검색할 수 있습니다. "스크립트 인덱스"에서는 샘플 스크립트를 편집 및 실행하고 명령에 대한 도움말을 볼 수도 있습니다.

JMP 팁 및 힌트 알아보기

JMP를 처음 시작할 때는 "오늘의 유익한 정보" 창이 표시됩니다. 이 창에서는 JMP를 사용하기 위한 팁을 제공합니다.

"오늘의 유익한 정보" 기능을 해제하려면 **시작할 때 정보 표시** 체크박스를 선택 해제하십시오. 이 창을 다시 보려면 **도움말 > 오늘의 유익한 정보**를 선택하십시오. 또는 "환경 설정" 창을 사용하여 이 기능을 해제할 수도 있습니다.

JMP 톨팁

JMP에서 다음과 같은 항목을 커서로 가리키면 설명 톨팁 (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 톨팁을 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 웹 > 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 [jmp.com/statisticalthinking](https://community.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 좀더 편하게 소프트웨어 사용 방법을 익히고 세계 최대 규모의 JMP 사용자 온라인 커뮤니티와 연결할 수 있습니다. 자세한 내용은 [jmp.com/welcome](https://community.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 [jmp.com/skp](https://community.jmp.com/skp) 에서 확인하십시오.

JMP 교육 과정

SAS에서는 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 [jmp.com/training](https://community.jmp.com/training) 에서 확인하십시오.

사용자가 작성한 JMP 설명서

JMP 웹 사이트에서는 JMP 사용자가 작성한 추가 JMP 사용 설명서가 제공됩니다. 자세한 내용은 [jmp.com/books](https://community.jmp.com/books) 에서 확인하십시오.

JMP 시작하기 창

JMP 또는 데이터 분석에 익숙하지 않다면 먼저 "JMP 시작하기" 창을 살펴보세요. 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.

- Windows 에서 JMP 를 열 때 자동으로 "JMP 시작하기 " 를 표시하려면 **파일 > 환경 설정 > 일반**을 선택한 후 " 초기 JMP 창 " 목록에서 **JMP 시작하기**를 선택합니다 . macOS 에서는 **JMP > 환경 설정 > 초기 JMP 시작하기 창**을 선택합니다 .

JMP 기술 지원

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

기술 지원 전화 번호를 포함한 많은 기술 지원 옵션이 jmp.com/support에서 제공됩니다 .

예측 및 전문 모델링 소개

모델링 기법 개요

예측 및 전문 모델링에서는 반응 변수 선별, 분할, 신경망 등의 좀 더 전문적인 모델링 기법에 대해 자세히 설명합니다.

- 신경망 플랫폼에서는 하나 또는 두 개의 층으로 완전히 연결된 다중 계층 퍼셉트론을 구현합니다. 신경망을 사용하면 입력 변수의 유연한 함수를 사용하여 하나 이상의 반응 변수를 예측할 수 있습니다. 자세한 내용은 "[신경망](#)"에서 확인하십시오.
- 파티션 플랫폼에서는 X 값과 Y 값 사이의 관계에 따라 데이터를 재귀적으로 분할하여 파티션의 결정 트리를 생성합니다. 자세한 내용은 "[파티션 모형](#)"에서 확인하십시오.
- **JMP PRO** 붓스트랩 포레스트 플랫폼을 사용하면 각각 훈련 데이터의 랜덤 부분집합에 적용되는 여러 결정 트리를 평균화하여 앙상블 모형을 적합시킬 수 있습니다. 자세한 내용은 "[붓스트랩 포레스트](#)"에서 확인하십시오.
- **JMP PRO** 부스티드 트리 플랫폼에서는 계층에 생성된 여러 개의 더 작은 결정 트리로 구성되는 가법 결정 트리 모형을 생성합니다. 각 계층의 트리는 소수의 분할로 구성됩니다 (보통 5개 이하). 각 계층은 재귀적 적합 방법론을 사용하여 적합됩니다. 자세한 내용은 "[부스티드 트리](#)"에서 확인하십시오.
- **JMP PRO** K 최근접 이웃 플랫폼에서는 지정된 관측값의 국소적 근방에 있는 관측값의 반응을 사용하여 해당 관측값에 대한 반응 값을 예측합니다. 분류를 위한 범주형 반응 및 예측을 위한 연속형 반응과 함께 사용할 수 있습니다. 자세한 내용은 "[K 최근접 이웃](#)"에서 확인하십시오.
- **JMP PRO** Naive Bayes 플랫폼에서는 관측값을 범주형 반응 변수의 수준에 의해 정의된 그룹으로 분류합니다. 분류에 사용되는 변수 또는 요인을 데이터 마이닝 문헌에서는 흔히 피처라고 합니다. 자세한 내용은 "[Naive Bayes](#)"에서 확인하십시오.
- **JMP PRO** 서포트 벡터 머신 플랫폼에서는 관측값을 범주형 반응 변수의 수준에 의해 정의된 그룹으로 분류합니다. 모형은 클래스를 구분하는 초평면을 최적화하여 데이터를 분류합니다. 자세한 내용은 "[서포트 벡터 머신](#)"에서 확인하십시오.
- **JMP PRO** 모형 선별 플랫폼을 사용하면 여러 예측 모형을 빠르게 실행하고 결과를 비교할 수 있습니다. 중첩된 진단 그림과 함께 각 모형에 대한 적합 측도가 제공됩니다. 자세한 내용은 "[모형 선별](#)"에서 확인하십시오.
- **JMP PRO** 모형 비교 플랫폼을 사용하면 여러 모형의 예측 능력을 비교할 수 있습니다. 중첩된 진단 그림과 함께 각 모형에 대한 적합 측도가 제공됩니다. 자세한 내용은 "[모형 비교](#)"에서 확인하십시오.
- **JMP PRO** 검증 열 생성 플랫폼을 사용하면 파티션을 생성하는 다섯 가지 방법 중 하나를 사용하여 데이터를 두 개 또는 세 개의 집합으로 분할할 수 있습니다. 자세한 내용은 "[검증 열 생성](#)"에서 확인하십시오.

- **JMP PRO** 계산식 저장소 플랫폼을 사용하면 배포할 모형을 구성, 비교 및 프로파일링하고 스코어링할 수 있습니다. 모형 탐색 작업의 경우 계산식 저장소를 사용하여 후보 모형을 JMP 데이터 테이블 외부에 저장할 수 있습니다. 자세한 내용은 "[계산식 저장소](#)"에서 확인하십시오.
- 곡선 적합 플랫폼에서는 다항식, 로지스틱, Gompertz, 지수, 정상점 및 약동학 모형과 같은 사전 정의된 모형을 제공합니다. 다양한 분석과 그래픽 기법을 사용하여 여러 그룹 또는 개체를 비교합니다. 자세한 내용은 "[곡선 적합](#)"에서 확인하십시오.
- 비선형 플랫폼을 사용하면 모형 계산식 및 추정할 모수가 포함된 사용자 비선형 모형을 적합시킬 수 있습니다. 자세한 내용은 "[비선형 회귀](#)"에서 확인하십시오.
- 가우스 과정 플랫폼에서는 연속형 반응과 하나 이상의 연속형 예측 변수 간의 관계를 모델링합니다. 이러한 모형은 유한 요소 코드의 출력과 같은 컴퓨터 시뮬레이션 실험 분야에서 흔히 볼 수 있으며 데이터를 완벽하게 보간하는 경우가 많습니다. 자세한 내용은 "[가우스 과정](#)"에서 확인하십시오.
- **JMP PRO** 함수 데이터 탐색기 플랫폼을 사용하면 함수 데이터를 다른 JMP 플랫폼에서 분석할 수 있는 형식으로 변환할 수 있습니다. 자세한 내용은 "[함수 데이터 탐색기](#)"에서 확인하십시오.
- 시계열 플랫폼을 사용하면 단변량 시계열을 탐색, 분석 및 예측할 수 있습니다. 자세한 내용은 "[시계열 분석](#)"에서 확인하십시오.
- 시계열 예측 플랫폼을 사용하면 다중 시계열을 모델링하고 예측할 수 있습니다. 최대 30 개의 지수 평활 모형 집합에서 최량 적합 모형이 자동으로 선택됩니다. 자세한 내용은 "[시계열 예측](#)"에서 확인하십시오.
- 매칭 쌍 플랫폼에서는 상관관계가 있는 두 개 이상의 변수 간 평균을 비교하고 차이를 평가합니다. 자세한 내용은 "[매칭 쌍 분석](#)"에서 확인하십시오.
- 이상치 탐색 플랫폼을 사용하면 단변량 및 다변량 데이터 둘 다에서 이상치를 식별, 탐색 및 관리할 수 있습니다. 자세한 내용은 "[이상치 탐색](#)"에서 확인하십시오.
- 결측값 탐색 플랫폼은 데이터에서 결측값을 식별, 파악 및 대처할 수 있는 여러 가지 방법을 제공합니다. 자세한 내용은 "[결측값 탐색](#)"에서 확인하십시오.
- 패턴 탐색 플랫폼은 데이터에서 비정상적이거나 예기치 않은 패턴을 감지하는 데 사용됩니다. 자세한 내용은 "[패턴 탐색](#)"에서 확인하십시오.
- 반응 변수 선별 플랫폼에서는 다수의 반응 변수에 대한 검정 수행 과정을 자동화합니다. 검정 결과와 요약 통계량이 보고서 대신 데이터 테이블에 제공되어 데이터 탐색이 가능합니다. 자세한 내용은 "[반응 변수 선별](#)"에서 확인하십시오.
- 예측 변수 선별 플랫폼을 사용하면 유의한 예측 변수에 대한 데이터 집합을 선별할 수 있습니다. 자세한 내용은 "[예측 변수 선별](#)"에서 확인하십시오.
- **JMP PRO** 연관성 분석 플랫폼을 사용하면 서로 선호도가 있는 항목을 식별할 수 있습니다. 트랜잭션에서 자주 함께 나타나는 항목을 식별하기 위해 트랜잭션 데이터 (장바구니라고도 함)를 분석하는 데 흔히 사용됩니다. 자세한 내용은 "[연관성 분석](#)"에서 확인하십시오.
- 공정 기록 탐색기 플랫폼을 사용하면 복잡한 공정 기록에서 문제를 일으키는 성분을 식별할 수 있습니다. 자세한 내용은 "[공정 기록 탐색기](#)"에서 확인하십시오.

3 장

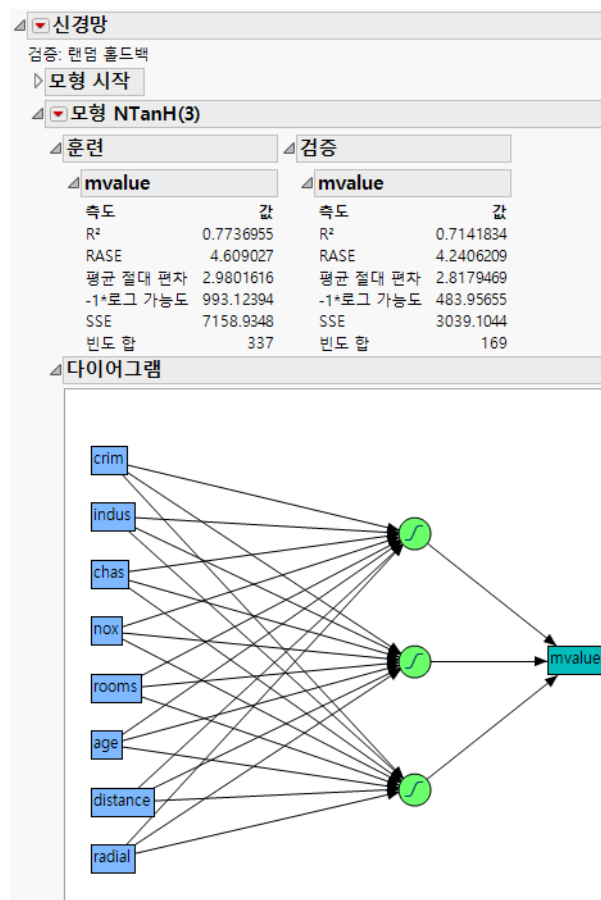
신경망

노드와 층을 사용하여 비선형 모형 적합

이 플랫폼에 있는 대부분의 기능은 JMP Pro 에서만 사용할 수 있으며 그러한 경우 이 아이콘이 표시됩니다.

신경망 플랫폼에서는 하나 또는 두 개의 층으로 완전히 연결된 다중 계층 퍼셉트론을 구현합니다. 신경망을 사용하면 입력 변수의 유연한 함수를 사용하여 하나 이상의 반응 변수를 예측할 수 있습니다. 신경망은 반응 표면의 함수 형태를 설명할 필요가 없거나 입력과 반응 사이의 관계를 설명할 필요가 없을 때 매우 좋은 예측 변수가 될 수 있습니다.

그림 3.1 신경망의 예



목차

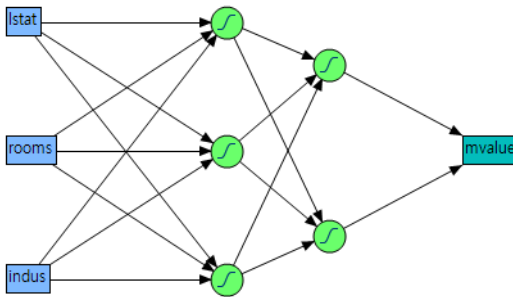
신경망 플랫폼 개요	33
신경망의 예	33
신경망 플랫폼 시작	36
신경망 시작 창	36
모형 시작 제어판	38
신경망의 검증 방법	39
은닉층 구조	40
부스트	40
신경망 적합 옵션	41
모형 보고서	43
훈련 및 검증 적합 측도	43
혼동 통계량	44
신경망 플랫폼 옵션	44
신경망 모형 옵션	45

신경망 플랫폼 개요

신경망은 은닉 노드라고 하는 파생 입력 집합의 함수입니다. 은닉 노드는 원래 입력의 비선형 함수입니다. 최대 두 개의 은닉 노드 층을 지정할 수 있으며 각 층은 은닉 노드를 원하는 만큼 포함할 수 있습니다.

그림 3.2에서는 세 개의 X 변수와 하나의 Y 변수로 구성된 2층 신경망을 보여 줍니다. 이 예에서 첫 번째 층에는 두 개의 노드가 있고 각 노드는 두 번째 층에 있는 세 가지 모든 노드의 함수입니다. 두 번째 층에는 세 개의 노드가 있고 모든 노드는 세 가지 X 변수의 함수입니다. 예측된 Y 변수는 첫 번째 층에 있는 두 노드의 함수입니다.

그림 3.2 신경망 다이어그램



은닉층 노드에 적용되는 함수를 활성화 함수라고 합니다. 활성화 함수는 X 변수의 선형 결합을 변환한 것입니다. 활성화 함수에 대한 자세한 내용은 "[은닉층 구조](#)"에서 확인하십시오. 반응에 적용되는 함수는 선형 결합(연속형 반응의 경우) 또는 로지스틱 변환(명목형 또는 순서형 반응의 경우)입니다.

신경망 모형의 주요 장점은 다양한 반응 표면을 효율적으로 모델링할 수 있다는 것입니다. 은닉 노드와 층이 충분하면 모든 표면을 원하는 정확도에 맞게 근사화할 수 있습니다. 신경망 모형의 주요 단점은 결과를 쉽게 해석할 수 없다는 것입니다. 정규 회귀 분석의 경우처럼 X 변수에서 Y 변수로의 직접 경로가 아니라 중간 층이 있기 때문입니다.

신경망의 예

주택 가격 중앙값을 여러 인구 통계적 특성의 함수로 예측하기 위한 모형을 생성하려는 경우를 가정해 보겠습니다. 아래 단계에 따라 신경망 모형을 생성합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Boston Housing.jmp**를 엽니다.
2. **분석 > 예측 모델링 > 신경망**을 선택하여 신경망 플랫폼을 시작합니다.
3. **mvalue**를 **Y, 반응** 역할에 할당합니다.
4. 다른 열 (**crim ~ lstat**)을 **X, 요인** 역할에 할당합니다.

5. **확인**을 클릭합니다.
6. **홀드백 비율**에 "0.2" 를 입력합니다.
7. **난수 시드값**에 "1234" 를 입력합니다.

참고 : 일반적으로 검증 데이터 집합을 선택하는 랜덤 특성 때문에 결과가 달라집니다. 위의 시드값을 입력하면 이 예에 표시된 결과를 재현할 수 있습니다.

8. 첫 번째 층의 TanH 노드 수에 "3" 을 입력합니다.
9. **공변량 변환** 옵션을 선택합니다.
10. **시작**을 클릭합니다.

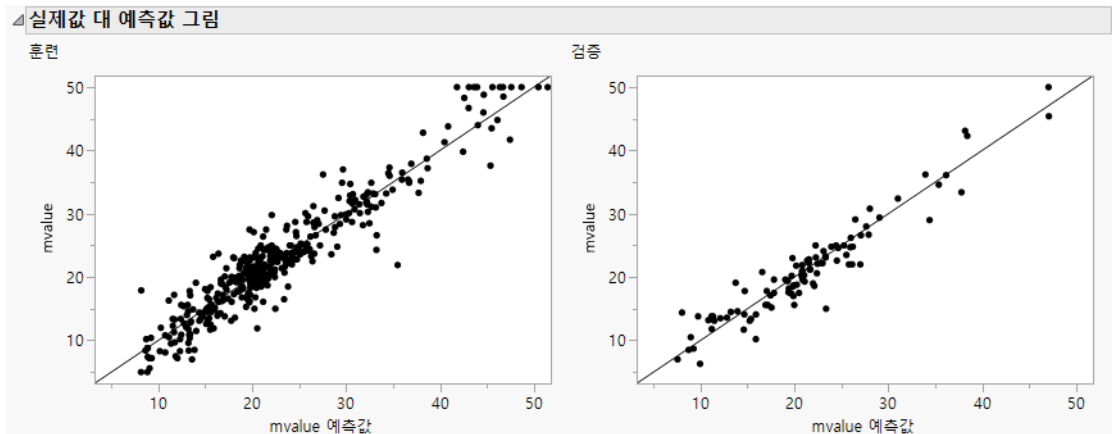
그림 3.3 신경망 보고서

모형 NTanH(3)			
훈련		검증	
mvalue		mvalue	
측도	값	측도	값
R ²	0.907698	R ²	0.9124801
RASE	2.8528693	RASE	2.4320654
평균 절대 편차	2.125991	평균 절대 편차	1.8299258
-1*로그 가능도	996.77457	-1*로그 가능도	235.3833
SSE	3288.1007	SSE	603.32408
빈도 합	404	빈도 합	102

훈련 데이터 집합과 검증 데이터 집합에 대한 보고서가 둘 다 제공됩니다. 검증 데이터 집합의 결과를 미래 관측값에 대한 모형의 예측 검정력을 나타내는 데 사용합니다.

검증 데이터 집합의 R² 통계량은 0.913이며, 이는 모형이 모형 훈련에 사용되지 않는 데이터에 대해 잘 예측하고 있음을 나타냅니다. 다른 방법으로 모형 적합을 평가하기 위해 "모형 NTanH(3)" 옆의 빨간색 삼각형을 클릭하고 **실제값 대 예측값 그림**을 선택합니다.

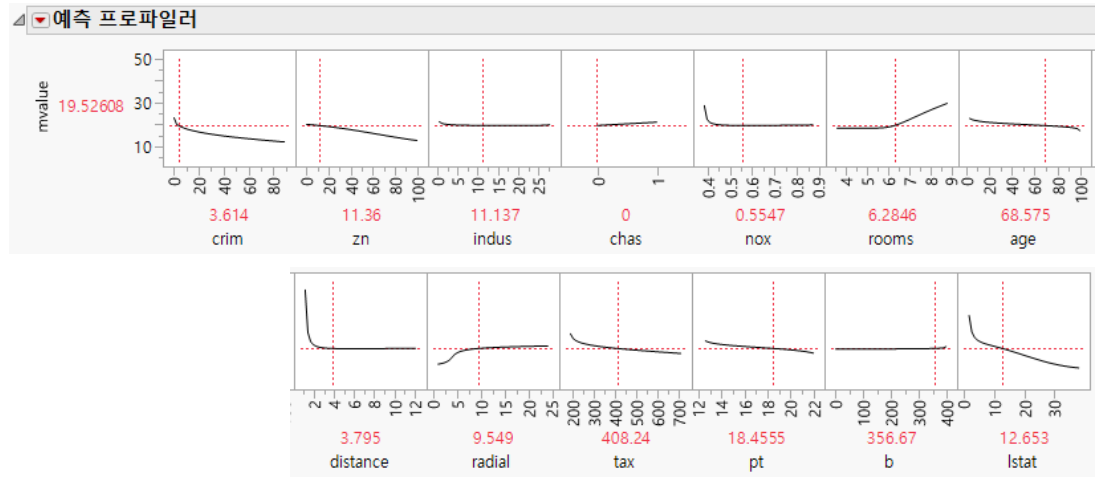
그림 3.4 실제값 대 예측값 그림



점이 선을 따라 분포되어 있으므로 예측값이 실제값과 유사하다는 것을 알 수 있습니다.

X 변수가 예측값에 미치는 영향에 대한 전반적인 이해를 위해 "모형 NTanH(3)" 옆의 빨간색 삼각형을 클릭하고 **프로파일러**를 선택합니다.

그림 3.5 프로파일러



변수 중 일부는 프로파일러의 기울기가 양수이고, 다른 변수는 프로파일러의 기울기가 음수입니다. 예를 들어 **rooms** 변수의 기울기는 양수입니다. 이는 집에 방이 많을수록 예측 중앙값이 더 높다는 것을 나타냅니다. **pt** 변수는 도시별 학생 교사 비율입니다. 이 변수의 기울기가 음수이므로 학생 대 교사 비율이 높을수록 중앙값이 낮다는 것을 나타냅니다.

신경망 플랫폼 시작

신경망 플랫폼을 시작하려면 **분석 > 예측 모델링 > 신경망**을 선택합니다.

신경망 플랫폼 시작은 2단계 프로세스입니다. 먼저 신경망 시작 창에 변수를 입력합니다. 그 다음 "모형 시작" 제어판에서 옵션을 지정합니다.

JMP PRO 신경망 시작 창

신경망 시작 창을 사용하여 X 변수, Y 변수 및 검증 열을 지정하고 결측값 정보화 값 코딩을 활성화할 수 있습니다.

그림 3.6 신경망 시작 창

입력 변수의 유연한 함수를 사용하여 하나 이상의 반응 변수를 예측합니다.

열 선택	선택한 열 역할 지정	작업
14개 열 - crim - zn - indus - chas - nox - rooms - age - distance - radial - tax - pt - b - lstat - mvalue	Y, 반응 X, 요인 빈도 검증 기준	확인 취소 제거 재호출 도움말

☐ 결측값 정보화
 난수 시드값 설정

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

Y, 반응 분석할 하나 이상의 반응 변수입니다. 반응을 여러 개 지정하면 반응에 대한 모형이 은닉층의 모든 모수 (반응에 연결되지 않은 모수)를 공유합니다. 입력 변수가 주어지면 반응 변수는 조건부로 독립적이지만 이러한 변수를 통해 약간의 상관관계를 유지하여 하나의 전체 신경망 모형을 생성합니다.

X, 요인 예측 변수입니다.

빈도 분석의 각 행에 빈도를 할당하는 숫자 값이 포함된 열입니다.

JMP PRO 검증 검증 데이터 집합을 정의하는 숫자 열입니다. 자세한 내용은 "**신경망의 검증 방법**"에서 확인하십시오. "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 "**검증 열 생성**"에서 확인하십시오.

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.

JMP PRO 결측값 정보화 결측값 정보화 코딩을 활성화하려면 이 상자를 선택합니다. 이 코딩을 사용하면 결측값이 있어도 예측 모형을 추정할 수 있습니다. 이 옵션은 결측 데이터에 정보가 포함된 경우 유용합니다. 이 옵션을 선택하지 않으면 결측값이 있는 행이 무시됩니다.

연속형 변수의 경우 결측값은 변수의 평균으로 대체됩니다. 또한 < 열 이름 > 결측됨이라는 결측값 표시자가 생성되어 모형에 포함됩니다. "모형 시작" 제어판의 "공변량 변환" 적합 옵션을 사용하여 변수를 변환하는 경우 결측값은 변환된 변수의 평균으로 대체됩니다.

범주형 변수의 경우 결측값은 해당 변수에 대한 별도의 수준으로 처리됩니다.

난수 시드값 설정 적합 절차에 사용되는 시작 값의 시드를 설정합니다. 검증 열을 지정하지 않는 경우 이 값은 "모형 시작" 제어판에서 난수 시드값으로 자동 할당되며 검증 할당에도 사용됩니다. "난수 시드값 설정"은 분석을 재현하려는 경우에 유용합니다. 난수 시드값을 설정하고 스크립트를 저장하면 시드값이 스크립트에 자동으로 저장됩니다.

모형 시작 제어판

신경망 플랫폼의 "모형 시작" 제어판을 사용하여 검증 방법, 은닉층 구조, 그래디언트 부스트 사용 여부 및 기타 적합 옵션을 지정할 수 있습니다.

그림 3.7 모형 시작 제어판

JMP Pro

신경망

모형 시작

검증 방법: **홀드백** (재현성: 0.3333, 난수 시드값: 0)

은닉층 구조: 각 활성 유형의 노드 수
 활성화 시그모이드 함수: TanH, 선형, 가우스
 첫 번째: 3, 0, 0
 두 번째: 0, 0, 0
 2층 모형에서 두 번째 층이 X에 더 가깝습니다.

부스트: 학습률로 적도화된 가법 모형 시퀀스를 적합시킵니다.
 모형 수: 0
 학습률: 0.1

적합 옵션: ☐ 공변량 변환, ☐ 로버스트 적합
 발점 방법: **제공**
 투여 수: 1

시작

JMP

신경망

모형 시작

검증 방법: **홀드백** (재현성: 0.3333, 난수 시드값: 0)

숨겨진 노드: 3

시작

검증 방법 모형 검증에 사용할 방법을 선택합니다. 자세한 내용은 "[신경망의 검증 방법](#)"에서 확인하십시오.

난수 시드값 이후의 신경망 플랫폼 실행을 위해 시작 값 및 검증 할당을 재현하려면 0 이 아닌 숫자 난수 시드값을 지정합니다. 기본적으로 난수 시드값은 0 으로 설정되어 재현 가능한 결과를 생성하지 않습니다. 분석을 스크립트에 저장하면 사용자가 입력한 난수 시드값이 스크립트에 저장됩니다.

은닉층 구조 또는 은닉 노드 각 층에서 각 유형의 은닉 노드 수를 지정합니다. 자세한 내용은 "[은닉층 구조](#)"에서 확인하십시오.

참고: JMP Standard Edition에서는 TanH 활성 함수만 사용되며 하나의 은닉층이 있는 신경망만 적합시킬 수 있습니다.

JMP Pro 부스트 중 그래디언트 부스트 옵션을 지정합니다. 자세한 내용은 "[부스트](#)"에서 확인하십시오.

JMP PRO 적합 옵션 변수 변환 및 모형 적합 옵션을 지정합니다. 자세한 내용은 "**신경망 적합 옵션**"에서 확인하십시오.

시작 신경망 모형을 적합시키고 모형 보고서를 표시합니다.

시작을 클릭하여 모형을 적합시킨 후 "모형 시작" 제어판을 다시 열고 다른 모형 적합을 위해 설정을 변경할 수 있습니다.

신경망의 검증 방법

신경망 플랫폼은 적합 루틴에 검증 방법을 활용합니다. 검증 방법 선택에는 홀드백, K 폴드 또는 검증 열 사용 (JMP Pro) 이 포함됩니다. 신경망 플랫폼에서는 다음을 수행하여 모형을 적합시킵니다.

- 모형 모수에 벌점 적용
- 검증 데이터 집합을 사용하여 모수에 대한 벌점 조정

다음 검증 방법 중 하나를 선택합니다.

제외된 행 홀드백 행 상태를 사용하여 데이터의 부분집합을 생성합니다. 제외되지 않은 행은 훈련 데이터 집합으로 사용되고 제외된 행은 검증 데이터 집합으로 사용됩니다.

행 상태 사용 및 행 제외 방법에 대한 자세한 내용은 **JMP 사용**에서 확인하십시오.

홀드백 원래 데이터를 무작위로 훈련 데이터 집합과 검증 데이터 집합으로 나눕니다. 원래 데이터에서 검증 데이터 집합으로 사용할 비율을 지정합니다 (홀드백). 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.

K 폴드 데이터를 K 개의 부분집합 또는 폴드로 나눕니다. 그러면 K 개의 각 폴드가 나머지 데이터에 대한 모형 적합을 검증하는 데 사용되어 총 K 개 모형을 적합시킵니다. 최종 모형은 최량 검증 통계량을 제공하는 모형을 생성한 폴드에 대한 모형입니다.

이 방법은 제한된 양의 데이터를 효율적으로 사용하기 때문에 소규모 데이터 집합에 유용합니다.

JMP PRO 검증 열 검증 데이터 집합을 정의하는 숫자 열을 사용합니다. 다음과 같이 열의 값에 따라 데이터 분할 방법이 결정됩니다.

- 검증 열에 두 개의 수준이 있는 경우 더 작은 값이 훈련 데이터 집합을 정의하고 더 큰 값이 검증 데이터 집합을 정의합니다.
- 검증 열에 세 개의 수준이 있는 경우 작은 값부터 순서대로 훈련 데이터 집합, 검증 데이터 집합, 테스트 데이터 집합을 각각 정의합니다.
- 검증 열에 4 개 이상의 고유 값이 있는 경우 K 폴드 검증이 수행됩니다. 폴드 수는 검증 열의 값 수에 따라 결정됩니다.

신경망 플랫폼에서는 검증 열을 사용하여 모형을 훈련 및 조정하거나 모형을 훈련, 조정 및 평가합니다. 검증에 대한 자세한 내용은 "**JMP 모델링의 검증**"에서 확인하십시오.

은닉층 구조

신경망 플랫폼은 층이 하나 또는 두 개인 신경망을 적합시킬 수 있습니다. 첫 번째 층의 노드 수를 늘리거나 두 번째 층을 추가하면 신경망이 더 유연해집니다. 각 층에 노드를 무제한으로 추가할 수 있습니다. 두 번째 층의 노드는 X 변수의 함수입니다. 첫 번째 층의 노드는 두 번째 층 노드의 함수입니다. Y 변수는 첫 번째 층 노드의 함수입니다.

참고 : JMP Standard Edition에서는 TanH 활성화 함수만 사용되며 하나의 은닉층이 있는 신경망만 적합시킬 수 있습니다.

주의 : 층이 두 개인 신경망에서는 부스트를 사용할 수 없습니다. 두 번째 층에 0 이 아닌 값을 지정하고 부스트도 지정하면 두 번째 층이 무시됩니다.

은닉층 노드에 적용되는 함수를 활성화 함수라고 합니다. 활성화 함수는 X 변수의 선형 결합을 변환한 것입니다. 다음 활성화 함수를 사용할 수 있습니다.

TanH 쌍곡선 탄젠트 함수는 시그모이드 함수입니다. TanH 는 값을 -1 에서 1 사이로 변환하며 로지스틱 함수의 중심화 및 척도화된 버전입니다. 쌍곡선 탄젠트 함수는 다음과 같습니다.

$$\frac{e^{2x} - 1}{e^{2x} + 1}$$

여기서 x 는 X 변수의 선형 결합입니다.

선형 항등 함수입니다. X 변수의 선형 결합이 변환되지 않습니다.

선형 활성화 함수는 비선형 활성화 함수 중 하나와 함께 사용되는 경우가 가장 흔합니다. 이 경우 선형 활성화 함수가 두 번째 층에 배치되고 비선형 활성화 함수가 첫 번째 층에 배치됩니다. 이 방법은 먼저 X 변수의 차원을 축소 한 후 Y 변수에 대한 비선형 모형을 사용하려는 경우에 유용합니다.

연속형 Y 변수의 경우 선형 활성화 함수만 사용하면 Y 변수의 모형이 X 변수의 선형 결합으로 축소됩니다. 명목형 또는 순서형 Y 변수의 경우에는 모형이 로지스틱 회귀로 축소됩니다.

가우스 가우스 함수입니다. 반응 표면이 가우스 (정규) 형상일 때 또는 방사 기저 함수 동작에 대해 이 옵션을 사용합니다. 가우스 함수는 다음과 같습니다.

$$e^{-x^2}$$

여기서 x 는 X 변수의 선형 결합입니다.

JMP[®] PRO 부스트

부스트는 일련의 작은 모형을 적합시킨 후 이를 누적하여 더 큰 신경망 모형을 생성하는 프로세스입니다. 작은 모형은 각각 이전 모형의 척도화된 잔차에 대해 적합됩니다. 모형이 결합되어 더 큰 최종 모형이 생성됩니다. 이 프로세스는 검증을 사용하여 지정된 모형 수를 초과하지 않고 적합시킬 성분 모형 수를 평가합니다.

부스트는 큰 모형 하나를 적합시키는 것보다 빠른 경우가 많습니다. 그러나 기준 모형은 1~2개 노드가 있는 단일층 모형이어야 합니다. 많은 수의 모형을 지정하면 적합 속도가 빠르다는 이점이 없어질 수 있습니다.

"모형 시작" 제어판의 "부스트" 패널을 사용하여 성분 모형 수와 학습률을 지정할 수 있습니다. "모형 시작" 제어판의 "은닉층 구조" 패널을 사용하여 기준 모형의 구조를 지정할 수 있습니다.

학습률은 $0 < r \leq 1$ 이어야 합니다. 학습률이 1에 가까우면 최종 모형에 대한 수렴이 빨라지지만 데이터가 과대적합되는 경향도 높습니다. 적은 수의 모형이 지정된 경우 1에 가까운 학습률을 사용하십시오.

부스트 작동 방식을 보여 주기 위해 하나의 층과 두 개의 노드로 구성된 기준 모형을 지정하고 모형 수는 8개라고 가정해 보겠습니다. 첫 번째 단계에서는 하나의 층과 두 개의 노드가 있는 모형을 적합시킵니다. 해당 모형의 예측값은 학습률로 척도화된 후 실제 값에서 빼서 척도화된 잔차를 구성합니다. 다음 단계에서는 하나의 층과 두 개의 노드가 있는 다른 모형을 적합시킵니다. 이때 반응 값은 이전 모형의 척도화된 잔차입니다. 이 프로세스는 8개의 모형이 적합될 때까지 또는 다른 모형을 추가해도 검증 통계량이 개선되지 않을 때까지 계속됩니다. 성분 모형이 결합되어 더 큰 최종 모형이 생성됩니다. 이 예에서 6개 모형이 적합된 후 중지되면 최종 모형은 하나의 층과 $2 \times 6 = 12$ 개의 노드로 구성됩니다.

주의: 층이 두 개인 신경망에서는 부스트를 사용할 수 없습니다. 두 번째 층에 0이 아닌 값을 지정하고 부스트도 지정하면 두 번째 층이 무시됩니다.

JMP[®] PRO 신경망 적합 옵션

다음과 같은 모형 적합 옵션을 사용할 수 있습니다.

공변량 변환 Johnson Su 또는 Johnson Sb 분포를 사용하여 모든 연속형 변수를 정규성에 가까운 변수로 변환합니다. 연속형 변수를 변환하면 이상치 또는 심하게 기울어진 분포가 미치는 부정적 영향을 줄이는 데 도움이 됩니다. 자세한 내용은 "[신경망 모형 옵션](#)"의 "변환된 공변량 저장" 옵션을 참조하십시오.

로버스트 적합 최소 제곱 대신 최소 절대 편차를 사용하여 모형을 훈련합니다. 이 옵션은 반응 이상치의 영향을 최소화하려는 경우에 유용합니다. 이 옵션은 연속형 반응에만 사용할 수 있습니다.

벌점 방법 벌점 방법을 선택합니다. 신경망에서 데이터가 과대적합되는 경향을 줄이기 위해 적합 프로세스에서 가능도에 대한 벌점을 통합합니다. 자세한 내용은 "[벌점 방법](#)"에서 확인하십시오.

투어 수 반복할 때마다 모수 추정값에 대한 랜덤 시작점을 다르게 사용하여 적합 프로세스를 다시 시작할 횟수를 지정합니다. 최량 검증 통계량을 제공하는 반복이 최종 모형으로 선택됩니다.

별점 방법

별점은 $\lambda p(\beta_i)$ 이며, 여기서 λ 는 별점 모수이고 $p()$ 는 별점 함수라고 하는 모수 추정값의 함수입니다. 별점 모수의 최적값을 찾기 위해 검증이 사용됩니다.

표 3.1 별점 방법 설명

방법	별점 함수	설명
제곱	$\sum \beta_i^2$	대부분의 X 변수가 모형의 예측 능력에 기여한다고 생각되면 이 방법을 사용합니다.
절대	$\sum \beta_i $	X 변수의 개수가 많고 이 중 몇 개가 다른 변수보다 모형의 예측 능력에 더 많이 기여한다고 생각되면 두 방법 중 하나를 사용합니다.
가중치 감소	$\sum \frac{\beta_i^2}{1 + \beta_i^2}$	
별점 없음	없음	별점을 사용하지 않습니다. 많은 양의 데이터가 있고 적합 프로세스를 빠르게 진행하려는 경우 이 옵션을 사용할 수 있습니다. 그러나 이 옵션을 사용하면 별점을 사용하는 모형보다 예측 성능이 낮은 모형이 생성될 수 있습니다.

모형 보고서

모든 신경망 모형에 대한 모형 보고서가 생성됩니다. 훈련 데이터 집합과 검증 데이터 집합에 대한 적합 측도가 나타납니다. 명목형 또는 순서형 반응의 경우 혼동 통계량도 나타납니다.

그림 3.8 신경망 모형 보고서의 예

모형 NTanH(3)			
훈련		검증	
chas		chas	
측도	값	측도	값
일반화 R^2	0.5326372	일반화 R^2	0.1796731
엔트로피 R^2	0.4705979	엔트로피 R^2	0.1459083
RASE	0.2006679	RASE	0.2498265
평균 절대 편차	0.0860496	평균 절대 편차	0.0920273
오분류 비율	0.0534125	오분류 비율	0.0828402
-1*log 가능성도	44.43917	-1*log 가능성도	36.985086
빈도 합	337	빈도 합	169
혼동 행렬		혼동 행렬	
실제값	예측 개수	실제값	예측 개수
chas	0 1	chas	0 1
0	311 3	0	155 2
1	15 8	1	12 0
혼동 비율		혼동 비율	
실제값	예측 비율	실제값	예측 비율
chas	0 1	chas	0 1
0	0.990 0.010	0	0.987 0.013
1	0.652 0.348	1	1.000 0.000

훈련 및 검증 적합 측도

신경망 모형 보고서에는 훈련 데이터 집합과 검증 데이터 집합에 대한 적합 측도 테이블이 포함되어 있습니다 (그림 3.8). 다음과 같은 적합 측도가 각 테이블에 포함됩니다.

일반화 R^2 일반 회귀 모형에 적용할 수 있는 측도입니다. 이 값은 가능성도 함수 L 을 기반으로 하며 최대값이 1 이 되도록 척도화됩니다. 값이 1 이면 완벽한 모형을 나타내고 0 이면 상수 모형과 같음을 나타냅니다. 일반화 R^2 측도는 표준 최소 제곱 설정 시 연속형 정규 반응에 대한 기존 R^2 로 단순화됩니다. 일반화 R^2 을 Nagelkerke 또는 Craig 와 Uhler R^2 이라고도 하는데, 이는 Cox-Snell 유사 R^2 을 정규화한 것입니다. 자세한 내용은 Nagelkerke 연구 자료 (1991) 에서 확인하십시오.

엔트로피 R^2 (반응이 명목형 또는 순서형인 경우에만 표시됨) 적합 모형과 상수 확률 모형의 로그 가능성도를 비교하는 적합 측도입니다. 엔트로피 R^2 범위는 0 ~ 1 이며, 값이 1 에 가까울수록 더 나은 적합을 나타냅니다. 자세한 내용은 "엔트로피 R^2 " 에서 확인하십시오.

R^2 모형에 대한 R^2 을 제공합니다.

RASE 제공된 평균 제곱 오차를 제공합니다. 반응이 명목형 또는 순서형인 경우 차이는 1 에서 p (실제로 발생한 반응 수준의 적합 확률) 사이입니다.

평균 절대 편차 반응과 예측 반응 사이의 차이에 대한 절대값의 평균입니다. 반응이 명목형 또는 순서형인 경우 차이는 1 에서 p (실제로 발생한 반응 수준의 적합 확률) 사이입니다.

오분류 비율 적합 확률이 가장 높은 반응 범주가 관측된 범주가 아닌 비율입니다. 반응이 명목형 또는 순서형인 경우에만 나타납니다.

-1* 로그 가능도 음의 로그 가능도를 제공합니다. 자세한 내용은 선형 모형 적합에서 확인하십시오.

SSE 오차 제곱합을 제공합니다. 반응이 연속형인 경우에만 사용할 수 있습니다.

빈도 합 사용된 관측값 수를 제공합니다. 신경망 시작 창에서 빈도 변수를 지정한 경우 빈도 합은 빈도 열의 합을 제공합니다.

반응이 여러 개인 경우 각 반응에 대한 적합 통계량과 전체 일반화 R^2 및 음의 로그 가능도가 제공됩니다.

혼동 통계량

신경망 플랫폼에서 명목형 또는 순서형 반응이 적합되면 혼동 행렬 보고서와 혼동 비율 보고서가 제공됩니다(그림 3.8). "혼동 행렬" 보고서에는 실제 반응 수준과 예측 반응 수준의 이원 분류가 표시됩니다. 범주형 반응의 경우 예측 수준은 예측 확률이 가장 높은 수준입니다. 혼동 비율 보고서는 값을 행 합계로 나눈 혼동 행렬 보고서와 동일합니다.

신경망 플랫폼 옵션

"신경망"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

다음 옵션에 대한 자세한 내용은 JMP 사용의 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.

신경망 모형 옵션

각 모형 보고서에는 추가 출력을 생성하거나 결과를 저장하기 위한 옵션이 포함된 빨간색 삼각형 메뉴가 있습니다. 모형 보고서의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

다이아그램 은닉층 구조를 나타내는 다이어그램을 표시합니다.

추정값 표시 보고서에 모수 추정값을 표시합니다.

프로파일러 예측 프로파일러를 시작합니다. 명목형 또는 순서형 반응의 경우 각 반응 수준은 예측 프로파일러에서 별도의 행으로 표시됩니다. 빨간색 삼각형 메뉴의 옵션에 대한 자세한 내용은 **Profilers** 에서 확인하십시오.

범주형 프로파일러 예측 프로파일러를 시작합니다. 모든 범주형 확률이 단일 프로파일러 행으로 결합된다는 점을 제외하면 "프로파일러" 옵션과 유사합니다. 명목형 또는 순서형 반응에만 사용할 수 있습니다. 빨간색 삼각형 메뉴의 옵션에 대한 자세한 내용은 **Profilers** 에서 확인하십시오.

등고선 프로파일러 등고선 프로파일러를 시작합니다. 이 옵션은 모형에 두 개 이상의 연속형 요인이 포함된 경우에만 사용할 수 있습니다. 빨간색 삼각형 메뉴의 옵션에 대한 자세한 내용은 **Profilers** 에서 확인하십시오.

표면 프로파일러 표면 프로파일러를 시작합니다. 이 옵션은 모형에 두 개 이상의 연속형 요인이 포함된 경우에만 사용할 수 있습니다. 빨간색 삼각형 메뉴의 옵션에 대한 자세한 내용은 **Profilers** 에서 확인하십시오.

ROC 곡선 ROC 곡선을 생성합니다. 명목형 또는 순서형 반응에만 사용할 수 있습니다. ROC 곡선에 대한 자세한 내용은 "**ROC 곡선**" 에서 확인하십시오.

향상도 곡선 향상도 곡선을 생성합니다. 명목형 또는 순서형 반응에만 사용할 수 있습니다. 향상도 곡선에 대한 자세한 내용은 "**향상도 곡선**" 에서 확인하십시오.

결정 임계값 (이항 범주형 반응에만 사용 가능) 훈련, 검증 및 테스트 (지정된 경우) 데이터 집합에 대한 "결정 임계값" 보고서를 표시하거나 숨깁니다. 각 보고서에는 모형 적합 비교를 위한 각 모형의 적합 확률 분포 그래프, 혼동 행렬 및 분류 그래프와 분류 정확성 측정 기준 테이블이 포함되어 있습니다. 자세한 내용은 "**결정 임계값 보고서**" 에서 확인하십시오.

실제값 대 예측값 그림 실제 반응 대 예측 반응 그림을 표시합니다. 연속형 반응에만 사용할 수 있습니다.

잔차 대 예측값 그림 잔차와 예측 반응을 비교하는 그림을 표시합니다. 연속형 반응에만 사용할 수 있습니다.

계산식 저장 예측 반응 및 은닉층 노드에 대한 계산식이 포함된 새 열을 데이터 테이블에 생성합니다.

프로파일 계산식 저장 예측 반응에 대한 계산식이 포함된 새 열을 데이터 테이블에 생성합니다. 은닉층 노드에 대한 계산식이 이 계산식에 포함되어 있습니다. 이 옵션은 대화식 HTML 버전의 프로파일러에서 사용할 수 있는 계산식을 생성합니다.

빠른 계산식 저장 예측 반응에 대한 계산식이 포함된 새 열을 데이터 테이블에 생성합니다. 은닉층 노드에 대한 계산식이 이 계산식에 포함되어 있습니다. 이 옵션은 다른 옵션보다 빠르게

실행되는 계산식을 생성하지만 대화식 버전의 프로파일러에서 이 계산식을 사용할 수 없습니다.

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 "[계산식 저장소](#)"에서 확인하십시오.

SAS Data 스텝 만들기 새 데이터 집합을 스코어링하는 데 사용할 수 있는 SAS 코드를 생성합니다.

검증 저장 훈련 데이터 집합과 검증 데이터 집합에 사용된 행을 식별하는 새 열을 데이터 테이블에 생성합니다. 신경망 시작 창에서 검증 열을 지정한 경우에는 이 옵션을 사용할 수 없습니다. 자세한 내용은 "[신경망 시작 창](#)"에서 확인하십시오.

JMP PRO **변환된 공변량 저장** 변환된 공변량을 보여 주는 새 열을 데이터 테이블에 생성합니다. 이 열에는 변환을 보여 주는 계산식이 포함됩니다. 이 옵션은 "모형 시작" 제어판에서 "공변량 변환" 옵션을 선택한 경우에만 사용할 수 있습니다. 자세한 내용은 "[신경망 적합 옵션](#)"에서 확인하십시오.

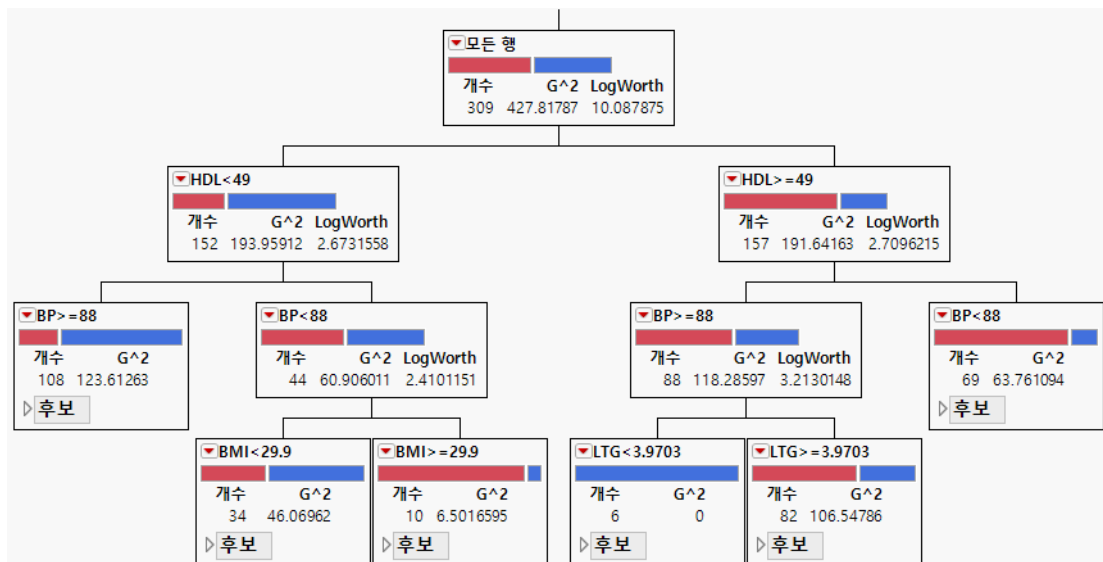
적합 제거 전체 모형 보고서를 제거합니다.

파티션 모형

결정 트리를 사용하여 데이터 탐색 및 모델링

파티션 플랫폼에서는 예측 변수와 반응 값 사이의 관계에 따라 데이터를 재귀적으로 분할하여 결정 트리를 생성합니다. 파티션 알고리즘은 반응을 가장 잘 예측하기 위해 예측 변수의 가능한 모든 분할을 검색합니다. 이러한 데이터 분할(파티션)은 재귀적으로 수행되어 결정 규칙 트리를 생성합니다. 원하는 적합을 얻을 때까지 분할이 계속됩니다. 파티션 알고리즘은 많은 수의 가능한 분할에서 최적 분할을 선택하므로 강력한 모델링 및 데이터 검색 도구라고 할 수 있습니다.

그림 4.1 결정 트리의 예



목차

파티션 플랫폼 개요	49
파티션 플랫폼의 예	49
파티션 플랫폼 시작	52
결측값 정보화	53
분할 보고서	54
분할 그림	54
분할 버튼	56
적합 통계량 요약	56
노드 보고서	57
파티션 플랫폼 옵션	58
적합 상세 정보 표시	61
수익 행렬 지정	63
ROC 곡선	64
향상도 곡선	65
파티션 검증	67
파티션 플랫폼의 추가 예	68
연속형 반응의 예	68
결측값 정보화의 예	71
수익 행렬 및 결정 행렬 보고서의 예	74
파티션 플랫폼에 대한 통계 상세 정보	77
반응 및 요인에 대한 통계 상세 정보	77
분할 기준에 대한 통계 상세 정보	77

파티션 플랫폼 개요

파티션 플랫폼에서는 예측 변수와 반응 값 사이의 관계에 따라 데이터를 재귀적으로 분할하여 결정 트리를 생성합니다. 분할은 결정 트리, CARTTM, CHAIDTM, C4.5, C5 등 다양한 이름 및 브랜드 이름에 따라 몇 가지 변형이 있습니다. 이 기법은 다음과 같은 이유 때문에 데이터 마이닝 기법으로 간주되는 경우가 많습니다.

- 적합한 사전 모형 없이 관계를 탐색하는 데 유용합니다.
- 많은 문제를 쉽게 처리합니다.
- 결과를 해석할 수 있습니다.

분할이 적용되는 대표적인 예는 질병에 대한 진단 휴리스틱을 만드는 것입니다. 여러 피실험자의 증상과 결과가 주어졌을 때 분할을 사용하여 새 환자를 진단하는 데 도움이 되는 질문 계층을 생성할 수 있습니다.

예측 변수는 연속형 또는 범주형(명목형 또는 순서형)일 수 있습니다. 예측 변수가 연속형이면 절단 값에 의해 분할이 생성됩니다. 표본은 이 절단 값보다 작은 값과 큰 값으로 구분됩니다. 예측 변수가 범주형이면 표본이 두 그룹의 수준으로 구분됩니다.

반응도 연속형 또는 범주형(명목형 또는 순서형)일 수 있습니다. 반응이 연속형이면 플랫폼에서 반응 값의 평균을 적합시키고, 오차 제곱합을 최소화하는 분할이 선택됩니다. 반응이 범주형이면 적합한 값은 반응 수준에 대한 확률이며, 잔차 로그 가능성도 카이제곱을 최소화하는 분할이 선택됩니다.

분할 기준에 대한 자세한 내용은 "[파티션 플랫폼에 대한 통계 상세 정보](#)"에서 확인하십시오.

재귀 분할에 대한 자세한 내용은 Hawkins and Kass(1982) 및 Kass(1980) 연구 자료에서 확인하십시오.

파티션 플랫폼의 예

이 예에서는 파티션 플랫폼을 사용하여 당뇨병 환자의 1년 질병 진행률(낮음 또는 높음)을 예측하는 결정 트리를 생성합니다.

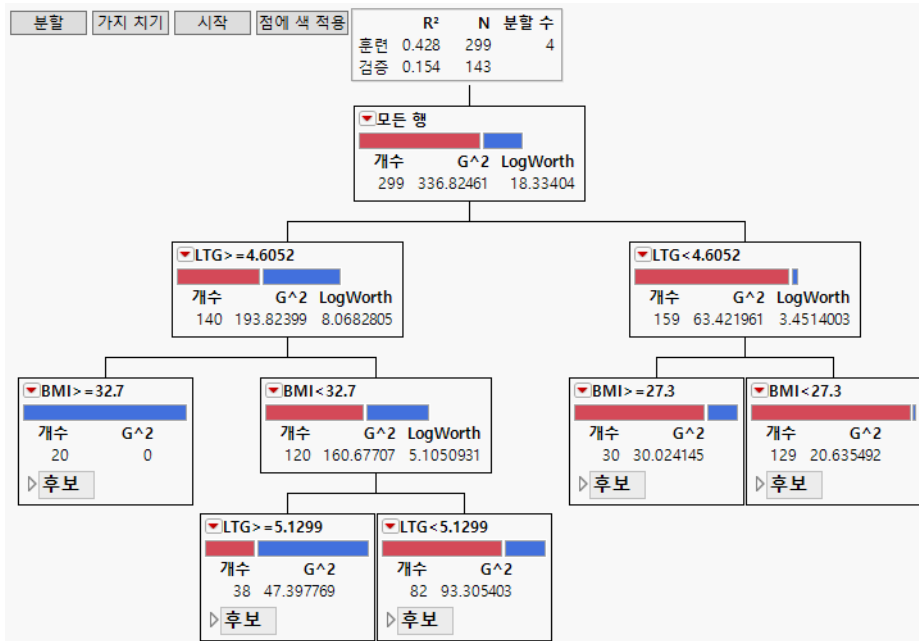
1. **도움말 > 샘플 데이터 폴더**를 선택하고 Diabetes.jmp 를 엽니다.
2. **분석 > 예측 모델링 > 파티션**을 선택합니다.
3. Y Binary 를 선택하고 Y, 반응을 클릭합니다.
4. Age ~ Glucose 를 선택하고 X, 요인을 클릭합니다.
5. **검증 데이터 비율**에 "0.33" 을 입력합니다.

참고 : JMP Pro에서는 검증을 위해 검증 열을 사용할 수 있습니다. Validation 을 선택하고 검증 을 클릭합니다. 그런 다음 검증 데이터 비율을 0 으로 설정합니다.

6. **확인**을 클릭합니다.
7. 플랫폼 보고서 창에서 **시작**을 클릭하여 자동 분할을 수행합니다.

참고 : 랜덤 검증 데이터 비율이 사용되므로 결과가 [그림 4.2](#)에 표시된 것과 다릅니다.

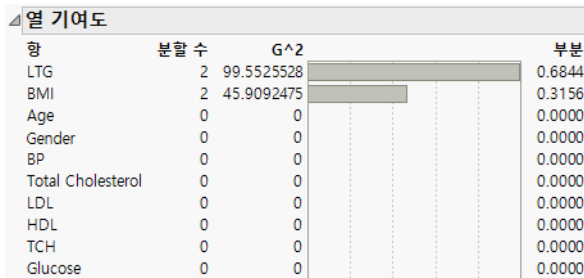
그림 4.2 Diabetes 에 대한 분할 보고서



자동 분할을 통해 네 개의 분할이 생성되었습니다. 검증 데이터 집합의 최종 R²은 0.154입니다. 결정 트리는 네 개의 분할과 각 분할의 관측값 수를 보여 줍니다.

8. "Y Binary 에 대한 분할" 옆의 빨간색 삼각형을 클릭하고 **열 기여도**를 선택합니다.

그림 4.3 열 기여도 보고서



"열 기여도" 보고서에서는 결정 트리 모형의 예측 변수가 LTG와 BMI 뿐임을 보여 줍니다. 각 열은 두 개의 분할에서 사용됩니다. 사용자가 실행한 결과는 다를 수 있습니다. 검증 데이터

비율을 사용하는 경우 데이터 테이블에서 검증 데이터 집합이 랜덤 선택됩니다. 분석을 다시 실행하면 새 검증 데이터 집합이 랜덤 선택되고 첫 번째 실행과 결과가 다를 수 있습니다.

9. "Y Binary 에 대한 분할" 옆의 빨간색 삼각형을 클릭하고 **프로파일러**를 선택합니다.

그림 4.4 파티션 모형에 대한 프로파일러



프로파일러를 사용하면 BMI 및 LTG 값을 변경하여 Y Binary 결과를 예측할 수 있습니다. 다른 요인은 파티션 모형에 분할이 없었습니다. 해당 요인의 프로파일은 편평한 일직선입니다.

10. "Y Binary 에 대한 분할" 옆의 빨간색 삼각형을 클릭하고 **열 저장 > 예측 계산식 저장**을 선택합니다.

Diabetes.jmp 데이터 테이블에 **확률 (Y Binary==Low)**, **확률 (Y Binary==High)** 및 최대 확률 분류 Y Binary 열이 추가됩니다. 이러한 반응 확률이 어떻게 계산되는지 보려면 "열" 패널에서 각 열 옆의 계산식 아이콘  을 두 번 클릭합니다.

파티션 플랫폼 시작

분석 > 예측 모델링 > 파티션을 선택하여 파티션 플랫폼을 시작합니다.

그림 4.5 파티션 시작 창

결정 트리를 생성하여 반응을 예측합니다.

열 선택

14개 열

- Y
- Y Binary
- Y Ordinal
- Age
- Gender
- BMI
- BP
- Total Cholesterol
- LDL
- HDL
- TCH
- LTG
- Glucose
- Validation

선택한 열 역할 지정

역할	지정
Y, 반응	필수 선택적
X, 요인	필수 선택적
가중치	선택적 숫자
빈도	선택적 숫자
검증	선택적 숫자
기준	선택적

작업

확인
취소
제거
재호출
도움말

옵션

방법: 결정 트리

검증 데이터 비율: 0.3

☒ 결측값 정보화

☒ 순서형의 순서를 지킴

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

Y, 반응 분석할 하나 이상의 반응 변수입니다.

X, 요인 예측 변수입니다.

가중치 분석의 각 행에 가중치를 할당하는 숫자 값이 포함된 열입니다.

빈도 분석의 각 행에 빈도를 할당하는 숫자 값이 포함된 열입니다.

JMP PRO 검증 검증 데이터 집합을 정의하는 숫자 열입니다. 자세한 내용은 "[파티션 검증](#)"에서 확인하십시오. "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 "[검증 열 생성](#)"에서 확인하십시오.

기준 개별 분석을 정의하는 수준이 포함된 하나 이상의 열입니다. 열의 각 수준에 대해 해당 행은 사용자가 지정하는 다른 변수를 사용하여 분석됩니다. 결과는 별도의 보고서에 나타납니다. 기준 변수가 둘 이상 할당되면 기준 변수의 가능한 각 수준 조합에 대해 개별 보고서가 생성됩니다.

JMP PRO 방법 분할 방법 (붓스트랩 포레스트, 부스티드 트리, K 최근접 이웃 또는 Naive Bayes) 을 선택할 수 있습니다.

이러한 방법에 대한 자세한 내용은 "붓스트랩 포레스트", "부스티드 트리", "K 최근접 이웃" 및 "Naive Bayes" 에서 확인하십시오.

검증 데이터 비율 검증 데이터 집합으로 사용할 데이터 비율입니다. 자세한 내용은 "파티션 검증" 에서 확인하십시오.

결측값 정보화 이 옵션을 선택하면 범주형 예측 변수에 대한 결측값 범주화 및 연속형 예측 변수에 대한 결측값 정보화 처리가 활성화됩니다. 자세한 내용은 "결측값 정보화" 에서 확인하십시오.

순서형의 순서를 지킴 이 옵션을 선택하면 순서를 유지하는 분할만 고려하도록 제한됩니다.

결측값 정보화

파티션 플랫폼의 "결측값 정보화" 옵션을 사용하면 예측 변수의 결측값을 정보로 처리할 수 있습니다. 적합 모형은 결정적 모형입니다. "결측값 정보화" 옵션은 시작 창에 있으며 기본적으로 선택되어 있습니다. "결측값 정보화" 옵션을 선택하면 범주형 예측 변수와 연속형 예측 변수가 다르게 처리됩니다.

- 범주형 예측 변수의 결측값을 포함하는 행은 해당 변수에 대한 별도의 수준으로 분석에 포함됩니다.
- 연속형 예측 변수의 결측값을 포함하는 행은 다음과 같이 분할에 할당됩니다. 연속형 예측 변수의 값이 정렬됩니다. 먼저 결측 행이 정렬된 값의 아래쪽 끝에 있다고 간주됩니다. 모든 분할이 생성됩니다. 그런 다음 결측 행이 정렬된 값의 위쪽 끝에 있다고 간주됩니다. 다시 모든 분할이 생성됩니다. 최적 분할은 LogWorth 기준을 사용하여 결정됩니다. 지정된 예측 변수에 대해 추가 분할을 수행하는 경우 알고리즘은 해당 예측 변수에 의해 유도된 첫 번째 분할에 따라 결측 행을 높은 값 또는 낮은 값으로 커밋합니다.

"결측값 정보화" 옵션을 선택하지 않으면 결측값이 다음과 같이 처리됩니다.

- 결측값이 있는 예측 변수가 분할 변수로 사용되면 해당 예측 변수에 대한 결측값이 있는 각 행이 분할의 양쪽 중 하나에 랜덤 할당됩니다.
- 결측값이 있는 예측 변수가 분할 변수로 처음 사용되면 결측값 대치 열이 요약 보고서에 추가되어 결측값 대치 수를 표시합니다. 결측값 대치가 추가로 발생하면 결측값 대치 열이 업데이트됩니다 (). 여기서는 결측값 대치가 5 번 수행되었습니다.

참고: 결측값 대치 수는 결측값이 포함된 행 수보다 클 수 있습니다. 결측값 대치는 각 분할에서 발생합니다. 결측값이 있는 행은 여러 번 랜덤 할당될 수 있습니다. 행이 랜덤 할당될 때마다 결측값 대치 수가 증가합니다.

분할 보고서

초기 분할 보고서에는 분할 그림, 컨트롤 버튼, 요약 패널 및 결정 트리가 표시됩니다. 분할 그림과 결정 트리는 분할 없이 초기화됩니다. 보고서 상세 정보는 범주형 반응과 연속형 반응에 대해서로 다릅니다.

- " 분할 그림 "
- " 분할 버튼 "
- " 적합 통계량 요약 "
- " 노드 보고서 "

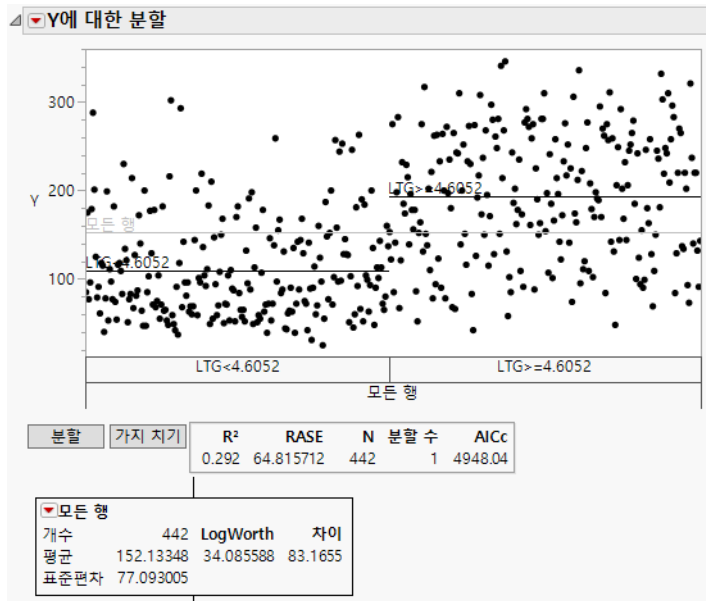
분할 그림

분할 그림의 각 점은 데이터 테이블의 관측값을 나타냅니다. 검증을 사용하는 경우 그림에는 훈련 데이터의 관측값만 표시됩니다. 초기 분할 그림에는 분할이 표시되지 않습니다. 분할 그림 모양은 연속형 반응과 범주형 반응에 대해서로 다릅니다.

연속형 반응에 대한 분할 그림

반응이 연속형인 경우 왼쪽 세로 축은 관측값을 나타냅니다. 그림에는 반응의 전체 평균을 나타내는 초기 가로선이 있습니다. 분할을 시작하면 결정 트리의 각 노드에 대한 평균 반응 값을 보여주는 다른 가로선이 추가됩니다. 분할은 텍스트 설명 및 세로선과 함께 가로 축 아래에 표시됩니다. 분할이 생성되거나 제거되면 관측값이 각 노드로 재구성됩니다. 가장 최근 분할은 가로 축 바로 아래에서 기존 분할 위에 나타납니다. 결정 트리의 분할 또는 가지 치기를 수행할 때마다 그림이 업데이트됩니다.

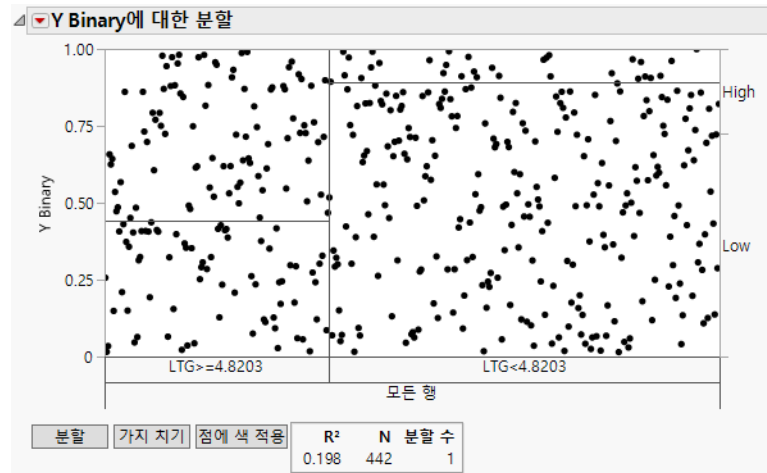
그림 4.6 연속형 반응에 대한 분할 보고서



범주형 반응에 대한 분할 그림

반응이 범주형인 경우 왼쪽 세로 축은 각 반응 결과의 비율을 나타내고 오른쪽 세로 축은 반응 수준이 그림에 표시되는 순서를 보여 줍니다. 그림에는 표시된 첫 번째 반응 수준의 전체 비율을 나타내는 초기 가로선이 있습니다. 분할을 시작하면 결정 트리의 각 노드에 대해 첫 번째 표시된 반응 수준의 비율을 보여 주는 다른 가로선이 추가됩니다. 분할은 텍스트 설명 및 그림의 관측값을 분할하는 세로선과 함께 가로 축 아래에 표시됩니다. 세로선은 그림까지 이어지고 각 노드의 경계를 나타냅니다. 가장 최근 분할은 가로 축 바로 아래에서 기존 분할 위에 나타납니다. 결정 트리의 분할 또는 가지 치기를 수행할 때마다 그림이 업데이트됩니다.

그림 4.7 범주형 반응에 대한 분할 보고서



분할 버튼

분할 보고서의 버튼을 사용하여 결정 트리와 상호 작용할 수 있습니다.

분할 최적 분할을 사용하여 데이터 파티션을 생성합니다. 여러 분할을 지정하려면 Shift 키를 누른 채 **분할**을 클릭합니다.

가지 치기 최악 분할을 제거합니다.

시작 (검증을 사용하는 경우에만 사용 가능) 검증 통계량이 최적화될 때까지 결정 트리에 분할을 자동으로 추가합니다. 자세한 내용은 "**파티션 검증**"에서 확인하십시오. 검증을 사용하지 않으면 파티션 모형에 사용할 분할 수만 결정하면 됩니다.

점에 색 적용 (범주형 반응에만 사용 가능) 반응 수준에 따라 관측값에 색상을 지정합니다. 이 색상은 데이터 테이블에 추가됩니다.

적합 통계량 요약

분할 보고서의 요약 섹션에는 훈련 데이터와 검증 및 테스트 데이터 (사용된 경우)에 대한 적합 통계량이 제공됩니다. 결정 트리의 분할을 추가하거나 가지 치기를 수행하면 요약 패널의 적합 통계량이 업데이트됩니다.

R² 현재 R² 값입니다.

N 관측값 수입니다.

분할 수 결정 트리의 현재 분할 수입니다.

RASE (연속형 반응에만 사용 가능) 제공된 평균 제곱 오차입니다. 자세한 내용은 선형 모형 적합에서 확인하십시오.

AICc (연속형 반응에만 사용 가능) 수정된 Akaike 정보 기준입니다. 자세한 내용은 선형 모형 적합에서 확인하십시오.

노드 보고서

분할 보고서 트리의 각 노드에는 보고서 및 추가 옵션이 포함된 빨간색 삼각형 메뉴가 있습니다. 자세한 내용은 "[노드 옵션](#)"에서 확인하십시오. 최종 노드에는 "후보" 보고서도 있습니다.

개수 노드로 특성화된 훈련 관측값 수입니다.

평균 (연속형 반응에만 사용 가능) 해당 노드의 모든 관측값에 대한 평균 반응입니다.

표준편차 (연속형 반응에만 사용 가능) 해당 노드의 모든 관측값에 대한 반응 표준편차입니다.

G² (범주형 반응에만 사용 가능) 연속형 반응에 사용되는 제곱합 대신 범주형 반응에 사용되는 적합 통계량입니다. 값이 작을수록 더 나은 적합을 나타냅니다. 자세한 내용은 "[파티션 플랫폼에 대한 통계 상세 정보](#)"에서 확인하십시오.

LogWorth $-\log_{10}(p \text{ 값})$ 으로 정의된 LogWorth 통계량입니다. 최적 분할은 LogWorth 를 최대화하는 분할입니다. 자세한 내용은 "[파티션 플랫폼에 대한 통계 상세 정보](#)"에서 확인하십시오.

차이 (연속형 반응에만 사용 가능) 두 노드 간 반응 평균의 차이입니다.

후보 각 열에 대해 "후보" 보고서는 해당 열의 최적 분할에 대한 상세 정보를 제공합니다. 모든 항에 대한 최적 분할은 별표로 표시됩니다.

항 후보 열을 표시합니다.

후보 G² (범주형 반응에만 사용 가능) 최선 분할에 대한 가능도비 카이제곱입니다. G² 값이 가장 큰 예측 변수에서 분할하면 모형 G² 의 축소가 최대화됩니다.

후보 SS (연속형 반응에만 사용 가능) 최선 분할에 대한 제곱합입니다.

LogWorth $-\log_{10}(p \text{ 값})$ 으로 정의된 LogWorth 통계량입니다. 최적 분할은 LogWorth 를 최대화하는 분할입니다. 자세한 내용은 "[파티션 플랫폼에 대한 통계 상세 정보](#)"에서 확인하십시오.

절단점 분할을 결정하는 예측 변수의 값입니다. 범주형 항의 경우 맨 왼쪽 분할의 수준이 나열됩니다.

최적 분할은 별표로 표시됩니다. 그러나 한 변수에 대해 후보 G² 또는 후보 SS가 더 높지만 다른 변수에 대해서는 LogWorth가 더 높은 경우가 있습니다. 이 경우 > 및 < 기호를 사용하여 각 변수의 최적 방향을 가리킵니다. 별표는 합치 상태에 해당합니다. 자세한 내용은 "[파티션 플랫폼에 대한 통계 상세 정보](#)"에서 확인하십시오.

노드 옵션

이 섹션에서는 각 노드의 빨간색 삼각형 메뉴에 있는 옵션에 대해 설명합니다.

최선 분할 이 노드 또는 그 아래에서 최선 분할을 찾아 실행합니다.

여기에서 분할 선택한 노드에서 가장 적절한 열을 기준으로 분할합니다.

세부 분할 분할이 실행되는 위치를 지정할 수 있습니다. 이 옵션은 기준을 절단점의 함수로 표시하고 사용자 절단점을 결정할 때 유용합니다. 분할 열을 지정할 때 분할 방법에 대해 다음 옵션을 선택할 수 있습니다.

최적값 선택한 변수의 최적값에서 분할합니다.

지정한 값 분할이 실행되는 수준을 지정할 수 있습니다.

출력 분할 테이블 가능한 모든 분할 및 연관된 분할 값을 표시하는 데이터 테이블을 생성합니다.

가지 치기 - 아래 선택한 노드 아래의 분할을 제거합니다.

가지 치기 - 최악 항목 선택한 노드 아래에서 최악 분할을 찾아 제거합니다.

행 선택하기 이 옆에 해당하는 데이터 테이블 행을 선택합니다. Shift 키를 누른 채 다른 노드에서 이 명령을 선택하여 선택 영역을 확장할 수 있습니다.

상세 정보 표시 선택한 변수에 대한 분할 기준을 표시하는 데이터 테이블을 생성합니다. 분할 구간 및 관련 기준 값으로 구성된 데이터 테이블에는 해당 기준에 대한 그래프를 생성하는 스크립트가 연결되어 있습니다.

잠금 해당 노드 또는 하위 노드를 분할 대상으로 선택할 수 없게 합니다. 이 옵션을 선택하면 노드 제목에 잠금 아이콘이 나타납니다.

파티션 플랫폼 옵션

"분할"의 빨간색 삼각형 메뉴 옵션을 사용하면 필요에 따라 보고서를 사용자 정의할 수 있습니다. 사용 가능한 옵션은 분석에 사용되는 데이터 유형에 따라 결정됩니다.

표시 옵션 보고서 요소를 표시하거나 숨기는 옵션이 포함되어 있습니다.

점 표시 점을 표시합니다. 범주형 반응의 경우 이 옵션을 선택하면 점 또는 색상이 지정된 패널이 표시됩니다.

트리 표시 큰 분할 트리를 표시합니다.

그래프 표시 분할 그래프를 표시합니다.

분할 막대 표시 (범주형 반응에만 사용 가능) 각 잎의 분할 비율을 나타내는 색상 막대를 표시합니다.

분할 통계량 표시 분할 통계량을 표시합니다. 범주형 분할 통계량 G^2 에 대한 자세한 내용은 "[파티션 플랫폼에 대한 통계 상세 정보](#)"에서 확인하십시오.

분할 확률 표시 (범주형 반응에만 사용 가능) 노드 보고서에 비율 및 확률 통계량을 표시합니다.

분할 개수 표시를 선택하면 비율 및 확률 통계량이 자동으로 표시됩니다. 비율 및 확률에 대한 자세한 내용은 "[파티션 플랫폼에 대한 통계 상세 정보](#)"에서 확인하십시오.

분할 개수 표시 (범주형 반응에만 사용 가능) 노드 보고서에 빈도 수를 표시합니다. 이 옵션을 선택하면 **분할 확률 표시**가 자동으로 선택됩니다. **분할 확률 표시**를 선택 취소하면 개수가 표시되지 않습니다.

분할 후보 표시 후보 보고서를 표시합니다.

분할 후보 정렬 후보 보고서를 통계량 또는 $\log(\text{worth})$ 를 기준으로 정렬합니다.

최선 분할 트리를 최적 분할 지점에서 분할합니다. 이 옵션은 **분할** 버튼을 클릭하는 것과 같습니다.

가지 치기 - 최악 항목 최소 판별 능력이 있는 최종 분할을 제거합니다. 이 옵션은 **가지 치기** 버튼을 클릭하는 것과 같습니다.

최소 크기 분할 숫자 또는 총 표본 크기의 소수 부분을 입력하여 허용되는 최소 크기 분할을 정의합니다. 숫자를 지정하려면 1 보다 크거나 같은 값을 입력합니다. 표본 크기의 비율을 지정하려면 1 보다 작은 값을 입력합니다. 기본값은 최대값 5로 설정되거나, 행 수를 10,000 으로 나눈 값 (내림) 으로 설정됩니다.

열 잠그기 열이 분할 대상에 포함되지 않도록 해당 열을 대화식으로 잠급니다. 개별 잠금에 영향을 주지 않고 표시를 해제하거나 다시 설정할 수 있습니다.

실제값 대 예측값 그림 (연속형 반응에만 사용 가능) 실제 반응 값 대 예측 반응 값 그림을 표시합니다. 결정 트리를 적합시키면 잎의 모든 관측값이 동일한 예측값을 가집니다. 잎이 n 개인 경우 실제값 대 예측값 그림에는 최대 n 개의 개별 예측값이 표시됩니다. 실제 값은 n 개의 세로선에서 각 잎 평균 주위에 점이 분산된 산점도를 구성합니다.

그림에는 $Y=X$ 선인 대각선이 포함됩니다. 모든 점이 이 대각선 위에 있을 때 완벽한 적합이 됩니다. 검증이 사용되면 훈련 데이터 집합과 검증 데이터 집합 둘 다에 대해 그림이 표시됩니다 ().

작은 트리 보기 분할 그래프 오른쪽에 소형 버전의 분할 트리를 표시합니다.

3D 트리 트리 구조의 3D 그림을 표시합니다. 이 옵션에 액세스하려면 Shift 키를 누른 채 빨간색 삼각형 메뉴를 클릭합니다.

잎 보고서 보고서의 맨 아래 수준 잎에 대한 평균과 개수 또는 비율을 표시합니다.

잎 기여도 적합에 대한 각 입력 열의 기여도를 나타내는 보고서를 표시합니다. 이 보고서에는 분할을 정의한 횟수와 해당 열에 기인한 총 G^2 또는 제곱합도 표시됩니다.

분할 기록 R^2 대 분할 수 그림을 표시합니다. 제외된 행 검증, 홀드백 검증 또는 검증 열을 사용하는 경우 훈련 및 검증 R^2 값에 대한 별도의 곡선이 표시됩니다. 훈련 데이터 집합의 R^2 곡선은 파란색이고 검증 데이터 집합의 R^2 곡선은 빨간색입니다. K 폴드 교차 검증을 선택하는 경우 모든 데이터에 대한 R^2 곡선은 파란색이고 교차 검증 R^2 곡선은 녹색입니다.

ROC 곡선 (범주형 반응에만 사용 가능) ROC(Receiver Operating Characteristic) 곡선은 반응 수준을 정렬하기 위한 모형 적합 확률의 효율성을 표시합니다. 자세한 내용은 "**ROC 곡선**"에서 확인하십시오.

향상도 곡선 (범주형 반응에만 사용 가능) 향상도 곡선은 파티션 모형의 예측 능력을 표시합니다. 자세한 내용은 "**향상도 곡선**"에서 확인하십시오.

결정 임계값 (이항 범주형 반응에만 사용 가능) 혼련, 검증 및 테스트 (지정된 경우) 데이터 집합에 대한 "결정 임계값" 보고서를 표시하거나 숨깁니다. 각 보고서에는 모형 적합 비교를 위한 각 모형의 적합 확률 분포 그래프, 혼동 행렬 및 분류 그래프와 분류 정확성 측정 기준 테이블이 포함되어 있습니다. 자세한 내용은 "[결정 임계값 보고서](#)"에서 확인하십시오.

적합 상세 정보 표시 (반응이 범주형인 경우에만 표시됨) 적합 상세 정보 보고서에는 여러 적합 측도가 표시되고 혼동 행렬 보고서가 제공됩니다. 자세한 내용은 "[적합 상세 정보 표시](#)"에서 확인하십시오.

열 저장 모형 및 트리 결과를 저장하고 SAS 코드를 생성하기 위한 옵션이 포함되어 있습니다.

잔차 저장 모형의 잔차 값을 데이터 테이블에 저장합니다.

예측값 저장 모형의 예측값을 데이터 테이블에 저장합니다.

잎 번호 저장 트리의 잎 번호를 데이터 테이블의 열에 저장합니다.

잎 라벨 저장 트리의 잎 라벨을 데이터 테이블에 저장합니다. 라벨은 트리를 따라 행이 추가하는 각 분지에 대한 내용을 기록합니다. 각 분지는 "&" 로 구분됩니다. 예를 들면 "size(Small,Medium)&size(Small)"로 라벨을 표시할 수 있습니다. 그러나 JMP에서는 반복되는 범주 라벨 형식으로 중복 정보를 포함하지 않습니다. 잎의 범주 라벨은 상위 트리 노드의 포괄적 범주 목록을 나타낼 수도 있습니다. 중복 라벨이 있는 트리 노드 위치에 캐럿 (^) 이 나옵니다. 따라서 "size(Small,Medium)&size(Small)"은 ^&size(Small)로 표시됩니다.

예측 계산식 저장 예측 계산식을 데이터 테이블의 하나 또는 여러 열에 저장합니다. 계산식은 트리 구조를 설명하는 내포된 조건부 절로 구성됩니다. 반응이 연속형이면 "예측" 특성이 포함된 열 하나가 추가됩니다. 반응이 범주형이면 반응의 각 수준에 대해 "반응 확률" 특성이 포함된 여러 열이 추가됩니다. 또한 각 관측값에 대해 발생 확률이 가장 높은 반응 수준이 포함된 최대 확률 분류 열이 추가됩니다.

허용 예측 계산식 저장 결측값이 있고 "결측값 정보화"를 선택하지 않은 경우에도 예측하는 계산식을 저장합니다. 예측 계산식은 결측 예측 변수에 대한 반응 값을 분할에 랜덤 할당하여 결측값을 허용합니다. 반응이 연속형이면 열에 "예측" 특성이 포함됩니다. 반응이 범주형이면 열에 "반응 확률" 특성이 포함됩니다. "결측값 정보화"를 선택한 경우 Shift 키를 누른 채 보고서의 빨간색 삼각형을 클릭하면 허용 예측 계산식을 저장할 수 있습니다.

잎 번호 계산식 저장 잎 번호 계산식이 포함된 열을 데이터 테이블에 저장합니다.

잎 라벨 계산식 저장 잎 라벨 계산식이 포함된 열을 데이터 테이블에 저장합니다.

SAS DATA 스텝 만들기 새 데이터 집합을 스코어링하기 위한 SAS 코드를 생성합니다.

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 "[계산식 저장소](#)"에서 확인하십시오.

JMP PRO **허용 예측 계산식 게시** 허용 예측 계산식을 생성하여 계산식 저장소 플랫폼에 계산식 열 스크립트로 저장합니다. "계산식 저장소" 보고서가 열려 있지 않으면 "계산식 저장소" 보고서가 생성됩니다. 자세한 내용은 "[계산식 저장소](#)"에서 확인하십시오. "결측값 정보화"를 선택한 경우 Shift 키를 누른 채 보고서의 빨간색 삼각형을 클릭하면 이 옵션을 사용할 수 있습니다.

수익 행렬 지정 (범주형 반응에만 사용 가능) 올바르게나 올바르게지 않은 분류 결정과 관련된 수익 또는 비용을 지정할 수 있습니다. 명목형 반응의 경우 확률 임계를 사용하여 수익 행렬 항목을 지정할 수 있습니다. 자세한 내용은 "[수익 행렬 지정](#)"에서 확인하십시오.

프로파일러 대화식 프로파일러 보고서를 표시합니다. 요인 값을 변경하면 추정된 분류 확률에 반영됩니다. 자세한 내용은 **Profilers**에서 확인하십시오.

점에 색 적용 (범주형 반응에만 사용 가능) 반응 수준에 따라 점에 색상을 적용합니다. 이 옵션은 "점에 색 적용"버튼을 클릭하는 것과 같습니다. 자세한 내용은 "[분할 보고서](#)"에서 확인하십시오.

다음 옵션에 대한 자세한 내용은 **JMP 사용의**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.

적합 상세 정보 표시

파티션 플랫폼의 "적합 상세 정보" 보고서에는 다음과 같은 적합 측도가 포함되어 있습니다.

엔트로피 R^2 적합 모형과 상수 확률 모형의 로그 가능도를 비교하는 적합 측도입니다. 엔트로피 R^2 범위는 0 ~ 1이며, 값이 1에 가까울수록 더 나은 적합을 나타냅니다. 자세한 내용은 "[엔트로피 \$R^2\$](#) "에서 확인하십시오.

일반화 R^2 일반 회귀 모형에 적용할 수 있는 측도입니다. 이 값은 가능도 함수 L을 기반으로 하며 최대값이 1이 되도록 척도화됩니다. 일반화 R^2 측도는 표준 최소 제곱 설정 시 연속형 정규 반응에 대한 기준 R^2 로 단순화됩니다. 일반화 R^2 을 Nagelkerke 또는 Craig 와 Uhler R^2 이라고도 하는데, 이는 Cox-Snell 유사 R^2 을 정규화한 것입니다. 자세한 내용은 Nagelkerke 연구 자료 (1991)에서 확인하십시오. 값이 1에 가까울수록 더 나은 적합을 나타냅니다.

평균 -Log p $-\log(p)$ 의 평균입니다. 여기서 p 는 발생한 사건과 연관된 적합 확률입니다. 값이 작을수록 더 나은 적합을 나타냅니다.

RASE 제공된 평균 제공 오차입니다. 여기서 차이는 반응과 p (실제로 발생한 사건의 적합 확률) 사이의 차이입니다. 값이 작을수록 더 나은 적합을 나타냅니다.

평균 절대 편차 반응과 p (실제로 발생한 사건의 적합 확률) 사이의 차이에 대한 절대값을 평균한 것입니다. 값이 작을수록 더 나은 적합을 나타냅니다.

오분류 비율 적합 확률이 가장 높은 반응 범주가 관측된 범주가 아닌 비율입니다. 값이 작을수록 더 나은 적합을 나타냅니다.

혼동 행렬 보고서

"혼동 행렬" 보고서에는 훈련 데이터 집합과 검증 및 테스트 데이터 집합(정의된 경우)에 대한 행렬이 표시됩니다. "혼동 행렬" 보고서에는 혼동 행렬과 혼동 비율 행렬이 포함됩니다. 혼동 행렬은 실제 반응과 예측 반응의 이원 분류입니다. 혼동 비율 행렬은 값을 행 합계로 나눈 혼동 행렬과 동일합니다.

결정 행렬 보고서

수익 행렬이 정의되면 파티션 알고리즘에서 행렬의 값을 사용하여 각 결정에 대한 수익을 계산합니다. "수익 행렬" 열 특성을 사용하거나, "수익 행렬 지정" 옵션으로 비용을 지정하여 수익 행렬을 정의할 수 있습니다. 자세한 내용은 "수익 행렬 지정"에서 확인하십시오.

"적합 상세 정보 표시"를 선택하면 "결정 행렬" 보고서가 나타납니다. "결정 행렬" 보고서에서 결정 개수는 수익 행렬의 가중치에 따라 수익성이 가장 높은 예측 결정을 반영합니다. 이 보고서는 훈련 데이터 집합과 검증 및 테스트 데이터 집합(정의된 경우)에 대한 결정 개수 및 결정 비율 행렬을 제공합니다. 참조를 위해 수익 행렬도 표시됩니다.

참고 : "수익 행렬 지정" 옵션을 사용하여 수익 행렬의 가중치를 변경하면 "결정 행렬" 보고서가 자동으로 업데이트되어 변경 내용을 반영합니다.

결정 개수 행렬 행에 실제 반응을 사용하고 열에 분류 개수를 사용하여 이원 분류를 표시합니다.

지정된 수익 행렬 수익 행렬을 정의하는 가중치를 제공합니다.

결정 비율 행렬 주어진 행에서 각 범주로 분류되는 관측값의 비율에 해당하는 비율 값을 표시합니다. 모든 관측값이 올바르게 분류되면 대각선에 나타나는 비율이 모두 1 이 됩니다.

팁 : 비용이 1 과 -1 인 기본 수익 행렬을 사용하여 반응에 대한 결정 비율 행렬을 얻을 수 있습니다. 빨간색 삼각형 메뉴에서 **수익 행렬 지정**을 선택한 후 기본값을 변경하지 않고 **확인**을 클릭합니다.

행렬은 다음과 같이 두 행으로 배열됩니다.

- 결정 개수 행렬은 첫 번째 행에 있습니다.
- 지정된 수익 행렬은 첫 번째 행의 오른쪽에 있습니다.

- 결정 비율 행렬은 두 번째 행에 있습니다.

수익 행렬 지정

범주형 반응의 경우 파티션 플랫폼에서 수익 행렬을 지정할 수 있습니다. 수익 행렬은 원하지 않는 결과에 비용을 할당하고 원하는 결과에 수익을 지정하는 데 사용됩니다.

그림 4.8 수익 행렬 지정 창

수익 행렬 지정

행의 반응이 실제 반응일 때 열의 반응을 예측할 경우 각 행렬 항목이 수익입니다.

대각에 올바른 결정에 대한 수익을 반영하는 값을 입력하십시오(양수).
비대각에 올바른지 않은 결정에 대한 손실을 반영하는 값을 입력하십시오(음수).
대체 결정에 대한 수익성을 반영하려면 결정되지 않음 열을 사용하십시오.

예측 계산식을 저장할 때 이러한 값을 사용하여 최선의 결정 열이 생성됩니다.
최선의 결정은 기대 수익이 가장 큰 결정입니다.

결정 또는 예측

	High	Low	결정되지 않음
Low	1	-1	.
High	-1	1	.

이항 반응에 대한 수익 행렬을 생성하려면 목표값 및 확률 임계를 입력하십시오.
예측 확률이 임계를 초과하면 최선의 결정은 목표값이 됩니다.

목표 수준:

확률 임계:

☐ 열에 특성으로 저장

실제 반응 범주와 예측 반응 범주의 각 조합에 수익 및 비용 값을 할당할 수 있습니다. 대체 범주로 분류하는 비용을 지정하려면 "결정되지 않음" 열에 값을 입력합니다. 할당을 반응 열에 특성으로 저장하려면 **열에 특성으로 저장**을 선택합니다. 이 옵션을 선택하지 않으면 수익 행렬이 현재 분할 보고서에만 적용됩니다.

수익 행렬의 확률 임계 지정

이항 반응의 경우 수익 행렬에 가중치를 입력하는 대신 "수익 행렬" 창에서 확률 임계를 지정할 수 있습니다. 수익 행렬에 대한 값을 계산하는 방법은 **JMP 사용**에서 확인하십시오.

목표값 확률을 모델링할 수준입니다.

확률 임계 목표 수준의 확률에 대한 임계입니다. 관측값이 목표 수준 이내가 될 확률이 확률 임계를 초과할 경우 해당 관측값은 해당 수준으로 분류됩니다.

"수익 행렬 지정" 옵션을 사용하여 비용을 정의한 후 "적합 상세 정보 표시"를 선택하면 "결정 행렬" 보고서가 나타납니다. 자세한 내용은 **"결정 행렬 보고서"**에서 확인하십시오.

수익 행렬을 지정하고 모형 예측 계산식을 저장하면 다음을 포함하는 계산식 열이 데이터 테이블에 저장됩니다.

- <level>에 대한 수익: 열의 각 반응 수준에 대해 각 관측값을 해당 수준으로 분류하기 위한 기대 수익을 지정합니다.
- <column name>에 대한 수익성이 가장 높은 예측: 각 관측값에 대해 기대 수익이 가장 높은 반응의 수준을 지정합니다.
- <column name>에 대한 기대 수익: 각 관측값에 대해 수익성이 가장 높은 예측 열에 의해 정의된 분류의 기대 수익을 지정합니다.
- <column name>에 대한 실제 수익: 각 관측값에 대해 해당 관측값을 수익성이 가장 높은 예측 열에 의해 지정된 수준으로 분류하기 위한 실제 수익을 지정합니다.

ROC 곡선

ROC(Receiver Operating Characteristic) 곡선은 반응 수준을 정렬하기 위한 모형 적합 확률의 효율성을 표시합니다. ROC 곡선은 범주형 반응에만 사용할 수 있습니다. ROC 곡선에 대한 소개는 기본 분석에서 확인하십시오.

파티션 모형의 각 관측값에 대한 예측 반응은 0과 1 사이의 값입니다. 예측 반응을 사용하여 관측값을 양수 또는 음수로 분류하려면 절단점이 사용됩니다. 예를 들어 절단점이 0.5인 경우 예측 반응이 0.5 이상인 관측값은 Positive로 분류되고 0.5 미만인 관측값은 Negative로 분류됩니다. 절단점이 달라지면 분류가 절충됩니다.

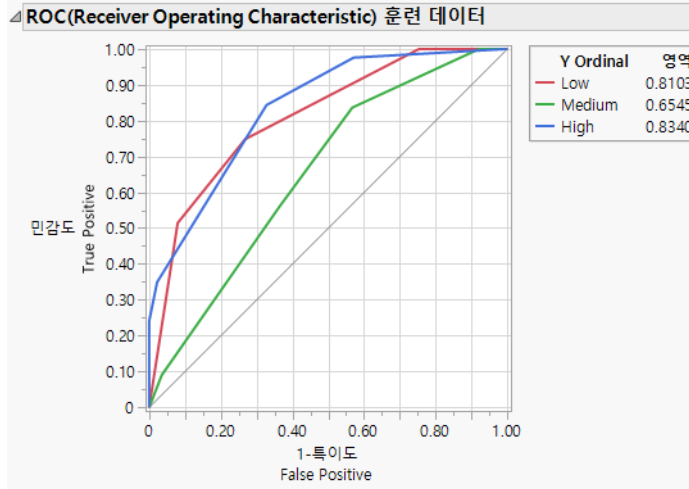
ROC 곡선을 생성하기 위해 각 예측 반응 수준을 가능한 절단점으로 간주하고 가능한 각 절단점에 대해 다음 값이 계산됩니다.

- 민감도는 True Positive 비율 또는 예측 반응이 절단점보다 큰 Positive 관측값의 비율입니다.
- 특이도는 True Negative 비율 또는 예측 반응이 절단점보다 작은 Negative 관측값의 비율입니다.

ROC 곡선은 민감도 대 1-특이도를 표시합니다. n 개의 분할이 있는 파티션 모형에는 $n+1$ 개의 예측값이 있습니다. 파티션 모형의 ROC 곡선에는 $n+1$ 개의 선분이 있습니다.

반응 수준이 세 개 이상인 경우 분할 보고서에는 각 반응 수준 대 다른 수준에 대한 개별 ROC 곡선이 포함됩니다. 각 곡선은 수준을 Positive 반응 수준으로 표현한 것입니다. 수준이 두 개뿐인 경우 한 곡선은 다른 곡선을 반영합니다.

그림 4.9 수준이 세 개인 반응의 ROC 곡선



모형이 반응 값의 순위를 완벽하게 지정하면 정렬된 데이터에 먼저 모든 양의 값이 포함된 후 다른 모든 값이 포함됩니다. 이 경우 곡선은 오른쪽으로 이동하기 전에 맨 위로 이동합니다. 모형이 잘 예측하지 못하면 곡선은 그림의 왼쪽 아래에서 오른쪽 위로 대각선을 따릅니다.

실제로 ROC 곡선은 대각선 위에 있습니다. 곡선 아래 면적은 모형의 적합도를 나타내는 표시자입니다. 값이 1이면 완벽한 적합을 나타내고 0.5에 가까우면 모형이 그룹을 구별할 수 없음을 나타냅니다.

반응 수준이 세 개 이상인 경우 ROC 곡선 그림을 통해 곡선 아래 면적이 가장 큰 반응 범주를 확인할 수 있습니다.

향상도 곡선

"향상도 곡선" 옵션은 파티션 모형의 예측 능력을 표시하는 다른 그림을 제공합니다. 향상도 곡선은 향상도 대 관측값 비율을 표시합니다. 고유한 각 예측 확률 값에 대한 점이 있습니다. 반응 수준의 각 예측 확률은 예측 확률이 고유 예측 확률 값보다 크거나 같은 관측값의 비율을 정의합니다. 특정 반응 수준에 대한 향상도 값은 관측 반응의 전체 비율 대비 해당 부분의 관측 반응 비율을 나타내는 값입니다.

참고: 작은 모형의 경우 대부분의 점이 동일한 예측 확률 값을 가질 수 있습니다. 이 확률이 반응 수준에 대한 가장 높은 예측 확률인 경우에는 향상도 곡선이 부분 = 0 에서 시작하지 않습니다. [그림 4.10](#)의 Low 향상도 곡선을 예로 들 수 있습니다.

그림 4.10 향상도 곡선

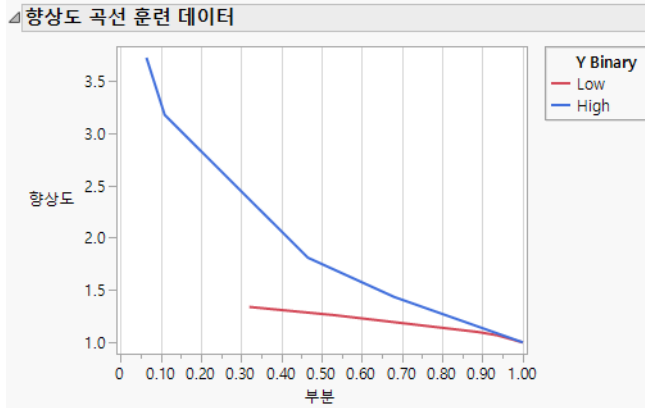


그림 4.11 향상도 곡선에 대한 향상도 테이블

Prob High	N > Prob High	Portion	N High in Portion	Portion High	Lift = portion high/ overall high of .27
0.97	20	0.06	20	1.00	3.72
0.77	44	0.14	39	0.89	3.30
0.33	68	0.22	47	0.69	2.57
0.31	168	0.54	78	0.46	1.73
0.04	309	1.00	83	0.27	1.00

그림 4.11에서는 그림 4.10에 나오는 High 향상도 곡선에 사용된 "향상도" 및 "부분"의 계산을 보여 주기 위한 값 테이블을 제공합니다. 반응 Y Binary를 예측하기 위해 5개의 분할이 있는 파티션 모형이 생성되었습니다. Y Binary에는 Low와 High라는 두 개의 수준이 있습니다. 향상도 곡선은 309개의 관측값을 기반으로 합니다. 전체 비율 0.27에 대해 관측된 High 반응은 83개입니다.

- Prob High: 파티션 모형에서 High 반응 수준에 대한 5개의 고유 예측 확률 값
- N > Prob High: 예측 확률 값이 Prob High 값보다 크거나 같은 관측값 수
- Portion: N > Prob High를 309, 즉 총 관측값 수로 나눈 값
- N High in Portion: 각 부분에서 High 반응이 관측된 관측값 수
- Portion High: N High in Portion을 N > Prob High로 나눈 값
- Lift: Portion High를 0.27, 즉 관측된 High 반응의 전체 비율로 나눈 값

향상도는 각 부분에 대한 기대 High 반응 수와 비교하여 해당 부분에 속하는 High 반응 수를 측정합니다. 데이터 집합의 처음 6%에 대한 향상도는 3.72입니다. 모형을 사용하여 예측값이 가장 높은 관측값의 6%를 선택한 결과, 해당 6%를 랜덤 선택한 경우보다 High 반응이 3.72개 더 많습니다.

파티션 검증

파티션 모형이 쉽게 과대적합된다는 점에서 파티션 모형과 함께 검증을 사용하는 것은 중요합니다. 이 경우 모형은 모형을 생성하는 데 사용된 데이터를 매우 잘 예측하지만 미래 관측값은 제대로 예측하지 못합니다. 검증은 데이터 집합의 일부를 사용하여 모형 모수를 추정하고 다른 부분을 사용하여 모형의 예측 능력을 평가하는 프로세스입니다. 검증에 대한 자세한 내용은 "[JMP 모델링의 검증](#)"에서 확인하십시오.

파티션에서 검증 방법을 사용할 때 **시작** 버튼이 나타납니다. **시작** 버튼을 사용하면 **분할** 버튼을 반복해서 클릭하지 않아도 분할을 반복할 수 있습니다. **시작** 버튼을 클릭하면 검증 R^2 이 다음 10개 분할을 통해 얻을 수 있는 값보다 나을 때까지 계속 분할합니다. 이 규칙은 해석하기 어렵지만 예측 검증력이 좋은 복잡한 트리를 생성할 수 있습니다.

시작 버튼을 사용하면 **분할 기록** 명령이 설정됩니다. **시작** 버튼을 사용하여 생성된 트리의 노드가 40개를 초과하는 경우 **트리 표시** 명령이 해제됩니다.

다음 검증 방법 중 하나를 선택합니다.

제외된 행 행 상태를 사용하여 데이터의 부분집합을 생성합니다. 제외되지 않은 행은 훈련 데이터 집합으로 사용되고 제외된 행은 검증 데이터 집합으로 사용됩니다.

행 상태 사용 및 행 제외 방법에 대한 자세한 내용은 **JMP 사용**에서 확인하십시오.

홀드백 원래 데이터를 무작위로 훈련 데이터 집합과 검증 데이터 집합으로 나눕니다. 플랫폼 시작 창의 검증 데이터 비율은 검증 데이터 집합으로 사용할 원래 데이터의 비율을 지정하는 데 사용됩니다 (홀드백). 검증 데이터 비율에 대한 자세한 내용은 "[파티션 플랫폼 시작](#)"에서 확인하십시오.

JMP PRO 검증 열 검증 데이터 집합을 정의하는 숫자 열을 사용합니다. 이 열에는 최대 세 개의 구분되는 값이 포함되어야 합니다.

- 검증 열에 두 개의 수준이 있는 경우 더 작은 값이 훈련 데이터 집합을 정의하고 더 큰 값이 검증 데이터 집합을 정의합니다.
- 검증 열에 세 개의 수준이 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.

"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 "[검증 열 생성](#)"에서 확인하십시오.

팁 : K 폴드 또는 내포 K 폴드 교차 검증을 사용하려면 모형 선별 플랫폼을 통해 파티션 모형을 적합시킵니다. 자세한 내용은 "[모형 선별](#)"에서 확인하십시오.

파티션 플랫폼의 추가 예

이 섹션에는 파티션 플랫폼을 사용하는 예가 포함되어 있습니다.

- "연속형 반응의 예"
- "결측값 정보화의 예"
- "수익 행렬 및 결정 행렬 보고서의 예"

연속형 반응의 예

이 예에서는 파티션 플랫폼을 사용하여 당뇨병 환자의 1년 질병 진행률(양적 척도로 측정)을 예측하는 결정 트리를 생성합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Diabetes.jmp**를 엽니다.
2. **분석 > 예측 모델링 > 파티션**을 선택합니다.
3. **Y**를 선택하고 **Y, 반응**을 클릭합니다.
4. **Age ~ Glucose**를 선택하고 **X, 요인**을 클릭합니다.
5. 설치된 JMP 유형에 따라 검증 절차를 선택합니다.
 - JMP Pro의 경우 **Validation**을 선택하고 **검증**을 클릭합니다.
 - JMP의 경우 **검증 데이터 비율**에 "0.3"을 입력합니다.

참고 : 검증 행의 랜덤 선택으로 인해 검증 비율을 사용한 결과는 여기에 표시된 것과 다를 수 있습니다.

그림 4.12 검증 데이터 비율 = 0.3 인 완료된 시작 창

결정 트리를 생성하여 반응을 예측합니다.

열 선택

14개 열

- Y
- Y Binary
- Y Ordinal
- Age
- Gender
- BMI
- BP
- Total Cholesterol
- LDL
- HDL
- TCH
- LTG
- Glucose
- Validation

선택한 열 역할 지정

Y, 반응 Y Binary
선택적

X, 요인 Age
Gender
BMI
BP

가중치 선택적 숫자

빈도 선택적 숫자

검증 선택적 숫자

기준 선택적

작업

확인

취소

제거

재호출

도움말

옵션

방법 결정 트리

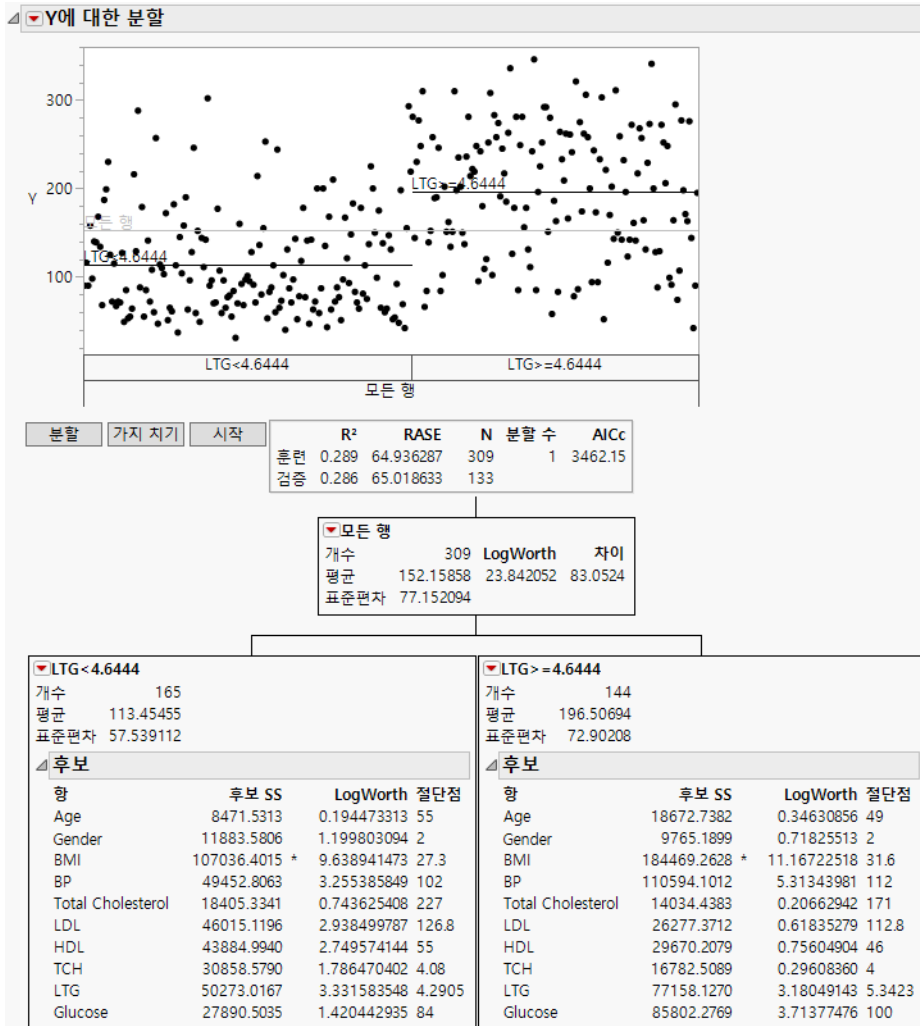
검증 데이터 비율 0.3

☒ 결측값 정보화

☒ 순서형의 순서를 지킴

6. **확인**을 클릭합니다.
7. 플랫폼 보고서 창에서 **분할**을 한 번 클릭하여 분할을 수행합니다.

그림 4.13 결정 트리가 숨겨진 첫 번째 분할 후 보고서



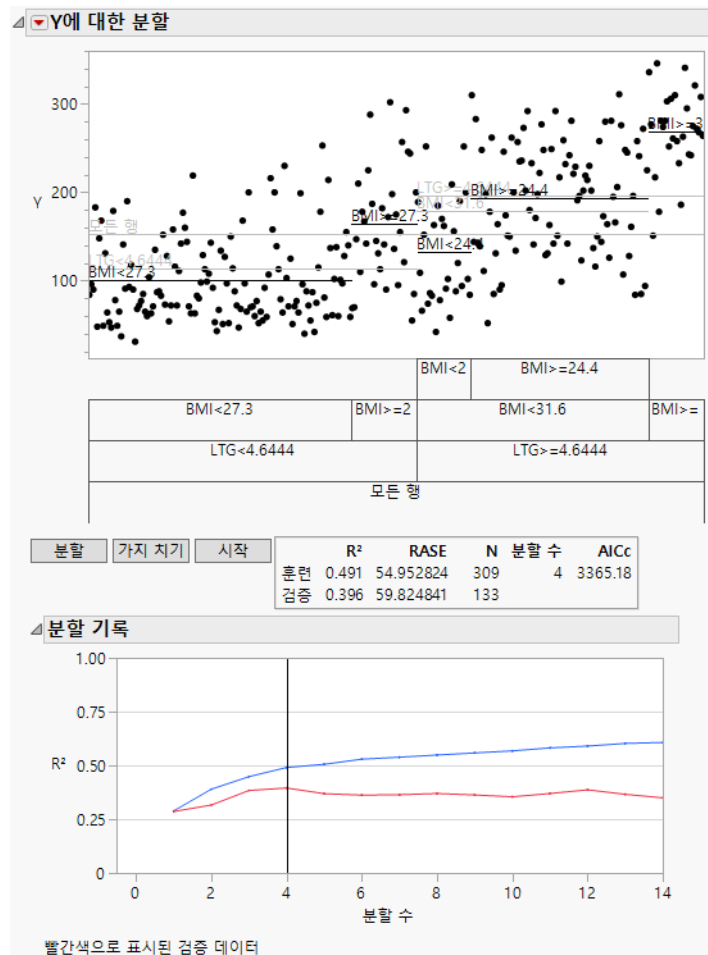
훈련 데이터 집합의 원래 값 309 개가 다음과 같이 두 부분으로 분할되었습니다.

- 왼쪽 옆 (LTG < 4.6444) 에 165 개의 관측값이 있습니다.
- 오른쪽 옆 (LTG >= 4.6444) 에 144 개의 관측값이 있습니다.

다음에는 오른쪽 옆과 왼쪽 옆 모두 BMI 에서 분할됩니다. BMI 에 대한 후보 SS 는 오른쪽 옆의 값이 왼쪽 옆의 값보다 높습니다. 따라서 다음 분할은 오른쪽 옆에서 수행됩니다.

8. 시작을 클릭하여 자동 분할을 사용합니다.

그림 4.14 검증을 사용한 자동 분할 후 보고서



네 개의 분할이 포함된 해를 찾았습니다."분할 기록" 그림에서는 4 번의 분할 후 검증 데이터 집합에서 더 이상 개선이 없음을 보여 줍니다. 검증 데이터의 R^2 값 0.39 는 이 모형이 질병 진행의 강력한 예측 변수로 도움 되지 않음을 나타냅니다. 분할 그림에서 각 파티션 간 산포는 이 모형이 Y 수준을 잘 분리하지 않음을 나타냅니다.

결측값 정보화의 예

이 예에서는 파티션 플랫폼을 사용하여 고객이 신용 위험 상태인지 예측하는 결정 트리 모형을 생성합니다. 데이터 집합에 결측값이 포함되어 있으므로 결측값 정보화 옵션의 유효성도 살펴 봅니다.

파티션 플랫폼 시작

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Equity.jmp 를 엽니다 .
2. **분석 > 예측 모델링 > 파티션**을 선택합니다 .
3. BAD 를 선택하고 **Y, 반응**을 클릭합니다 .
4. LOAN ~ DEBTINC 를 선택하고 **X, 요인**을 클릭합니다 .
5. **확인**을 클릭합니다 .

결측값 정보화를 사용하여 결정 트리 및 ROC 곡선 생성

1. Shift 키를 누른 채 **분할**을 클릭합니다 .
2. 분할 수에 "5" 를 입력하고 **확인**을 클릭합니다 .
3. "BAD 에 대한 분할 " 옆의 빨간색 삼각형을 클릭하고 **ROC 곡선**을 선택합니다 .
4. "BAD 에 대한 분할 " 옆의 빨간색 삼각형을 클릭하고 **열 저장 > 예측 계산식 저장**을 선택합니다 .

확률 (BAD==Good Risk) 및 확률 (BAD==Bad Risk) 열에 결측값 정보화 유틸리티에서 미래 대출 신청인의 신용 위험을 분류하기 위해 사용하는 계산식이 포함됩니다. 결측값 정보화를 사용하지 않는 모형과 비교하여 이 모형이 어떻게 수행되는지에 관심이 있습니다 .

결측값 정보화를 사용하지 않고 결정 트리 및 ROC 곡선 생성

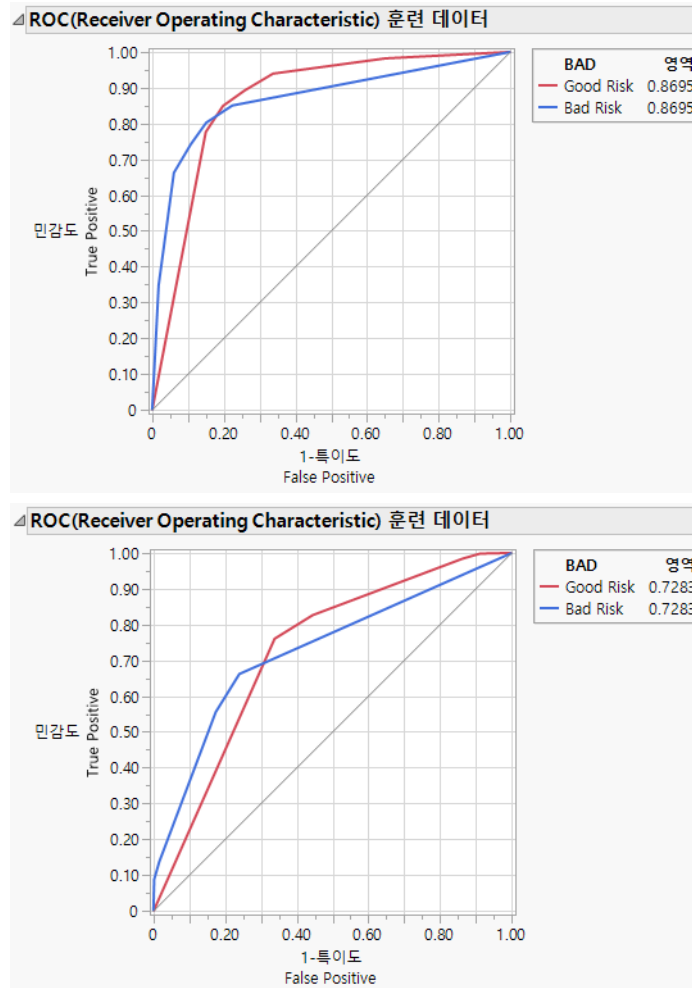
1. "BAD 에 대한 분할 " 옆의 빨간색 삼각형을 클릭하고 **다시 실행 > 분석 다시 시작**을 선택합니다 .
2. **결측값 정보화**를 선택 취소합니다 .
3. " 확인 " 을 클릭하고 " **결측값 정보화를 사용하여 결정 트리 및 ROC 곡선 생성** " 의 단계를 반복합니다 .

확률 (BAD==Good Risk) 2 및 확률 (BAD==Bad Risk) 2 에 결측값 정보화 유틸리티를 사용하지 않는 계산식이 포함됩니다 .

ROC 곡선 비교

두 모형의 ROC 곡선을 시각적으로 비교해 봅니다. 왼쪽 모형은 결측값 정보화를 사용한 모형이고 오른쪽 모형은 결측값 정보화를 사용하지 않은 모형입니다 .

그림 4.15 결측값 정보화를 사용한 모형의 ROC 곡선 (왼쪽) 과 결측값 정보화를 사용하지 않은 모형의 ROC 곡선 (오른쪽)



결측값 정보화를 사용한 모형의 AUC(곡선 아래 면적) 값 0.8695가 결측값 정보화를 사용하지 않은 모형의 AUC 값 0.7283 보다 높습니다 . 반응 수준이 두 개뿐이므로 각 모형의 ROC 곡선은 서로 다른 모형을 반영하고 AUC 가 같습니다 .

참고 : AUC 는 결측값 정보화를 사용하지 않은 모형에 대해 표시되는 것과 다를 수 있습니다 . 결측값 정보화를 사용하지 않는 경우 결측 행이 분할 측면에 랜덤 할당됩니다 . 분석을 다시 실행하면 결과에 약간 차이가 있을 수 있습니다 .

모형 비교 플랫폼 사용

다음으로, 모형 비교 플랫폼에서 모형을 비교하여 4 단계 및 3 단계에서 생성한 두 계산식 집합을 비교합니다.

1. 분석 > 예측 모델링 > 모형 비교를 선택합니다.
2. 확률 (BAD==Good Risk), 확률 (BAD==Bad Risk), 확률 (BAD==Good Risk) 2 및 확률 (BAD==Bad Risk) 2 를 선택하고 **Y, 예측 변수**를 클릭합니다.

첫 번째 계산식 열 쌍에는 결측값 정보화를 사용하는 모형의 계산식이 포함됩니다. 두 번째 계산식 열 쌍에는 결측값 정보화를 사용하지 않는 모형의 계산식이 포함됩니다.

3. **확인**을 클릭합니다.

그림 4.16 모형 비교의 적합 측도

BAD에 대한 적합 측도										
생성자	.2.4.6.8	엔트로피	R ²	일반화 R ²	평균 -Log p	RASE	평균 절대 편차	오분류 비율	N	
파티션		0.3813	0.5015	0.3092	0.3013	0.1817	0.1158	0.1158	5960	
파티션		0.1225	0.1824	0.4385	0.3689	0.2762	0.1768	0.1768	5960	

적합 측도 보고서에서는 결측값 정보화를 사용하여 적합된 첫 번째 모형이 결측값 정보화를 사용하지 않고 적합된 두 번째 모형보다 우수하다는 것을 보여 줍니다. 첫 번째 모형은 R² 값이 더 높고 RMSE 값과 오분류 비율이 더 낮습니다. 이러한 비교는 ROC 곡선 비교를 정렬합니다.

참고: 마찬가지로 결측값 정보화가 사용되지 않을 때는 랜덤 차이로 인해 결과가 다를 수 있습니다.

수익 행렬 및 결정 행렬 보고서의 예

이 예에서는 파티션 플랫폼의 "결정 행렬" 보고서를 사용하여 수익 행렬의 비용과 관련하여 분류 비율을 평가합니다. 간암 환자 연구에 대한 수익 행렬이 정의됩니다. 다양한 측정값 및 표식을 기반으로 각 환자의 질병 심각도 (높음 또는 낮음)에 따라 환자를 분류하려고 합니다. 환자를 분류할 때 심각도가 높은 환자를 낮은 그룹으로 분류하거나 심각도가 낮은 환자를 높은 그룹으로 분류하는 두 가지 오류가 발생할 수 있습니다. 임상적으로 심각도가 높은 환자를 낮은 환자로 오분류하면 해당 환자가 자신에게 필요한 공격적 치료를 받지 못할 수 있으므로 비용이 많이 발생합니다. 심각도가 낮은 환자를 심각도가 높은 그룹으로 분류하는 오류는 비용이 더 적게 듭니다. 해당 환자는 필요 이상으로 더 공격적인 치료를 받을 수 있지만 이것은 큰 문제가 아닙니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Liver Cancer.jmp 를 엽니다.
2. **분석 > 예측 모델링 > 파티션**을 선택합니다.
3. Severity 를 선택하고 **Y, 반응**을 클릭합니다.
4. BMI ~ Jaundice 를 선택하고 **X, 요인**을 클릭합니다.
5. 설치된 JMP 유형에 따라 검증 절차를 선택합니다.

- JMP Pro 의 경우 Validation 을 선택하고 **검증**을 클릭합니다 .
- JMP 의 경우 **검증 비율**에 "0.3" 을 입력합니다 .

참고 : 검증 행의 랜덤 선택으로 인해 검증 비율을 사용한 결과는 여기에 표시된 것과 다를 수 있습니다 .

그림 4.17 검증 데이터 비율 = 0.3 인 완료된 시작 창

결정 트리를 생성하여 반응을 예측합니다.

The screenshot shows the JMP Partitioning Platform Start Window. It is divided into several sections:

- 열 선택 (Column Selection):** A list of 9 columns is shown: Node Count, Severity, BMI, Age, Time, Markers, Hepatitis, Jaundice, and Validation. The 'Validation' column is selected.
- 선택한 열 역할 지정 (Selected Column Role Assignment):** The 'Y, 반응' (Y, Response) role is assigned to 'Severity'. The 'X, 요인' (X, Factor) role is assigned to 'BMI', 'Age', 'Time', and 'Markers'.
- 작업 (Action):** Buttons for '확인' (OK), '취소' (Cancel), '제거' (Remove), '재호출' (Reopen), and '도움말' (Help) are visible.
- 옵션 (Options):** The '방법' (Method) is set to '결정 트리' (Decision Tree). The '검증 데이터 비율' (Validation Data Ratio) is set to 0.3. The checkboxes for '결측값 정보화' (Impute Missing Values) and '순서형의 순서를 지킴' (Keep Order of Ordinal) are checked.

6. **확인**을 클릭합니다 .
7. Shift 키를 누른 채 **분할**을 클릭합니다 .
8. 분할 수에 "10" 을 입력하고 **확인**을 클릭합니다 .
그림 아래의 패널에서 "분할 수" 가 10 인지 확인합니다 .
9. "Severity 에 대한 분할" 옆의 빨간색 삼각형을 클릭하고 **수익 행렬 지정**을 선택합니다 .
10. 항목을 다음 값으로 변경합니다 .
 - High, High 상자에 1 입력
 - High, Low 상자에 -5 입력
 - Low, High 상자에 -3 입력
 - Low, Low 상자에 1 입력

그림 4.18 완료된 수익 행렬

		결정 또는 예측		
실제값		Low	High	결정되지 않음
	Low	1	-5	.
	High	-3	1	.

팁 : 이후 분석에 사용할 수 있도록 이 수익 행렬을 열 특성으로 저장할 수 있습니다. 수익 행렬 창 맨 아래의 " 열에 특성으로 저장 " 체크박스를 선택합니다.

다음 사항에 유의하십시오 .

- 1 값은 각각 올바른 결정을 내릴 때 수익을 반영합니다 .
- -3 값은 심각도가 낮은 환자 (Low) 를 심각도가 높은 환자 (High) 로 분류할 경우 손실이 올바른 결정의 수익보다 3 배 많음을 나타냅니다 .
- -5 값은 심각도가 높은 환자 (High) 를 심각도가 낮은 환자 (Low) 로 분류할 경우 손실이 올바른 결정의 수익보다 5 배 많음을 나타냅니다 .

11. **확인**을 클릭합니다 .

12. "Severity 에 대한 분할 " 옆의 빨간색 삼각형을 클릭하고 **적합 상세 정보 표시**를 선택합니다 .

그림 4.19 혼동 행렬 및 결정 행렬 보고서

적합 상세 정보					
측도	훈련	검증	정의		
엔트로피 R ²	0.3788	0.0367	1-Loglike(model)/Loglike(0)		
일반화 R ²	0.5382	0.0647	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))		
평균 -Log p	0.4134	0.6402	$\sum -\log(p[j])/n$		
RASE	0.3649	0.4694	$\sqrt{\sum (y[j]-p[j])^2/n}$		
평균 절대 편차	0.2854	0.3861	$\sum y[j]-p[j] /n$		
오분류 비율	0.1915	0.3810	$\sum (p[j] \neq pMax)/n$		
N	94	42	n		

혼동 행렬					
훈련			검증		
실제값	예측 개수		실제값	예측 개수	
Severity	Low	High	Severity	Low	High
Low	30	6	Low	11	5
High	12	46	High	11	15
실제값	예측 비율		실제값	예측 비율	
Severity	Low	High	Severity	Low	High
Low	0.833	0.167	Low	0.688	0.313
High	0.207	0.793	High	0.423	0.577

결정 행렬					
훈련			검증		지정된 수익 행렬
실제값	결정 개수		실제값	결정 개수	결정
Severity	Low	High	Severity	Low	High
Low	33	3	Low	12	4
High	16	42	High	12	14
실제값	결정 비율		실제값	결정 비율	
Severity	Low	High	Severity	Low	High
Low	0.917	0.083	Low	0.750	0.250
High	0.276	0.724	High	0.462	0.538
오분류 비율			오분류 비율		
0.2021			0.3810		

" 혼동 행렬 " 보고서와 " 결정 행렬 " 보고서는 " 적합 상세 정보 " 보고서의 " 측도 " 목록을 따릅니다. " 혼동 행렬 " 보고서와 " 결정 행렬 " 보고서의 혼동 행렬은 개수가 다르게 나타납니다. 이것은 수익 행렬의 가중치가 가중치 없이 예측된 확률과 다른 결정을 내리기 때문입니다.

13. "Severity에 대한 분할" 옆의 빨간색 삼각형을 클릭하고 **열 저장 > 예측 계산식 저장**을 선택합니다.

8 개의 열이 데이터 테이블에 추가됩니다.

팁 : 데이터 테이블로 빨리 돌아가려면 보고서 창 오른쪽 아래의 " 연결된 데이터 보기 " 아이콘 (Windows) 또는 도구 모음 메뉴의 " 데이터 테이블 표시 " 아이콘 (macOS) 을 클릭하십시오 .

- 처음 세 열은 예측 확률만 포함합니다 . 혼동 행렬 수는 예측 확률이 가장 높은 수준으로 환자를 분류하는 **최대 확률 분류 Severity** 열을 기반으로 합니다 . 이러한 확률은 **확률 (Severity == High)** 및 **확률 (Severity == Low)** 열에 제공됩니다 .
- 마지막 5 개 열은 수익 행렬 가중치를 포함합니다 . **Severity**에 대한 수익성이 가장 높은 예측 열에는 수익 행렬을 기반으로 한 결정이 포함됩니다 . 환자의 결정은 수익성이 가장 높은 수준입니다 . 이익은 **High**에 대한 수익 및 **Low**에 대한 수익 열에 제공됩니다 .

파티션 플랫폼에 대한 통계 상세 정보

이 섹션에는 파티션 플랫폼에 대한 통계 상세 정보가 포함되어 있습니다.

- [" 반응 및 요인에 대한 통계 상세 정보 "](#)
- [" 분할 기준에 대한 통계 상세 정보 "](#)

반응 및 요인에 대한 통계 상세 정보

파티션 플랫폼에서 반응은 연속형 또는 범주형 (명목형 또는 순서형) 일 수 있습니다 .

- 반응이 범주형이면 반응 수준에 대해 추정된 확률을 적합시켜 잔차 로그 가능도 카이제곱[2* 엔트로피]을 최소화합니다 .
- 반응이 연속형이면 플랫폼에서 평균을 적합시켜 오차 제곱합을 최소화합니다 .

요인은 연속형 또는 범주형 (명목형 또는 순서형) 일 수 있습니다 .

- 요인이 연속형이면 요인에 대한 분할 " 절단값 "에 따라 분할이 수행됩니다 .
- 요인이 범주형이면 X 범주를 두 개의 수준 그룹으로 나누고 가능한 모든 그룹을 두 수준으로 고려합니다 .

분할 기준에 대한 통계 상세 정보

파티션 플랫폼의 노드 분할은 노드에 대한 " 후보 " 보고서에 나오는 LogWorth 통계량을 기반으로 합니다 . LogWorth 는 다음과 같이 계산됩니다 .

$$-\log_{10}(p \text{ 값})$$

여기서 조정 p 값은 분할이 발생할 수 있는 다양한 방법을 고려하는 복합적 형식으로 계산됩니다. 이 계산은 수준이 많은 X 를 우선으로 하는 비조정 p 값 및 수준이 적은 X 를 우선으로 하는 Bonferroni p 값에 비해 매우 공정합니다. 이 방법에 대한 자세한 내용은 Sall(2002) 연구 자료에서 확인하십시오.

연속형 반응의 경우 SS(제곱합)가 노드 보고서에 보고됩니다. 이 값은 분할로 인한 오차 제곱합의 변화입니다.

선택된 후보 SS는 다음과 같습니다.

$$SS_{\text{test}} = SS_{\text{parent}} - (SS_{\text{right}} + SS_{\text{left}}) \text{ 여기서 노드의 SS는 } s^2(n-1) \text{ 입니다.}$$

연속형 반응의 경우 차이 통계량도 보고됩니다. 이 값은 상위 노드의 두 하위 노드에 대한 예측값의 차이입니다.

범주형 반응의 경우 G^2 (가능도비 카이제곱)이 보고서에 나타납니다. 실제로 이 값은 [자연 로그] 엔트로피의 두 배 또는 엔트로피 변화의 두 배입니다. 엔트로피는 각 관측값에 대한 $\sum -\log(p)$ 이며, 여기서 p 는 발생한 반응에 기인한 확률입니다.

선택된 후보 G^2 은 다음과 같습니다.

$$G^2_{\text{test}} = G^2_{\text{parent}} - (G^2_{\text{left}} + G^2_{\text{right}})$$

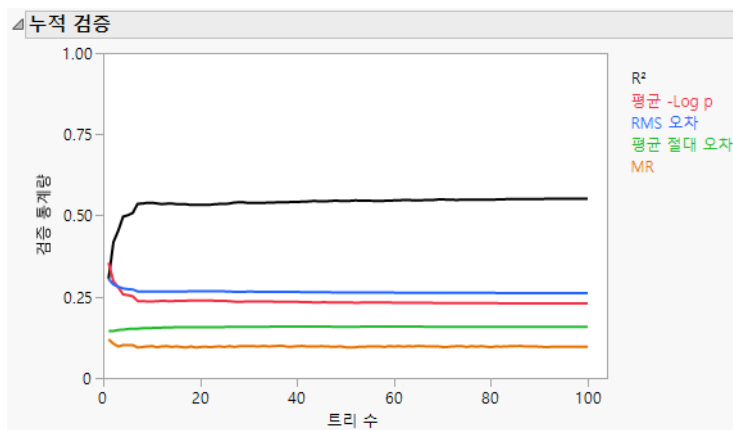
파티션에는 실제로 두 가지 비율이 있습니다. 하나는 훈련에 사용되는 일반적인 개수 대 합계 비율이며, 다른 하나는 0에서 약간 벗어난 비율입니다. 귀속 확률이 0이 아니므로 엔트로피 R^2 에 사용되는 검증 또는 제외된 데이터 집합에 대해 확률 로그를 계산할 수 있습니다.

JMP Pro 붓스트랩 포레스트 여러 트리를 평균화하여 모형 적합

붓스트랩 포레스트 플랫폼은 JMP Pro 에서만 사용할 수 있습니다.

붓스트랩 포레스트 플랫폼은 각각 훈련 데이터의 붓스트랩 표본에 적합되는 여러 결정 트리를 평균화하여 앙상블 모형을 적합시킵니다. 각 트리의 각 분할은 예측 변수의 랜덤 부분집합을 고려합니다. 이런 방식으로 여러 약한 모형을 결합하여 더 강력한 모형을 생성합니다. 관측값에 대한 최종 예측은 모든 결정 트리에서 해당 관측값의 예측값을 평균한 것입니다.

그림 5.1 누적 검증 보고서의 예



목차

붓스트랩 포레스트 플랫폼 개요.....	81
범주형 반응이 있는 붓스트랩 포레스트의 예.....	81
연속형 반응이 있는 붓스트랩 포레스트의 예.....	84
붓스트랩 포레스트 플랫폼 시작.....	86
규격 창.....	88
붓스트랩 포레스트 보고서.....	90
모형 검증 집합 요약.....	91
규격.....	91
전체 통계량.....	91
누적 검증.....	93
트리별 요약.....	93
붓스트랩 포레스트 플랫폼 옵션.....	94

JMP PRO 붓스트랩 포레스트 플랫폼 개요

붓스트랩 포레스트 플랫폼은 많은 결정 트리에서 예측된 반응 값의 평균을 구하여 반응 값을 예측합니다. 각 트리는 훈련 데이터의 붓스트랩 표본에서 성장합니다. 붓스트랩 표본은 복원 추출법을 사용한 관측값의 랜덤 표본입니다. 또한 예측 변수는 결정 트리의 각 분할에서 표집됩니다. 결정 트리는 "[파티션 모형](#)"에 설명된 재귀 분할 방법을 사용하여 적합됩니다.

다음은 훈련 데이터 집합에 대한 적합 프로세스입니다.

1. 각 트리에 대해 관측값의 붓스트랩 표본을 선택합니다.
2. 재귀 분할을 사용하여 각 결정 트리를 적합시킵니다.
 - 각 분할에 대한 랜덤 예측 변수 집합을 선택하고 집합에서 예측 변수의 순서를 랜덤화합니다.
 - "붓스트랩 포레스트 규격" 창에 지정된 중지 규칙이 충족될 때까지 분할을 계속합니다.
3. "붓스트랩 포레스트 규격" 창에 지정된 트리 수에 도달하거나 조기 중지가 발생할 때까지 [1 단계](#)와 [2 단계](#)를 반복합니다.

개별 트리의 경우 트리를 적합시키는 데 사용되는 관측값의 붓스트랩 표본은 복원 추출됩니다. 표집할 관측값 비율을 지정할 수 있습니다. 관측값이 복원 추출되므로 관측값의 100%를 표집하도록 지정하는 경우 사용되지 않은 관측값의 기대 비율은 $1/e$ 또는 약 36.8%입니다. 각 개별 트리에 대해 사용되지 않은 이러한 관측값을 Out-Of-Bag 관측값이라고 합니다. 트리를 적합시키는 데 사용된 관측값은 In-Bag 관측값이라고 합니다. 연속형 반응의 경우 붓스트랩 포레스트 플랫폼은 Out-Of-Bag 관측값의 오차율에 대한 측도를 제공합니다(Out-Of-Bag 오차).

연속형 반응의 경우 관측값의 예측값은 개별 트리 모음에 대한 예측값의 평균입니다. 범주형 반응의 경우 관측값의 예측 확률은 개별 트리 모음에 대한 예측 확률의 평균입니다. 관측값은 예측 확률이 가장 높은 수준으로 분류됩니다.

붓스트랩 포레스트에 대한 자세한 내용은 Hastie et al. (2009) 연구 자료에서 확인하십시오.

JMP PRO 범주형 반응이 있는 붓스트랩 포레스트의 예

이 예에서는 고객의 신용 위험이 낮은지 여부를 예측하기 위해 붓스트랩 포레스트 모형을 생성합니다. 그런 다음 예측 변수 값의 예측 정도를 탐색합니다.

붓스트랩 포레스트 모형

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Equity.jmp 를 엽니다.
2. **분석 > 예측 모델링 > 붓스트랩 포레스트**를 선택합니다.
3. BAD 를 선택하고 **Y, 반응**을 클릭합니다.
4. LOAN ~ DEBTINC 를 선택하고 **X, 요인**을 클릭합니다.
5. Validation 을 선택하고 **검증**을 클릭합니다.

6. **확인**을 클릭합니다.
7. "분할당 표집되는 항 수" 옆에 "4"를 입력합니다.
8. "트리당 최대 분할 수" 옆에 "30"을 입력합니다.
9. **여러 항에 대한 다중 적합**을 선택하고 "최대 항 수" 옆에 "10"을 입력합니다.
10. (선택 사항) **멀티스레딩 제한**을 선택하고 "난수 시드값" 옆에 "123"을 입력합니다.

붓스트랩 포레스트 방법에는 랜덤 표집이 포함되므로 이렇게 하면 아래 표시된 결과와 정확히 동일한 결과를 얻을 수 있습니다.

11. **확인**을 클릭합니다.

그림 5.2 전체 통계량 보고서

BAD에 대한 붓스트랩 포레스트			
규격			
전체 통계량			
측도	훈련	검증	테스트 정의
엔트로피 R ²	0.5118	0.4770	0.4731 1-Loglike(model)/Loglike(0)
일반화 R ²	0.6297	0.6096	0.5976 (1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
평균 -Log p	0.2357	0.2825	0.2665 $\sum -\log(p[j])/n$
RASE	0.2592	0.2907	0.2784 $\sqrt{\sum (y[j]-p[j])^2/n}$
평균 절대 편차	0.1577	0.1795	0.1687 $\sum y[j]-p[j] /n$
오분류 비율	0.0867	0.1141	0.0990 $\sum (p[j] \neq pMax)/n$
N	3576	1192	1192 n
혼동 행렬			
훈련			
실제값	예측 개수		
BAD	Good Risk	Bad Risk	
Good Risk	2830	75	
Bad Risk	235	436	
실제값	예측 비율		
BAD	Good Risk	Bad Risk	
Good Risk	0.974	0.026	
Bad Risk	0.350	0.650	
검증			
실제값	예측 개수		
BAD	Good Risk	Bad Risk	
Good Risk	889	28	
Bad Risk	108	167	
실제값	예측 비율		
BAD	Good Risk	Bad Risk	
Good Risk	0.969	0.031	
Bad Risk	0.393	0.607	
테스트			
실제값	예측 개수		
BAD	Good Risk	Bad Risk	
Good Risk	926	23	
Bad Risk	95	148	
실제값	예측 비율		
BAD	Good Risk	Bad Risk	
Good Risk	0.976	0.024	
Bad Risk	0.391	0.609	

"여러 항에 대한 다중 적합" 옵션이 지정되었으므로 4, 5, 6, 8, 10을 각 분할의 예측 변수 수로 사용하여 모형이 생성되었습니다. "모형 검증 집합 요약" 보고서에서는 검증 데이터 집합의 엔트로피 R²이 가장 높은 모형이 5항 모형임을 보여 줍니다. 이 모형은 오분류 비율이 가장 작은 모형이기도 합니다. 이 모형이 최적 모형으로 결정되고 이 모형에 대한 결과가 전체 보고서에 표시됩니다.

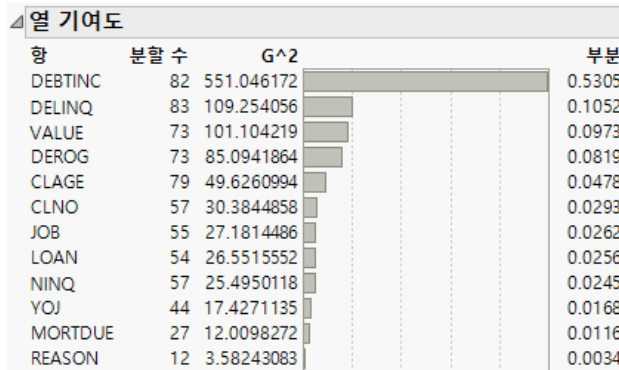
전체 보고서에서는 검증 데이터 집합과 테스트 데이터 집합의 오분류 비율이 각각 11.4%와 9.9%임을 보여 줍니다. 혼동 행렬은 Bad Risk 고객을 Good Risk 고객으로 분류한 것이 오분류의 가장 큰 원인임을 나타냅니다.

테스트 데이터 집합의 결과는 모형이 독립적 관측값으로 얼마나 잘 확장되는지를 나타냅니다. 검증 데이터 집합은 붓스트랩 포레스트 모형을 선택하는 데 사용되었습니다. 따라서 검증 데이터 집합의 결과는 모형이 독립 데이터로 일반화되는 방식을 나타내는 데 더 관심을 둡니다.

모형에 가장 많이 기여한 예측 변수를 확인하려고 합니다.

12. "BAD에 대한 붓스트랩 포레스트" 옆의 빨간색 삼각형을 클릭하고 **열 기여도**를 선택합니다.

그림 5.3 열 기여도 보고서



"열 기여도" 보고서에 따르면 신용 위험에 대해 가장 강력한 예측 변수는 총 부채 상환 비율을 나타내는 DEBTINC입니다. 다음으로 모형에 가장 많이 기여하는 변수는 연체 신용 한도 수를 나타내는 DELINQ와 고객의 사정 평가액을 나타내는 VALUE입니다.

결측값

1. 분석 > 선별 > 결측값 탐색을 선택합니다.

2. Bad ~ DEBTINC를 선택하고 **Y, 열**을 클릭합니다.

3. 경고가 나타나면 **확인**을 클릭합니다.

REASON 및 JOB 열은 문자 데이터 유형이 있으므로 "Y, 열" 목록에 추가되지 않습니다. 분포"(이 예에서 설명하지 않음)를 사용하여 이 두 열에 대해 결측값 수를 확인할 수 있습니다.

4. **확인**을 클릭합니다.

그림 5.4 결측값 보고서

결측값 탐색

명령

- 결측값 보고서** 각 열의 결측값 수를 나열합니다.
- 결측값 군집화** 열 그룹에 유사한 결측 패턴이 있는지 여부를 판별하기 위해 결측 데이터의 계층적 군집화 분석을 제공합니다.
- 결측값 스냅샷** 그래픽 맵을 사용하여 결측값 패턴을 찾습니다.
- 다변량 정규 대치법** 비결측 열에서 최소 제곱 예측을 사용하여 결측값을 대치합니다. 열의 수가 수백 개인 경우에는 이 방법을 사용하지 마십시오. 축소 옵션을 사용하여 공분산 행렬 추정을 향상시킬 수 있습니다.
- 다변량 SVD 대치법** 반복된 낮은 계수 SVD 행렬 완성 방법을 사용하여 큰 문제에 대해 빠르게 결측값을 대치합니다.
- 다변량 RPCA 대치법** 로버스트 주성분을 사용하여 결측값을 대치합니다. 즉, 이상치에 로버스트한 낮은 계수 행렬 분해(SVD)를 사용하여 결측값을 대체합니다. 이 방법은 변수의 수가 많은 와이드 문제에 유용합니다.
- 자동 데이터 결측값 대치** 낮은 계수 행렬 근사 방법을 사용하여 결측값을 대치합니다. 이 방법은 데이터를 기반으로 낮은 계수 근사에 대한 최선의 차원을 자동으로 선택합니다.

자동 데이터 결측값 대치 컨트롤

결측 열

☐ 결측값을 포함하는 열만 표시

열을 선택하고 작업을 선택하십시오.

열	결측 수
BAD	0
LOAN	0
MORTDUE	518
VALUE	112
YOJ	515
DEROG	708
DELINQ	580
CLAGE	308
NINQ	510
CLNO	222
DEBTINC	1267

DEBTINC 열에는 관측값의 약 21%에 해당하는 1267 개의 결측값이 있습니다. 붓스트랩 포레스트 분석에 포함된 대부분의 다른 열에도 결측값이 포함되어 있습니다. 시작 창의 "결측값 정보화" 옵션을 사용할 경우 결측값은 해당 값이 전달하는 정보가 인식되도록 처리됩니다. 자세한 내용은 "ROC 곡선"에서 확인하십시오.

JMP PRO 연속형 반응이 있는 붓스트랩 포레스트의 예

이 예에서는 남성 피실험자의 체지방량을 예측하기 위해 붓스트랩 포레스트 모델을 생성합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Body Fat.jmp를 엽니다.
2. **분석 > 예측 모델링 > 붓스트랩 포레스트**를 선택합니다.
3. Percent body fat 을 선택하고 **Y, 반응**을 클릭합니다.
4. Age (years) ~ Wrist circumference (cm) 를 선택하고 **X, 요인**을 클릭합니다.

5. Validation 을 선택하고 **검증**을 클릭합니다 .
6. **확인**을 클릭합니다 .
7. (선택 사항) **멀티스레딩 제한**을 선택하고 " 난수 시드값 " 옆에 "123" 을 입력합니다 .
붓스트랩 포레스트 방법에는 랜덤 표집이 포함되므로 이렇게 하면 아래 표시된 결과와 정확히 동일한 결과를 얻을 수 있습니다 .
8. **확인**을 클릭합니다 .

그림 5.5 전체 통계량

Percent body fat에 대한 붓스트랩 포레스트			
규격			
목표값	Percent body fat	훈련 행:	180
검증 열:	Validation	검증 행:	72
		테스트 행:	0
포레스트의 트리 수:	100	행 수:	13
분할당 표집되는 행 수:	10	붓스트랩 표본:	180
		트리당 최소 분할 수:	10
		최소 크기 분할:	5
전체 통계량			
개별 트리	RASE		
In Bag	2.000681		
Out of Bag	6.069647		
	R ²	RASE	N
훈련	0.894	2.6707272	180
검증	0.654	5.1170211	72

" 전체 통계량 " 보고서에는 " 검증 R² " 이 0.654 로 표시됩니다 .

모형과 별개로 가장 중요한 예측 변수를 나타내려고 합니다 .

9. "Percent body fat 에 대한 붓스트랩 포레스트 " 옆의 빨간색 삼각형을 클릭하고 **열 기여도**를 선택합니다 .

그림 5.6 열 기여도

열 기여도			
항	분할 수	SS	부분
Abdomen circumference (cm)	101	4185.96516	0.6194
Chest circumference (cm)	39	850.607271	0.1259
Knee circumference (cm)	33	254.906105	0.0377
Height (inches)	50	242.813772	0.0359
Age (years)	45	239.973342	0.0355
Ankle circumference (cm)	45	167.897611	0.0248
Thigh circumference (cm)	25	162.416449	0.0240
Weight (lbs)	25	145.612142	0.0215
Biceps (extended) circumference (cm)	44	142.127633	0.0210
Wrist circumference (cm)	38	117.953731	0.0175
Neck circumference (cm)	32	111.864186	0.0166
Hip circumference (cm)	24	90.0947826	0.0133
Forearm circumference (cm)	25	46.3944108	0.0069

"열 기여도" 보고서에 따르면 Abdomen circumference (cm), Chest circumference (cm) 및 Age (years) 가 Percent body fat 에 대한 가장 강력한 예측 변수입니다 .

JMP PRO 붓스트랩 포레스트 플랫폼 시작

분석 > 예측 모델링 > 붓스트랩 포레스트를 선택하여 붓스트랩 포레스트 플랫폼을 시작합니다 .

그림 5.7 붓스트랩 포레스트 시작 창

랜덤 표집을 사용하여 여러 개의 결정 트리를 생성하고 그 결과의 평균을 구해 반응을 예측합니다.

열 선택

14개 열

- BAD
- LOAN
- MORTDUE
- VALUE
- REASON
- JOB
- YOJ
- DEROG
- DELINQ
- CLAGE
- NINQ
- CLNO
- DEBTINC
- Validation

선택한 열 역할 지정

Y, 반응	필수 선택적
X, 요인	필수 선택적
가중치	선택적 숫자
빈도	선택적 숫자
검증	선택적 숫자
기준	선택적

작업

확인
취소
제거
재호출
도움말

옵션

방법 붓스트랩 포레스트

검증 데이터 비율 0.3

☒ 결측값 정보화

☒ 순서형의 순서를 지킴

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

붓스트랩 포레스트 플랫폼 시작 창에는 다음과 같은 옵션이 있습니다 .

Y, 반응 분석할 하나 이상의 반응 변수입니다 .

X, 요인 예측 변수입니다 .

가중치 분석의 각 행에 가중치를 할당하는 숫자 값이 포함된 열입니다 .

빈도 분석의 각 행에 빈도를 할당하는 숫자 값이 포함된 열입니다 .

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

- 검증 열에 두 개의 수준이 있는 경우 더 작은 값이 훈련 데이터 집합을 정의하고 더 큰 값이 검증 데이터 집합을 정의합니다.
- 검증 열에 세 개의 수준이 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.

붓스트랩 포레스트 플랫폼에서는 검증 열을 사용하여 모델을 훈련 및 조정하거나 모델을 훈련, 조정 및 평가합니다. 검증에 대한 자세한 내용은 "[JMP 모델링의 검증](#)"에서 확인하십시오.

"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 "[검증 열 생성](#)"에서 확인하십시오.

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.

방법 분할 방법 (결정 트리, 붓스트랩 포레스트, 부스티드 트리, K 최근접 이웃 또는 Naive Bayes)을 선택할 수 있습니다. 결정 트리를 제외한 이 대체 방법은 JMP Pro에서 사용할 수 있습니다.

이러한 방법에 대한 자세한 내용은 "[파티션 모델](#)", "[부스티드 트리](#)", "[K 최근접 이웃](#)" 및 "[Naive Bayes](#)"에서 확인하십시오.

검증 데이터 비율 검증 데이터 집합으로 사용할 데이터 비율입니다.

결측값 정보화 이 옵션을 선택하면 범주형 예측 변수에 대한 결측값 범주화 및 연속형 예측 변수에 대한 결측값 정보화 처리가 활성화됩니다. 자세한 내용은 "[ROC 곡선](#)"에서 확인하십시오.

순서형의 순서를 지킴 이 옵션을 선택하면 순서를 유지하는 분할만 고려하도록 제한됩니다.

JMP^{PRO} 규격 창

"붓스트랩 포레스트 규격" 창에서 행과 항 수, 포레스트 상세 정보, 다중 적합 및 재현성을 지정할 수 있습니다. "붓스트랩 포레스트" 시작 창에서 "확인" 을 선택하면 "붓스트랩 포레스트 규격" 창이 나타납니다.

그림 5.8 붓스트랩 포레스트 규격 창

붓스트랩 포레스트 규격

행 수:	252
항 수:	13

포레스트

포레스트의 트리 수:

분할당 표집되는 항 수:

붓스트랩 표본 비율:

트리당 최소 분할 수:

트리당 최대 분할 수:

최소 크기 분할:

☒ 조기 중지

다중 적합

☐ 여러 항에 대한 다중 적합

최대 항 수:

☐ 설계 조정 테이블 사용

재현성

☐ 멀티스레딩 제한

난수 시드값:

JMP^{PRO} 규격 패널

행 수 데이터 테이블의 행 수입니다.

항 수 예측 변수로 지정된 열 수입니다.

JMP^{PRO} 포레스트 패널

포레스트의 트리 수 증가 후 평균을 구할 트리 수입니다.

분할당 표집되는 항 수 각 분할에서 분할 후보로 고려할 예측 변수의 수입니다. 각 분할에 대해 예측 변수의 새 랜덤 표본이 후보 집합으로 사용됩니다.

붓스트랩 표본 비율 각 트리를 증가시키기 위해 표집할 관측값의 비율입니다 (복원 추출법 사용). 각 트리에 대해 새 랜덤 표본이 생성됩니다.

트리당 최소 분할 수 각 트리의 최소 분할 수입니다.

트리당 최대 분할 수 각 트리의 최대 분할 수입니다.

최소 크기 분할 후보 분할에 필요한 최소 관측값 수입니다.

조기 중지 (검증을 사용하는 경우에만 사용 가능) 이 옵션을 선택하면 트리를 추가해도 검증 통계량이 개선되지 않을 경우 추가 트리 증가가 중지됩니다. 검증 통계량은 검증 데이터 집합

의 엔트로피 R^2 값 (범주형 반응의 경우) 및 R^2 값 (연속형 반응의 경우) 입니다. 이 옵션을 선택하지 않으면 지정된 트리 수에 도달할 때까지 프로세스가 계속됩니다.

JMP PRO 다중 적합 패널

여러 항에 대한 다중 적합 이 옵션을 선택하면 분할당 표집된 항 수의 여러 값에 대한 붓스트랩 포레스트가 생성됩니다. 결과가 표시되는 모형은 검증 데이터 집합의 엔트로피 R^2 값 (범주형 반응의 경우) 또는 R^2 값 (연속형 반응의 경우) 이 가장 큰 모형입니다.

하한은 "분할당 표집되는 항 수" 규격 값입니다. 상한은 다음 옵션으로 지정됩니다.

최대 항 수 분할에 대해 고려할 최대 항 수입니다.

설계 조정 테이블 사용 "포레스트" 패널 조정 모수의 값이 포함된 데이터 테이블 (설계 조정 테이블) 을 선택할 수 있는 창을 엽니다. 설계 조정 테이블에는 지정할 각 옵션에 대한 열과 각각 단일 붓스트랩 포레스트 모형 설계를 나타내는 하나 이상의 행이 있습니다. 설계 조정 테이블에 옵션이 지정되지 않으면 기본값이 사용됩니다.

JMP에서는 지정된 조정 모수를 사용하여 테이블의 각 행에 대해 붓스트랩 포레스트 모형을 생성합니다. 설계 조정 테이블에 둘 이상의 모형이 지정된 경우 "모형 검증 집합 요약" 보고서에 각 모형의 R^2 값이 나열됩니다. "붓스트랩 포레스트" 보고서에는 R^2 값이 가장 큰 모형의 적합 통계량이 표시됩니다.

실험 설계 기능을 사용하여 설계 조정 테이블을 생성할 수 있습니다. 붓스트랩 포레스트 설계 조정 테이블에는 다음 열이 임의의 순서로 포함될 수 있습니다.

- 트리 수
- 항 수
- 부분 붓스트랩
- 트리당 최소 분할 수
- 트리당 최대 분할 수
- 최소 크기 분할

JMP PRO 재현성 패널

멀티스레딩 제한 이 옵션을 선택하면 모든 계산이 단일 스레드에서 수행됩니다.

난수 시드값 이후 플랫폼 실행을 위해 결과를 재현하는 데 사용할 0 이 아닌 숫자 난수 시드값을 지정합니다. 기본적으로 난수 시드값은 0 으로 설정되어 재현 가능한 결과를 생성하지 않습니다. 분석을 스크립트에 저장하면 사용자가 입력한 난수 시드값이 스크립트에 저장됩니다.

JMP PRO 붓스트랩 포레스트 보고서

"붓스트랩 포레스트 규격" 창에서 확인을 클릭하면 JMP에서 모형을 적합시키는 데 걸리는 대략적인 시간을 알려 주는 창이 나타납니다. 예정 시간이 너무 길면 **현재 모형 채택**을 선택하여 현재 모형을 볼 수도 있습니다. 전체 모형이 적합되거나 현재 모형이 채택된 후 "붓스트랩 포레스트" 보고서가 나타납니다.

그림 5.9 현재 모형 채택 대화상자

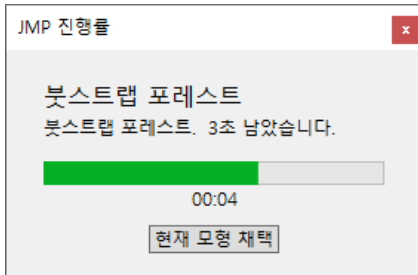


그림 5.10 범주형 반응에 대한 붓스트랩 포레스트 보고서

BAD에 대한 붓스트랩 포레스트				
규격				
전체 통계량				
측도	훈련	검증	테스트	정의
엔트로피 R ²	0.5118	0.4770	0.4731	1-Loglike(model)/Loglike(0)
일반화 R ²	0.6297	0.6096	0.5976	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
평균 -Log p	0.2357	0.2825	0.2665	$\sum -\log(p[j])/n$
RASE	0.2592	0.2907	0.2784	$\sqrt{\sum (y[j]-p[j])^2/n}$
평균 절대 편차	0.1577	0.1795	0.1687	$\sum y[j]-p[j] /n$
오분류 비율	0.0867	0.1141	0.0990	$\sum (p[j] \neq pMax)/n$
N	3576	1192	1192	n
혼동 행렬				
훈련			검증	
실제값	예측 개수		실제값	예측 개수
BAD	Good Risk	Bad Risk	BAD	Good Risk Bad Risk
Good Risk	2830	75	Good Risk	889 28
Bad Risk	235	436	Bad Risk	108 167
테스트			테스트	
실제값	예측 개수		실제값	예측 개수
BAD	Good Risk	Bad Risk	BAD	Good Risk Bad Risk
Good Risk	926	23	Good Risk	926 23
Bad Risk	95	148	Bad Risk	95 148
비율			비율	
실제값	예측 비율		실제값	예측 비율
BAD	Good Risk	Bad Risk	BAD	Good Risk Bad Risk
Good Risk	0.974	0.026	Good Risk	0.969 0.031
Bad Risk	0.350	0.650	Bad Risk	0.393 0.607
비율			비율	
실제값	예측 비율		실제값	예측 비율
BAD	Good Risk	Bad Risk	BAD	Good Risk Bad Risk
Good Risk	0.976	0.024	Good Risk	0.976 0.024
Bad Risk	0.391	0.609	Bad Risk	0.391 0.609

그림 5.11 연속형 반응에 대한 붓스트랩 포레스트 보고서

Percent body fat에 대한 붓스트랩 포레스트			
규격			
목표값	Percent body fat	훈련 행:	180
검증 열:	Validation	검증 행:	72
		테스트 행:	0
포레스트의 트리 수:	100	항 수:	13
분할당 표집되는 항 수:	10	붓스트랩 표본:	180
		트리당 최소 분할 수:	10
		최소 크기 분할:	5
전체 통계량			
개별 트리	RASE		
In Bag	2.000681		
Out of Bag	6.069647		
	R ²	RASE	N
훈련	0.894	2.6707272	180
검증	0.654	5.1170211	72

반응이 범주형인지 아니면 연속형인지에 따라 다음과 같은 보고서가 제공됩니다.

- "모형 검증 집합 요약"
- "규격"
- "전체 통계량"
- "누적 검증"
- "트리별 요약"

JMP PRO 모형 검증 집합 요약

붓스트랩 포레스트 플랫폼의 "모형 검증 집합 요약" 보고서에는 모든 적합 모형에 대한 적합 통계량이 표시됩니다. 이 보고서는 "붓스트랩 포레스트 규격" 창에서 "여러 항에 대한 다중 적합" 옵션을 선택한 경우에만 사용할 수 있습니다. 자세한 내용은 [그림 5.10](#) 및 "[다중 적합 패널](#)"에서 확인하십시오.

JMP PRO 규격

붓스트랩 포레스트 플랫폼의 "규격" 보고서에는 모형 적합에 사용된 설정이 표시됩니다.

JMP PRO 전체 통계량

붓스트랩 포레스트 플랫폼의 "전체 통계량" 보고서에는 훈련 데이터 집합과 검증 및 테스트 데이터 집합(지정된 경우)에 대한 적합 통계량이 제공됩니다. 보고서의 특정 형식은 반응의 모델링 유형에 따라 다릅니다.

"붓스트랩 포레스트 규격" 창의 "여러 항에 대한 다중 적합" 옵션을 사용하여 여러 모형이 적합된다고 가정해 보겠습니다. 그러면 "전체 통계량" 및 "누적 검증" 보고서에 결과가 표시되는 모형은 검증 데이터 집합의 엔트로피 R^2 값(범주형 반응의 경우) 또는 R^2 값(연속형 반응의 경우)이 가장 큰 모형입니다.

JMP[®] PRO 범주형 반응

측도 보고서

훈련 데이터 집합과 검증 및 테스트 데이터 집합(지정된 경우)에 대한 다음 통계량을 제공합니다.

참고 : 엔트로피 R^2 및 일반화 R^2 의 경우 값이 1에 가까울수록 더 나은 적합을 나타냅니다. 평균 $-\log p$, RASE, 평균 절대 편차 및 오분류 비율의 경우 값이 작을수록 더 나은 적합을 나타냅니다.

엔트로피 R^2 적합 모형과 상수 확률 모형의 로그 가능도를 비교하는 적합 측도입니다. 값 범위는 0에서 1 사이입니다. 자세한 내용은 "[엔트로피 \$R^2\$](#) "에서 확인하십시오.

일반화 R^2 일반 회귀 모형에 적용할 수 있는 측도입니다. 이 값은 가능도 함수 L 을 기반으로 하며 최대값이 1이 되도록 척도화됩니다. 값이 1이면 완벽한 모형을 나타내고 0이면 상수 모형과 같음을 나타냅니다. 일반화 R^2 측도는 표준 최소 제곱 설정 시 연속형 정규 반응에 대한 기존 R^2 로 단순화됩니다. 일반화 R^2 을 Nagelkerke 또는 Craig 와 Uhler R^2 이라고도 하는데, 이는 Cox-Snell 유사 R^2 을 정규화한 것입니다.

평균 $-\log p$ 음의 $\log(p)$ 의 평균입니다. 여기서 p 는 발생한 사건과 연관된 적합 확률입니다.

RASE 제곱근 평균 제곱 예측 오차입니다. 차이는 1에서 p (실제로 발생한 반응 수준의 적합 확률) 사이입니다.

평균 절대 편차 반응과 예측 반응 사이의 차이에 대한 절대값의 평균입니다. 차이는 1에서 p (실제로 발생한 반응 수준의 적합 확률) 사이입니다.

오분류 비율 적합 확률이 가장 높은 반응 범주가 관측된 범주가 아닌 비율입니다.

N 관측값 수입니다.

혼동 행렬 보고서

(범주형 반응에만 사용 가능) 훈련 데이터 집합과 검증 및 테스트 데이터 집합(지정된 경우)에 대한 분류 통계량을 표시합니다. "혼동 행렬" 보고서에는 혼동 행렬과 혼동 비율 행렬이 포함됩니다. 혼동 행렬은 실제 반응과 예측 반응의 이원 분류입니다. 혼동 비율 행렬은 값을 행 합계로 나눈 혼동 행렬과 동일합니다.

결정 행렬

(범주형 반응을 사용하고 반응에 "수익 행렬" 열 특성이 있거나 "수익 행렬 지정" 옵션을 사용하여 비용을 지정하는 경우에만 사용 가능) 훈련 데이터 집합과 검증 및 테스트 데이터 집합(지정

된 경우)에 대한 결정 개수 및 결정 비율 행렬을 제공합니다. 자세한 내용은 "[파티션 플랫폼의 추가 예](#)"에서 확인하십시오.

JMP PRO 연속형 반응

개별 트리 보고서

모든 트리에 대해 In Bag 관측값과 Out of Bag 관측값의 평균인 RASE 값을 제공합니다. 트리를 생성하는 데 사용된 훈련 데이터 집합 관측값을 In-Bag 관측값이라고 합니다. 트리를 생성하는 데 사용되지 않는 훈련 관측값을 OOB(Out-Of-Bag) 관측값이라고 합니다.

각 트리에 대해 Out of Bag RASE는 오차 제공합의 제곱근을 OOB 관측값 수로 나눈 값입니다. 각 트리에 대한 Out of Bag RASE 제공은 "트리별 요약" 보고서에 "OOB SSE/N"으로 제공됩니다.

R² 및 RASE 보고서

훈련 데이터 집합과 검증 및 테스트 데이터 집합(정의된 경우)에 대한 R², 제곱근 평균 제곱 예측 오차 및 관측값 수를 제공합니다.

JMP PRO 누적 검증

붓스트랩 포레스트 플랫폼의 "누적 검증" 보고서에는 검증 데이터 집합 대 트리 수에 대한 적합 통계량 그림이 표시됩니다. 이 보고서는 검증을 사용하는 경우에만 사용할 수 있습니다.

연속형 반응의 단일 적합 통계량은 R²입니다. 범주형 반응의 적합 통계량은 아래에 나열되어 있으며 "[측도 보고서](#)"에서 설명합니다.

- R²(엔트로피 R²)
- 평균 -Log p(평균 -Log p)
- 제곱근 평균 제곱 예측 오차 (RASE)
- 평균 절대 오차 (평균 절대 편차)
- MR(오분류 비율)

"누적 검증" 그림 아래의 "누적 상세 정보" 보고서에는 그림에 사용된 값이 제공됩니다.

JMP PRO 트리별 요약

붓스트랩 포레스트 플랫폼의 " 트리별 요약 " 보고서에는 In-Bag 및 Out-Of-Bag 관측값에 대한 개념이 포함됩니다. 개별 트리의 경우 트리를 적합시키는 데 사용되는 관측값의 붓스트랩 표본은 복원 추출됩니다. 관측값이 복원 추출되므로 관측값의 100% 를 표집하도록 지정해도 사용되지 않은 관측값의 기대 비율은 $1/e$ 입니다. 각 개별 트리에 대해 사용되지 않은 관측값을

Out-Of-Bag 관측값이라고 합니다 . 트리를 적합시키는 데 사용된 관측값은 In-Bag 관측값이라고 합니다 .

"트리별 요약" 보고서에는 각 트리에 대해 다음과 같은 요약 통계량이 표시됩니다 .

분할 결정 트리의 분할 수입니다 .

순위 트리의 OOB 손실 순위입니다(오름차순). OOB 손실이 가장 작은 트리의 순위가 1입니다 .

OOB 손실 트리가 Out Of Bag 행에 적용될 경우 총 예측 부정확성을 나타내는 척도입니다 . 값이 작을수록 예측 정확도가 높다는 것을 나타냅니다 .

OOB 손실 /N OOB 손실을 OOB 행 수 , 즉 OOB N 으로 나눈 값입니다 .

R² (연속형 반응에만 사용 가능) 트리의 R² 값입니다 .

IB SSE (연속형 반응에만 사용 가능) In Bag 행에 대한 오차 제곱합입니다 .

IB SSE/N (연속형 반응에만 사용 가능) In Bag 행에 대한 오차 제곱합을 In Bag 관측값 수로 나눈 값입니다 . In Bag 관측값 수는 훈련 데이터 집합의 관측값 수에 "붓스트랩 포레스트 규격" 창에 지정한 붓스트랩 표집 비율을 곱한 값과 같습니다 .

OOB N (연속형 반응에만 사용 가능) Out Of Bag 행 수입니다 .

OOB SSE (연속형 반응에만 사용 가능) 트리가 Out Of Bag 행에 적용될 경우 오차 제곱합입니다 .

OOB SSE/N (연속형 반응에만 사용 가능) OOB SSE 를 OOB 행 수 , 즉 OOB N 으로 나눈 값입니다 .

JMP PRO 붓스트랩 포레스트 플랫폼 옵션

"붓스트랩 포레스트"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다 .

실제값 대 예측값 그림 (연속형 반응에만 사용 가능) 실제값 대 예측값 그림을 제공합니다 .

열 기여도 적합에 대한 각 입력 열의 기여도를 보여 주는 보고서를 표시합니다 . 다음 항목도 보고서에 표시됩니다 .

- 지정된 열을 사용하여 데이터를 분할할 때 모든 트리에 대한 총 인스턴스 수
- 열에 기인한 총 G^2 (범주형 반응의 경우) 또는 SS, 즉 제곱합 (연속형 반응의 경우)
- G^2 또는 SS 의 막대 차트
- 열에 기인한 G^2 또는 SS 의 비율

트리 표시 "트리 보기" 보고서에 트리를 표시하기 위한 다양한 옵션을 제공합니다 . 이 보고서에는 붓스트 프로세스의 각 계층에서 적합된 트리 그림이 제공됩니다 . "이름 범주 추정값 표시" 옵션에 의해 표시되는 확률 열에 대한 설명은 "[트리 방법의 예측 확률](#)"에서 확인하십시오 .

ROC 곡선 (범주형 반응에만 사용 가능) 자세한 내용은 "[ROC 곡선](#)"에서 확인하십시오 .

향상도 곡선 (범주형 반응에만 사용 가능) 자세한 내용은 "[향상도 곡선](#)"에서 확인하십시오 .

결정 임계값 (이항 범주형 반응에만 사용 가능) 훈련, 검증 및 테스트 (지정된 경우) 데이터 집합에 대한 "결정 임계값" 보고서를 표시하거나 숨깁니다. 각 보고서에는 모형 적합 비교를 위한 각 모형의 적합 확률 분포 그래프, 혼동 행렬 및 분류 그래프와 분류 정확성 측정 기준 테이블이 포함되어 있습니다. 자세한 내용은 "[결정 임계값 보고서](#)"에서 확인하십시오.

열 저장 모형 및 트리 결과를 저장하고 SAS 코드를 생성하기 위한 옵션이 포함되어 있습니다.

예측값 저장 모형의 예측값을 데이터 테이블에 저장합니다.

예측 계산식 저장 예측 계산식을 데이터 테이블의 열에 저장합니다. 계산식은 트리 구조를 설명하는 내포된 조건부 절로 구성됩니다. 반응이 연속형이면 열에 "예측" 특성이 포함됩니다. 반응이 범주형이면 열에 "반응 확률" 특성이 포함됩니다.

허용 예측 계산식 저장 (이 옵션 대신 "예측 계산식 저장" 옵션을 사용해야 합니다.) "예측 계산식 저장"을 사용할 수 없을 때만 이 옵션을 사용하십시오.) 결측값이 있고 "결측값 정보화"를 선택하지 않은 경우에도 예측하는 계산식을 저장합니다. 예측 계산식은 결측 예측 변수에 대한 반응 값을 분할에 랜덤 할당하여 결측값을 허용합니다. 반응이 연속형이면 열에 "예측" 특성이 포함됩니다. 반응이 범주형이면 열에 "반응 확률" 특성이 포함됩니다. "결측값 정보화"를 선택한 경우 Shift 키를 누른 채 보고서의 빨간색 삼각형을 클릭하면 허용 예측 계산식을 저장할 수 있습니다.

잔차 저장 (연속형 반응에만 사용 가능) 잔차를 데이터 테이블에 저장합니다.

누적 상세 정보 저장 (검증을 사용하는 경우에만 사용 가능) 각 트리에 대한 적합 통계량이 포함된 데이터 테이블을 생성합니다.

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 "[계산식 저장소](#)"에서 확인하십시오.

허용 예측 계산식 게시 (이 옵션 대신 "예측 계산식 게시" 옵션을 사용해야 합니다.) "예측 계산식 게시"를 사용할 수 없을 때만 이 옵션을 사용하십시오.) 허용 예측 계산식을 생성하여 계산식 저장소 플랫폼에 계산식 열 스크립트로 저장합니다. "계산식 저장소" 보고서가 열려 있지 않으면 "계산식 저장소" 보고서가 생성됩니다. 자세한 내용은 "[계산식 저장소](#)"에서 확인하십시오. "결측값 정보화"를 선택한 경우 Shift 키를 누른 채 보고서의 빨간색 삼각형을 클릭하면 이 옵션을 사용할 수 있습니다.

SAS DATA 스템 만들기 새 데이터 집합을 스코어링하기 위한 SAS 코드를 생성합니다.

수익 행렬 지정 (범주형 반응에만 사용 가능) 올바르게나 올바르게지 않은 분류 결과와 관련된 수익 또는 비용을 지정할 수 있습니다. 자세한 내용은 "[적합 상세 정보 표시](#)"에서 확인하십시오.

프로파일러 예측 프로파일러를 표시합니다. 자세한 내용은 **Profilers**에서 확인하십시오.

다음 옵션에 대한 자세한 내용은 **JMP 사용**의 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.

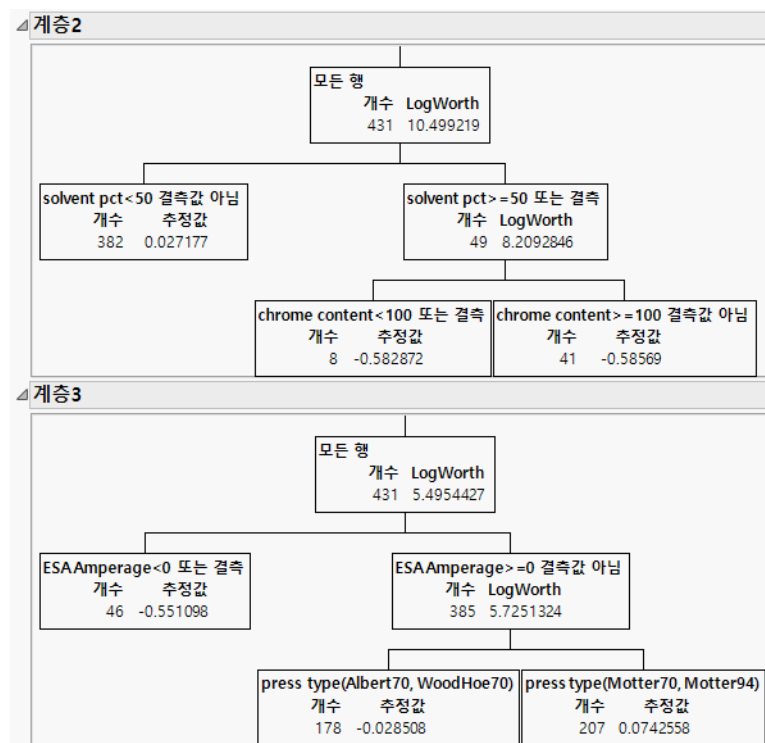
그룹별 스크립트 저장 기준 변수의 모든 수준에 대한 플랫폼 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다. 시작 창에서 기준 변수를 지정한 경우에만 사용할 수 있습니다.

각각 이전 계층에 기반한 여러 트리 계층 적합

부스티드 트리 플랫폼은 JMP Pro에서만 사용할 수 있습니다.

부스트는 계층이라고 하는 일련의 작은 결정 트리를 적합시킨 후 이를 누적하여 더 큰 결정 트리를 생성하는 프로세스입니다. 각 계층의 트리는 적은 수의 분할로 구성됩니다. 트리는 이전 계층의 잔차를 기반으로 적합되므로 각 계층에서 이전 계층의 잘못된 적합 데이터에 대한 적합을 수정할 수 있습니다. 관측값에 대한 최종 예측은 모든 계층에서 해당 관측값의 예측을 합한 것입니다.

그림 6.1 부스티드 트리 계층의 예



목차

부스티드 트리 플랫폼 개요.....	99
범주형 반응이 있는 부스티드 트리의 예.....	99
연속형 반응이 있는 부스티드 트리의 예.....	101
부스티드 트리 플랫폼 시작.....	103
규격 창.....	105
부스티드 트리 보고서.....	107
모형 검증 집합 요약.....	109
규격.....	109
전체 통계량.....	109
누적 검증.....	110
부스티드 트리 플랫폼 옵션.....	110
부스티드 트리 플랫폼에 대한 통계 상세 정보.....	112
과대적합 벌점에 대한 통계 상세 정보.....	112

JMP PRO 부스티드 트리 플랫폼 개요

부스티드 트리 플랫폼에서는 계층에 생성된 여러 개의 더 작은 결정 트리를 기반으로 하는 가법 결정 트리 모형을 생성합니다. 각 계층의 트리는 소수의 분할로 구성됩니다(보통 5개 이하). 각 계층은 "**파티션 모형**"에 설명된 재귀적 적합 방법론을 사용하여 적합됩니다. 차이는 지정된 수의 분할이 수행되면 적합이 중지된다는 것입니다. 트리가 주어지면 앞에 있는 관측값의 예측값은 해당 앞에 있는 모든 관측값의 평균입니다.

적합 프로세스는 다음과 같습니다.

1. 초기 계층을 적합시킵니다.
2. 잔차를 계산합니다. 이 값은 앞 내의 관측값에 대한 실제 값에서 예측 평균을 뺀 값입니다.
3. 계층을 잔차에 적합시킵니다.
4. 가법 트리를 생성합니다. 주어진 관측값에 대해 전체 계층에서 예측값의 합계를 계산합니다.
5. 지정된 계층 수에 도달할 때까지 또는 검증을 사용하는 경우 추가 계층을 적합시켜도 검증 통계량이 더 이상 개선되지 않을 때까지 **2 단계 ~ 4 단계**를 반복합니다.

최종 예측은 모든 계층에서 관측값에 대한 예측을 합한 것입니다.

이전 계층의 잔차에 대해 연속적으로 계층을 적합시키면 각 계층에서 적합성을 향상시킬 수 있습니다.

범주형 반응의 경우 반응 수준이 두 개인 반응만 지원됩니다. 범주형 반응을 사용할 때 각 계층에서 적합된 잔차는 선형 로짓의 오프셋입니다. 최종 예측은 모든 계층에 대한 선형 로짓 합계의 로지스틱 변환입니다.

부스티드 트리에 대한 자세한 내용은 Hastie et al. (2009) 연구 자료에서 확인하십시오.

JMP PRO 범주형 반응이 있는 부스티드 트리의 예

이 예에서는 밴딩 결함의 영향을 받는 인쇄 작업을 예측하기 위해 부스티드 트리 모형을 생성합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Bands Data.jmp 를 엽니다.
2. **분석 > 예측 모델링 > 부스티드 트리**를 선택합니다.
3. **Banding?** 을 선택하고 **Y, 반응**을 클릭합니다.
4. **Predictors** 열 그룹을 선택하고 **X, 요인**을 클릭합니다.
5. **검증 데이터 비율**에 "0.2" 를 입력합니다.
6. **확인**을 클릭합니다.

부스티드 트리 규격 창이 나타납니다.

7. (선택 사항) " 재현성 " 패널에서 **멀티스레딩 제한**을 선택하고 " 난수 시드값 " 에 "123" 을 입력합니다.

부스티드 트리 적합에는 랜덤 성분이 포함되므로 이렇게 하면 아래 표시된 것과 정확히 동일한 결과를 얻을 수 있습니다.

8. **확인**을 클릭합니다.

그림 6.2 명목형 반응에 대한 전체 통계량

전체 통계량

측도	훈련	검증	정의
엔트로피 R^2	0.7103	0.3299	$1 - \text{Loglike}(\text{model}) / \text{Loglike}(0)$
일반화 R^2	0.8340	0.4803	$(1 - (L(0)/L(\text{model}))^{2/n}) / (1 - L(0)^{2/n})$
평균 $-\log p$	0.1985	0.4383	$\sum -\text{Log}(p[j]) / n$
RASE	0.2196	0.3801	$\sqrt{\sum (y[j] - p[j])^2 / n}$
평균 절대 편차	0.1650	0.2828	$\sum y[j] - p[j] / n$
오분류 비율	0.0232	0.1852	$\sum (p[j] \neq p\text{Max}) / n$
N	431	108	n

혼동 행렬

훈련			검증		
실제값	예측 개수		실제값	예측 개수	
Banding?	band	noband	Banding?	band	noband
band	179	9	band	22	17
noband	1	242	noband	3	66

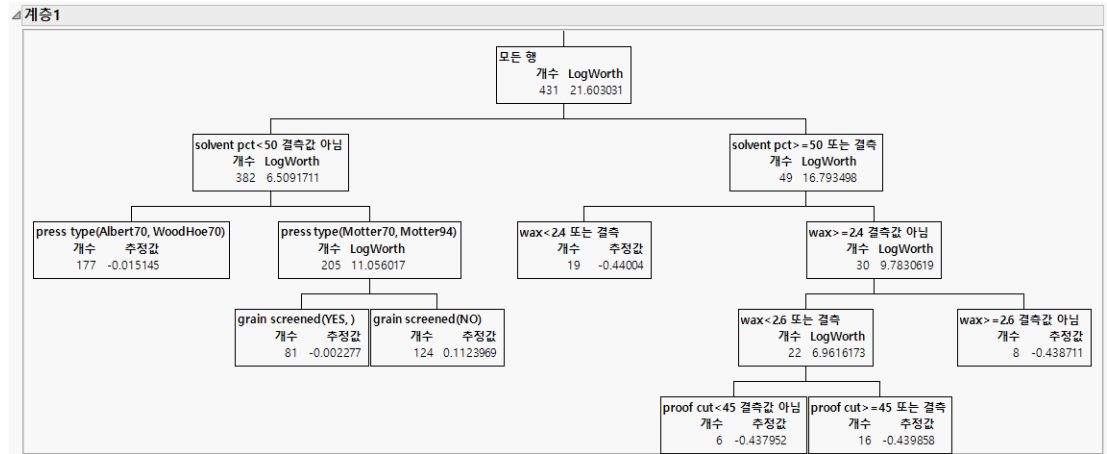
실제값	예측 비율		실제값	예측 비율	
Banding?	band	noband	Banding?	band	noband
band	0.952	0.048	band	0.564	0.436
noband	0.004	0.996	noband	0.043	0.957

반응 Banding? 이 범주형이므로 부스티드 트리 분석을 수행한 결과 " 측도 " 아래의 " 오분류 비율 " 과 " 혼동 행렬 " 보고서가 제공됩니다. 검증 데이터 집합의 오분류 비율은 0.1852(약 19%) 입니다.

9. "Banding? 에 대한 부스티드 트리 " 옆의 빨간색 삼각형을 클릭하고 **트리 표시 > 이름 범주 추정값 표시**를 선택합니다.

계층의 개요가 포함된 " 트리 보기 " 보고서가 나타납니다. 계층을 검토하여 적합한 트리과 예측값을 확인할 수 있습니다.

그림 6.3 부스티드 트리의 계층 1



10. "Banding? 에 대한 부스티드 트리 " 옆의 빨간색 삼각형을 클릭하고 **열 저장 > 예측 계산식 저장**을 선택합니다 .

확률(Banding?==noband), 확률(Banding?==band) 및 최대 확률 분류 Banding? 열이 데이터 테이블에 추가됩니다 . 확률 (Banding?==noband) 열을 검토하여 계층에서 모형 예측이 계산되는 방식을 확인할 수 있습니다 .

JMP PRO 연속형 반응이 있는 부스티드 트리의 예

이 예에서는 명목형 요인과 연속형 요인의 조합이 주어졌을 때 체지방률을 예측하기 위해 부스티드 트리 모형을 생성합니다 .

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Body Fat.jmp** 샘플 데이터 테이블을 엽니다 .
2. **분석 > 예측 모델링 > 부스티드 트리**를 선택합니다 .
3. Percent body fat 을 선택하고 **Y, 반응**을 클릭합니다 .
4. Age (years) ~ Wrist circumference (cm) 를 선택하고 **X, 요인**을 클릭합니다 .
5. Validation 을 선택하고 **검증**을 클릭합니다 .
6. **확인**을 클릭합니다 .
7. **확인**을 클릭합니다 .

그림 6.4 연속형 반응에 대한 전체 통계량

전체 통계량

	R ²	RASE	N
훈련	0.847	3.2116281	180
검증	0.611	5.4301412	72

" 전체 통계량 " 보고서에는 부스티드 트리 모형에 대한 R^2 및 RASE 가 제공됩니다 . 검증 데이터 집합의 R^2 은 0.611 입니다 . 검증 데이터 집합의 RASE 는 약 5.43 입니다 .

모형과 별개로 Percent body fat 에 대해 중요한 예측 변수를 나타내려고 합니다 .

8. "Percent body fat 에 대한 부스티드 트리 " 옆의 빨간색 삼각형을 클릭하고 **프로파일러**를 선택합니다 .
9. " 예측 프로파일러 " 옆의 빨간색 삼각형을 클릭하고 **변수 중요도 평가 > 독립 균등 입력**을 선택합니다 .

참고 : 변수 중요도 평가에는 랜덤화가 사용되므로 [그림 6.5](#) 의 결과와 정확히 일치하지 않을 수 있습니다 .

그림 6.5 변수 중요도의 요약 보고서

▼ 변수 중요도: 독립 균등 입력

▲ 요약 보고서

열	주요과	총 효과	.2	.4	.6	.8
Abdomen circumference (cm)	0.865	0.887				
Age (years)	0.063	0.085				
Wrist circumference (cm)	0.007	0.013				
Biceps (extended) circumference (cm)	0.003	0.006				
Hip circumference (cm)	0.002	0.005				
Chest circumference (cm)	0.002	0.004				
Height (inches)	0.002	0.003				
Ankle circumference (cm)	0.001	0.003				
Thigh circumference (cm)	0.001	0.002				
Weight (lbs)	0.001	0.002				
Knee circumference (cm)	3e-4	0.001				
Neck circumference (cm)	8e-18	2e-17				
Forearm circumference (cm)	8e-18	2e-17				

" 요약 보고서 " 에서는 Abdomen circumference (cm) 가 Percent body fat 의 가장 중요한 예측 변수임을 보여 줍니다 .

JMP PRO 부스티드 트리 플랫폼 시작

분석 > 예측 모델링 > 부스티드 트리를 선택하여 부스티드 트리 플랫폼을 시작합니다.

그림 6.6 Body Fat.jmp 를 사용한 부스티드 트리 시작 창

일련의 작은 트리로 구성된 결정 트리를 생성하여 반응을 예측합니다.

The screenshot shows the JMP Pro Boosted Tree platform start dialog. It is divided into three main sections: 'Select Y', 'Select X', and 'Options'.

- Select Y:** A list of 26 variables is shown, including 'Percent body fat', 'Age (years)', 'Weight (lbs)', 'Height (inches)', 'Neck circumference (cm)', 'Chest circumference (cm)', 'Abdomen circumference (cm)', 'Hip circumference (cm)', 'Thigh circumference (cm)', 'Knee circumference (cm)', 'Ankle circumference (cm)', 'Biceps (extended) circumference (cm)', 'Forearm circumference (cm)', 'Wrist circumference (cm)', 'Prediction Formulas(11/0)', and 'Validation'.
- Select X:** A list of variables is shown, including '가중치', '빈도', '검증', and '기준'.
- Options:** A section for configuring the model, including 'Method: Boosted Tree', 'Check data balance: 0.3', and checkboxes for 'Save model information' and 'Save model order'.

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

부스티드 트리 플랫폼 시작 창에는 다음과 같은 옵션이 있습니다.

Y, 반응 분석할 하나 이상의 반응 변수입니다.

X, 요인 예측 변수입니다.

가중치 분석의 각 행에 가중치를 할당하는 숫자 값이 포함된 열입니다.

빈도 분석의 각 행에 빈도를 할당하는 숫자 값이 포함된 열입니다.

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

- 검증 열에 두 개의 수준이 있는 경우 더 작은 값이 훈련 데이터 집합을 정의하고 더 큰 값이 검증 데이터 집합을 정의합니다.
- 검증 열에 세 개의 수준이 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.

부스티드 트리 플랫폼에서는 검증 열을 사용하여 모형을 훈련 및 조정하거나 모형을 훈련, 조정 및 평가합니다. 검증에 대한 자세한 내용은 "[JMP 모델링의 검증](#)"에서 확인하십시오.

"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 "[검증 열 생성](#)"에서 확인하십시오.

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.

방법 분할 방법 (결정 트리, 붓스트랩 포레스트, 부스티드 트리, K 최근접 이웃 또는 Naive Bayes)을 선택할 수 있습니다. 결정 트리를 제외한 이 대체 방법은 JMP Pro에서 사용할 수 있습니다.

이러한 방법에 대한 자세한 내용은 "[파티션 모형](#)", "[붓스트랩 포레스트](#)", "[K 최근접 이웃](#)" 및 "[Naive Bayes](#)"에서 확인하십시오.

검증 데이터 비율 검증 데이터 집합으로 사용할 데이터 비율입니다.

결측값 정보화 이 옵션을 선택하면 범주형 예측 변수에 대한 결측값 범주화 및 연속형 예측 변수에 대한 결측값 정보화 처리가 활성화됩니다. 자세한 내용은 "[ROC 곡선](#)"에서 확인하십시오.

순서형의 순서를 지킴 이 옵션을 선택하면 순서를 유지하는 분할만 고려하도록 제한됩니다.

JMP^{PRO} 규격 창

"부스티드 트리" 시작 창에서 "확인" 을 선택하면 "그래디언트 부스티드 트리 규격" 창이 나타납니다.

그림 6.7 부스티드 트리 규격 창

그래디언트 부스티드 트리 규격

부스트 계층 수: 152 트리당 분할 수: 2 학습률: 0.066 최소 크기 분할: 5		다중 적합 ☐ 분할 및 학습률에 대한 다중 적합 트리당 최대 분할 수: 3 최대 학습률: 0.1 ☐ 설계 조정 테이블 사용	
Stochastic 부스트 행 표집 비율: 1.0000 열 표집 비율: 1.0000		재현성 ☒ 멀티스레딩 제한 난수 시드값: 0	

☒ 조기 중지

JMP^{PRO} 부스트 패널

계층 수 최종 트리에 포함할 최대 계층 수입니다.

트리당 분할 수 각 계층의 분할 수입니다.

학습률 $0 < r \leq 1$ 의 숫자입니다. 학습률이 1 에 가까우면 최종 트리에 대한 수렴이 빨라지지만 데이터가 과대적합되는 경향도 높습니다. 적은 수의 계층이 지정된 경우 1 에 가까운 학습률을 사용하십시오. 학습률은 모형의 수렴을 늦추는 작은 소수이며 보통 0.01 에서 0.1 사이입니다. 이렇게 하면 이후 계층이 이전 계층과 다른 분할을 사용할 수 있도록 합니다.

과대적합 벌점 (범주형 반응에만 사용 가능) 적합 확률이 0 이 되지 않도록 하는 데 도움이 되는 편향 모수입니다. 자세한 내용은 "[과대적합 벌점에 대한 통계 상세 정보](#)"에서 확인하십시오.

최소 크기 분할 후보 분할에 필요한 최소 관측값 수입니다.

JMP^{PRO} 다중 적합 패널

분할 및 학습률에 대한 다중 적합 이 옵션을 선택하면 트리당 분할 수 (정수 증분) 와 학습률 (0.1 증분) 의 모든 조합에 대한 부스티드 트리가 생성됩니다.

조합의 하한은 "트리당 분할 수" 및 "학습률" 옵션으로 지정됩니다. 조합의 상한은 다음 옵션으로 지정됩니다.

트리당 최대 분할 수 트리당 분할 수의 상한입니다.

최대 학습률 학습률의 하한입니다.

설계 조정 테이블 사용 일부 조정 모수의 값이 포함된 데이터 테이블 (설계 조정 테이블) 을 선택할 수 있는 창을 엽니다. 설계 조정 테이블에는 지정할 각 옵션에 대한 열과 각각 단일 부스티드 트리 모형 설계를 나타내는 하나 이상의 행이 있습니다. 설계 조정 테이블에 옵션이 지정되지 않으면 기본값이 사용됩니다.

JMP에서는 지정된 조정 모수를 사용하여 테이블의 각 행에 대해 부스티드 트리 모형을 생성합니다. 설계 조정 테이블에 둘 이상의 모형이 지정된 경우 "모형 검증 집합 요약" 보고서에 각 모형의 R^2 값이 나열됩니다. "부스티드 트리" 보고서에는 R^2 값이 가장 큰 모형의 적합 통계량이 표시됩니다.

실험 설계 기능을 사용하여 설계 조정 테이블을 생성할 수 있습니다. 부스티드 트리 설계 조정 테이블에는 다음 열이 임의의 순서로 포함될 수 있습니다.

- 계층 수
- 트리당 분할 수
- 학습률
- 최소 크기 분할
- 행 표집 비율
- 열 표집 비율

JMP[®] PRO Stochastic 부스트 패널

행 표집 비율 각 계층에 대해 표집할 훈련 행의 비율입니다.

참고 : 반응이 범주형이면 층별 랜덤 표집을 사용하여 훈련 행이 표집됩니다.

열 표집 비율 각 계층에 대해 표집할 예측 변수 열의 비율입니다.

JMP[®] PRO 재현성 패널

멀티스레딩 제한 이 옵션을 선택하면 모든 계산이 단일 스레드에서 수행됩니다.

난수 시드값 이후 플랫폼 실행을 위해 결과를 재현하는 데 사용할 0 이 아닌 숫자 난수 시드값을 지정합니다. 기본적으로 난수 시드값은 0 으로 설정되어 재현 가능한 결과를 생성하지 않습니다. 분석을 스크립트에 저장하면 사용자가 입력한 난수 시드값이 스크립트에 저장됩니다.

JMP[®] PRO 조기 중지

조기 중지 이 옵션을 선택하면 계층을 추가해도 검증 통계량이 개선되지 않을 경우 부스트 프로세스에서 추가 계층 적합을 중지합니다. 이 옵션을 선택하지 않으면 지정된 계층 수에 도달할 때까지 부스트 프로세스가 계속됩니다. 이 옵션은 검증을 사용하는 경우에만 나타납니다.

JMP^{PRO} 부스티드 트리 보고서

"그래디언트 부스티드 트리 규격" 창에서 "확인"을 클릭하면 부스티드 트리 보고서가 열립니다.

그림 6.8 연속형 반응에 대한 부스티드 트리 보고서

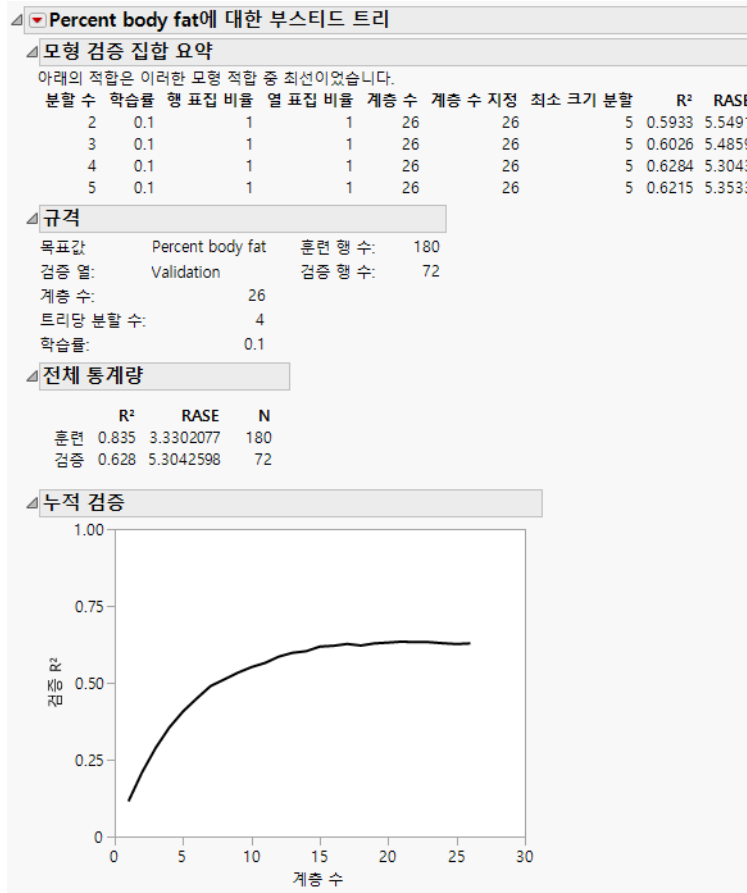
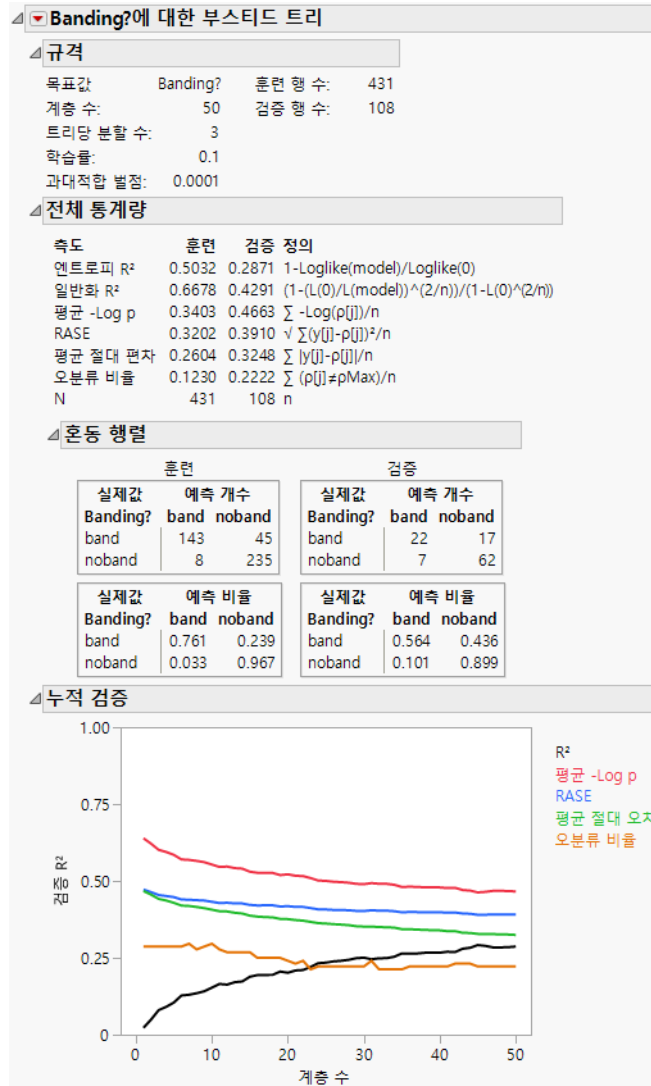


그림 6.9 범주형 반응에 대한 부스티드 트리 보고서



반응이 범주형인지 아니면 연속형인지에 따라 다음과 같은 보고서가 제공됩니다.

- "모형 검증 집합 요약"
- "규격"
- "전체 통계량"
- "누적 검증"

JMP PRO 모형 검증 집합 요약

부스티드 트리 플랫폼의 "규격" 창에서 "분할 및 학습률에 대한 다중 적합" 옵션을 "선택한 경우" "모형 검증 집합 요약" 보고서에 모든 모형 적합에 대한 적합 통계량이 표시됩니다. 자세한 내용은 [그림 6.8](#) 및 "다중 적합 패널"에서 확인하십시오.

JMP PRO 규격

부스티드 트리 플랫폼의 "규격" 보고서에는 모형 적합에 사용된 설정이 표시됩니다.

JMP PRO 전체 통계량

부스티드 트리 플랫폼의 "전체 통계량" 보고서에는 훈련 데이터 집합과 검증 및 테스트 데이터 집합(지정된 경우)에 대한 적합 통계량이 표시됩니다.

부스티드 트리 규격 창의 "분할 및 학습률에 대한 다중 적합" 옵션을 사용하여 여러 모형을 적합시키는 경우를 가정해 보겠습니다. 그러면 "전체 통계량" 및 "누적 검증" 보고서에 결과가 표시되는 모형은 검증 데이터 집합의 엔트로피 R^2 값(범주형 반응의 경우) 또는 R^2 값(연속형 반응의 경우)이 가장 큰 모형입니다.

JMP PRO 측도 보고서

(범주형 반응에만 사용 가능) 훈련 데이터 집합과 검증 및 테스트 데이터 집합(지정된 경우)에 대한 다음 통계량을 제공합니다.

참고 : 엔트로피 R^2 및 일반화 R^2 의 경우 값이 1에 가까울수록 더 나은 적합을 나타냅니다. 평균 $-\text{Log } p$, RASE, 평균 절대 편차 및 오분류 비율의 경우 값이 작을수록 더 나은 적합을 나타냅니다.

엔트로피 R^2 적합 모형과 상수 확률 모형의 로그 가능도를 비교하는 적합 측도입니다. 값 범위는 0에서 1 사이입니다. 자세한 내용은 "[엔트로피 \$R^2\$](#) "에서 확인하십시오.

일반화 R^2 일반 회귀 모형에 적용할 수 있는 측도입니다. 이 값은 가능도 함수 L 을 기반으로 하며 최대값이 1이 되도록 적도화됩니다. 값이 1이면 완벽한 모형을 나타내고 0이면 상수 모형과 같음을 나타냅니다. 일반화 R^2 측도는 표준 최소 제곱 설정 시 연속형 정규 반응에 대한 기준 R^2 으로 단순화됩니다. 일반화 R^2 을 Nagelkerke 또는 Craig 와 Uhler R^2 이라고도 하는데, 이는 Cox-Snell 유사 R^2 을 정규화한 것입니다.

평균 $-\text{Log } P$ 음의 $\log(p)$ 의 평균입니다. 여기서 p 는 발생한 사건과 연관된 적합 확률입니다.

RASE 제곱근 평균 제곱 예측 오차입니다. 차이는 1에서 p (실제로 발생한 반응 수준의 적합 확률) 사이입니다.

평균 절대 편차 반응과 예측 반응 사이의 차이에 대한 절대값의 평균입니다. 차이는 1에서 p (실제로 발생한 반응 수준의 적합 확률) 사이입니다.

오분류 비율 적합 확률이 가장 높은 반응 범주가 관측된 범주가 아닌 비율입니다.

N 관측값 수입니다.

JMP^{PRO} 혼동 행렬 보고서

(범주형 반응에만 사용 가능) 훈련 데이터 집합과 검증 및 테스트 데이터 집합(지정된 경우)에 대한 분류 통계량을 표시합니다. "혼동 행렬" 보고서에는 혼동 행렬과 혼동 비율 행렬이 포함됩니다. 혼동 행렬은 실제 반응과 예측 반응의 이원 분류입니다. 혼동 비율 행렬은 값을 행 합계로 나눈 혼동 행렬과 동일합니다.

JMP^{PRO} 결정 행렬

(범주형 반응을 사용하고 반응에 "수익 행렬" 열 특성이 있거나 "수익 행렬 지정" 옵션을 사용하여 비용을 지정하는 경우에만 사용 가능) 훈련 데이터 집합과 검증 및 테스트 데이터 집합(지정된 경우)에 대한 결정 개수 및 결정 비율 행렬을 제공합니다. 자세한 내용은 "[파티션 플랫폼의 추가 예](#)"에서 확인하십시오.

JMP^{PRO} 누적 검증

부스티드 트리 플랫폼의 "누적 검증" 보고서에는 검증 데이터 집합 대 계층 수에 대한 적합 통계량 그림이 표시됩니다. 이 보고서는 검증을 사용하는 경우에만 사용할 수 있습니다.

연속형 반응의 단일 적합 통계량은 R^2 입니다. 범주형 반응의 적합 통계량은 아래에 나열되어 있으며 "[측도 보고서](#)"에서 설명합니다.

- R^2 (엔트로피 R^2)
- 평균 -Log p(평균 -Log p)
- 제공된 평균 제공 예측 오차 (RASE)
- 평균 절대 오차 (평균 절대 편차)
- MR(오분류 비율)

"누적 검증" 그림 아래의 "누적 상세 정보" 보고서에는 그림에 사용된 값이 제공됩니다.

JMP^{PRO} 부스티드 트리 플랫폼 옵션

"부스티드 트리"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

트리 표시 "트리 보기" 보고서에 트리를 표시하기 위한 옵션을 제공합니다. 이 보고서에는 부스트 프로세스의 각 계층에서 적합된 트리 그림이 제공됩니다.

실제값 대 예측값 그림 (연속형 반응에만 사용 가능) 실제값 대 예측값 그림을 제공합니다.

열 기여도 적합에 대한 각 입력 열의 기여도를 보여 주는 보고서를 표시합니다. 다음 항목도 보고서에 표시됩니다.

- 지정된 열을 사용하여 데이터를 분할할 때 모든 트리에 대한 총 인스턴스 수
- 열에 기인한 총 G^2 (범주형 반응의 경우) 또는 SS, 즉 제곱합 (연속형 반응의 경우)
- G^2 또는 SS의 막대 차트
- 열에 기인한 G^2 또는 SS의 비율

ROC 곡선 (범주형 반응에만 사용 가능) 자세한 내용은 "[ROC 곡선](#)"에서 확인하십시오.

향상도 곡선 (범주형 반응에만 사용 가능) 자세한 내용은 "[향상도 곡선](#)"에서 확인하십시오.

결정 임계값 (이항 범주형 반응에만 사용 가능) 훈련, 검증 및 테스트 (지정된 경우) 데이터 집합에 대한 "결정 임계값" 보고서를 표시하거나 숨깁니다. 각 보고서에는 모형 적합 비교를 위한 각 모형의 적합 확률 분포 그래프, 혼동 행렬 및 분류 그래프와 분류 정확성 측정 기준 테이블이 포함되어 있습니다. 자세한 내용은 "[결정 임계값 보고서](#)"에서 확인하십시오.

열 저장 모형 및 트리 결과를 저장하고 SAS 코드를 생성하기 위한 옵션이 포함되어 있습니다.

예측값 저장 모형의 예측값을 데이터 테이블에 저장합니다.

예측 계산식 저장 예측 계산식을 데이터 테이블의 열에 저장합니다. 계산식은 트리 구조를 설명하는 내포된 조건부 절로 구성됩니다. 반응이 연속형이면 열에 "예측" 열 특성이 포함됩니다. 반응이 범주형이면 열에 "반응 확률" 열 특성이 포함됩니다.

허용 예측 계산식 저장 (이 옵션 대신 "예측 계산식 저장" 옵션을 사용해야 합니다.) "예측 계산식 저장"을 사용할 수 없을 때만 이 옵션을 사용하십시오.) 결측값이 있고 "결측값 정보화"를 선택하지 않은 경우에도 예측하는 계산식을 저장합니다. 예측 계산식은 결측 예측 변수에 대한 반응 값을 분할에 랜덤 할당하여 결측값을 허용합니다. 반응이 연속형이면 열에 "예측" 열 특성이 포함됩니다. 반응이 범주형이면 열에 "반응 확률" 열 특성이 포함됩니다. "결측값 정보화"를 선택한 경우 Shift 키를 누른 채 보고서의 빨간색 삼각형을 클릭하면 허용 예측 계산식을 저장할 수 있습니다.

잔차 저장 (연속형 반응에만 사용 가능) 잔차를 데이터 테이블에 저장합니다.

오프셋 추정값 저장 (범주형 반응에만 사용 가능) 선형 성분의 합을 저장합니다. 이는 적합 확률의 로짓입니다.

트리 상세 정보 저장 각 계층에 대한 분할 상세 정보 및 추정값이 포함된 데이터 테이블을 생성합니다.

누적 상세 정보 저장 (검증을 사용하는 경우에만 사용 가능) 각 계층에 대한 적합 통계량이 포함된 데이터 테이블을 생성합니다.

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 "[계산식 저장소](#)"에서 확인하십시오.

허용 예측 계산식 게시 (이 옵션 대신 "예측 계산식 게시" 옵션을 사용해야 합니다.) "예측 계산식 게시"를 사용할 수 없을 때만 이 옵션을 사용하십시오.) 허용 예측 계산식을 생성하여 계산식 저장소 플랫폼에 계산식 열 스크립트로 저장합니다. "계산식 저장소" 보고서가 열려 있지 않으면 "계산식 저장소" 보고서가 생성됩니다. 자세한 내용은 "[계산식 저장소](#)"에서 확인하십시오. "결측값 정보화"를 선택한 경우 Shift 키를 누른 채 보고서의 빨간색 삼각형을 클릭하면 이 옵션을 사용할 수 있습니다.

SAS DATA 스텝 만들기 새 데이터 집합을 스코어링하기 위한 SAS 코드를 생성합니다.

수익 행렬 지정 (범주형 반응에만 사용 가능) 올바르게나 올바르게지 않은 분류 결정과 관련된 수익 또는 비용을 지정할 수 있습니다. 자세한 내용은 "[적합 상세 정보 표시](#)"에서 확인하십시오.

프로파일러 예측 프로파일러를 표시합니다. 자세한 내용은 **Profilers**에서 확인하십시오.

다음 옵션에 대한 자세한 내용은 **JMP 사용**의 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.

JMP PRO 부스티드 트리 플랫폼에 대한 통계 상세 정보

이 섹션에서는 부스티드 트리 플랫폼과 관련된 상세 정보를 설명합니다. 재귀 결정 트리에 대한 통계 상세 정보는 "[파티션 플랫폼에 대한 통계 상세 정보](#)"에서 확인하십시오.

JMP PRO 과대적합 벌점에 대한 통계 상세 정보

반응이 범주형이면 모수 벌점이 부과됩니다. 각 계층에서 추정값은 음의 로그 가능도 + 벌점 값과 각 관측값에 대한 추정값의 제곱합을 곱한 값을 최소화합니다. 이 벌점은 각 새 계층이 훈련 데이터를 과대적합하지 않도록 합니다.

인접한 관측값을 사용하여 반응 값 예측

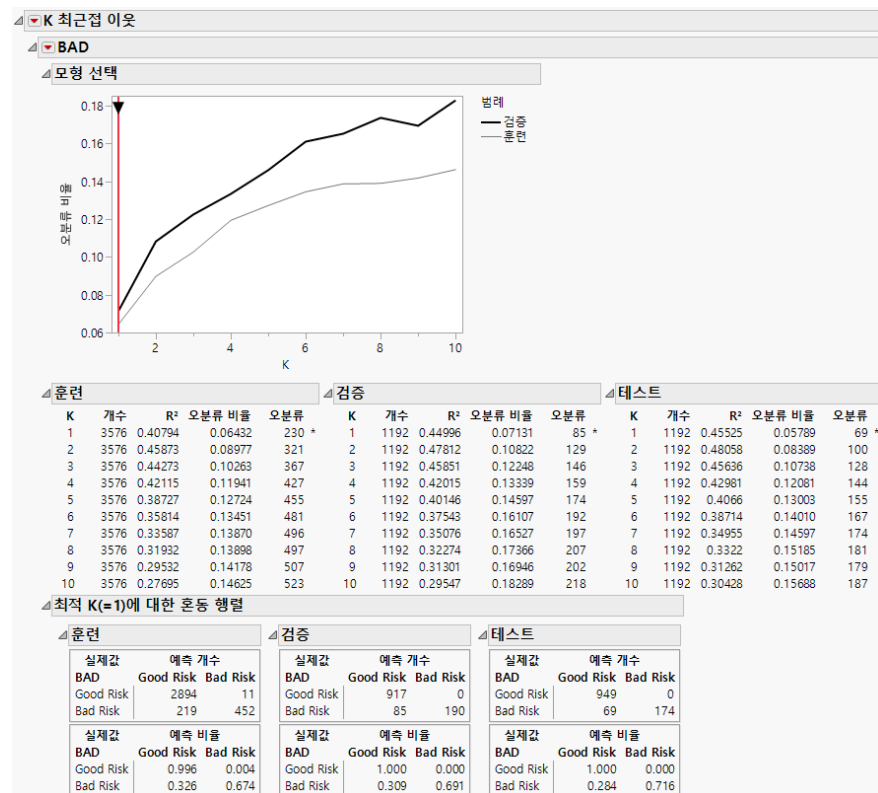
K 최근접 이웃 플랫폼은 JMP Pro에서만 사용할 수 있습니다.

K 최근접 이웃 플랫폼에서는 지정된 관측값의 국소적 근방에 있는 관측값의 반응을 사용하여 해당 관측값에 대한 반응 값을 예측합니다. 이 플랫폼은 범주형 반응을 분류하고 연속형 반응을 예측하는 데 사용할 수 있습니다.

K 최근접 이웃은 인접한 관측값까지의 거리를 기반으로 하는 비모수 방법입니다. 이 때문에 K 최근접 이웃 방법은 불규칙한 예측 변수 값 경계를 사용하여 관측값을 분류할 수 있습니다. 그러나 이 알고리즘은 관련 없는 예측 변수의 영향을 받으므로 예측 변수를 선택하면 결과에 영향을 줄 수 있습니다.

K 최근접 이웃은 위성 이미지와 EKG(심전도) 패턴 등 다양한 응용 분야에서 사용되고 있습니다.

그림 7.1 K 최근접 이웃 플랫폼의 예



목차

K 최근접 이웃 플랫폼 개요.....	115
범주형 반응이 있는 K 최근접 이웃의 예.....	115
연속형 반응이 있는 K 최근접 이웃의 예.....	117
K 최근접 이웃 플랫폼 시작.....	119
K 최근접 이웃 보고서	120
연속형 반응.....	121
범주형 반응.....	121
K 최근접 이웃 플랫폼 옵션.....	122
모형 적합 옵션	123

JMP PRO K 최근접 이웃 플랫폼 개요

K 최근접 이웃 플랫폼은 k 개의 최근접 이웃의 반응을 기반으로 반응 값을 예측합니다. 주어진 관측값에 대한 k 개의 최근접 이웃은 해당 관측값에 대한 예측 변수 값과 다른 각 관측값에 대한 예측 변수 값 사이의 최소 유클리드 거리 k 개를 식별하여 결정됩니다. K 최근접 이웃 플랫폼에서는 연속형 반응과 범주형 반응을 모두 모델링합니다.

k 최근접 이웃 방법은 대규모 문제의 경우 예측 계산식이 복잡하고 해석하기 어려워 유용성이 제한될 수 있다는 단점을 가지고 있습니다. 또한 K 최근접 이웃 방법은 범주형 반응에 대한 확률을 계산하지 않습니다. k 최근접 이웃 방법에 대한 자세한 내용은 Hastie et al. (2009), Hand et al. (2001) 및 Shmueli et al. (2017) 연구 자료에서 확인하십시오.

연속형 반응

연속형 반응의 경우 예측값은 k 개의 최근접 이웃에 대한 반응의 평균입니다. 각 연속형 예측 변수는 표준편차로 척도화됩니다. 이 척도를 사용하면 범위가 큰 단일 예측 변수가 거리 계산에 과도하게 영향을 미치지 않습니다. 연속형 예측 변수의 예측값은 해당 예측 변수의 평균으로 대체됩니다. 자세한 내용은 "[연속형 반응이 있는 K 최근접 이웃의 예](#)"에서 확인하십시오.

범주형 반응

범주형 반응의 경우 예측값은 k 개의 최근접 이웃에 대해 가장 빈도가 높은 반응 수준입니다. 둘 이상의 수준이 가장 빈도가 높은 수준으로 묶여 있으면 이러한 수준 중 하나를 랜덤 선택하여 예측 반응이 할당됩니다.

참고: 범주형 반응의 경우 가장 빈도가 높은 수준에 대한 동점 연결이 무작위로 끊어지기 때문에 각 플랫폼 실행 결과가 다를 수 있습니다. 재현 가능한 결과를 얻으려면 시작 창에서 "난수 시드 값 설정" 옵션을 사용하거나 JSL 스크립트에 `Set Random Seed()` 함수를 포함하십시오.

범주형 예측 모형에서 각 범주형 예측 변수는 표시 변수로 표현되며, 이때 하나의 표시 변수가 각 수준을 나타냅니다. 범주형 예측 변수의 예측값이 있는 행은 해당 예측 변수에 대한 모든 표시 변수에서 값이 0으로 표시됩니다.

JMP PRO 범주형 반응이 있는 K 최근접 이웃의 예

이 예에서는 주택 담보 대출을 신청하는 고객의 신용 위험을 분류하는 모형을 생성하려고 합니다. 5,960 명의 고객을 대상으로 각 고객이 Good Risk(안전한 위험) 또는 Bad Risk(불안한 위험)로 분류된 과거 재무 데이터가 있습니다. 많은 예측 변수에 대해 예측값이 있습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Equity.jmp**를 엽니다.
2. **분석 > 예측 모델링 > K 최근접 이웃**을 선택합니다.
3. **BAD**를 선택하고 **Y, 반응**을 클릭합니다.

4. LOAN ~ CLNO 를 선택하고 **X, 요인**을 클릭합니다 .

잠재적 예측 변수 중 하나인 DEBTINC 에는 결측값이 많이 있으므로 모형에 포함하지 않습니다 . 연속형 예측 변수의 결측값은 해당 예측 변수의 평균으로 대체됩니다 . 이 절차는 주로 결측값이 랜덤인 경우에 잘 맞습니다 . DEBTINC 의 높은 결측 비율은 결측값이 정보를 제공할 수 있음을 나타내지만 이 예에서는 조사하지 않습니다 .

5. Validation 을 선택하고 **검증**을 클릭합니다 .6. **확인**을 클릭합니다 .

그림 7.2 K 최근접 이웃 보고서



JMP에서는 각 K 값에 대해 훈련 데이터 집합 관측값만 사용하여 모형을 생성합니다 . 이러한 각 모형은 검증 데이터 집합 관측값을 분류하는 데 사용됩니다 . 검증 데이터 집합 결과는

최적 모형을 선택하는 데 사용됩니다. 이 예에서는 단일 최근접 이웃 ($K=1$) 에 기반한 모형의 오분류 비율이 가장 작습니다. 테스트 데이터 집합은 단일 최근접 이웃 모형이 독립 데이터에 대해 최상의 성능을 발휘하는지 확인합니다.

7. "BAD" 의 빨간색 삼각형을 클릭하고 **예측 계산식 게시**를 선택합니다.
8. **이웃 수, K** 옆의 기본값 1 을 그대로 사용합니다.
9. **확인**을 클릭합니다.

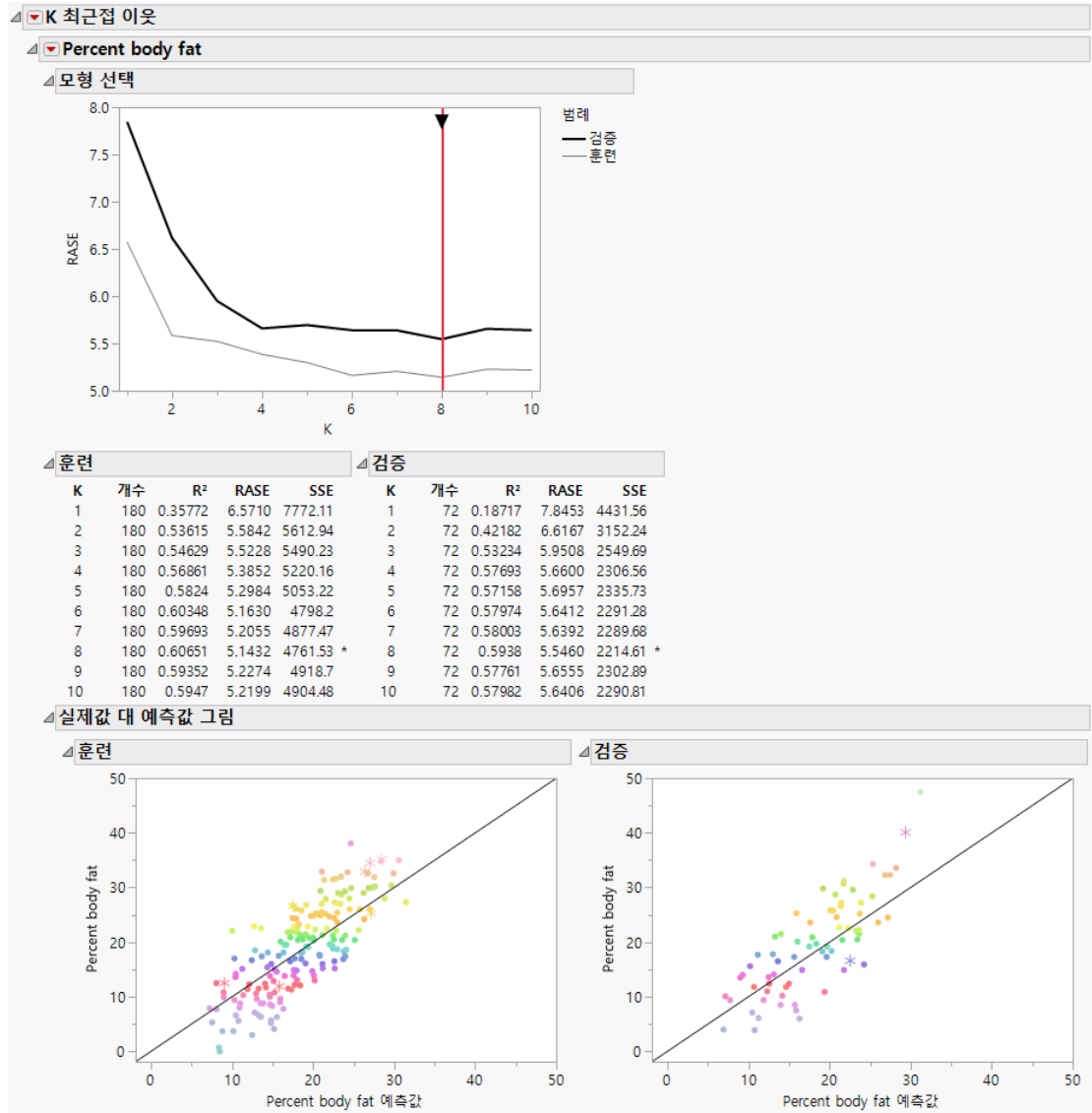
예측 방정식이 계산식 저장소에 저장됩니다. "계산식 저장소" 의 "모형 비교" 옵션을 사용하여 $K=1$ 최근접 이웃 모형의 성능과 계산식 저장소에 게시된 대체 모형의 성능을 비교할 수 있습니다. 자세한 내용은 "[계산식 저장소](#)" 에서 확인하십시오.

JMP PRO 연속형 반응이 있는 K 최근접 이웃의 예

이 예에서는 13 개의 예측 변수를 사용하여 남성의 체지방률을 예측하려고 합니다. 샘플 데이터 테이블에는 수중체중법 및 다양한 신체 둘레 측정을 기반으로 하는 체지방률 추정값이 포함되어 있습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Body Fat.jmp 를 엽니다.
2. **분석 > 예측 모델링 > K 최근접 이웃**을 선택합니다.
3. Percent body fat 을 선택하고 **Y, 반응**을 클릭합니다.
4. Age (years) ~ Wrist circumference (cm) 를 선택하고 **X, 요인**을 클릭합니다.
5. Validation 을 선택하고 **검증**을 클릭합니다.
6. **확인**을 클릭합니다.
7. "Percent body fat" 의 빨간색 삼각형을 클릭하고 **실제값 대 예측값 그림**을 선택합니다.

그림 7.3 K 최근접 이웃 보고서



K=8 모형이 검증 데이터 집합에 대한 RASE 가 가장 낮았습니다. k 최근접 이웃 모형 중에서 8 개의 최근접 이웃에 기반한 모형의 성능이 가장 좋은 것으로 보입니다. 훈련 데이터 집합의 실제값 대 예측값 그림을 보면 점이 선을 따라 분포되어 있으므로 예측값이 실제값과 유사하다는 것을 알 수 있습니다. 검증 데이터 집합에 대한 그림에서는 대부분의 점이 선을 따라 분포되며 약간 멀리 떨어진 상단에 몇몇 관측값이 있습니다.

JMP PRO K 최근접 이웃 플랫폼 시작

분석 > 예측 모델링 > K 최근접 이웃을 선택하여 K 최근접 이웃 플랫폼을 시작합니다.

그림 7.4 K 최근접 이웃 시작 창

K 최근접 이웃의 반응을 기반으로 반응을 예측합니다.

열 선택

14개 열

- BAD
- LOAN
- MORTDUE
- VALUE
- REASON
- JOB
- YOJ
- DEROG
- DELINQ
- CLAGE
- NINQ
- CLNO
- DEBTINC
- Validation

선택한 열 역할 지정

Y, 반응: 선택적

X, 요인: 필수, 선택적

검증: 선택적 숫자

기준: 선택적

작업

확인

취소

제거

재호출

도움말

검증 데이터 비율: 0

K: 10

범주 편향: 0.5

난수 시드값 설정: .

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

K 최근접 이웃 시작 창에는 다음과 같은 옵션이 있습니다.

Y, 반응 분석할 하나 이상의 반응 변수입니다.

참고: K 최근접 이웃 플랫폼은 반응 변수가 없는 경우에도 인접 관측값 사이의 거리를 결정하는 유틸리티로 사용할 수 있습니다. 반응 변수를 지정하지 않으면 빈 보고서가 나타납니다. 그러나 빨간색 삼각형 메뉴의 "근접 이웃 행 저장" 및 "근접 이웃 거리 저장" 옵션을 사용할 수 있습니다.

X, 요인 예측 변수입니다.

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

- 검증 열에 두 개의 수준이 있는 경우 더 작은 값이 훈련 데이터 집합을 정의하고 더 큰 값이 검증 데이터 집합을 정의합니다.
- 검증 열에 세 개의 수준이 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.

K 최근접 이웃 플랫폼에서는 검증 열을 사용하여 모형을 훈련 및 조정하거나 모형을 훈련, 조정 및 평가합니다. 검증에 대한 자세한 내용은 "[JMP 모델링의 검증](#)"에서 확인하십시오. "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 "[검증 열 생성](#)"에서 확인하십시오.

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.

검증 데이터 비율 검증 데이터 집합으로 사용할 데이터 비율입니다.

이웃 수, K 분석할 최근접 이웃의 최대 수입니다. 모형은 K에 지정한 값까지 하나의 최근접 이웃에 대해 적합됩니다.

참고: 최대 이웃 수 K는 훈련 데이터 테이블의 행 수에서 1을 뺀 값 이하여야 합니다. 허용 가능한 최대 K보다 큰 K를 지정하면 경고가 나타납니다.

범주 편향 범주형 반응의 적합 확률이 항상 양수가 되도록 보증하는 조정 모수를 지정합니다. 개수가 0이 되지 않도록 하기 위해 관측값 수에 편향 값의 비율이 추가됩니다. 이 절차에서는 적합 확률이 항상 양수가 되도록 0의 로그를 계산하지 않습니다. 범주 편향 값은 0에서 1 사이여야 합니다. 기본값은 0.5입니다.

난수 시드값 설정 명목형 및 순서형 반응에 대한 우선 순위 결정에 사용되는 랜덤화 프로세스의 시드값을 설정합니다. "검증 데이터 비율"을 지정하는 경우 이 옵션은 검증에 사용되는 행의 시드값도 설정합니다. "난수 시드값 설정"은 분석을 재현하려는 경우에 유용합니다. 난수 시드값을 설정하고 스크립트를 저장하면 시드값이 스크립트에 자동으로 저장됩니다.

JMP PRO K 최근접 이웃 보고서

"K 최근접 이웃" 보고서에는 각 반응 변수에 대한 개별 보고서가 포함되어 있습니다. 각 반응 변수 보고서에는 해당 반응에 대한 적합 모형 정보가 포함되어 있습니다. 이 정보는 "모형 선택" 보고서 및 적합한 각 k개 모형에 대한 요약 정보를 포함합니다. 보고서에는 훈련 데이터 집합에 대한 테이블과 검증 및 테스트 데이터 집합(검증을 사용하여 정의한 경우)에 대한 테이블이 표시됩니다.

"모형 선택" 보고서에는 범주형 반응의 오분류 비율 또는 연속형 반응의 RASE를 기반으로 K에 대한 해 경로 그림이 표시됩니다. 기본적으로 최적 성능 모형에 해당하는 K 값에 슬라이더가 배치됩니다. 슬라이더를 드래그하여 보고서에서 K 값을 변경할 수 있습니다.

반응의 모델링 유형에 따라 보고되는 통계량이 달라집니다. 요약 테이블의 각 행은 k개의 최근접 이웃으로 정의된 모형에 해당하며, K의 범위는 1 ~ 시작 창의 "이웃 수, K"에 지정한 값입니다.

JMP PRO 연속형 반응

기본적으로 연속형 반응의 "K 최근접 이웃" 보고서에는 "모형 선택" 그래프 외에 요약 테이블이 포함됩니다.

요약 테이블

별표는 RASE 가 가장 작은 K 값에 대한 모형을 표시합니다. 연속형 반응의 보고서에는 다음 열이 포함됩니다.

K 모형에 사용된 최근접 이웃 수입니다. K의 범위는 1 ~ 시작 창의 "이웃 수, K"에 지정한 값입니다.

개수 관측값 수입니다.

R² 모형의 R² 값입니다.

RASE 모형에 대한 제곱근 평균 제곱 예측 오차입니다. RASE가 가장 작은 모형에 별표가 표시됩니다. 동점 RASE 값이 있으면 K가 가장 작은 모형에 별표가 표시됩니다.

SSE 모형에 대한 오차 제곱합입니다.

JMP PRO 범주형 반응

기본적으로 범주형 반응의 "K 최근접 이웃" 보고서에는 "모형 선택" 그래프 외에 요약 테이블, 혼동 행렬 및 모자이크 그림이 포함됩니다.

요약 테이블

별표는 오분류 비율이 가장 작은 K 값에 대한 모형을 표시합니다. 범주형 반응의 보고서에는 다음 열이 포함됩니다.

K 모형에 사용된 최근접 이웃 수입니다. K의 범위는 1 ~ 시작 창의 "이웃 수, K"에 지정한 값입니다.

개수 관측값 수입니다.

R² 모형의 엔트로피 R²입니다.

오분류 비율 모형에 의해 잘못 분류된 관측값의 비율입니다. 이 값은 "오분류"를 "개수"로 나누어 계산됩니다. 오분류 비율이 가장 작은 모형에 별표가 표시됩니다. 동점 오분류 비율 값이 있으면 K가 가장 작은 모형에 별표가 표시됩니다.

오분류 모형에서 잘못 예측한 관측값 수입니다.

혼동 행렬

기본적으로 오분류 비율이 가장 작은 모형에 대한 "혼동 행렬" 보고서가 표시됩니다. 가장 작은 오분류 비율에 동점이 있으면 K가 가장 작은 모형에 대한 보고서가 표시됩니다. "혼동 행렬" 보고서에는 혼동 행렬과 혼동 비율 행렬이 포함되어 있습니다. 혼동 행렬은 실제 반응과 예측 반응의 이원 분류입니다. 혼동 비율 행렬은 값을 행 합계로 나눈 혼동 행렬과 동일합니다. 검증율 사

용하는 경우 검증 데이터 집합과 테스트 데이터 집합에 대한 혼동 행렬 및 혼동 비율 행렬이 나타납니다. "혼동 행렬" 보고서와 오분류 비율을 사용하여 모형을 평가할 수 있습니다.

팁 : 해 경로 그림에서 슬라이더 위치를 변경하면 선택한 K 값에 대한 추가 "혼동 행렬" 보고서가 표시됩니다. 이 보고서를 사용하여 대체 모형을 기본 최적 모형과 비교할 수 있습니다.

모자이크 그림

기본적으로 오분류 비율이 가장 작은 모형에 대한 모자이크 그림이 표시됩니다. 가장 작은 오분류 비율에 동점이 있으면 K 가 가장 작은 모형에 대한 모자이크 그림이 표시됩니다. 모자이크 그림은 각 세그먼트가 해당 그룹의 빈도 수에 비례하는 누적 막대 차트입니다. 모자이크 그림에 대한 자세한 내용은 기본 분석에서 확인하십시오. 검증을 사용하는 경우 검증 데이터 집합과 테스트 데이터 집합에 대한 모자이크 그림이 표시됩니다.

팁 : 해 경로 그림에서 슬라이더 위치를 변경하면 모자이크 그림이 업데이트되어 선택한 K 값의 결과가 표시됩니다.

JMP PRO K 최근접 이웃 플랫폼 옵션

"K 최근접 이웃"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

근접 이웃 행 저장 K 개 열을 데이터 테이블에 저장합니다. 열 이름은 이웃 행 k 입니다. 주어진 행의 k 번째 열에는 k 번째 최근접 이웃의 행 번호가 포함됩니다.

주의 : 데이터 테이블에서 행을 재정렬할 때 이웃 행 k 열의 행 번호는 업데이트되지 않습니다. 행을 재정렬하면 해당 열의 값이 올바르게 표시되지 않습니다.

근접 이웃 거리 저장 K 개 열을 데이터 테이블에 저장합니다. 열 이름은 거리 k 입니다. 주어진 행의 k 번째 열에는 해당 행의 k 번째 최근접 이웃까지의 거리가 포함됩니다.

다음 옵션에 대한 자세한 내용은 JMP 사용의 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.

모형 적합 옵션

"K 최근접 이웃" 보고서의 "모형 적합" 섹션에는 다음과 같은 빨간색 삼각형 메뉴 옵션이 있습니다.

모자이크 그림 (명목형 또는 순서형 반응에만 사용 가능) 모자이크 그림을 표시하거나 숨깁니다. 자세한 내용은 "[모자이크 그림](#)"에서 확인하십시오.

실제값 대 예측값 그림 (연속형 반응에만 사용 가능) RASE가 가장 작은 모형에 대한 실제 반응 대 예측 반응 그림을 표시합니다. 가장 작은 RASE에 동점이 있는 경우 그림은 K가 가장 작은 모형을 기반으로 합니다.

팁: 해 경로 그림의 슬라이더 위치를 다른 K로 변경하면 "실제값 대 예측값 그림"이 업데이트되어 선택한 K 값에 대한 모형이 반영됩니다.

잔차 대 예측값 그림 (연속형 반응에만 사용 가능) RASE가 가장 작은 모형에 대한 잔차 대 예측 반응 그림을 표시합니다. 가장 작은 RASE에 동점이 있는 경우 그림은 K가 가장 작은 모형을 기반으로 합니다.

팁: 해 경로 그림의 슬라이더 위치를 다른 K로 변경하면 "잔차 대 예측값 그림"이 업데이트되어 선택한 K 값에 대한 모형이 반영됩니다.

예측값 저장 K개의 예측값 열을 데이터 테이블에 저장합니다. 열 이름은 예측 <반응> k입니다. k 번째 열에는 k 번째 최근접 이웃을 기반으로 하는 모형에 대한 예측이 포함됩니다. 여기서 반응은 반응 열 이름입니다. 플랫폼에 나타나는 보고서 통계량은 이러한 원시 예측을 사용하여 생성됩니다.

예측 계산식 저장 k 개의 특정 최근접 이웃 모형에 대한 예측 계산식이 포함된 열을 저장합니다. 메시지가 표시되면 K 값을 입력합니다. 예측 계산식에는 모든 훈련 데이터가 포함되므로 큰 데이터 테이블의 경우 이 옵션을 사용하면 실용성이 떨어질 수 있습니다. 이 옵션은 새 관측 값을 스코어링하거나 결측 반응 값을 예측하는 데 유용합니다.

주의: "예측 계산식 저장"과 "예측값 저장"을 사용하여 얻은 훈련 데이터 행의 예측값이 일치하지 않습니다. "예측 계산식 저장" 옵션을 사용하여 얻은 예측값은 예측값 행을 포함하여 훈련 데이터 집합의 모든 행을 사용합니다. "예측값 저장" 옵션을 사용하여 얻은 예측값은 훈련 데이터 집합의 예측값 행을 제외한 다른 모든 행만 사용합니다. "예측 계산식 저장" 옵션은 모든 훈련 데이터를 사용하므로 각 훈련 데이터 행에 첫 번째 이웃에 대한 행 자체가 사용되고 $k=1$ 에 대한 완전 예측을 제공합니다. 즉, 훈련 데이터에 대한 예측이 너무 정확하여 모형 통계량을 계산하는 데 사용할 경우 모형 정확도 추정값이 부풀려질 수 있습니다. 따라서 플랫폼에 나타나는 보고서 통계량은 예측 계산식의 예측을 사용하여 생성되지 않습니다.

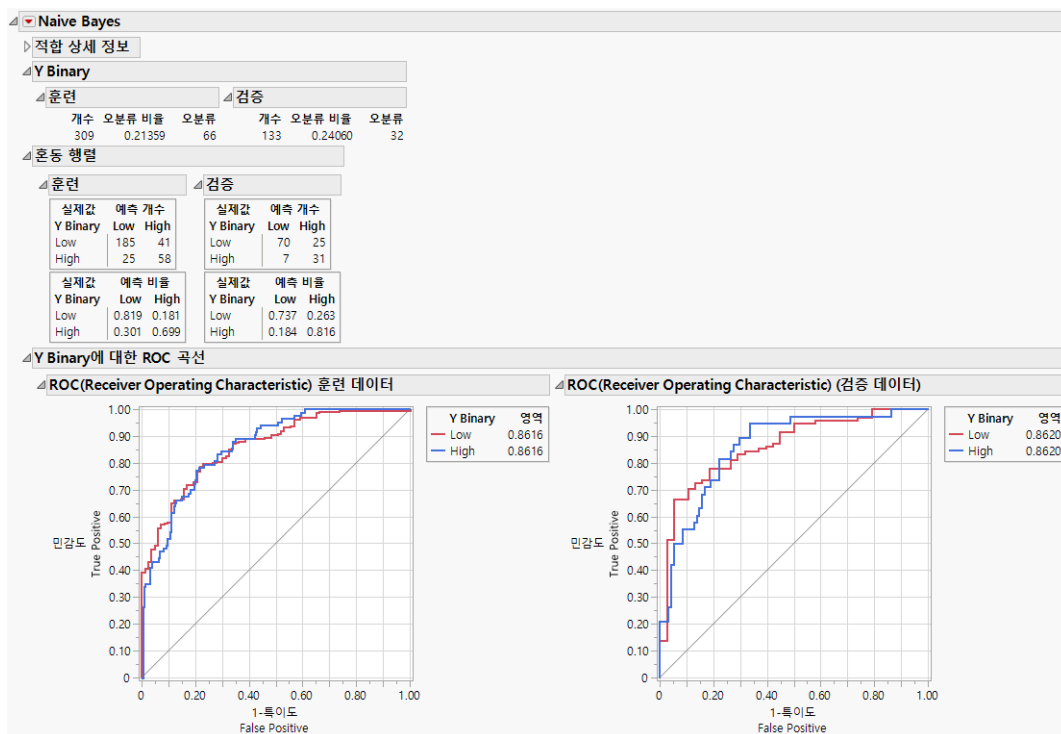
예측 계산식 게시 지정된 k 최근접 이웃 모형에 대한 예측 계산식을 생성하여 계산식 저장소 플랫폼에 계산식 열 스크립트로 저장합니다. "계산식 저장소" 보고서가 열려 있지 않으면 "계산식 저장소" 보고서가 생성됩니다. 자세한 내용은 "[계산식 저장소](#)"에서 확인하십시오.

확률적 가정을 사용하여 관측값 분류

Naive Bayes 플랫폼은 JMP Pro 에서만 사용할 수 있습니다.

Naive Bayes 플랫폼은 범주형 변수의 값을 예측하기 위해 모형을 적합시킵니다. Naive Bayes 는 빠르고 계산이 단순한 방법입니다. 특히 예측 변수의 수가 많은 경우에 알맞습니다.

그림 8.1 Naive Bayes 분석의 예



목차

Naive Bayes 플랫폼 개요	127
Naive Bayes 의 예	127
Naive Bayes 플랫폼 시작	130
Naive Bayes 보고서	131
적합 상세 정보 보고서	131
반응 열 보고서	132
혼동 행렬 보고서	132
ROC 곡선	132
Naive Bayes 플랫폼 옵션	132
Naive Bayes 의 추가 예	134
Naive Bayes 플랫폼에 대한 통계 상세 정보	135
Naive Bayes 알고리즘에 대한 통계 상세 정보	135
저장된 확률 계산식에 대한 통계 상세 정보	136

JMP PRO Naive Bayes 플랫폼 개요

Naive Bayes 플랫폼에서는 관측값을 범주형 반응 변수의 수준에 의해 정의된 클래스로 분류합니다. 분류에 사용되는 변수 또는 요인을 데이터 마이닝 문헌에서는 흔히 피처라고 합니다.

Naive Bayes 알고리즘은 각 클래스에 대해 각 피처 값이 발생하는 조건부 확률을 계산합니다. 피처가 연속형이면 조건부 주변 밀도가 추정됩니다. Naive Bayes 기법은 클래스 내에서 피처가 서로 독립적이라고 가정합니다. 이런 이유 때문에 이 기법을 "Naive"라고 합니다. 분류는 특정 클래스 내에서 피처 값의 조건부 확률이 높은 관측값이 해당 클래스에 속할 확률이 높다는 개념을 기반으로 합니다. 자세한 내용은 Hastie et al. (2009) 연구 자료에서 확인하십시오.

이 알고리즘은 1차원 밀도 또는 분포만 추정하므로 속도가 매우 빠릅니다. 따라서 큰 데이터 집합, 특히 피처 수가 많은 데이터 집합에 알맞습니다. 관측값에 대한 모든 비결측 피처 값은 조건부 확률을 계산하는 데 사용됩니다.

각 관측값에는 각 클래스에 대한 Naive 스코어가 할당됩니다. 지정된 클래스에 대한 관측값의 Naive 스코어는 해당 클래스에 속한 훈련 관측값의 비율에 관측값의 조건부 확률 곱을 곱한 값입니다. 관측값이 클래스에 속할 Naive 확률은 해당 클래스의 Naive 스코어를 모든 클래스의 Naive 스코어 합으로 나눈 값입니다. 관측값은 Naive 확률이 가장 높은 클래스에 할당됩니다.

주의: 클래스 소속의 조건부 확률은 독립적이라고 가정되므로 Naive Bayes 추정 확률은 비효율적입니다.

Naive Bayes를 사용하려면 모든 예측 변수 값과 클래스를 표현하기 위해 많은 수의 훈련 관측값이 필요합니다. 검증 데이터 집합의 관측값을 분류하는 중이고 해당 관측값에 훈련 데이터 집합에서 결측인 범주형 예측 변수 값이 있는 경우 플랫폼은 비결측 피처를 사용하여 예측합니다. 모든 예측 변수 값에 대해 관측값이 결측인 경우 예측 반응은 가장 빈도가 높은 반응입니다. 예측 계산식은 결측값을 관측값 스코어에 전혀 기여하지 않는 것으로 설정하여 결측값을 처리합니다.

Naive Bayes 기법에 대한 자세한 내용은 Hand et al. (2001) 및 Shmueli et al. (2010) 연구 자료에서 확인하십시오.

JMP PRO Naive Bayes의 예

미래 환자의 질병 경과를 High 또는 Low 로 예측하는 데 사용할 분류 모형을 생성하려고 합니다. 442 명의 당뇨병 환자에 대한 기준 의료 데이터가 있습니다. 또한 각 환자가 병원에 처음 내원한 후 1년 동안 얻은 당뇨병 진행 경과의 이진 측정값도 있습니다. 이 측정값은 질병 경과를 Low 또는 High 로 정량화합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Diabetes.jmp 를 엽니다.
2. **분석 > 예측 모델링 > Naive Bayes** 를 선택합니다.
3. Y Binary 를 선택하고 **Y, 반응**을 클릭합니다.
4. Age ~ Glucose 를 선택하고 **X, 요인**을 클릭합니다.

5. Validation 을 선택하고 **검증**을 클릭합니다 .
6. **확인**을 클릭합니다 .

그림 8.2 Naive Bayes 보고서

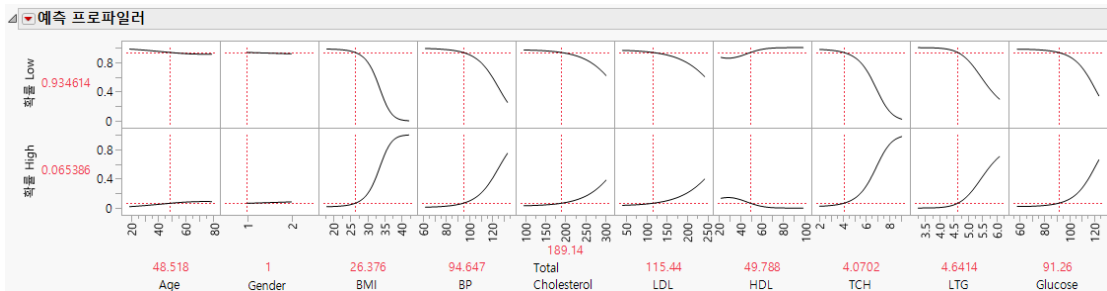
Naive Bayes						
적합 상세 정보						
Y Binary						
훈련			검증			
개수	오분류 비율	오분류	개수	오분류 비율	오분류	
309	0.21359	66	133	0.24060	32	
혼동 행렬						
훈련			검증			
실제값		예측 개수	실제값		예측 개수	
Y Binary		Low High	Y Binary		Low High	
Low		185 41	Low		70 25	
High		25 58	High		7 31	
실제값		예측 비율	실제값		예측 비율	
Y Binary		Low High	Y Binary		Low High	
Low		0.819 0.181	Low		0.737 0.263	
High		0.301 0.699	High		0.184 0.816	

훈련 데이터 집합의 오분류 비율은 약 21% 이고 검증 데이터 집합의 오분류 비율은 약 24% 입니다. " 혼동 행렬 " 을 보면 훈련 데이터 집합과 검증 데이터 집합 둘 다 질병 경과가 Low 인 환자를 High 로 분류한 것이 오분류의 큰 원인임을 알 수 있습니다. 검증 데이터 집합 결과는 모형이 독립적 관측값으로 확장하는 방식을 나타냅니다.

Naive Bayes 분류에 가장 큰 영향을 미치는 개별 예측 변수에 대해 알아보려고 합니다.

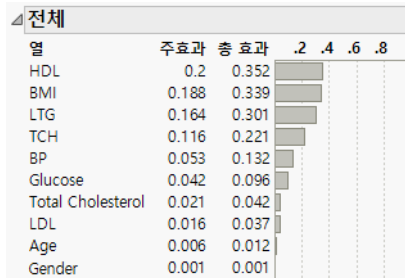
7. "Naive Bayes" 의 빨간색 삼각형을 클릭하고 **프로파일러**를 선택합니다 .

그림 8.3 질병 경과에 대한 예측 프로파일러



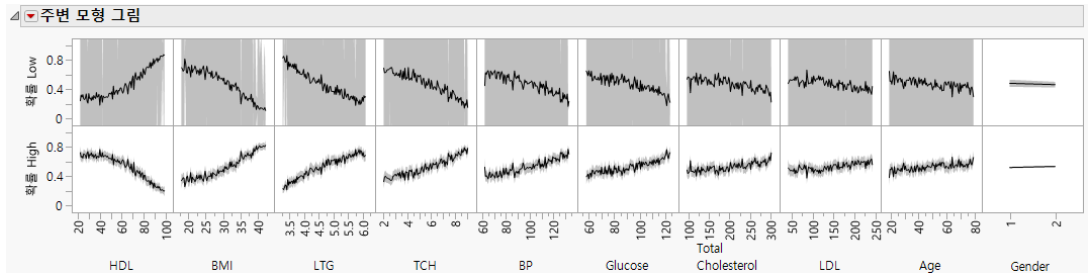
8. " 예측 프로파일러 " 의 빨간색 삼각형을 클릭하고 **변수 중요도 평가 > 독립 균등 입력**을 선택합니다 .

그림 8.4 변수 중요도



" 요약 보고서 " 에서는 HDL, BMI 및 LTG 가 추정 확률에 가장 큰 영향을 미친다는 것을 나타냅니다.

그림 8.5 주변 모형 그림 보고서



" 주변 모형 그림 " 보고서의 두 번째 행 그림에서는 HDL 값이 높을수록 환자를 High 로 분류할 확률이 낮음을 보여 줍니다. 또한 BMI 와 LTG 값이 높을수록 환자를 High 로 분류할 확률이 높습니다.

JMP^{PRO} Naive Bayes 플랫폼 시작

분석 > 예측 모델링 > Naive Bayes 를 선택하여 Naive Bayes 플랫폼을 시작합니다.

그림 8.6 Naive Bayes 시작 창

범주형 변수의 소속 그룹을 예측합니다.

The screenshot shows the Naive Bayes Start Window. On the left, under '열 선택' (Select Columns), there is a list of 14 columns: Y, Y Binary, Y Ordinal, Age, Gender, BMI, BP, Total Cholesterol, LDL, HDL, TCH, LTG, Glucose, and Validation. The '검증 데이터 비율' (Validation Data Ratio) is set to 0. In the center, under '선택한 열 역할 지정' (Assign Selected Column Roles), there are six rows with buttons: 'Y, 반응' (Y, Response) with a '필수' (Required) button, 'X, 요인' (X, Factor) with a '필수' (Required) button and a '선택적' (Optional) button, '가중치' (Weights) with a '선택적 숫자' (Optional Number) button, '빈도' (Frequencies) with a '선택적 숫자' (Optional Number) button, '검증' (Validation) with a '선택적 숫자' (Optional Number) button, and '기준' (Criteria) with a '선택적' (Optional) button. On the right, under '작업' (Actions), there are buttons for '확인' (OK), '취소' (Cancel), '제거' (Remove), '재호출' (Reopen), and '도움말' (Help).

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

Naive Bayes 시작 창에는 다음과 같은 옵션이 있습니다.

Y, 반응 관심 클래스가 값인 범주형 반응 열입니다.

X, 요인 범주형 또는 연속형 예측 변수 열입니다.

가중치 분석의 각 행에 가중치를 할당하는 숫자 값이 포함된 열입니다.

빈도 분석의 각 행에 빈도를 할당하는 숫자 값이 포함된 열입니다.

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

- 열의 고유 값이 세 개인 경우 다음과 같습니다.
 - 가장 작은 값이 있는 행이 훈련 데이터 집합에 사용됩니다.
 - 중간 값이 있는 행이 검증 데이터 집합에 사용됩니다.
 - 가장 큰 값이 있는 행이 테스트 데이터 집합에 사용됩니다.
- 열의 고유 값이 두 개인 경우 훈련 데이터 집합과 검증 데이터 집합만 사용됩니다.
- 검증 열에 네 개 이상의 수준이 있는 경우 가장 작은 세 개의 값을 포함하는 행이 검증 데이터 집합을 정의합니다. 다른 모든 행은 분석에서 제외됩니다.

형 반응에 대한 적합 측도"에서 확인하십시오. 기본적으로 "Naive Bayes" 보고서 창의 "적합 상세 정보" 보고서는 닫혀 있습니다.

JMP^{PRO} 반응 열 보고서

Naive Bayes 플랫폼에서 반응 열에 대한 보고서 이름은 반응 열 이름을 따라 지정됩니다. 이 보고서에는 훈련 데이터 집합과 검증 및 테스트 데이터 집합 (지정된 경우)에 대한 Naive Bayes 분류의 성능 통계를 보여주는 요약 테이블이 있습니다. 요약 테이블에는 다음 열이 포함됩니다.

개수 테이블에 해당하는 데이터 집합(훈련, 검증 또는 테스트 데이터 집합)의 관측값 수입니다.

오분류 비율 해당 데이터 집합에서 모형에 의해 잘못 분류된 관측값의 비율입니다. 이 값은 "오분류"를 "개수"로 나누어 계산됩니다.

오분류 해당 데이터 집합에서 잘못 분류된 관측값 수입니다.

JMP^{PRO} 혼동 행렬 보고서

Naive Bayes 플랫폼의 "혼동 행렬" 보고서에는 훈련 데이터 집합과 검증 및 테스트 데이터 집합 (지정된 경우)에 대한 혼동 행렬과 혼동 비율 행렬이 표시됩니다. 혼동 행렬은 실제 반응과 예측 반응의 이원 분류입니다. 혼동 비율 행렬은 값을 행 합계로 나눈 혼동 행렬과 동일합니다.

JMP^{PRO} ROC 곡선

이 보고서에는 훈련 데이터 집합과 검증 및 테스트 데이터 집합 (지정된 경우)에 대한 ROC(Receiver Operating Characteristic) 곡선이 표시됩니다. ROC 곡선은 반응 수준을 올바르게 분류하기 위한 적합 확률 능력을 측정합니다. 대각과 곡선의 거리가 멀수록 적합도가 높습니다. ROC 곡선에 대한 소개는 기본 분석에서 확인하십시오.

ROC 곡선에 대한 자세한 내용은 "[ROC 곡선](#)"에서 확인하십시오.

JMP^{PRO} Naive Bayes 플랫폼 옵션

"Naive Bayes"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

ROC 곡선 ROC 그림을 표시하거나 숨깁니다. ROC 그림은 기본적으로 표시됩니다. 자세한 내용은 "[ROC 곡선](#)"에서 확인하십시오.

향상도 곡선 모형에 대한 향상도 곡선을 표시하거나 숨깁니다. 검증을 사용한 경우 훈련 데이터 집합과 검증 및 테스트 데이터 집합 (지정된 경우)에 대해 각각 향상도 곡선이 표시됩니다.

향상도 곡선은 적합 확률이 감소할 때 반응 수준이 얼마나 효과적으로 분류되는지 보여줍니다. 적합 확률은 가로 축을 따라 내림차순으로 표시됩니다. 적합 확률의 수직 좌표는 해당 확률 이상에 대한 올바른 분류 비율을 전체 올바른 분류 비율로 나눈 값입니다. 적합 확률이 임

계값을 초과하는 관측값만 선택한 경우 항상도 곡선을 사용하여 대부분의 관측값을 올바르게 분류할 수 있는지 확인할 수 있습니다.

반응 수준이 두 개이면 항상도 곡선 그림에 반응의 첫 번째 수준에 대한 항상도 곡선만 표시됩니다. 반응 수준이 세 개 이상이면 항상도 곡선 그림에 각 반응 수준에 대한 곡선의 하위 개요가 표시됩니다. 항상도 곡선에 대한 자세한 내용은 "[항상도 곡선](#)"에서 확인하십시오.

예측값 저장 예측 분류를 데이터 테이블의 Naive 예측 <Y, 반응> 열에 저장합니다.

예측 계산식 저장 Naive 예측 계산식 <Y, 반응> 열을 데이터 테이블에 저장합니다. 이 열에는 분류에 대한 예측 계산식이 포함됩니다.

확률 계산식 저장 각 관측값을 분류하는 데 사용되는 계산식이 포함된 열을 데이터 테이블에 저장합니다. 다음과 같은 세 개의 열 그룹이 저장됩니다.

Naive 스코어 <클래스>, 합 클래스를 나타내는 각 열에 대해 이 열은 지정된 클래스의 소속 강도를 측정하는 스코어 계산식을 제공합니다. "Naive 스코어 합" 열에서 이러한 스코어는 전체 클래스에 대해 합산됩니다. 자세한 내용은 "[저장된 확률 계산식에 대한 통계 상세 정보](#)"에서 확인하십시오.

Naive 확률 <클래스> 각 클래스에 대해 이 열은 관측값이 해당 클래스에 속할 조건부 확률 계산식을 제공합니다. 자세한 내용은 "[저장된 확률 계산식에 대한 통계 상세 정보](#)"에서 확인하십시오.

Naive 예측 계산식 <Y, 반응> 예측 클래스에 대한 계산식을 제공합니다.

확률 계산식 게시 확률 계산식을 생성하여 계산식 저장소 플랫폼에 계산식 열 스크립트로 저장합니다. "계산식 저장소" 보고서가 열려 있지 않은 경우 이 옵션을 선택하면 "계산식 저장소" 창이 열립니다. 자세한 내용은 "[계산식 저장소](#)"에서 확인하십시오.

프로파일러 대화식 프로파일러 보고서를 표시하거나 숨깁니다. 요인 값을 변경하면 추정된 분류 확률에 반영됩니다. 자세한 내용은 **Profilers**에서 확인하십시오.

다음 옵션에 대한 자세한 내용은 **JMP 사용**의 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.

JMP^{PRO} Naive Bayes의 추가 예

Naive Bayes를 사용하여 대부분의 예측 변수에 걸쳐 데이터가 있는 데이터 집합으로 분류 모형을 생성합니다. 주택 담보 대출을 신청한 5,960 명의 고객에 대한 과거 재무 데이터가 있습니다. 각 고객은 Good Risk(안전한 위험) 또는 Bad Risk(불안한 위험)로 분류되었습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Equity.jmp를 엽니다.
2. **분석 > 예측 모델링 > Naive Bayes**를 선택합니다.
3. BAD를 선택하고 **Y, 반응**을 클릭합니다.

잠재적 예측 변수 중 하나인 DEBTINC에는 정보가 될 수 있는 결측값이 많이 있습니다. 그러나 Naive Bayes는 많은 수의 결측값을 잘 처리할 준비가 되어 있지 않으므로 모형에 DEBTINC를 포함하지 않습니다.

4. LOAN ~ CLNO를 선택하고 **X, 요인**을 클릭합니다.
5. Validation을 선택하고 **검증**을 클릭합니다.
6. **확인**을 클릭합니다.

그림 8.8 BAD에 대한 Naive Bayes 보고서

BAD											
훈련				검증				테스트			
개수	오분류 비율	오분류		개수	오분류 비율	오분류		개수	오분류 비율	오분류	
3576	0.18009	644		1192	0.19128	228		1192	0.18960	226	
혼동 행렬											
훈련				검증				테스트			
실제값		예측 개수		실제값		예측 개수		실제값		예측 개수	
BAD	Good Risk	Bad Risk		BAD	Good Risk	Bad Risk		BAD	Good Risk	Bad Risk	
Good Risk	2691	214		Good Risk	859	58		Good Risk	885	64	
Bad Risk	430	241		Bad Risk	170	105		Bad Risk	162	81	
실제값		예측 비율		실제값		예측 비율		실제값		예측 비율	
BAD	Good Risk	Bad Risk		BAD	Good Risk	Bad Risk		BAD	Good Risk	Bad Risk	
Good Risk	0.926	0.074		Good Risk	0.937	0.063		Good Risk	0.933	0.067	
Bad Risk	0.641	0.359		Bad Risk	0.618	0.382		Bad Risk	0.667	0.333	

훈련, 검증 및 테스트 데이터 집합의 오분류 비율은 18%에서 19% 사이입니다. 모든 데이터 집합의 혼동 행렬은 Bad Risk 고객을 Good Risk 고객으로 분류한 것이 오분류의 가장 큰 원인임을 나타냅니다.

특정 재무 배경 값을 가진 고객이 High Risk로 분류될 확률에 관심이 있습니다.

7. "Naive Bayes"의 빨간색 삼각형을 클릭하고 **확률 계산식 저장**을 선택합니다.

3개의 열 집합이 데이터 테이블에 추가됩니다.

- 세 개의 Naive 스코어 열에는 Good Risk, Bad Risk 및 이 둘의 합에 대한 Naive 스코어 계산식이 포함됩니다.
- 두 개의 Naive 확률 열에는 Good Risk와 Bad Risk에 대한 확률 계산식이 포함됩니다.

- Naive 예측 계산식 Bad 열에는 관측값의 Naive 확률이 가장 높은 클래스에 관측값을 할당하는 계산식이 포함됩니다.

이 계산식을 사용하여 새 고객을 스코어링할 수 있습니다. 계산식 열에 대한 자세한 내용은 "[저장된 확률 계산식에 대한 통계 상세 정보](#)"에서 확인하십시오.

JMP PRO Naive Bayes 플랫폼에 대한 통계 상세 정보

이 섹션에는 Naive Bayes 플랫폼에 대한 통계 상세 정보가 포함되어 있습니다.

- "[Naive Bayes 알고리즘에 대한 통계 상세 정보](#)"
- "[저장된 확률 계산식에 대한 통계 상세 정보](#)"

JMP PRO Naive Bayes 알고리즘에 대한 통계 상세 정보

Naive Bayes 방법은 피쳐 값이 주어졌을 때 관측값을 소속 확률이 가장 높은 클래스로 분류합니다. 이 방법에서는 피쳐가 각 클래스 내에서 조건부로 독립적이라고 가정합니다.

가능한 분류를 $C_1, \dots, C_K$ 로 나타냅니다. 피쳐 또는 예측 변수를 $X_1, X_2, \dots, X_p$ 로 나타냅니다.

예측 변수 값이 $x_1, x_2, \dots, x_p$ 인 관측값이 C_k 클래스에 속할 조건부 확률은 다음과 같이 계산됩니다.

$$P(C_k | (x_1, \dots, x_p)) = \frac{\left(P(C_k) \prod_{j=1}^p [P(x_j | C_k)] \right) / (\mathbf{R})}{\sum_{k=1}^K \left(P(C_k) \left(\prod_{j=1}^p [P(x_j | C_k)] \right) / \mathbf{R} \right)},$$

여기서 $\mathbf{R}$ 은 정규화 상수입니다. 위의 계산식에서 $X_j = x_j$ 인 관측값이 C_k , $P(x_j | C_k)$ 클래스에 속할 조건부 확률은 다음과 같이 결정됩니다.

- X_j 가 범주형인 경우

$$P(x_j | C_k) = \frac{\text{다음 요소의 관측값 수} - x_j | C_k + (0.5/K)}{\text{다음 요소의 관측값 수} - C_k + 0.5}$$

- X_j 가 연속형인 경우

$$P(x_j | C_k) = \frac{1}{s_{jk}} \phi((x_j - m_{jk})/s_{jk})$$

여기서 ϕ 는 표준 정규 밀도 함수이고, m 과 s 는 각각 C_k 클래스 내 예측 변수 값의 평균과 표준편차입니다.

관측값이 C_k 클래스에 속할 무조건부 확률은 다음과 같이 계산됩니다.

$$P(C_k) = \frac{\text{다음 요소의 관측값 수} - C_k + (0.5/K)}{\text{총 관측값 수} + 0.5}$$

관측값은 조건부 확률이 가장 큰 클래스로 분류됩니다.

참고 : $P(C_k)$ 계산식에서 0.5 는 사전 편향 요인입니다 . 이 값이 기본값입니다 . 사전 기본 요인을 변경하려면 파일 > 환경 설정 > 플랫폼 > Naive Bayes 로 이동한 후 " 사전 편향 " 체크박스를 선택하고 값을 변경하십시오 .

JMP PRO 저장된 확률 계산식에 대한 통계 상세 정보

이 섹션에서는 "Naive Bayes" 의 빨간색 삼각형 메뉴에 있는 "확률 계산식 저장" 옵션을 사용하여 저장된 계산식에 대해 설명합니다. 예측 변수 값이 $x_1, x_2, \dots, x_p$ 인 관측값이 C_k 클래스에 속할 조건부 확률은 "Naive Bayes 알고리즘에 대한 통계 상세 정보" 섹션에 표시된 $P(C_k | (x_1, \dots, x_p))$ 에서 제공하는 확률과 약간 다릅니다. 이는 계산 효율성을 위해서입니다.

Naive 스코어 계산식

지정된 클래스 C_k 에 대한 Naive 스코어 계산식 $S(C_k)$ 는 $P(C_k | (x_1, \dots, x_p))$ 의 표현식에서 분자가 변형된 것으로, 다음과 같이 계산됩니다.

$$S(C_k) = \exp[\ln\{P(C_k)\} + \text{연속형} + \text{범주형} + \ln(\mathbf{R})]$$

Naive 스코어 계산식은 연속형 예측 변수와 범주형 예측 변수의 스코어 조합입니다. $\mathbf{R}$ 은 정규화 상수입니다. 계산식의 연속형 부분은 다음과 같이 계산됩니다.

$$\text{연속형} = \sum_{j=1}^{p_1} \left\{ \text{Normal Log Density} \left(\frac{x_j - m_{jk}}{s_{jk}} \right) + \ln(1/s_{jk}) \right\}$$

다음은 각 요소에 대한 설명입니다.

$j = 1, \dots, p_1$ 연속형 예측 변수

계산식의 범주형 부분은 다음과 같이 계산됩니다.

$$\text{범주형} = \sum_{r=1}^{p_2} \left\{ \sum_{l=1}^{L_r} 1_{rl} \ln(P(x_{rl} | C_k)) \right\}$$

다음은 각 요소에 대한 설명입니다.

$r = 1, \dots, p_2$ 범주형 변수

$l = r$ 번째 범주형 변수의 $1, \dots, L_r$ 수준

$1_{rl} = x_{rl}$ 이 r 번째 범주형 예측 변수의 l 번째 수준인 경우 1 이고, 그렇지 않은 경우 0 인 표시 변수

Naive 스코어 합 계산식

Naive 스코어 합 계산식 S 는 모든 클래스에 대해 Naive 스코어 계산식을 합한 것입니다. 이 계산식은 $P(C_k | (x_1, \dots, x_p))$ 의 표현식에서 분모가 변형되었습니다.

$$S = \sum_{k=1}^K S(C_k)$$

Naive 확률 계산식

지정된 클래스 C_k 의 Naive 확률 계산식은 $P(C_k | (x_1, \dots, x_p))$ 입니다. JMP 계산식에서는 다음과 같습니다.

$$P(C_k | (x_1, \dots, x_p)) = \frac{S(C_k)}{S}$$

Naive 예측 계산식

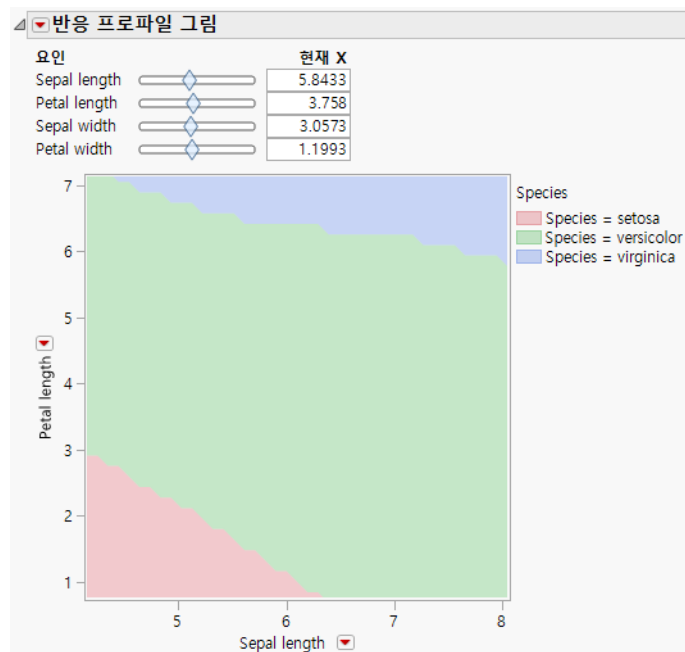
관측값의 Naive 예측 계산식은 해당 관측값을 $P(C_k | (x_1, \dots, x_p))$ 가 가장 큰 클래스로 분류합니다. 이것은 관측값을 Naive 스코어 계산식이 가장 큰 클래스 분류하는 것과 동일합니다.

서포트 벡터 머신 플랫폼은 JMP Pro 에서만 사용할 수 있습니다.

SVM(서포트 벡터 머신) 플랫폼에서는 비선형 회귀와 분류를 위한 머신 학습 도구를 제공합니다. SVM 알고리즘은 훈련 데이터를 사용하여 새 관측값을 예측하거나 분류하는 모형을 생성하는 지도 학습 알고리즘입니다.

범주형 반응을 사용하는 경우 반응 프로파일 그림에서 모형 분류를 시각화할 수 있습니다. 또한 SVM 플랫폼에서는 모형이 데이터를 얼마나 잘 적합시키는지 확인할 수 있도록 오분류 비율, 혼동 행렬, ROC 곡선 및 향상도 곡선을 제공합니다. 연속형 반응을 사용하는 경우 모형이 데이터를 얼마나 잘 적합시키는지 확인하기 위해 적합 통계량 및 실제값 대 예측값 그림을 사용합니다.

그림 9.1 반응 프로파일 그림



목차

서포트 벡터 머신 개요.....	141
서포트 벡터 머신의 예.....	141
서포트 벡터 머신 플랫폼 시작.....	144
모형 시작 제어판.....	144
서포트 벡터 머신 보고서.....	147
모형 비교 보고서.....	147
서포트 벡터 머신 모형 보고서.....	148
서포트 벡터 머신 플랫폼 옵션.....	150
서포트 벡터 머신 모형 보고서 옵션.....	151
서포트 벡터 머신의 추가 예.....	152

JMP PRO 서포트 벡터 머신 개요

SVM(서포트 벡터 머신) 모형은 새 관측값을 예측하거나 분류하는 데 사용되는 지도 학습 알고리즘입니다. 모형은 반응이 알려진 훈련 데이터 집합에 대해 적합됩니다. 그런 다음 모형을 사용하여 새 관측값의 반응을 예측합니다.

반응이 범주형인 경우 SVM 모형은 클래스를 구분하는 초평면을 최적화하여 데이터를 분류합니다. 이것은 클래스 간의 마진을 최대화하는 초평면을 찾는 것으로 볼 수도 있습니다. 단순한 문제에서 이 초평면은 선형입니다. 그러나 더 복잡한 데이터는 선형으로 구분할 수 없는 경우가 많습니다. 이러한 경우 SVM 플랫폼에서는 방사 기저 함수 커널을 사용하는 옵션을 제공하여 클래스를 더 쉽게 분리할 수 있는 비선형 차원에 점을 매핑합니다.

반응이 연속형인 경우 적합한 모형을 SVR(서포트 벡터 회귀) 모형이라고 합니다. 일반적인 회귀 분석 문제의 목표는 예측 반응과 실제 반응 사이의 오차를 최소화하는 모형을 적합시키는 것입니다. SVR 문제의 목표는 예측 반응과 실제 반응 사이의 오차가 $-\epsilon \sim \epsilon$ 범위 내에 속하도록 모형을 적합시키는 것입니다. 이렇게 하면 더 유연한 적합을 얻게 됩니다. JMP Pro에서 ϵ 는 0.1과 같습니다. SVR 알고리즘은 $Y + \epsilon$ 와 $Y - \epsilon$ 라는 두 클래스를 생성하여 데이터를 두 배로 늘립니다. 그런 다음 분류 문제에 사용된 동일한 알고리즘이 예측(SVR) 문제에도 사용됩니다.

SVM 알고리즘의 최대화는 2차 계획법 문제의 해를 구하여 수행됩니다. JMP Pro에서 SVM 플랫폼에 사용되는 알고리즘은 1998년 John Platt이 소개한 SMO(Sequential Minimal Optimization) 알고리즘을 기반으로 합니다. 일반적으로 SVM 2차 계획법 문제는 상당히 큰 규모입니다. SMO 알고리즘은 전체 2차 계획법 문제를 일련의 더 작은 2차 계획법 문제로 분할합니다. 더 작은 2차 계획법 문제는 수치적 계산 대신 분석적으로 해를 구합니다. 즉, 닫힌 형식의 해가 생성됩니다. 따라서 SMO 알고리즘은 전체 2차 계획법 문제의 해를 구하는 것보다 더 효율적입니다(Platt 1998).

JMP PRO 서포트 벡터 머신의 예

이 예에서는 "설계 조정" 옵션을 사용하여 커널 함수 및 모수 값이 서로 다른 여러 서포트 벡터 머신 모형을 적합시킵니다. 당뇨병 환자의 질병 경과를 예측하기 위한 최적 분류 모형을 찾으려고 합니다. 각 환자가 병원에 처음 내원한 후 1년 동안 얻은 당뇨병 진행 경과의 이진 측정값 및 기준 의료 데이터가 있습니다. 이 측정값은 질병 경과를 Low 또는 High로 정량화합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Diabetes.jmp를 엽니다.
2. **분석 > 예측 모델링 > 서포트 벡터 머신**을 선택합니다.
3. Y Binary를 선택하고 **Y, 반응**을 클릭합니다.
4. Age ~ Glucose를 선택하고 **X, 요인**을 클릭합니다.
5. Validation을 선택하고 **검증**을 클릭합니다.
6. **확인**을 클릭합니다.

7. "모형 시작" 제어판에서 커널 함수가 "방사 기저 함수" 인지 확인하고 "설계 조정"을 선택합니다.
8. "런 수" 옆에 "10"을 입력합니다.
9. 시작을 클릭합니다.
10. "모형 시작" 옆의 회색 삼각형을 클릭하여 "모형 시작" 제어판을 엽니다.
11. 커널 함수를 "선형" 함수로 변경하고 "설계 조정"을 선택합니다.
12. "런 수" 옆에 "10"을 입력합니다.
13. 시작을 클릭합니다.

그림 9.2 모형 비교 보고서

모형 비교

표시	방법	커널 함수	비용	감마	SV 수	훈련 오분류 비율	검증 오분류 비율	검증 일반화 R ²	확률 임계	검증 오분류 비율 임계	최선
☐	모형 1	방사 기저 함수	2.57061	0.43877	251	0.00324	0.25564	0.00838	0.39079	0.25564	
☐	모형 2	방사 기저 함수	3.04271	0.31899	218	0.01294	0.28571	-0.4206	0.23223	0.28571	
☐	모형 3	방사 기저 함수	3.93906	0.46899	255	0.00324	0.28571	-0.0081	0.53372	0.28571	
☐	모형 4	방사 기저 함수	1.073	0.27968	208	0.0712	0.21805	0.22158	0.17372	0.21805	
☐	모형 5	방사 기저 함수	0.41677	0.0281	156	0.15534	0.18797	0.50485	0.36015	0.18797	
☐	모형 6	방사 기저 함수	0.23369	0.39898	246	0.26861	0.28571	0.22323	6e-8	0.28571	
☐	모형 7	방사 기저 함수	4.8013	0.15144	167	0.05178	0.2782	-0.4245	0.26219	0.2782	
☐	모형 8	방사 기저 함수	3.51202	0.20295	181	0.04207	0.28571	-0.454	0.22753	0.28571	
☐	모형 9	방사 기저 함수	4.2693	0.06093	140	0.11974	0.19549	0.42841	0.32812	0.19549	
☐	모형 10	방사 기저 함수	1.87641	0.0919	150	0.13269	0.21053	0.43499	0.30598	0.21053	
☐	모형 11	선형	4.42735	.	122	0.17476	0.17293	0.51944	0.47722	0.17293	
☐	모형 12	선형	4.99046	.	123	0.16828	0.17293	0.52037	0.47551	0.17293	
☐	모형 13	선형	3.8272	.	123	0.17476	0.17293	0.51925	0.47763	0.17293	
☐	모형 14	선형	2.74788	.	124	0.17152	0.17293	0.51723	0.48019	0.17293	
☐	모형 15	선형	3.22207	.	123	0.17476	0.17293	0.5185	0.47842	0.17293	
☐	모형 16	선형	1.29304	.	124	0.17152	0.17293	0.51523	0.4819	0.17293	
☐	모형 17	선형	1.74671	.	124	0.16828	0.17293	0.51556	0.48143	0.17293	
☐	모형 18	선형	0.83575	.	123	0.16505	0.15789	0.51498	0.48207	0.15789	
☐	모형 19	선형	2.22	.	125	0.17152	0.17293	0.5158	0.48255	0.17293	
☒	모형 20	선형	0.04975	.	138	0.15534	0.14286	0.52302	0.41946	0.14286	최소 오분류 비율 최대 R ²

"모형 비교" 보고서에서는 오분류 비율과 R^2 을 기준으로 한 결과 모형 20이 최적 모형임을 보여 줍니다. 이 모형에는 비용 모수가 0.04975인 선형 커널 함수가 있습니다. 이 모형으로 추가 분석을 수행합니다.

JMP PRO 서포트 벡터 머신 플랫폼 시작

분석 > 예측 모델링 > 서포트 벡터 머신을 선택하여 SVM(서포트 벡터 머신) 플랫폼을 시작합니다.

그림 9.4 서포트 벡터 머신 시작 창

X 변수의 공간에서 서포트 벡터를 기반으로 반응을 예측합니다.

열 선택	선택한 열 역할 지정	작업												
14개 열 - Y - Y Binary - Y Ordinal - Age - Gender - BMI - BP - Total Cholesterol - LDL - HDL - TCH - LTG - Glucose - Validation	<table border="1"> <tr> <td>Y, 반응</td> <td>필수</td> </tr> <tr> <td>X, 요인</td> <td>필수</td> </tr> <tr> <td></td> <td>선택적</td> </tr> <tr> <td>빈도</td> <td>선택적 숫자</td> </tr> <tr> <td>검증</td> <td>선택적 숫자</td> </tr> <tr> <td>기준</td> <td>선택적</td> </tr> </table>	Y, 반응	필수	X, 요인	필수		선택적	빈도	선택적 숫자	검증	선택적 숫자	기준	선택적	확인 취소 제거 재로출 도움말
Y, 반응	필수													
X, 요인	필수													
	선택적													
빈도	선택적 숫자													
검증	선택적 숫자													
기준	선택적													

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

서포트 벡터 머신 시작 창에는 다음과 같은 옵션이 있습니다.

Y, 반응 분석할 반응 변수입니다.

X, 요인 예측 변수입니다.

빈도 분석의 각 행에 빈도를 할당하는 숫자 값이 포함된 열입니다.

검증 검증 데이터 집합을 정의하는 숫자 열입니다. 자세한 내용은 "**검증 방법**"에서 확인하십시오. "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 "**검증 열 생성**"에서 확인하십시오.

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.

JMP PRO 모형 시작 제어판

서포트 벡터 머신 플랫폼을 시작하면 모형 적합을 위한 "모형 시작" 제어판이 나타납니다. "모형 시작" 제어판을 사용하여 커널 함수 및 관련 모수 값과 검증 방법을 지정할 수 있습니다.

그림 9.5 모형 시작 제어판

"모형 시작" 제어판에는 다음 옵션이 포함되어 있습니다.

커널 함수 모형에 사용되는 커널 함수를 지정합니다. 다음 커널 함수 중 하나를 선택할 수 있습니다.

방사 기저 함수 클래스를 구분하는 비선형 초평면을 생성하려면 방사 기저 함수 커널을 선택합니다.

- "비용" 모수는 훈련 데이터 집합의 관측값을 잘못 분류하는 것과 관련된 벌점입니다. 비용 모수가 높을수록 훈련 데이터 집합의 점을 잘못 분류할 가능성이 낮은 알고리즘이 구현되는 반면, 비용 모수가 낮을수록 마진이 더 넓어집니다. "비용" 모수는 0 보다 커야 하며 기본값은 1 입니다.
- "감마" 모수는 커널 함수의 모수입니다. 이 모수는 결정선에 대한 곡률의 정도를 결정합니다. 감마 값이 높을수록 곡률이 크다는 것을 나타냅니다. 비선형 결정선은 더 유연한 적합을 제공하지만 곡률이 너무 크면 과대적합이 발생할 수 있습니다. "감마" 모수는 0 보다 커야 하며 기본값은 $1/(\text{예측 변수의 수})$ 입니다.

선형 클래스를 구분하는 선형 초평면을 생성하려면 선형 커널 함수를 선택합니다.

- "비용" 모수는 훈련 데이터 집합의 관측값을 잘못 분류하는 것과 관련된 벌점입니다. 비용 모수가 높을수록 훈련 데이터 집합의 점을 잘못 분류할 가능성이 낮은 알고리즘이 구현되는 반면, 비용 모수가 낮을수록 마진이 더 넓어집니다. "비용" 모수는 0 보다 커야 하며 기본값은 1 입니다.

참고 : 범위를 벗어난 모수 값을 지정하면 기본값이 사용됩니다.

팁 : 최량 적합 모형을 찾으려면 다양한 커널 함수 및 모수 값을 적합시키고 "모형 비교" 보고서 사용하십시오.

설계 조정 지정된 커널에 대한 모수 값 범위를 적합시킬 수 있습니다. R^2 이 가장 크고 오분류 비율 또는 RASE 가 가장 작은 모형이 "모형 비교" 보고서에서 식별됩니다. "설계 조정"을 선택한 후에는 모수의 최소값과 최대값을 지정해야 합니다. 데이터를 기반으로 한 기본값이 제공되며 최소값은 0 보다 커야 합니다. "런 수"의 값도 지정해야 합니다. SVM 플랫폼은 최소값과 최대값에 의해 결정된 모수 값 격자에서 많은 모형을 적합시킵니다.

검증 방법 모형 검증 방법을 지정합니다. "시작" 버튼을 처음 클릭하면 지정된 검증 방법을 사용하여 첫 번째 SVM 모형이 적합됩니다. 그런 다음 SVM 창 내에서 적합된 모든 SVM 모형에 이 검증 방법이 사용됩니다. 이렇게 하면 보고서 창의 모든 모형이 동일한 검증 방법 및 검증 데이터 집합을 사용하여 적합됩니다.

홀드백 원래 데이터를 무작위로 훈련 데이터 집합과 검증 데이터 집합으로 나눕니다. 원래 데이터에서 검증 데이터 집합으로 사용할 비율을 지정할 수 있습니다 (홀드백).

K 폴드 (Y가 연속형 또는 명목형인 경우에만 사용 가능) 원래 데이터를 무작위로 K 개의 부분집합으로 나눕니다. 그러면 K 개의 각 집합이 나머지 데이터에 대한 모형 적합을 검증하는 데 사용되며 총 K 개 모형을 적합시킵니다. Y가 연속형이면 최량 검증 RASE 통계량을 제공하는 모형이 최종 모형으로 선택됩니다. Y가 명목형이면 최량 검증 오분류 비율을 제공하는 모형이 최종 모형으로 선택됩니다.

검증 열 (시작 창에서 "검증" 열을 지정한 경우에만 사용 가능) 지정된 "검증" 열의 값을 사용하여 데이터를 여러 부분으로 나눕니다. 열의 값은 데이터가 분할되는 방법과 검증에 사용되는 방법을 결정합니다.

- 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 "검증 열 K 폴드"가 사용됩니다.

SVM 플랫폼에서는 "설계 조정"이 사용되는 경우 외에는 검증 열을 사용하여 모형을 훈련하고 평가합니다. "설계 조정" 옵션을 선택하는 경우 SVM 플랫폼에서는 검증 열을 사용하여 모형을 훈련 및 조정하거나 모형을 훈련, 조정 및 평가합니다. 검증에 대한 자세한 내용은 ["JMP 모델링의 검증"](#)에서 확인하십시오.

참고: 검증 열을 사용하여 데이터가 올바르게 분할되지 않으면 홀드백 검증 방법이 대신 사용됩니다.

검증 열 K 폴드 ("Y, 반응" 열에 정확히 두 개의 수준이 있고 시작 창에서 "검증" 열을 지정한 경우에만 사용 가능) 지정된 "검증" 열의 값을 사용하여 데이터를 K 개의 집합으로 나눕니다. 여기서 K는 열의 고유 값 수입니다. 그런 다음 K 폴드 검증이 수행됩니다.

없음 검증이 사용되지 않습니다.

시작 지정된 SVM 모형을 적합시키고 모형 보고서를 표시합니다.

참고 : 데이터 테이블이 큰 경우 데이터에 적합되는 각 모형에 대해 진행률 표시줄이 나타납니다. 총 적합 모형 수는 $k!/2(k-2)!$ 이며, 여기서 k 는 반응 변수의 수준 수입니다. 각 진행률 표시줄에는 **현재 추정값 채택** 버튼이 있습니다. 적합 알고리즘을 조기에 중지하고 현재 추정값을 채택하려면 이 버튼을 클릭합니다. 이 버튼을 클릭한 후 예측 계산이 수행되므로 보고서가 나타나는 데 시간이 걸릴 수 있습니다.

결측값

예측 변수에 대해 결측값이 있는 행은 SVM 모델링 절차에 포함되지 않습니다. 따라서 데이터 테이블에 저장된 모든 열은 해당 행에 결측값을 포함합니다. 결측값이 있는 데이터를 SVM 모형에 포함하려면 전처리 작업이 필요합니다. 자세한 내용은 "[결측값 탐색](#)"에서 확인하십시오.

JMP^{PRO} 서포트 벡터 머신 보고서

"모형 시작" 제어판에서 **시작**을 클릭하면 제어판 위에 "모형 비교" 보고서가 나타나고 제어판 아래에 모형 보고서가 나타납니다. 모형 보고서에는 순차적으로 번호가 지정됩니다. 새 모형이 적합될 때마다 "모형 비교" 보고서에 추가되고 새 모형 보고서가 창에 추가됩니다. 자세한 내용은 "[모형 비교 보고서](#)" 및 "[서포트 벡터 머신 모형 보고서](#)"에서 확인하십시오. "모형 비교" 보고서의 열을 기준으로 모형을 정렬하려면 원하는 열 머리글을 클릭합니다. 처음 클릭하면 오름차순으로 정렬되고 해당 열 머리글을 두 번째 클릭하면 내림차순으로 정렬됩니다.

JMP^{PRO} 모형 비교 보고서

서포트 벡터 머신 플랫폼의 "모형 비교" 보고서에는 다음 정보가 포함되어 있습니다.

표시 해당 적합에 대한 "서포트 벡터 머신 모형" 보고서를 보고서 창에 표시할지 여부를 나타내는 체크박스입니다.

방법 모형 번호입니다.

커널 함수 해당 모형의 커널 함수입니다.

비용 해당 모형의 비용 모수 값입니다.

감마 해당 모형의 감마 모수 값입니다. 모형에 선형 커널이 사용되는 경우 이 값은 결측입니다.

SV 수 해당 모형에 사용된 서포트 벡터 수입니다.

훈련 오분류 비율 (반응이 범주형인 경우에만 표시됨) 훈련 데이터 집합의 관측값에 대한 오분류 비율입니다. 이 비율은 플랫폼에서 계산되는 분류 결정 규칙을 기반으로 합니다.

검증 오분류 비율 (반응이 범주형이고 검증을 사용하는 경우에만 표시됨) 검증 데이터 집합의 관측값에 대한 오분류 비율입니다. 이 비율은 플랫폼에서 계산되는 분류 결정 규칙을 기반으로 합니다.

테스트 오분류 비율 (반응이 범주형이고 검증 열을 사용하여 테스트 데이터 집합을 지정한 경우에만 표시됨) 테스트 데이터 집합의 관측값에 대한 오분류 비율입니다. 이 비율은 플랫폼에서 계산되는 분류 결정 규칙을 기반으로 합니다.

훈련 RASE (반응이 연속형인 경우에만 표시됨) 훈련 데이터 집합의 관측값에 대한 평균 제공 예측 오차의 제곱근 (제곱근 평균 제공 오차)입니다. 자세한 내용은 "[RASE](#)"에서 확인하십시오.

검증 RASE (반응이 연속형이고 검증을 사용하는 경우에만 표시됨) 검증 데이터 집합의 관측값에 대한 평균 제공 예측 오차의 제곱근 (제곱근 평균 제공 오차)입니다. 자세한 내용은 "[RASE](#)"에서 확인하십시오.

테스트 RASE (반응이 연속형이고 검증 열을 사용하여 테스트 데이터 집합을 지정한 경우에만 표시됨) 테스트 데이터 집합의 관측값에 대한 평균 제공 예측 오차의 제곱근 (제곱근 평균 제공 오차)입니다. 자세한 내용은 "[RASE](#)"에서 확인하십시오.

검증 일반화 R^2 (검증을 사용하는 경우에만 표시됨) 검증 데이터 집합의 관측값에 대한 일반화 R^2 값입니다. 자세한 내용은 "[적합 상세 정보](#)"에서 확인하십시오.

확률 임계 (반응이 이항이고 검증을 사용하는 경우에만 표시됨) 플랫폼에서 계산되는 분류 결정 규칙에 의해 결정된 경계 확률입니다. 관측값은 예측 확률이 이 값을 초과할 경우 목표 수준으로 분류됩니다.

팁: "혼동 행렬" 보고서의 "확률 임계 설정" 옵션을 사용하여 이 값을 변경할 수 있습니다. 자세한 내용은 "[혼동 행렬](#)"에서 확인하십시오.

조건부 검증 오분류 비율 (반응이 이항이고 검증을 사용하는 경우에만 표시됨) 확률 임계값을 조건으로 하는 검증 데이터 집합의 관측값에 대한 오분류 비율입니다.

최적 오분류 비율이 가장 작은 모형 적합을 나타냅니다. 테스트 데이터 집합 없이 검증을 사용하는 경우 검증 오분류 비율이 가장 작은 모형입니다. 테스트 데이터 집합과 함께 검증을 사용할 경우에는 테스트 오분류 비율이 가장 작은 모형입니다.

모형 비교 그림

전체 모수 값 격자에서 검증 데이터 집합의 모형 성능을 나타내는 등고선 그림을 표시합니다. 이 보고서는 "모형 시작" 제어판에서 "설계 조정"을 지정한 경우에만 사용할 수 있습니다.

JMP PRO 서포트 벡터 머신 모형 보고서

적합된 각 모형은 다음 섹션을 포함하는 모형 보고서를 생성합니다.

반응 프로파일 그림 (반응이 범주형인 경우에만 사용 가능) 분류 모형의 시각적 표현을 제공합니다. 그림의 점은 실제 데이터 관측값이며 정확히 두 개의 연속형 변수가 있는 경우에만 기본적으로 표시됩니다. 표시된 두 변수의 경우 음영 등고선은 나머지 모형 요인의 고정 값에 의해 결정된 예측 공간의 평면을 나타냅니다. 예측은 플랫폼에서 계산되는 분류 결정 규칙을 기반으로 합니다. 고정 값에 대한 컨트롤은 그림 위에 있습니다. 슬라이더 또는 숫자 상자를 사용하여 요인의 고정 값을 변경하면 표시된 변수의 예측 공간이 자동으로 업데이트

됩니다. 각 축의 빨간색 삼각형을 사용하면 그림의 축을 변경하여 연속형 요인을 표시할 수도 있습니다.

"반응 프로파일 그림"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

격자 밀도 음영 등고선의 기반이 되는 예측 격자의 정밀도를 결정합니다. 격자 밀도가 높을수록 더 매끄러운 결정선이 표시됩니다.

점 표시 그림에 점을 표시하거나 숨깁니다. 변수가 두 개뿐인 경우 기본적으로 선택되어 있습니다.

모형 요약 반응 열 이름, 검증 방법 및 모형 적합에 사용된 커널 함수 유형을 제공합니다. "모형 요약" 테이블에는 훈련, 검증 및 테스트 데이터 집합에 대한 모형 적합 정보도 포함되어 있습니다. 각 집합에 대해 관측값 수와 서포트 벡터 수가 보고됩니다. 반응이 범주형이면 각 집합에 대해 오분류 비율이 보고됩니다. 오분류 비율은 모형에 의해 잘못 분류된 관측값의 비율입니다. 이 값은 오분류 수를 총 관측값 수로 나누어 계산됩니다. 반응이 연속형이면 각 집합에 대해 RASE 및 R^2 값이 보고됩니다.

추정 상세 정보 모형에 사용된 모수의 값을 제공합니다.

적합 상세 정보 (반응이 범주형인 경우에만 사용 가능) 훈련 데이터 집합과 검증 및 테스트 데이터 집합(지정된 경우)에 대한 다음 통계량을 제공합니다.

엔트로피 R^2 적합 모형과 상수 확률 모형의 로그 가능도를 비교하는 적합 측도입니다. 엔트로피 R^2 범위는 0~1이며, 값이 1에 가까울수록 더 나은 적합을 나타냅니다. 자세한 내용은 "[엔트로피 \$R^2\$](#) "에서 확인하십시오.

일반화 R^2 일반 회귀 모형에 적용할 수 있는 측도입니다. 이 값은 가능도 함수 L 을 기반으로 하며 최대값이 1이 되도록 척도화됩니다. 일반화 R^2 측도는 표준 최소 제곱 설정 시 연속형 정규 반응에 대한 기존 R^2 로 단순화됩니다. 일반화 R^2 을 Nagelkerke 또는 Craig 와 Uhler R^2 이라고도 하는데, 이는 Cox-Snell 유사 R^2 을 정규화한 것입니다. 자세한 내용은 Nagelkerke 연구 자료 (1991)에서 확인하십시오. 값이 1에 가까울수록 더 나은 적합을 나타냅니다.

평균 -Log p $-\log(p)$ 의 평균입니다. 여기서 p 는 발생한 사건과 연관된 적합 확률입니다. 값이 작을수록 더 나은 적합을 나타냅니다.

RASE 평균 제곱 예측 오차의 제곱근입니다(제곱근 평균 제곱 오차). RASE는 다음과 같이 계산되며, 여기서 **Source**는 훈련, 검증 또는 테스트 데이터 집합을 나타냅니다.

$$RASE_{Source} = \sqrt{\frac{SSE_{Source}}{n}}$$

평균 절대 편차 반응과 예측 반응 사이의 차이에 대한 절대값의 평균입니다. 차이는 1에서 p (실제로 발생한 반응 수준의 적합 확률)사이입니다. 값이 작을수록 더 나은 적합을 나타냅니다.

오분류 비율 적합 확률이 가장 높은 반응 범주가 관측된 범주가 아닌 비율입니다.

참고 : "적합 상세 정보" 보고서의 오분류 비율이 "혼동 행렬" 보고서의 오분류 비율과 일치하지 않을 수 있습니다. 이항 반응의 경우 "적합 상세 정보" 보고서의 비율은 확률 경계 0.5 를 사용하지만 "혼동 행렬" 보고서의 비율은 확률 임계값을 경계로 사용합니다.

N 관측값 수입니다.

혼동 행렬 (반응이 범주형인 경우에만 사용 가능) 모형에 지정된 각 데이터 집합 (훈련, 검증 및 테스트)에 대한 혼동 행렬이 표시됩니다. 혼동 행렬은 실제 반응과 예측 반응의 이원 분류입니다. 혼동 행렬과 오분류 비율을 사용하여 모형을 평가할 수 있습니다.

혼동 행렬과 오분류 비율은 확률 임계 상자의 값을 경계 값으로 사용합니다. 기본적으로 이 값은 플랫폼에서 계산되는 분류 결정 규칙을 기반으로 합니다. 슬라이더를 드래그하거나 "확률 임계" 옆의 상자에 새 값을 입력하여 경계 값을 변경할 수 있습니다. 확률 임계를 변경하면 혼동 행렬과 오분류 비율이 자동으로 업데이트됩니다. "모형 비교" 보고서의 확률 임계 및 조건부 검증 오분류 비율 열도 업데이트됩니다.

실제값 대 예측값 그림 (반응이 연속형인 경우에만 사용 가능) 실제 반응 대 예측 반응 그림을 표시합니다. 대각선은 $Y = X$ 선입니다. 점이 이 선에 가까울수록 모형이 데이터를 더 잘 적합시킵니다. 검증이 사용되면 훈련 및 검증 데이터 집합과 테스트 데이터 집합 (지정된 경우)에 대해 그림이 표시됩니다.

JMP[®] PRO 서포트 벡터 머신 플랫폼 옵션

"서포트 벡터 머신"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

다음 옵션에 대한 자세한 내용은 JMP 사용의 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.

JMP^{PRO} 서포트 벡터 머신 모형 보고서 옵션

각 "서포트 벡터 머신 모형" 보고서에는 다음 옵션이 포함되어 있습니다.

반응 프로파일 그림 (반응이 범주형인 경우에만 사용 가능) 반응 프로파일 그림을 표시하거나 숨깁니다.

서포트 벡터 계수 서포트 벡터 계수 테이블을 표시하거나 숨깁니다. 이 테이블에서 행을 선택하면 원래 데이터 테이블과 반응 프로파일 그림에서 서포트 벡터가 강조 표시됩니다.

혼동 행렬 (반응이 범주형인 경우에만 사용 가능) 혼동 행렬을 표시하거나 숨깁니다. 자세한 내용은 "**혼동 행렬**"에서 확인하십시오.

ROC 곡선 (반응이 범주형인 경우에만 사용 가능) ROC 곡선을 표시하거나 숨깁니다. 검증을 사용하는 경우 훈련 데이터 집합에 대한 곡선 외에 검증 데이터 집합 및 테스트 데이터 집합 (지정된 경우)에 대한 ROC 곡선이 나타납니다. 자세한 내용은 "**ROC 곡선**"에서 확인하십시오.

향상도 곡선 (반응이 범주형인 경우에만 사용 가능) 향상도 곡선을 표시하거나 숨깁니다. 검증을 사용하는 경우 훈련 데이터 집합에 대한 곡선 외에 검증 데이터 집합 및 테스트 데이터 집합 (지정된 경우)에 대한 향상도 곡선이 나타납니다. 자세한 내용은 "**향상도 곡선**"에서 확인하십시오.

실제값 대 예측값 그림 (반응이 연속형인 경우에만 사용 가능) 실제값 대 예측값 그림을 표시하거나 숨깁니다.

잔차 대 예측값 그림 (반응이 연속형인 경우에만 사용 가능) 잔차 대 예측값 그림을 표시하거나 숨깁니다.

프로파일러 예측 프로파일러를 표시하거나 숨깁니다. 각 반응 수준은 예측 프로파일러에서 별도의 행으로 표시됩니다. 빨간색 삼각형 메뉴의 옵션에 대한 자세한 내용은 **Profilers**에서 확인하십시오.

등고선 프로파일러 (모형에 두 개 이상의 연속형 요인이 포함된 경우에만 사용 가능) 등고선 프로파일러를 표시하거나 숨깁니다. 빨간색 삼각형 메뉴의 옵션에 대한 자세한 내용은 **Profilers**에서 확인하십시오.

표면 프로파일러 (모형에 하나 이상의 연속형 요인이 포함된 경우에만 사용 가능) 표면 프로파일러를 표시하거나 숨깁니다. 빨간색 삼각형 메뉴의 옵션에 대한 자세한 내용은 **Profilers**에서 확인하십시오.

예측값 저장 임계 확률 경계 값에 기반한 예측 분류를 데이터 테이블의 <Y, 반응> 예측값 열에 저장합니다.

예측 계산식 저장 임계 확률 경계 값에 기반한 예측 계산식을 데이터 테이블에 저장합니다. 반응이 이항이면 단일 예측 계산식 열이 데이터 테이블에 추가됩니다. 반응 수준이 세 개 이상이면 최종 예측에 대한 계산식 열 외에 각 쌍별 모형에 대한 예측 계산식 열이 데이터 테이블에 추가됩니다.

예측 계산식 게시 임계 확률 경계 값에 기반한 확률 계산식을 생성하여 계산식 저장소 플랫폼에 계산식 열 스크립트로 저장합니다. "계산식 저장소" 보고서가 열려 있지 않은 경우 이 옵션

을 선택하면 " 계산식 저장소 " 창이 열립니다. 자세한 내용은 " [계산식 저장소](#) "에서 확인하십시오.

확률 저장 최대 확률 분류 <Y, 반응> 열 및 각 반응 수준의 확률을 데이터 테이블에 개별 열로 저장합니다.

확률 계산식 저장 (반응이 이항인 경우에만 사용 가능) 다음 계산식 열을 데이터 테이블에 저장합니다.

결정 값 결정 값에 대한 계산식을 포함하는 숨겨진 열입니다.

확률 이항 반응의 각 수준에 대해 하나씩 총 두 개의 확률 계산식 열입니다.

최대 확률 분류 확률 값이 더 높은 예측 분류입니다.

임계 예측 분류 혼동 행렬에서 얻은 임계 확률 경계 값에 기반한 예측 분류입니다. 기본적으로 이 값은 플랫폼에서 계산되는 분류 결정 규칙을 기반으로 합니다. 임계값을 0.5로 변경할 경우 임계 예측 분류 열은 최대 확률 분류 열과 동일하므로 데이터 테이블에 저장되지 않습니다.

확률 계산식 게시 (반응이 이항인 경우에만 사용 가능) 확률 계산식을 생성하여 계산식 저장소 플랫폼에 계산식 열 스크립트로 저장합니다. " 계산식 저장소 " 보고서가 열려 있지 않은 경우 이 옵션을 선택하면 " 계산식 저장소 " 창이 열립니다. 자세한 내용은 " [계산식 저장소](#) "에서 확인하십시오.

검증 저장 (" 홀드백 " 또는 "K 폴드 " 검증 방법을 사용하는 경우에만 사용 가능) 훈련 데이터 집합과 검증 데이터 집합에 사용된 행을 식별하는 새 열을 데이터 테이블에 생성합니다.

적합 제거 전체 모형 보고서를 제거합니다.

서포트 벡터 머신의 추가 예

이 예에서는 SVR(서포트 벡터 회귀) 을 사용하여 연속형 반응을 모델링합니다. SVR 모형은 SVM 알고리즘을 사용하여 연속형 반응 데이터를 예측합니다. " [서포트 벡터 머신의 예](#) "에서 사용된 것과 동일한 당뇨병 환자 의료 데이터를 사용합니다. 그러나 당뇨병 진행 경과의 이진 측정값 대신 연속형 측정값이 제공됩니다. 병이 많이 진행될수록 값이 높습니다. 방사 기저 함수의 비용 모수를 변경하여 최량 적합 모형을 찾을 수 있습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Diabetes.jmp 를 엽니다.
2. **분석 > 예측 모델링 > 서포트 벡터 머신**을 선택합니다.
3. Y 를 선택하고 **Y, 반응**을 클릭합니다.
4. Age ~ Glucose 를 선택하고 **X, 요인**을 클릭합니다.
5. Validation 을 선택하고 **검증**을 클릭합니다.
6. **확인**을 클릭합니다.
7. " 모형 시작 " 제어판에서 커널 함수가 " 방사 기저 함수 " 이고 " 비용 " 모수가 1, " 감마 " 모수가 0.1 인지 확인합니다.

8. **시작**을 클릭합니다.
9. "모형 시작" 옆의 회색 삼각형을 클릭하여 "모형 시작" 제어판을 엽니다.
10. 비용 모수를 0.1 로 변경합니다.
11. **시작**을 클릭합니다.
12. "모형 시작" 옆의 회색 삼각형을 클릭하여 "모형 시작" 제어판을 엽니다.
13. 비용 모수를 2 로 변경합니다.
14. **시작**을 클릭합니다.

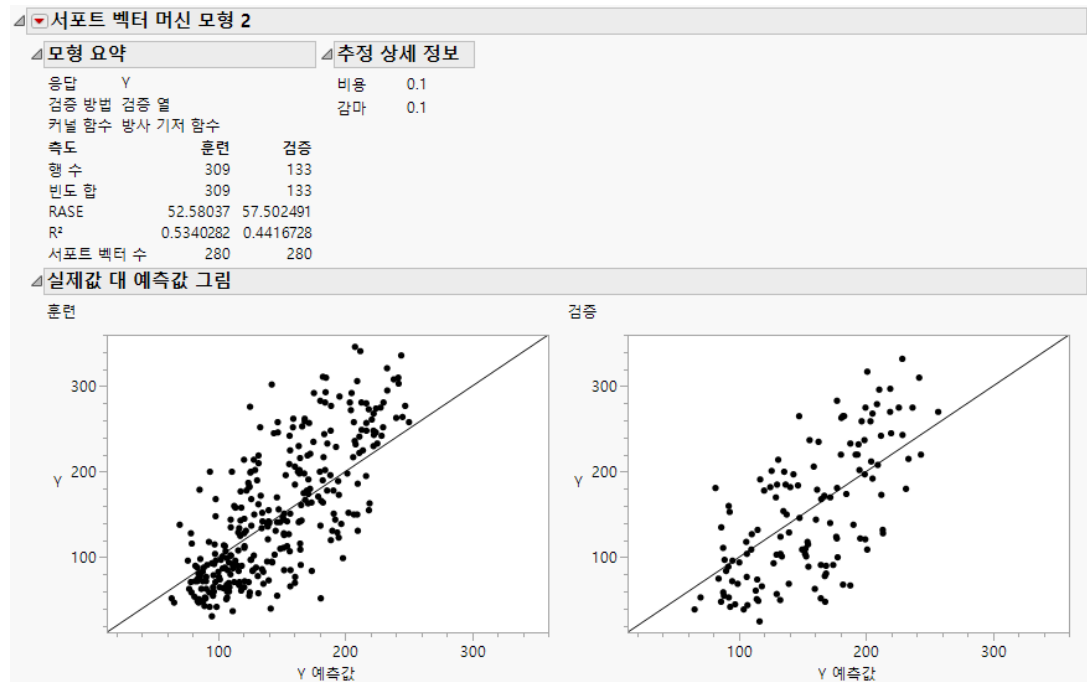
그림 9.6 모형 비교 보고서

모형 비교							
표시 방법	커널 함수	비용	감마	SV 수	훈련 RASE	검증 RASE	검증 일반화 R ² 최선
☐ 모형 1 방사 기저 함수		1	0.1	266	43.8147	58.3285	0.42552
☒ 모형 2 방사 기저 함수		0.1	0.1	280	52.5804	57.5025	0.44167 최소 RASE 최대 R ²
☐ 모형 3 방사 기저 함수		2	0.1	265	40.553	60.7801	0.37621

"모형 비교" 보고서에는 서로 다른 비용 모수를 사용한 모형의 결과가 포함되어 있습니다. 비용 모수가 높을수록 점을 잘못 분류할 가능성이 낮은 알고리즘이 구현되는 반면, 비용 모수가 낮을수록 더 유연한 알고리즘이 구현됩니다. 이 시나리오에서는 더 유연한 모형, 즉 비용 모수가 0.1 인 모형이 최적 모형 적합을 생성합니다.

15. "모형 비교" 보고서의 "표시" 열에서 "모형 1" 과 "모형 3" 을 선택 취소합니다.

그림 9.7 연속형 반응에 대한 모형 보고서

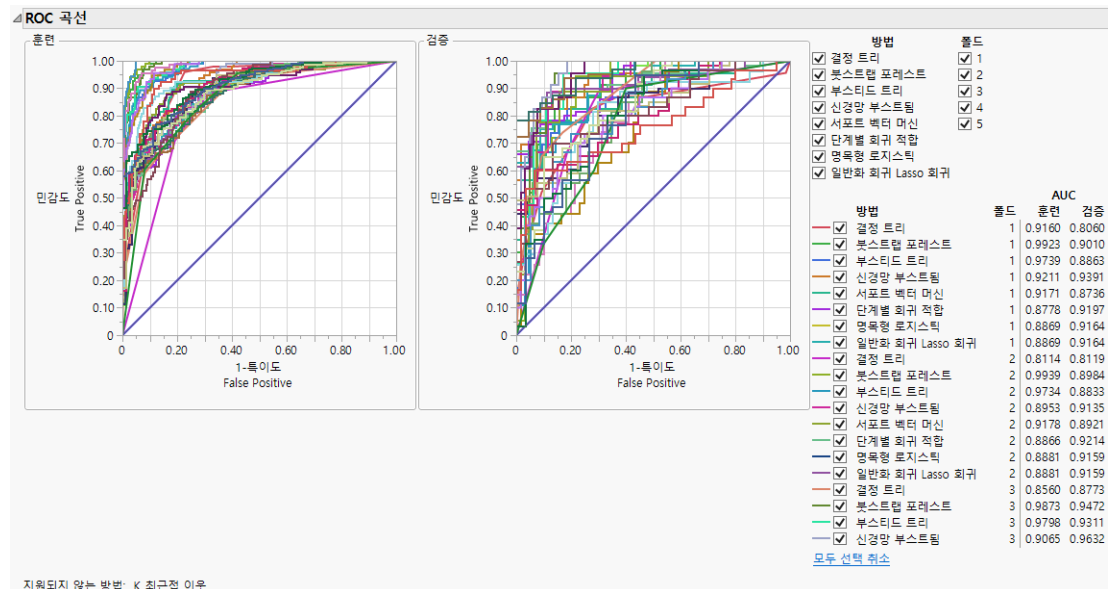


연속형 반응에 대한 모형 보고서에는 "모형 요약", "추정 상세 정보" 및 "실제값 대 예측값 그림"이 포함됩니다. "모형 요약" 보고서에서 훈련 데이터 집합의 RASE 는 52.58, 검증 데이터 집합의 RASE 는 57.50 입니다.

모형 선별 플랫폼은 JMP Pro에서만 사용할 수 있습니다.

JMP Pro의 모형 선별 플랫폼을 사용하면 하나의 시작 창에서 여러 예측 모델링 플랫폼을 실행하고 다양한 방법의 요약을 조합할 수 있습니다. 이렇게 하면 개별 모형 플랫폼을 시작하지 않고도 최적 성능 모형을 선택할 수 있습니다. "모형 선별" 보고서 창에서 원하는 모형의 개별 플랫폼을 손쉽게 실행하여 보다 세분화할 수도 있습니다.

그림 10.1 모형 선별 보고서



목차

모형 선별의 예 157

모형 선별 플랫폼 시작..... 159

모형 선별 보고서 162

 폴드 간 요약 보고서..... 162

 훈련, 검증 및 테스트 적합 측도..... 163

모형 선별 플랫폼 옵션..... 164

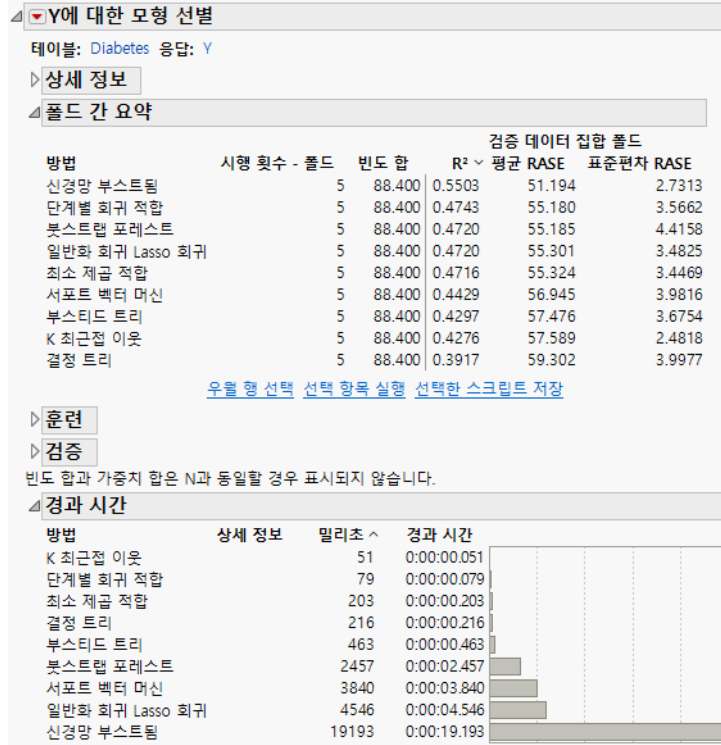
결정 임계값 보고서 167

JMP[®] PRO 모형 선별의 예

당뇨병 진행 경과를 예측하기 위해 다양한 모형의 성능을 평가하려고 합니다. 데이터 집합에는 442 명의 당뇨병 환자에 대한 기준 의료 데이터 및 각 환자가 병원에 처음 내원한 후 1 년 동안 얻은 당뇨병 진행 경과의 연속형 측정값이 포함되어 있습니다. 병이 많이 진행될수록 값이 높습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Diabetes.jmp 를 엽니다.
2. **분석 > 예측 모델링 > 모형 선별**을 선택합니다.
3. Y 를 선택하고 **Y, 반응**을 클릭합니다.
4. Age ~ Glucose 를 선택하고 **X, 요인**을 클릭합니다.
5. "폴드 교차 검증" 섹션에서 **K 폴드 교차 검증** 옆의 상자를 선택합니다.
6. (선택 사항) "작업 옵션" 섹션에서 **난수 시드값 설정** 옆에 "42920" 을 입력합니다.
예제 결과와 일치시키려면 난수 시드값을 사용하십시오.
7. **확인**을 클릭합니다.
8. "Y 에 대한 모형 선별" 옆의 빨간색 삼각형을 클릭하고 **선택적 보고서 > 경과 시간**을 선택합니다.

그림 10.2 연속형 반응에 대한 모형 보고서



전체 폴드에서 평균화된 검증 R^2 을 기반으로 하는 최적 성능 모형은 "부스트된 신경망"입니다. "부스트된 신경망"의 평균 검증 R^2 은 0.5503입니다. "경과 시간" 보고서에 따르면 "부스트된 신경망" 모형이 적합 시간도 가장 오래 걸렸습니다. 이 경우 경과 시간은 5초에 불과하지만 데이터 집합이 크거나 모형이 복잡한 경우에는 이 정보가 중요할 수 있습니다.

팁: 개별 모형의 성능 측정 기준을 보려면 "검증" 옆의 회색 표시 아이콘을 클릭하십시오.

JMP PRO 모형 선별 플랫폼 시작

모형 선별 플랫폼을 시작하려면 **분석 > 예측 모델링 > 모형 선별**을 선택합니다.

그림 10.3 모형 선별 시작 창

다양한 예측 모델을 적당시키고 적당 측도에 대한 요약물을 제공합니다.

열 선택

15개 열

- Y
- Y Binary
- Y Ordinal
- Age
- Gender
- BMI
- BP
- Total Cholesterol
- LDL
- HDL
- TCH
- LTG
- Glucose
- Validation
- Hidden Validation

선택한 열 역할 지정

Y, 반응 Y

X, 요인 Age, Gender, BMI, BP

가중치 선택적 숫자

빈도 선택적 숫자

검증 선택적 숫자

기준 선택적

작업

확인

취소

제거

재호출

도움말

방법

☒ 결정 트리

☒ 붓스트랩 포레스트

☒ 부스티드 트리

☒ K 최근접 이웃

☐ Naive Bayes

☒ 신경망

☒ 서포트 벡터 머신

☒ 판별

☒ 최소 제곱 적합

☒ 단계별 회귀 적합

☒ 로지스틱 회귀

☒ 일반화 회귀

☐ 추가 방법

☐ 부분 최소 제곱

☐ XGBoost

모델링 옵션

☐ 이원 교호작용 추가

☐ 2차 추가

☐ 결측값 정보화

작업 옵션

난수 시도값 설정

개별 시간 한계

☐ Live 보고서 제거

☐ 로그에 방법 표시

폴드 교차 검증

연속 폴드를 사용하여 반복적으로 적합시킵니다.

☒ K 폴드 교차 검증 K 5

☐ 내부 교차 검증 K 4 L 5

반복 K 폴드 0

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

Y, 반응 분석할 하나 이상의 반응 변수입니다.

X, 요인 예측 변수입니다.

가중치 (K 최근접 이웃, 서포트 벡터 머신 또는 신경망 모델링 플랫폼에는 적용할 수 없음) 분석의 각 행에 가중치를 할당하는 숫자 값이 포함된 열입니다.

빈도 (K 최근접 이웃 모델링 플랫폼에는 적용할 수 없음) 분석의 각 행에 빈도를 할당하는 숫자 값이 포함된 열입니다.

검증 (시작 창에서 "교차 검증" 옵션을 선택한 경우에는 적용할 수 없음) 검증 데이터 집합을 정의하는 숫자 열입니다. "열 선택" 목록에서 아무 열도 선택하지 않은 상태로 "검증" 버튼

을 클릭하면 데이터 테이블에 검증 열을 추가할 수 있습니다. "검증 열 생성" 유틸리티에 대한 자세한 내용은 "검증 열 생성"에서 확인하십시오.

참고 : 네 개 이상의 수준이 있는 검증 열을 지정하는 경우 이 열은 K 폴드 교차 검증을 수행하는 데 사용됩니다.

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.

방법 원하는 모델링 플랫폼을 선택할 수 있습니다. 기본적으로 적용되는 모델링 플랫폼은 "결정 트리 (파티션)", "붓스트랩 포레스트", "부스티드 트리", "K 최근접 이웃", "신경망", "서포트 벡터 머신", "판별", "최소 제곱 적합", "단계별 회귀 적합", "로지스틱 회귀" 및 "일반화 회귀"입니다. "Naive Bayes", "부분 최소 제곱" 및 "XGBoost"도 사용할 수 있습니다.

참고 :

- XGBoost는 JMP에서 지원되지 않으며 XGBoost 추가기능이 설치된 경우에만 사용할 수 있습니다. XGBoost에 대한 자세한 내용은 community.jmp.com에서 확인하십시오.
- 결정 트리 (파티션), 판별 및 부분 최소 제곱 방법 모두 모형을 적합시키려면 특정 유형의 검증 데이터 집합이 필요합니다.
- 검증 데이터 집합의 관측값이 20 개 미만인 경우 결정 트리 (파티션) 모형을 적합시킬 수 없습니다.
- 모델링 플랫폼에서는 기본 옵션 및 조정 모수를 사용하여 모형을 적합시킵니다. 플랫폼을 직접 호출하고 다른 옵션을 선택하여 기본 설정을 사용할 때보다 적합성을 향상시킬 수 있습니다.
- 일반화 회귀의 "추가 방법" 옵션은 일반화 회귀 플랫폼에서 능형 회귀, Elastic Net 및 Lasso 회귀와 같은 몇 가지 추가 방법을 호출합니다. Lasso 회귀 방법의 경우 관측값이 1000 개 미만이고 변수가 100 개 미만이면 "조기 중지"가 비활성화됩니다. 자세한 내용은 선형 모형 적합에서 확인하십시오.

주의 : 이로 인해 추가 모형 적합이 생성됩니다.

모델링 옵션

모델링 플랫폼에 대한 추가 옵션을 제공합니다.

이원 교호작용 추가 모든 이원 교호작용 효과를 선형 모형에 추가합니다.

2차 추가 연속형 변수의 제곱에 대한 효과를 선형 모형에 추가합니다.

결측값 정보화 모든 플랫폼에 대해 결측값 정보화를 활성화합니다.

작업 옵션

추가 옵션을 제공합니다.

난수 시드값 설정 모형 적합 루틴의 랜덤 성분에 사용되는 난수 시드값을 설정합니다. 이렇게 하면 플랫폼을 다시 실행하고 동일한 모형 적합을 얻을 수 있습니다.

개별 시간 한계 각 적합에 대한 시간 한계 (초) 를 지정합니다. 조기 중지를 지원하는 플랫폼의 경우 해당 시점까지의 최적 추정값이 제공됩니다. 조기 중지를 지원하지 않는 플랫폼의 경우 결과가 제공되지 않습니다.

Live 보고서 제거 " 모형 선별 " 보고서 창에 개별 모형 플랫폼 보고서를 포함하지 않습니다.

팁 : 많은 방법과 적합이 포함된 큰 문제의 경우 이 옵션을 선택하여 메모리를 확보할 수 있습니다.

실행 시 로그에 방법 표시 적합 플랫폼이 호출될 때마다 진행 메시지를 로그에 기록합니다.

폴드 교차 검증

다양한 유형의 교차 검증에 대한 옵션을 제공합니다.

K 폴드 교차 검증 데이터를 무작위로 K 개의 부분 또는 폴드로 나눕니다. 모형 생성을 위해 $K-1$ 개의 폴드를 사용하여 모형이 데이터에 적합되고, 나머지 폴드는 교차 검증에 사용됩니다. 총 K 개 모형에 대해 이 과정이 K 회 반복됩니다. K 의 기본값은 5 입니다.

- K 는 K 폴드 교차 검증을 위한 폴드 수를 지정합니다. 기본값은 5 이고 K 는 1 보다 커야 합니다.
- 최적 모형에 대한 결과가 제공됩니다.

내포 교차 검증 교차 검증을 위해 데이터를 내포된 폴드로 나눕니다. 먼저 데이터를 $k=1, \dots, K$ 개의 동일한 부분 또는 폴드로 나눕니다. 각 폴드에 대해 k 번째 폴드는 테스트 데이터 집합으로 사용되고, 나머지 데이터는 L 개의 동일한 부분으로 추가 분할됩니다. 이러한 L 개의 부분을 내측 폴드라고 합니다. 그런 다음 $L-1$ 개의 내측 폴드를 사용하여 모형이 데이터에 적합되고, 나머지 내측 폴드는 매번 교차 검증 집합으로 홀드아웃됩니다. 그런 다음 L 개의 모형이 k 번째 폴드를 공통 테스트 데이터 집합으로 사용합니다. 총 $K*L$ 개 모형이 적합됩니다. K 의 기본값은 4 이고 L 의 기본값은 5 입니다.

예를 들어 $K=2, L=3$ 으로 설정합니다. 처음에는 데이터가 두 개의 폴드로 구분됩니다. 첫 번째 폴드는 테스트 데이터 집합으로 홀드아웃되고 두 번째 폴드는 3 개의 내측 폴드로 구분됩니다. 세 개의 모형이 데이터에 적합되고, 매번 교차 검증 집합으로 홀드아웃된 서로 다른 내측 폴드가 사용됩니다. 그러면 세 개 모형이 모두 첫 번째 폴드에서 테스트됩니다.

그런 다음 두 번째 폴드가 테스트 데이터 집합으로 홀드아웃되고 첫 번째 폴드는 3 개의 내측 폴드로 구분됩니다. 세 개의 모형이 데이터에 적합되고, 매번 교차 검증 집합으로 홀드아웃된 서로 다른 내측 폴드가 사용됩니다. 그러면 세 개 모형이 모두 두 번째 폴드에서 테스트됩니다.

- K 는 내포 교차 검증을 위한 폴드 수를 지정합니다. 기본값은 4 이고 K 는 1 보다 커야 합니다.

- L 은 내포 교차 검증을 위한 내측 폴드 수를 지정합니다. 기본값은 5 이고 L 은 1 보다 커야 합니다.

참고 : "K 폴드 교차 검증" 과 "내포 교차 검증" 을 모두 선택하면 "내포 교차 검증" 이 수행됩니다.

반복 K 폴드 K 폴드 교차 검증 또는 내포 교차 검증 프로세스의 반복 횟수를 지정합니다.

"확인" 을 클릭하면 지정된 모형이 적합되고 진행률 표시줄 집합이 표시됩니다. 위쪽 진행률 표시줄은 모든 적합에 대한 진행률을 보고합니다. 아래쪽 진행률 표시줄은 현재 개별 모형 적합에 대한 진행률을 보고합니다. 아래쪽 진행률 표시줄은 조기 중지를 사용하여 중지할 수 있으며 위쪽 진행률 표시줄은 계속 실행됩니다.

JMP PRO 모형 선별 보고서

"모형 선별" 보고서의 내용은 구현된 교차 검증 유형과 반응 변수의 모델링 유형에 따라 달라집니다. "모형 선별" 보고서 맨 위에 데이터 테이블 이름과 반응 변수의 열 이름이 표시됩니다. 검증 열을 지정한 경우 해당 열 이름도 표시됩니다.

기본적으로 "상세 정보" 보고서가 다음에 나타납니다. 이 보고서에는 각 폴드 및 시행에 대한 모든 모형의 개별 모형 보고서가 포함되어 있습니다. 이러한 보고서는 각 플랫폼의 실제 출력이며 모든 보고서 및 명령 기능을 포함합니다. 예를 들어 최적 모형을 결정한 경우 "상세 정보" 보고서의 해당 플랫폼 보고서로 이동하여 원하는 예측 계산식 열을 저장할 수 있습니다. 시작 창에서 "Live 보고서 제거" 옵션을 선택하여 "상세 정보" 보고서를 숨길 수 있습니다.

시작 창에서 "K 폴드 교차 검증" 옵션 또는 "내포 교차 검증" 옵션을 지정한 경우 "폴드 간 요약" 보고서가 나타납니다. 이 보고서는 테스트 폴드에서 평균화된 적합 측도를 요약한 것입니다.

각 개별 모형 적합, 훈련 데이터 집합, 검증 데이터 집합 및 테스트 데이터 집합(해당하는 경우)에 대한 적합 측도가 표시됩니다. 시작 창에서 검증 열을 지정하여 교차 검증이 사용된 경우 개별 모형 적합 측도만 보고서에 표시됩니다.

폴드 간 요약 보고서

모형 선별 플랫폼의 "폴드 간 요약" 보고서에는 전체 폴드에 대한 적합 측도 및 전체 시행에 대한 적합 측도 ("반복 K 폴드"가 지정된 경우)가 요약되어 있습니다. "K 폴드 교차 검증" 옵션을 지정한 경우 검증 데이터 집합 전체에 대해 적합 측도가 요약됩니다. "내포 교차 검증" 옵션을 지정한 경우에는 테스트 데이터 집합 전체에 대해 적합 측도가 요약됩니다. 이 보고서에는 다음 열을 포함하는 테이블이 있습니다.

방법 모형을 적합시키는 데 사용된 방법의 이름입니다.

시행 횟수 - 폴드 모든 폴드 및 시행 (해당하는 경우)에 대해 적합된 총 모형 수입니다.

빈도 합 검증 데이터 집합 또는 테스트 데이터 집합의 평균 관측값 수입니다. 테스트 데이터 집합은 모형 성능을 추정하는 데 사용됩니다.

R² 모든 검증 데이터 집합 폴드 또는 테스트 데이터 집합 폴드에 대한 평균 R² 입니다. 반응이 범주형인 경우 엔트로피 R² 이 이 열에 포함됩니다.

평균 RASE 모든 검증 데이터 집합 폴드 또는 테스트 데이터 집합 폴드에 대한 평균 RASE(제곱근 평균 제곱 오차) 입니다.

표준편차 RASE 모든 검증 데이터 집합 또는 테스트 데이터 집합에 대한 RASE의 표준편차입니다.

평균 AUC (범주형 반응에만 사용 가능) 모든 검증 데이터 집합 또는 테스트 데이터 집합에 대한 평균 AUC(ROC 곡선 아래 면적) 입니다.

평균 MR (범주형 반응에만 사용 가능) 모든 검증 데이터 집합 또는 테스트 데이터 집합에 대한 평균 MR(오분류 비율) 입니다.

테이블 아래에 다음과 같은 옵션이 있습니다.

우월 행 선택 모형 적합 기준 조합에 따라 다른 모든 모형보다 우수하거나 동일한 각 모형을 선택합니다. 이를 파레토 프론티어 선택이라고도 합니다. 연속형 반응의 경우 우월 모형을 결정할 때 R² 과 빈도 합을 고려합니다. 범주형 반응의 경우 우월 모형을 결정할 때 엔트로피 R², 오분류 비율, AUC 및 빈도 합을 고려합니다.

선택 항목 실행 선택한 각 행에 대해 개별 모형을 실행합니다. 시작 창에서 폴드 교차 검증 유형을 지정한 경우 최종 모형에 해당하는 폴드, 내측 폴드 및 시행 조합을 사용하여 모형이 실행됩니다. 최종 모형은 가중 평균 R² 이 가장 높은 모형입니다. 가중 평균 R² 은 훈련 R², 검증 R² 및 테스트 R²(테스트 데이터 집합이 사용된 경우)의 관측값 수에 따라 가중치가 적용된 평균입니다. "선택 항목 실행"을 클릭하면 필요한 폴드, 내측 폴드 및 시행 조합에 대한 검증 열이 데이터 테이블에 생성됩니다.

선택한 스크립트 저장 선택한 각 행의 스크립트 창에 모형 스크립트를 저장합니다. 시작 창에서 폴드 교차 검증 유형을 지정한 경우 스크립트에 지정된 검증 데이터 집합은 최종 모형에 해당하는 폴드, 내측 폴드 및 시행 조합을 사용합니다. 최종 모형은 가중 평균 R² 이 가장 높은 모형입니다. 가중 평균 R² 은 훈련 R², 검증 R² 및 테스트 R²(테스트 데이터 집합이 사용된 경우)의 관측값 수에 따라 가중치가 적용된 평균입니다. "선택한 스크립트 저장"을 클릭하면 요청된 폴드, 내측 폴드 및 시행 조합에 대한 검증 열이 데이터 테이블에 생성됩니다.

훈련, 검증 및 테스트 적합 척도

모형 선별 플랫폼에는 지정된 각 모형 데이터 집합에 대한 적합 척도 보고서가 있습니다. 훈련 데이터 집합, 훈련 및 검증 데이터 집합 또는 훈련, 검증 및 테스트 데이터 집합이 될 수 있습니다. 각 보고서에는 다음 열을 포함하는 테이블이 있습니다.

방법 모형을 적합시키는 데 사용된 방법의 이름입니다.

N 데이터 집합의 관측값 수입니다.

가중치 합 가중치의 합입니다.

R² (연속형 반응에만 사용 가능) 적합 모형의 R² 값입니다.

엔트로피 R² (범주형 반응에만 사용 가능) 적합 모형과 상수 확률 모형의 로그 가능도를 비교하는 적합 척도입니다. 엔트로피 R² 범위는 0~1이며, 값이 1에 가까울수록 더 나은 적합을 나타냅니다. 자세한 내용은 "**엔트로피 R²**"에서 확인하십시오.

오분류 비율 (범주형 반응에만 사용 가능) 모형에 의해 잘못 분류된 관측값의 비율입니다. 값이 작을수록 더 나은 적합을 나타냅니다.

참고 : 이러한 테이블에서 오분류 비율은 항상 확률 임계 0.5를 사용하여 계산됩니다.

AUC (범주형 반응에만 사용 가능) ROC 곡선 아래 면적입니다. 값이 1에 가까울수록 더 나은 적합을 나타냅니다.

RASE 평균 제곱 예측 오차의 제곱근입니다 (제곱근 평균 제곱 오차). RASE는 다음과 같이 계산되며, 여기서 **Source**는 훈련, 검증 또는 테스트 데이터 집합을 나타냅니다.

$$RASE_{Source} = \sqrt{\frac{SSE_{Source}}{n}}$$

일반화 R² (범주형 반응에만 사용 가능) 일반 회귀 모형에 적용할 수 있는 척도입니다. 이 값은 가능도 함수 L 을 기반으로 하며 최대값이 1이 되도록 척도화됩니다. 값이 1이면 완벽한 모형을 나타내고 0이면 상수 모형과 같음을 나타냅니다. 일반화 R² 척도는 표준 최소 제곱 설정 시 연속형 정규 반응에 대한 기준 R²로 단순화됩니다. 일반화 R²을 Nagelkerke 또는 Craig와 Uhler R²이라고도 하는데, 이는 Cox-Snell 유사 R²을 정규화한 것입니다. 자세한 내용은 Nagelkerke 연구 자료 (1991)에서 확인하십시오.

폴드 (시작 창에서 "K 폴드 교차 검증" 옵션 또는 "내포 교차 검증" 옵션을 지정한 경우에만 사용 가능) 해당 행에서 모형 적합을 위해 홀드아웃된 폴드를 식별합니다.

내측 폴드 (시작 창에서 "내포 교차 검증" 옵션을 지정한 경우에만 사용 가능) 해당 행에서 모형 적합을 위해 홀드아웃된 내측 폴드를 식별합니다.

시행 (시작 창에서 "반복 K 폴드" 옵션을 지정한 경우에만 사용 가능) 해당 행에서 모형 적합에 대한 시행 번호를 식별합니다.

각 테이블 아래에 다음과 같은 옵션이 있습니다.

우월 행 선택 모형 적합 기준 조합에 따라 다른 모든 모형보다 우수하거나 동일한 각 모형을 선택합니다. 연속형 반응의 경우 우월 모형을 결정할 때 R²과 빈도 합을 고려합니다. 범주형 반응의 경우 우월 모형을 결정할 때 엔트로피 R², 오분류 비율, AUC 및 빈도 합을 고려합니다.

선택 항목 실행 선택한 각 행에 지정된 개별 모형을 실행합니다.

선택한 스크립트 저장 선택한 각 행의 스크립트 창에 모형 스크립트를 저장합니다.

실제값 대 예측값 그림 (연속형 반응에만 사용 가능) 모든 모형 적합의 실제값 대 예측값 점을 중첩하는 그림 보고서를 표시하거나 숨깁니다. 훈련, 검증 및 테스트 데이터 집합에 대한 개별 그림이 있습니다. 이 보고서에는 방법, 폴드, 내측 폴드 및 시행 조합을 기반으로 그림에 나타나는 적합을 지정할 수 있는 체크박스가 포함된 "방법" 범례가 있습니다. 범례 맨 아래에 있는 "모두 선택 취소"를 클릭하면 그림에서 모든 모형 적합이 제거됩니다.

참고 :

- 훈련, 검증 또는 테스트 보고서에서 모형의 부분집합을 선택한 후 "실제값 대 예측값 그림"을 선택하면 선택한 모형에 대한 그림만 표시됩니다.
- 일부 모델링 플랫폼에서는 실제값 대 예측값 그림이 지원되지 않습니다. 지원되지 않는 플랫폼 목록이 그림 아래에 표시됩니다.
- 시작 창에서 "Live 보고서 제거"를 선택한 경우 이 옵션을 사용할 수 없습니다.

프로파일러 (연속형 반응에만 사용 가능) 각 모형 적합 유형에 대한 예측 프로파일러를 표시하거나 숨깁니다. 자세한 내용은 **Profilers**에서 확인하십시오. 시작 창에서 폴드 교차 검증 유형을 지정한 경우 예측 프로파일러의 선택된 모형은 최종 모형에 해당하는 폴드, 내측 폴드 및 시행 조합을 사용합니다. 최종 모형은 가중 평균 R^2 이 가장 높은 모형입니다. 가중 평균 R^2 은 훈련 R^2 , 검증 R^2 및 테스트 R^2 (테스트 데이터 집합이 사용된 경우)의 관측값 수에 따라 가중치가 적용된 평균입니다.

참고 :

- 표준 최소 제곱 모형 및 일반화 회귀 모형의 예측 프로파일러에 신뢰 구간이 표시됩니다.
- K 최근접 이웃 모형에는 예측 프로파일러를 사용할 수 없습니다.
- 시작 창에서 "Live 보고서 제거"를 선택한 경우 이 옵션을 사용할 수 없습니다.

ROC 곡선 (범주형 반응에만 사용 가능) 모든 모형 적합에 대한 ROC 곡선을 중첩하는 그림 보고서를 표시하거나 숨깁니다. 훈련, 검증 및 테스트 데이터 집합에 대한 개별 그림이 있습니다. 이 보고서에는 그림에 나타나는 적합을 지정할 수 있는 체크박스가 포함된 두 개의 범례가 있습니다. 범례 맨 아래에 있는 "모두 선택 취소"를 클릭하면 그림에서 모든 모형 적합이 제거됩니다. ROC 곡선에 대한 자세한 내용은 "[ROC 곡선](#)"에서 확인하십시오.

참고 :

- 훈련, 검증 또는 테스트 보고서에서 모형의 부분집합을 선택한 후 "ROC 곡선"을 선택하면 선택한 모형에 대한 그림만 표시됩니다.
- 일부 모델링 플랫폼에서는 ROC 그림이 지원되지 않습니다. 지원되지 않는 플랫폼 목록이 그림 아래에 표시됩니다.
- 시작 창에서 "Live 보고서 제거"를 선택한 경우 이 옵션을 사용할 수 없습니다.

수익 표시 ("수익 행렬" 열 특성이 있는 이항 범주형 반응에만 사용 가능) 각 모형에 대한 기대 수익을 표시하거나 숨깁니다.

결정 임계값 (이항 범주형 반응에만 사용 가능) 훈련, 검증 및 테스트 (지정된 경우) 데이터 집합에 대한 "결정 임계값" 보고서를 표시하거나 숨깁니다. 각 보고서에는 임계값을 조정하고 모형 적합을 비교할 수 있도록 각 모형에 대한 적합 확률 분포 그래프, 각 모형에 대한 혼동

행렬 및 분류 그래프가 포함되어 있습니다. 자세한 내용은 "[결정 임계값 보고서](#)"에서 확인하십시오.

참고 :

- 훈련, 검증 또는 테스트 보고서에서 모형의 부분집합을 선택한 후 "[결정 임계값](#)"을 선택하면 선택한 모형에 대한 정보만 표시됩니다.
- 시작 창에서 폴드 교차 검증 유형을 지정한 경우 "[결정 임계값](#)" 보고서에 표시되는 확률은 모든 폴드에 대한 평균 확률입니다.

결과 테이블 저장 검증 보고서의 정보를 "모형 선별 통계량 검증 데이터 집합"이라는 새 데이터 테이블에 저장합니다. "내포 교차 검증" 옵션이 지정된 경우와 같이 테스트 데이터 집합이 있는 경우 테스트 보고서의 정보도 "모형 선별 통계량 테스트 데이터 집합"이라는 새 데이터 테이블에 저장됩니다.

K 폴드 결과 테이블 저장 (시작 창에서 교차 검증 옵션을 지정한 경우에만 사용 가능) "폴드 간 요약" 보고서의 정보를 "K 폴드 결과"라는 새 데이터 테이블에 저장합니다.

예측 계산식 저장 (Live 보고서가 모형 보고서에 포함된 경우에만 사용 가능) 선택한 모형에 대한 예측 계산식을 데이터 테이블에 저장합니다. "폴드 간 요약" 보고서에서 모형을 선택하고 시작 창에 폴드 교차 검증 유형이 지정된 경우 최종 모형에 해당하는 폴드, 내측 폴드 및 시행 조합에 대한 예측 계산식이 저장됩니다. 최종 모형은 가중 평균 R^2 이 가장 높은 모형입니다. 가중 평균 R^2 은 훈련 R^2 , 검증 R^2 및 테스트 R^2 (테스트 데이터 집합이 사용된 경우)의 관측값 수에 따라 가중치가 적용된 평균입니다.

참고 :

- 단계별 모형의 경우 예측 계산식이 저장되지 않습니다.
- 순서형 반응을 사용하는 SVM 모형의 경우 예측 계산식이 저장되지 않습니다.
- 시작 창에서 "Live 보고서 제거"를 선택한 경우 이 옵션을 사용할 수 없습니다.

선택적 보고서 추가 보고서 옵션 하위 메뉴를 표시합니다.

경과 시간 각 모형 유형을 적합시키는 데 소요된 총 경과 시간을 포함하는 보고서를 표시하거나 숨깁니다.

예측 변수 카디널리티 각 범주형 예측 변수의 선형 모형 적합에 사용되는 수준 수와 모수 수에 대한 보고서를 표시하거나 숨깁니다.

다음 옵션에 대한 자세한 내용은 [JMP 사용](#)의 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.

결정 임계값 보고서

모형 선별 플랫폼에서 "결정 임계값" 보고서를 사용하면 이항 분류 모형에 대한 임계값을 탐색할 수 있습니다. 검증 방법으로 지정된 각 모형 데이터 집합에 대한 "결정 임계값" 보고서가 있습니다. 훈련 데이터 집합, 훈련 및 검증 데이터 집합 또는 훈련, 검증 및 테스트 데이터 집합이 될 수 있습니다. 각 "결정 임계값" 보고서에는 적합 확률 분포 그래프, 분류 막대 차트 및 혼동 행렬이 포함되어 있습니다. 모두 모형 적합, 폴드, 시행 및 클래스 수준별로 구성됩니다. 보고서에는 분류 정확성 측도에 대한 탭 형식 섹션과 수익 행렬 설정 옵션도 포함되어 있습니다. 확률 임계를 조정하면 보고서가 업데이트됩니다.

적합 확률 분포

적합 확률 분포 또는 모형 스코어를 통해 각 개별 모형 적합이 두 클래스를 어떻게 구분하는지 확인할 수 있습니다. 그래프의 세로선은 각 관측값의 분류를 결정하는 확률 임계를 나타냅니다. 기본적으로 확률 임계는 0.5입니다. 세로선을 드래그하거나, "확률 임계" 값을 클릭하고 새 값을 입력하여 확률 임계를 변경할 수 있습니다. 그러면 전체 "결정 임계값" 보고서에서 확률 임계값이 변경됩니다. 확률 임계값은 0에서 1 사이여야 합니다.

분류 개수

분류 막대 차트에는 현재 임계에서 반응 변수의 각 수준에 대한 분류 개수가 표시됩니다. 녹색 막대는 올바르게 분류된 관측값을 나타내고 빨간색 막대는 잘못 분류된 관측값을 나타냅니다.

혼동 행렬

분할표라고도 하는 혼동 행렬은 각 개별 모형 적합에 대한 실제 반응과 예측 반응의 이원 분류를 보여 줍니다. 혼동 비율 행렬도 표시됩니다. 비율은 혼동 행렬의 값을 행 합계로 나눈 것과 같습니다.

분류 정확성 측도

임계별 분류 (False) 오분류 수 대 확률 임계 그림과 오분류 비율 대 확률 임계 그림을 표시합니다. 각 그림에는 각 개별 모형 적합에 대한 두 개의 곡선이 포함됩니다. 낮은 반응 범주의 곡선은 실선이고 높은 반응 범주의 곡선은 파선입니다. 곡선은 각 반응 수준에 대해 동일한 오분류 수 또는 비율을 생성하는 임계값에서 교차합니다. 또한 각 그래프에는 현재 확률 임계값을 나타내는 세로선이 있습니다. 이 세로선을 드래그하여 확률 임계값을 변경할 수 있습니다. 그러면 전체 보고서에서 확률 임계값이 변경됩니다.

비율별 분류 (False) 오분류 수 대 순위별 스코어 비율 그림 또는 오분류 비율 대 순위별 스코어 비율 그림을 표시합니다. 각 그림에는 각 개별 모형 적합에 대한 두 개의 곡선이 포함됩니다. 낮은 반응 범주의 곡선은 실선이고 높은 반응 범주의 곡선은 파선입니다.

임계별 분류 (True) True 수 대 확률 임계 그림과 True 비율 대 확률 임계 그림을 표시합니다. 각 그림에는 각 개별 모형 적합에 대한 두 개의 곡선이 포함됩니다. 낮은 반응 범주의 곡선은 파선이고 높은 반응 수준의 곡선은 실선입니다. 곡선은 각 반응에 대해 동일한 올바른 분류 수 또는 비율을 생성하는 임계값에서 교차합니다. 또한 각 그래프에는 현재 확률 임계값을 나타내는 세로선이 있습니다. 이 세로선을 드래그하여 확률 임계값을 변경할 수 있습니다. 그러면 전체 보고서에서 확률 임계값이 변경됩니다.

비율별 분류 (True) True 수 대 순위별 스코어 비율 그림 또는 True 비율 대 순위별 스코어 비율 그림을 표시합니다. 각 그림에는 각 개별 모형 적합에 대한 두 개의 곡선이 포함됩니다. 낮은 반응 범주의 곡선은 파선이고 높은 반응 범주의 곡선은 실선입니다.

임계별 수익 (수익 행렬을 지정한 경우에만 사용 가능) 평균 수익 대 확률 임계 그림을 표시합니다. 각 개별 모형 적합에 대한 곡선과 현재 확률 임계값을 나타내는 세로선이 있습니다. 지정된 수익 행렬도 그림 옆에 표시됩니다.

측정 기준 각 모형에 대한 분류 정확성 측정 기준 테이블을 표시합니다. 각 열의 측정 기준이 계산되는 방법을 설명하는 범례가 제공됩니다.

참고 : 덜 일반적인 두 가지 분류 정확성 측정 기준은 F1 과 MCC 입니다. F1 스코어는 정밀도와 재호출 또는 민감도의 조합입니다. F1 을 계산하는 또 다른 방법은 $2(\text{정밀도} \times \text{민감도}) / (\text{정밀도} + \text{민감도})$ 입니다. MCC(Mathews 상관계수) 는 두 이항 변수에 대해 추정된 Pearson 상관계수와 동일합니다. 자세한 내용은 **다변량 방법**에서 확인하십시오.

수익 행렬 설정

원하지 않는 결과에 비용을 할당하고 원하는 결과에 수익을 할당할 수 있습니다. 자세한 내용은 **"수익 행렬 지정"**에서 확인하십시오. 수익 행렬 창에서 확률 임계를 변경하고 "확인"을 클릭하면 해당 값을 확률 임계로 사용하여 "결정 임계값" 보고서가 업데이트됩니다.

11 장

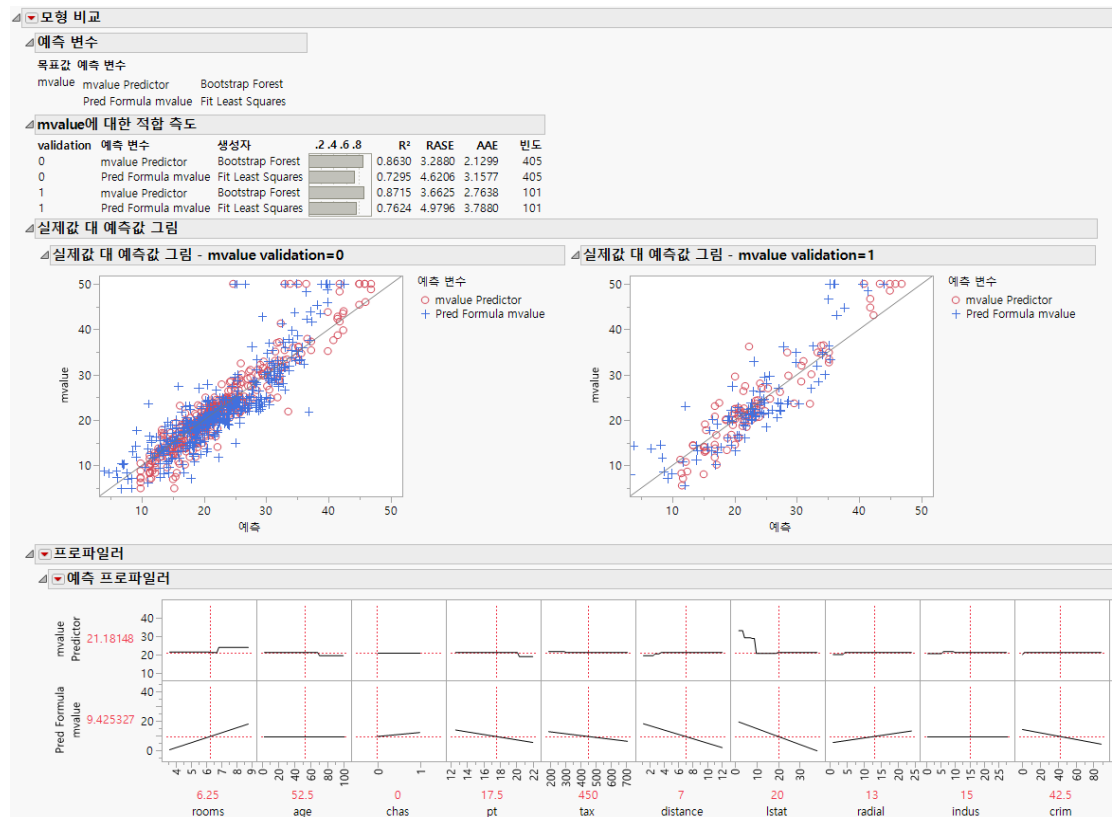
JMP Pro 모형 비교

적합 모형의 예측 능력 비교

모형 비교 플랫폼은 JMP Pro 에서만 사용할 수 있습니다.

모형 비교 플랫폼을 사용하면 여러 모형의 예측 능력을 비교할 수 있습니다. 중첩된 진단 그림과 함께 각 모형에 대한 적합 측도가 제공됩니다.

그림 11.1 모형 비교의 예



목차

모형 비교의 예 171

모형 비교 플랫폼 시작..... 174

모형 비교 보고서 175

모형 비교 플랫폼 옵션..... 176

 연속형 반응..... 176

 범주형 반응..... 176

모형 비교의 추가 예..... 177

JMP PRO 모형 비교의 예

이 예에서는 회귀 모형과 붓스트랩 포레스트 모형을 비교합니다. 인구 통계학 데이터를 사용하며, 중위수 주택 가격에 대한 모형을 생성하는 것이 목표입니다.

도움말 > 샘플 데이터 폴더를 선택하고 Boston Housing.jmp를 열어 시작합니다.

검증 열 생성

1. **분석 > 예측 모델링 > 검증 열 생성**을 선택합니다.
2. 시작 창에서 열을 선택하지 않습니다.
이는 플랫폼에서 단순 랜덤 검증 열을 생성한다는 것을 나타냅니다.
3. **확인**을 클릭합니다.
4. "새 열 이름" 옆의 상자에 "Create Validation"을 입력합니다.
5. "난수 시드값" 옆의 상자에 "1234"를 입력합니다.
6. **시작**을 클릭합니다.
새 검증 열이 생성됩니다. 0이 할당된 행은 훈련 데이터 집합입니다. 1이 할당된 행은 검증 데이터 집합입니다.

회귀 모형 생성 및 열에 예측 계산식 저장

1. **분석 > 모형 적합**을 선택합니다.
2. mvalue를 선택하고 **Y**를 클릭합니다.
3. crim ~ lstat를 선택하고 **추가**를 클릭합니다.
4. Create Validation을 선택하고 **검증**을 클릭합니다.
5. "분석법" 목록에서 **단계별**을 선택합니다.
6. **실행** 버튼을 클릭합니다.
7. "중지 규칙" 목록에서 **p 값 임계**를 선택합니다.
8. **시작** 버튼을 클릭합니다.
9. **모형 실행** 버튼을 클릭합니다.

그림 11.2 모형 적합 보고서

반응 mvalue				
검증: Create Validation				
▶ 효과 요약				
▲ 적합 요약				
R ²		0.753984		
Adj-R ²		0.74663		
제곱근 평균 제곱 오차		4.787344		
반응 평균		23.02868		
관측값 수 (또는 가중치 합)		380		
▲ 분산 분석				
소스	DF	제곱합	평균 제곱	F 비
모형	11	25848.508	2349.86	102.5306
오차	368	8434.069	22.92	Prob > F
수정 합계	379	34282.577		<.0001*
▲ 변수 추정값				
항	추정값	표준 오차	t 비	Prob> t
절편	38.500819	5.676983	6.78	<.0001*
crim	-0.128236	0.03545	-3.62	0.0003*
zn	0.0434609	0.015555	2.79	0.0055*
chas[0]	-1.215423	0.490413	-2.48	0.0136*
nox	-15.38954	4.180964	-3.68	0.0003*
rooms	3.8274522	0.445164	8.60	<.0001*
distance	-1.501159	0.218407	-6.87	<.0001*
radial	0.3364509	0.073112	4.60	<.0001*
tax	-0.012471	0.003826	-3.26	0.0012*
pt	-0.994528	0.146259	-6.80	<.0001*
b	0.0088057	0.003206	2.75	0.0063*
lstat	-0.59558	0.054166	-11.00	<.0001*

10. 예측 계산식을 열에 저장하려면 "반응"의 빨간색 삼각형을 클릭하고 **열 저장 > 예측 계산식**을 선택합니다.

붓스트랩 포레스트 모형 생성 및 열에 예측 계산식 저장

1. 분석 > 예측 모델링 > 붓스트랩 포레스트를 선택합니다.
2. mvalue를 선택하고 **Y, 반응**을 클릭합니다.
3. crim ~ lstat를 선택하고 **X, 요인**을 클릭합니다.
4. Create Validation을 선택하고 **검증**을 클릭합니다.
5. **확인**을 클릭합니다.
6. **조기 중지** 체크박스를 선택합니다.
7. **난수 시드값** 옆의 상자에 "617"을 입력합니다.
8. **확인**을 클릭합니다.

그림 11.3 붓스트랩 포레스트 모형

mvalue에 대한 붓스트랩 포레스트			
규격			
목표값:	mvalue	훈련 행:	380
검증 열:	Create Validation	검증 행:	126
		테스트 행:	0
포레스트의 트리 수:	100	행 수:	13
분할당 표집되는 행 수:	10	붓스트랩 표본:	380
		트리당 최소 분할 수:	10
		최소 크기 분할:	5
전체 통계량			
개별 트리	RASE		
In Bag	2.051425		
Out of Bag	5.131707		
	R ²	RASE	N
훈련	0.937	2.3838256	380
검증	0.845	3.1440678	126

9. 예측 계산식을 열에 저장하려면 "붓스트랩 포레스트"의 빨간색 삼각형을 클릭하고 **열 저장 > 예측 계산식 저장**을 선택합니다.

모형 비교

1. 분석 > 예측 모델링 > 모형 비교를 선택합니다.
2. 예측 계산식 열 두 개를 선택하고 **Y, 예측 변수**를 클릭합니다.
3. Create Validation 을 선택하고 **그룹**을 클릭합니다.

팁 : 그룹 열을 지정하지 않으면 JMP 에서 동일한 검증 열이 모든 예측 변수에 사용될 때 이를 자동으로 인식하여 그룹화 변수로 추가하라는 메시지가 표시됩니다.

4. 확인을 클릭합니다.

그림 11.4 모형 비교 보고서

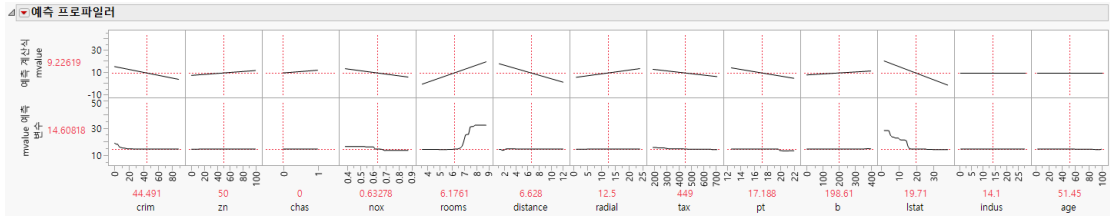
모형 비교									
예측 변수									
mvalue에 대한 적합 측도									
Create Validation	예측 변수	생성자	.2 .4 .6 .8	R ²	RASE	AAE	빈도		
훈련	예측 계산식 mvalue	최소 제곱 적합		0.7540	4.7111	3.3770	380		
훈련	mvalue 예측 변수	붓스트랩 포레스트		0.9370	2.3838	1.4362	380		
검증	예측 계산식 mvalue	최소 제곱 적합		0.6550	4.6972	3.3155	126		
검증	mvalue 예측 변수	붓스트랩 포레스트		0.8454	3.1441	1.9536	126		

훈련 데이터 집합의 행은 모형을 생성하는 데 사용되었으므로 Create Validation = 훈련에 대한 R² 통계량이 인위적으로 부풀려질 수 있습니다. 이 경우 통계량은 모형의 미래 예측 능력을 대표하지 못합니다. 붓스트랩 포레스트 모형의 경우 특히 그렇습니다.

Create Validation = 검증에 대한 통계량을 사용하여 모형을 비교합니다. 이 경우 붓스트랩 포레스트 모형이 회귀 모형보다 더 잘 예측합니다.

5. "모형 비교"의 빨간색 삼각형을 클릭하고 "프로파일러"를 선택합니다.

그림 11.5 모든 모형에 대한 예측 프로파일러



예측 프로파일러를 사용하면 서로 다른 모형에서 각 요인의 영향을 비교할 수 있습니다. 프로파일러는 회귀 모형과 파티션 모형을 사용하는 경우와 같이 여러 유형의 모형을 비교할 때 특히 유용합니다.

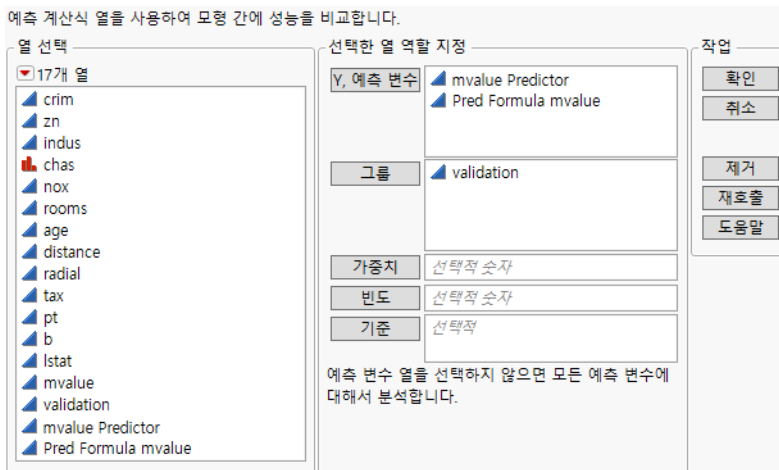
관련 정보

- 선형 모형 적합
- "파티션 모형"

JMP PRO 모형 비교 플랫폼 시작

모형 비교 플랫폼을 시작하려면 **분석 > 예측 모델링 > 모형 비교**를 선택합니다.

그림 11.6 모형 비교 시작 창



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

Y, 예측 변수 비교할 모형에 대한 예측이 포함된 열입니다. 계산식 열 또는 데이터 열일 수 있습니다. JMP 플랫폼에서 생성된 예측 계산식 열에는 "예측" 또는 "반응 확률" 열 특성이 있습니다.

니다. 이러한 특성 중 하나가 포함되지 않은 열을 지정하면 플랫폼에서 지정된 Y 열에 의해 예측되는 열을 지정하라는 메시지가 표시됩니다.

수준이 k 개 있는 범주형 반응의 경우 대부분의 모형 적합 플랫폼은 각각 수준의 확률을 예측하는 k 개 열을 데이터 테이블에 저장합니다. k 개 열은 모두 **Y, 예측 변수**로 지정해야 합니다. k 개의 확률 열을 저장하지 않는 플랫폼의 경우 예측 반응 수준이 포함된 열을 **Y, 예측 변수** 열로 지정할 수 있습니다.

Y, 예측 변수 열을 지정하지 않으면 데이터 테이블에서 "예측" 또는 "반응 확률" 열 특성이 있는 예측 계산식 열이 사용됩니다.

그룹 데이터를 별도로 평가되는 그룹으로 구분하는 열입니다. **그룹** 열을 지정하지 않으면 JMP에서 동일한 검증 열이 모든 예측 변수에 사용될 때 이를 자동으로 인식하여 그룹화 변수로 추가하라는 메시지가 표시됩니다.

다른 역할 버튼은 JMP 플랫폼 간에 공통적입니다. 자세한 내용은 JMP 사용에서 확인하십시오.

JMP PRO 모형 비교 보고서

초기 "모형 비교" 보고서에는 "예측 변수" 보고서와 "적합 척도" 보고서가 포함되어 있습니다. "예측 변수" 보고서에는 모든 반응 및 각 반응에 대해 비교되는 모든 모형이 표시됩니다. 예측 변수 열을 생성한 적합 플랫폼도 나열됩니다. "적합 척도" 보고서에는 각 모형에 대한 적합 척도가 표시됩니다. 연속형 반응과 범주형 반응의 열은 서로 다릅니다.

연속형 반응에 대한 적합 척도

R² R² 통계량입니다. 결측값이 없는 데이터 테이블에서는 "모형 비교" 보고서의 R² 통계량이 원래 모형과 일치합니다. 그러나 결측값이 있으면 R² 통계량이 다릅니다.

RASE 평균 제곱 예측 오차의 제곱근입니다. 이 값은 다음과 같이 계산됩니다.

- 예측 오차 (실제 반응과 예측 반응의 차이) 를 제공하고 합하여 SSE 를 구합니다.
- 관측값 수를 n 으로 나타냅니다.
- RASE 는 다음과 같습니다.

$$RASE = \sqrt{\frac{SSE}{n}}$$

AAE 평균 절대 오차입니다.

빈도 각 행의 빈도 수를 포함하는 열입니다.

범주형 반응에 대한 적합 척도

엔트로피 R² 1 에서 적합 모형과 상수 확률 모형의 음의 로그 가능도 비율을 뺀 값입니다. 값 범위는 0 에서 1 사이입니다.

일반화 R² 일반 회귀 모형에 적용할 수 있는 척도입니다. 이 값은 가능도 함수 L 을 기반으로 하며 최대값이 1 이 되도록 척도화됩니다. 값이 1 이면 완벽한 모형을 나타내고 0 이면 상수 모

형과 같음을 나타냅니다. 일반화 R^2 측도는 표준 최소 제곱 설정 시 연속형 정규 반응에 대한 기존 R^2 로 단순화됩니다. 일반화 R^2 을 Nagelkerke 또는 Craig 와 Uhler R^2 이라고도 하는데, 이는 Cox-Snell 유사 R^2 을 정규화한 것입니다. 자세한 내용은 Nagelkerke 연구 자료 (1991) 에서 확인하십시오.

평균 -Log p $-\log(p)$ 의 평균입니다. 여기서 p 는 발생한 사건과 연관된 적합 확률입니다.

RASE 제곱근 평균 제곱 예측 오차입니다. 범주형 반응의 경우 차이는 1 에서 p (실제로 발생한 반응 수준의 적합 확률) 사이입니다.

평균 절대 편차 반응과 예측 반응 사이의 차이에 대한 절대값의 평균입니다. 범주형 반응의 경우 차이는 1 에서 p (실제로 발생한 반응 수준의 적합 확률) 사이입니다.

오분류 비율 적합 확률이 가장 높은 반응 범주가 관측된 범주가 아닌 비율입니다.

N 관측값 수입니다.

관련 정보

"[훈련 및 검증 적합 측도](#)"에서는 범주형 반응의 적합 측도에 대한 자세한 정보를 제공합니다.

JMP PRO 모형 비교 플랫폼 옵션

"모형 비교"의 빨간색 삼각형 메뉴에 포함되는 옵션은 반응이 연속형인지 아니면 범주형인지에 따라 다릅니다.

- "연속형 반응"
- "범주형 반응"

JMP PRO 연속형 반응

반응이 연속형이면 "모형 비교"의 빨간색 삼각형 메뉴에 다음 옵션이 포함됩니다.

모형 평균화 전체 모형에서 예측값의 산술평균에 대한 새 열을 생성합니다.

실제값 대 예측값 그림 실제값 대 예측값의 산점도를 표시하거나 숨깁니다. 여러 모형에 대한 그림이 중첩됩니다.

잔차 대 행 번호 그림 잔차 대 행 번호 그림을 표시하거나 숨깁니다. 여러 모형에 대한 그림이 중첩됩니다.

프로파일러 데이터의 예측 계산식 열을 기반으로 각 반응에 대한 프로파일러를 표시하거나 숨깁니다. 프로파일러에는 비교되는 각 모형에 대한 행이 있습니다.

JMP PRO 범주형 반응

반응이 범주형이면 "모형 비교"의 빨간색 삼각형 메뉴에 다음 옵션이 포함됩니다.

모형 평균화 전체 모형에서 예측 확률의 산술평균에 대한 새 열을 생성합니다.

ROC 곡선 반응 변수의 각 수준에 대한 ROC 곡선을 표시하거나 숨깁니다. 여러 모형에 대한 곡선이 중첩됩니다.

AUC 비교 각 모형에서 ROC AUC(곡선 아래 면적) 비교를 표시하거나 숨깁니다. 곡선 아래 면적은 적합도를 나타내는 표시자이며 값이 1 이면 완벽한 적합을 나타냅니다.

이 보고서에는 다음과 같은 정보가 포함됩니다.

- 각 AUC 에 대한 표준 오차 및 신뢰 구간
- 각 AUC 쌍 간의 차이에 대한 표준 오차, 신뢰 구간 및 가설 검정
- 모든 AUC 가 동일한지 여부를 검정하기 위한 전체 가설 검정

향상도 곡선 반응 변수의 각 수준에 대한 향상도 곡선을 표시하거나 숨깁니다. 여러 모형에 대한 곡선이 중첩됩니다.

누적 이득 곡선 반응 변수의 각 수준에 대한 누적 이득 곡선을 표시하거나 숨깁니다. 누적 이득 곡선은 모든 반응의 비율 대비 모형에서 식별된 반응 수준의 비율을 보여 주는 그림입니다. 완벽한 모형의 누적 이득 곡선은 반응 수준의 전체 비율에서 1.0 에 도달합니다. 여러 모형에 대한 곡선이 중첩됩니다.

혼동 행렬 각 모형에 대한 혼동 행렬을 표시하거나 숨깁니다. 혼동 행렬은 실제 반응과 예측 반응의 이원 분류입니다. 개수 및 비율 혼동 행렬이 표시됩니다. 그룹 변수의 각 수준에 대해 별도의 혼동 행렬이 생성됩니다.

반응에 "수익 행렬" 열 특성이 있으면 "실제값 대 결정 개수" 및 "실제값 대 결정 비율" 행렬이 혼동 행렬 오른쪽에 표시됩니다. 이러한 행렬에 대한 자세한 내용은 ["파티션 플랫폼의 추가 예"](#) 에서 확인하십시오.

프로파일러 데이터의 예측 계산식 열을 기반으로 각 반응에 대한 프로파일러를 표시하거나 숨깁니다. 프로파일러에는 비교되는 각 모형에 대한 행이 있습니다.

결정 임계값 (이항 범주형 반응에만 사용 가능) 훈련, 검증 및 테스트 (지정된 경우) 데이터 집합에 대한 "결정 임계값" 보고서를 표시하거나 숨깁니다. 각 보고서에는 모형 적합 비교를 위한 각 모형의 적합 확률 분포 그래프, 혼동 행렬 및 분류 그래프와 분류 정확성 측정 기준 테이블이 포함되어 있습니다. 자세한 내용은 ["결정 임계값 보고서"](#) 에서 확인하십시오.

관련 정보

- ["ROC 곡선"](#)
- ["향상도 곡선"](#)

JMP PRO 모형 비교의 추가 예

이 예에서는 로지스틱 회귀 모형과 결정 트리 모형을 비교합니다. 자동차 데이터를 사용하며, 구매한 차량의 크기를 예측하는 것이 목표입니다.

도움말 > 샘플 데이터 폴더를 선택하고 Car Physical Data.jmp를 열어 시작합니다.

로지스틱 회귀 모형 생성

1. 분석 > 모형 적합을 선택합니다.
2. Type 을 선택하고 Y 를 클릭합니다.
3. Country, Weight, Turning Cycle, Displacement 및 Horsepower 열을 선택하고 추가를 클릭합니다.
4. 실행을 클릭합니다.
명목형 로지스틱 적합 보고서가 나타납니다.
5. 예측 계산식을 열에 저장하려면 "명목형 로지스틱"의 빨간색 삼각형을 클릭하고 확률 계산식 저장을 선택합니다.

결정 트리 모형 생성 및 열에 예측 계산식 저장

1. 분석 > 예측 모델링 > 파티션을 선택합니다.
2. Type 을 선택하고 Y, 반응을 클릭합니다.
3. Country, Weight, Turning Cycle, Displacement 및 Horsepower 열을 선택하고 X, 요인을 클릭합니다.
4. "방법" 목록에서 결정 트리가 선택되어 있는지 확인합니다.
5. 확인을 클릭합니다.
분할 보고서가 나타납니다.
6. Shift 키를 누른 채 분할을 클릭합니다.
7. "분할 수 입력" 옆에 "10" 을 입력합니다.
8. 확인을 클릭합니다.
9. 예측 계산식을 열에 저장하려면 "분할"의 빨간색 삼각형을 클릭하고 열 저장 > 예측 계산식 저장을 선택합니다.

모형 비교

1. 분석 > 예측 모델링 > 모형 비교를 선택합니다.
2. 확률로 시작하는 모든 열을 선택하고 Y, 예측 변수를 클릭합니다.
3. 확인을 클릭합니다.

그림 11.7 초기 모형 비교 보고서

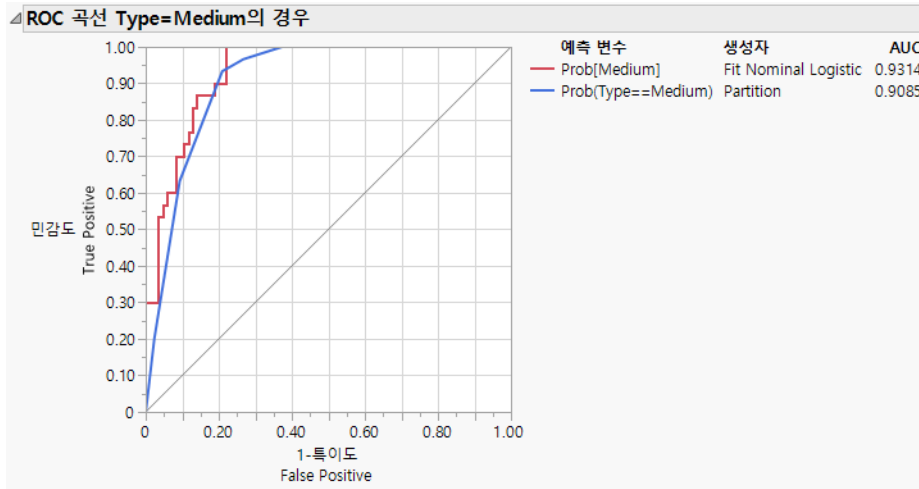
모형 비교								
예측 변수								
Type에 대한 적합 측도								
생성자	.2 .4 .6 .8	엔트로피 R ²	일반화 R ²	평균 -Log p	RASE	평균 절대 편차	오분류 비율	N
Fit Nominal Logistic		0.5821	0.8798	0.6656	0.4780	0.3900	0.3103	116
Partition		0.6248	0.9006	0.5976	0.4575	0.3986	0.2759	116

보고서에 따르면 파티션 모형의 엔트로피 R^2 및 일반화 R^2 값이 약간 더 높고 오분류 비율 값이 약간 더 낮습니다.

4. "모형 비교"의 빨간색 삼각형을 클릭하고 **ROC 곡선**을 선택합니다.

각 Type에 대한 ROC 곡선이 나타나며, 그 중 하나가 [그림 11.8](#)에 표시되어 있습니다.

그림 11.8 Medium에 대한 ROC 곡선



모든 ROC 곡선을 검토하면 두 모형의 예측 능력이 비슷하다는 것을 알 수 있습니다.

5. "모형 비교"의 빨간색 삼각형을 클릭하고 **AUC 비교**를 선택합니다.

각 Type에 대한 AUC 비교 보고서가 나타나며, 그 중 하나가 [그림 11.9](#)에 표시되어 있습니다.

그림 11.9 Medium에 대한 AUC 비교

AUC 비교 Type=Medium의 경우

예측 변수	AUC	표준 오차	95% 하한	95% 상한
Prob[Medium]	0.9314	0.0218	0.8742	0.9637
Prob(Type==Medium)	0.9085	0.0255	0.8448	0.9477

예측 변수	대 예측 변수	AUC 차이	표준 오차	95% 하한	95% 상한	카이제곱	Prob>ChiSq
Prob[Medium]	Prob(Type==Medium)	0.0229	0.0234	-0.023	0.0687	0.9553	0.3284

검정	카이제곱	DF	Prob>ChiSq
모든 AUC가 동일함	0.95534	1	0.3284

이 보고서에는 AUC 값(ROC 곡선 아래 면적)의 차이에 대한 가설 검정 결과가 표시됩니다. 결과를 검토하면 모든 Type 수준의 값 사이에 통계적 차이가 없음을 알 수 있습니다.

다음과 같은 이유로 두 모형의 예측 능력에 큰 차이가 없다는 결론을 내릴 수 있습니다.

- R^2 값 및 ROC 곡선이 비슷합니다.
- AUC 값 사이에 통계적으로 유의한 차이가 없습니다.

12 장

JMP[®] PRO 검증 열 생성

데이터를 훈련 , 검증 및 테스트 데이터 집합으로 나누기

검증 열 생성 플랫폼은 JMP Pro 에서만 사용할 수 있습니다 .

검증은 데이터 집합의 일부를 사용하여 모형 모수를 추정하고 다른 부분을 사용하여 모형의 예측 능력을 평가하는 프로세스입니다 . 데이터를 두 개 또는 세 개의 집합으로 분할하거나 , K 폴드 교차 검증을 위해 네 개 이상의 범주를 생성할 수 있습니다 . 검증 열 생성 플랫폼에서는 이러한 분할을 생성하는 여러 가지 방법을 제공합니다 .

목차

검증 열 생성 플랫폼 개요	183
검증 열 생성 플랫폼의 예	183
검증 열 생성 플랫폼 시작	185
검증 열 생성 보고서	188
비율 또는 상대 비율 지정	188
절단점 설정	188
폴드 수 설정	189
옵션	189
시작	189
검증 열 생성 플랫폼 옵션	189
검증 열 생성 플랫폼의 추가 예	190

JMP PRO 검증 열 생성 플랫폼 개요

검증은 데이터 집합의 일부를 사용하여 모형 모수를 추정하고 다른 부분을 사용하여 모형의 예측 능력을 평가하는 프로세스입니다. 복잡한 데이터의 경우 이 방법을 사용하면 모형 과대적합 위험을 줄일 수 있습니다.

검증 열의 한 가지 용도는 데이터를 두 개 또는 세 개의 부분으로 분할하는 것입니다.

- 훈련 데이터 집합은 모형 모수를 추정하는 데 사용됩니다.
- 검증 데이터 집합은 예측 능력이 좋은 모형을 선택하는 데 사용됩니다.
- 테스트 데이터 집합은 모형이 선택된 후 해당 모형의 예측 능력을 확인합니다.

검증 열의 또 다른 용도는 데이터를 K 폴드 교차 검증에 사용할 네 개 이상의 폴드로 분할하는 것입니다.

검증 열은 많은 JMP 플랫폼에서 검증 방법으로 사용될 수 있지만 검증 열을 통한 K 폴드 교차 검증은 일부 플랫폼에서만 지원됩니다. 자세한 내용은 "[JMP 모델링의 검증](#)"에서 확인하십시오.

검증 열 생성 플랫폼에서는 다양한 방법을 사용하여 훈련, 검증 및 테스트 데이터 집합을 생성할 수 있습니다. 층화, 그룹화 또는 절단점 열을 지정하여 검증 열 생성에 사용되는 방법을 결정할 수 있습니다.

JMP PRO 검증 열 생성 플랫폼의 예

검증 열 생성 플랫폼에서 층화 검증 열을 생성하여 훈련 데이터 집합과 검증 데이터 집합 전체에서 특정 변수의 수준 균형을 맞춥니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Lipid Data.jmp 를 엽니다.
2. **분석 > 분포**를 선택합니다.
3. Gender 를 선택하고 **Y, 열**을 클릭합니다.
4. **확인**을 클릭합니다.

그림 12.1 Lipid Data.jmp 의 Gender 분포

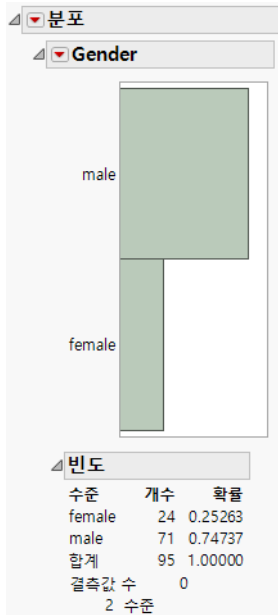


그림 12.1에서는 데이터 집합의 Gender 분포를 보여 줍니다. 남성과 여성의 비율이 같지 않다는 점에 유의하십시오. 데이터 내에 여성 수가 더 적으므로 검증 데이터 집합과 훈련 데이터 집합에서 성별 균형을 유지하려고 합니다.

5. **분석 > 예측 모델링 > 검증 열 생성**을 선택합니다.

6. Gender 를 선택하고 **증화 열**을 클릭합니다.

7. **확인**을 클릭합니다.

선택한 검증 방법에 대한 설명이 포함된 "검증 열 생성" 보고서가 나타납니다. 비율, 열 유형을 변경하거나 시드값을 설정하는 옵션도 있습니다.

8. (선택 사항) 보고서의 "옵션" 섹션에서 **난수 시드값** 옆에 "1234" 를 입력합니다.

9. **시작**을 클릭합니다.

"검증" 열이 데이터 테이블에 추가됩니다. 모자이크 그림을 생성하여 검증 데이터 집합과 훈련 데이터 집합의 분포를 탐색할 수 있습니다.

10. **분석 > X로 Y 적합**을 선택합니다.

11. 검증을 **Y, 반응**에 할당하고 Gender 를 **X, 요인**에 할당합니다.

12. **확인**을 클릭합니다.

그림 12.2 검증 데이터 집합과 훈련 데이터 집합의 Gender 분포

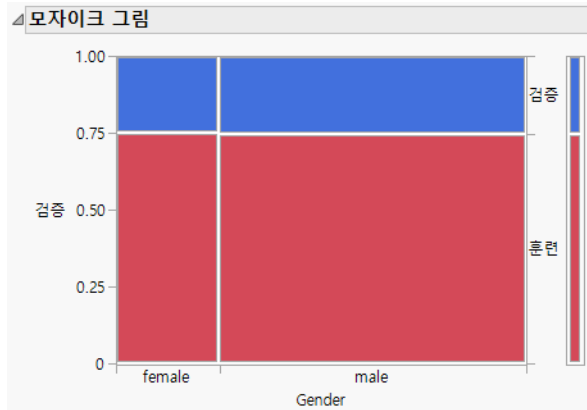
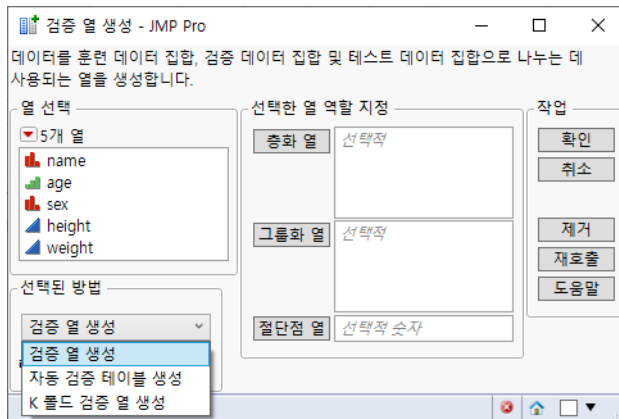


그림 12.2에서는 각 검증 데이터 집합과 훈련 데이터 집합에 대한 Gender 분포를 보여 줍니다. 여성과 남성의 약 75%가 훈련 데이터 집합에 있고 약 25%가 검증 데이터 집합에 있습니다.

JMP PRO 검증 열 생성 플랫폼 시작

분석 > 예측 모델링 > 검증 열 생성을 선택하여 검증 열 생성 플랫폼을 시작합니다.

그림 12.3 검증 열 생성 시작 창



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

검증 열 생성 시작 창에는 다음과 같은 옵션이 있습니다.

증화 열 하나 이상의 증화 열을 할당합니다.

그룹화 열 하나 이상의 범주형 그룹화 열을 할당합니다.

절단점 열 숫자 절단점 열을 할당합니다.

절단점 배치 ID 절단점 열이 할당되면 절단점 배치 ID 열도 할당할 수 있습니다. 이렇게 하면 "절단점 배치 ID" 열의 각 수준 내에서 절단점 값을 결정할 수 있습니다.

선택된 방법

세 가지 검증 방법을 제공합니다.

검증 열 생성 지정된 층화, 그룹화 및 절단점 열을 기반으로 검증 열을 생성합니다. 지정된 층화, 그룹화 및 절단점 열에 의해 결정된 검증 열 방법이 상자 아래에 설명됩니다. 방법을 선택하고 "확인"을 클릭한 후 "검증 열 생성" 보고서에서 각 집합에 대한 할당을 지정합니다. 자세한 내용은 "**비율 또는 상대 비율 지정**" 및 "**절단점 설정**"에서 확인하십시오. 홀드백 집합을 구성하는 방법에는 5 가지가 있습니다. 절단점 검증을 제외한 모든 방법은 K 폴드 검증을 위한 폴드를 생성하는 데도 사용됩니다. 자세한 내용은 "**K 폴드 검증 열 생성**"에서 확인하십시오.

랜덤 검증 열 시작 창에서 열을 할당하지 않은 경우 기본 방법입니다. 이 방법은 "검증 열 생성" 보고서에 입력한 할당 방식에 따라 데이터를 집합 또는 폴드로 분할합니다.

층화 검증 열 하나 이상의 층화 열을 할당한 경우 선택된 방법입니다. 이 방법은 지정된 층화 열의 수준에 따라 균일한 비율로 데이터 집합을 분할합니다. "랜덤 검증 열" 방법과 마찬가지로 "검증 열 생성" 보고서에 입력한 할당 방식에 따라 행이 홀드백 집합 또는 폴드에 랜덤 할당됩니다. 그러나 이 방법은 층화 열의 각 수준 또는 수준 조합에서 수행됩니다. 각 훈련, 검증 및 테스트 데이터 집합 또는 K 폴드 교차 검증의 각 폴드에서 열 수준을 균일한 비율로 나타내려면 이 방법을 사용합니다.

그룹화 검증 열 하나 이상의 그룹화 열을 지정한 경우 선택된 방법입니다. 이 방법은 지정된 열의 전체 수준 또는 둘 이상 열의 수준 조합이 동일한 집합이나 폴드에 배치되도록 데이터 집합을 분할합니다. 이 때문에 결과 집합의 크기는 사용자가 지정한 크기와 약간 다릅니다. 홀드백 집합 또는 폴드 간에 수준을 분할하는 것이 적절하지 않은 경우 이 옵션을 사용합니다.

그룹별 층화 검증 열 층화 열과 그룹화 열을 둘 다 지정한 경우 선택된 방법입니다. 이 방법은 층화 열 간에 수준 균형을 맞추면서 지정된 그룹이 동일한 홀드백 집합 또는 폴드에 함께 유지되는 방식으로 데이터를 분할합니다. "그룹화 검증 열"과 마찬가지로 그룹을 지정된 열의 수준 또는 둘 이상 열의 수준 조합으로 생성할 수 있습니다. 결과 집합의 크기는 사용자가 지정한 크기와 약간 다릅니다.

절단점 검증 열 절단점 열을 지정한 경우 선택된 방법입니다. 이 방법은 시계열 절단점에 따라 데이터 집합을 분할합니다. 기간에 따라 홀드백 집합에 데이터를 할당하려는 경우 이 옵션을 사용합니다. 훈련 데이터 집합은 첫 번째 절단점과 두 번째 절단점 사이의 행으로 구성됩니다. 검증 데이터 집합은 두 번째 절단점과 세 번째 절단점 사이의 행으로 구성됩니다. 테스트 데이터 집합은 나머지 행으로 구성됩니다. 이러한 집합은 "절단점 설정" 보고서의 옵션을 기반으로 선택됩니다.

자동 검증 테이블 생성 원래 데이터 테이블의 행에 연결된 원래 데이터 테이블의 복제 행을 포함하는 새 데이터 테이블을 생성합니다. 교차 검증에 사용할 수 있는 새 데이터 테이블에는 다음과 같은 네 개의 추가 열이 있습니다.

Valid Set 원래 데이터에 0 을 할당하고 복제 데이터에 1 을 할당합니다. 이 열의 값은 훈련 데이터 집합과 검증 데이터 집합을 지정합니다. 분석 시작 창의 " 검증 " 역할에 이 열을 사용합니다.

Valid ID 원래 관측값의 행 번호를 할당합니다. 이렇게 하면 각 원래 관측값에 대한 훈련 데이터 집합 행과 검증 데이터 집합 행을 일치시킬 수 있습니다.

Valid Weight 분석 시작 창의 " 빈도 " 역할에 사용할 자동 검증 가중치를 할당합니다. Valid ID의 각 값에 대해 훈련 관측값과 검증 관측값에 대한 동일한 균등 난수가 생성됩니다. 훈련 데이터 집합의 경우 Valid Weight 가 다음과 같이 계산됩니다.

$$\text{Valid Weight} = -\log(1 - \text{Valid Uniform})$$

검증 데이터 집합의 경우 Valid Weight 가 다음과 같이 계산됩니다.

$$\text{Valid Weight} = -\log(\text{Valid Uniform})$$

Valid Weight 열은 훈련 데이터 가중치가 검증 데이터 가중치와 음의 상관관계를 갖도록 구성됩니다. 이렇게 하면 검증 데이터의 적합 차이로 인해 적합 방법에 대한 효과적인 교차 검증이 가능합니다.

Null Factor Valid ID 의 각 값에 대해 동일한 정규 난수를 할당합니다.

팁 : 부분집합을 훈련 데이터로 사용하면 추정 문제가 발생할 수 있는 작은 데이터 테이블의 경우 " 자동 검증 테이블 생성 " 을 사용하십시오.

K 폴드 검증 열 생성 지정된 층화 및 그룹화 열을 기반으로 네 개 이상의 범주가 있는 검증 열을 생성합니다. 각 범주는 K 폴드 교차 검증에 사용할 폴드를 나타냅니다. Y 열은 행 순서를 지정하는 데 사용되며 각 행은 폴드에 순차적으로 할당됩니다. 지정된 층화 및 그룹화 열에 의해 결정된 검증 열 방법이 상자 아래에 설명됩니다. 이러한 방법은 " 검증 열 생성 " 에서 설명한 방법과 동일합니다. 방법을 선택하고 " 확인 " 을 클릭한 후 " 검증 열 생성 " 보고서에서 폴드 수 K 를 지정합니다. 자세한 내용은 " 폴드 수 설정 " 에서 확인하십시오.

결측값

층화, 그룹화 또는 절단점 열의 결측값은 해당 행에 대한 검증 열의 결측값이 됩니다.

JMP^{PRO} 검증 열 생성 보고서

"검증 열 생성" 보고서에는 시작 창에서 선택한 검증 방법에 대한 설명과 비율, 절단점, 열 유형 및 난수 시드값 설정을 위한 관련 옵션이 포함되어 있습니다.

JMP^{PRO} 비율 또는 상대 비율 지정

"검증 열 생성" 보고서의 이 섹션에서는 훈련, 검증 및 테스트 데이터 집합에 대한 할당을 지정할 수 있습니다. "훈련 데이터 집합", "검증 데이터 집합" 및 "테스트 데이터 집합" 옆의 상자에 각 집합에 포함할 행의 수 또는 비율을 나타내는 값을 입력합니다. 기본값은 행의 약 75%를 포함하는 훈련 데이터 집합과 약 25%를 포함하는 검증 데이터 집합을 구성합니다.

데이터 테이블의 행 수와 선택한 검증 방법에 따라 결과 집합의 실제 크기가 사용자가 지정한 크기와 약간 다를 수 있습니다. "조정 비율", "행 수" 및 "그룹 수" 열은 집합의 실제 크기를 제공합니다.

JMP^{PRO} 절단점 설정

"검증 열 생성" 보고서의 이 섹션에서는 절단점을 결정하는 방법을 지정할 수 있습니다. 이 섹션은 시작 창에서 절단점 열을 지정한 경우에만 나타납니다. 절단점은 다음 네 가지 방법 중 하나에 의해 결정됩니다.

비율 각 집합에 대해 지정된 행의 비율에 따라 절단점을 결정합니다. "훈련 데이터 집합", "검증 데이터 집합" 및 "테스트 데이터 집합" 옆의 상자에 각 집합에 포함할 비율을 나타내는 값을 입력합니다. 행 수와 비율에 따라 결과 집합의 실제 크기가 사용자가 지정한 크기와 약간 다를 수 있습니다.

행 수 각 집합에 대해 지정된 행 수에 따라 절단점을 결정합니다. "훈련 데이터 집합", "검증 데이터 집합" 및 "테스트 데이터 집합" 옆의 상자에 각 집합에 포함할 행 수를 나타내는 값을 입력합니다. 이 옵션을 사용하면 집합을 정확하게 지정할 수 있습니다.

고정 시간 또는 날짜 지정된 절단점 열의 고정 데이터 점에 따라 절단점을 결정합니다. 이 옵션을 선택하면 절단점 열의 최소값 및 최대값이 참조용으로 표시됩니다. "훈련 데이터 집합", "검증 데이터 집합" 및 "테스트 데이터 집합" 옆의 상자에 각 집합에 포함할 최소값을 나타내는 값을 입력합니다.

경과 시간 절단점 열의 첫 번째 타임스탬프 이후 경과한 시간을 기준으로 절단점을 결정합니다. 이 옵션을 선택하면 첫 번째 값에서 마지막 값까지 총 경과 시간이 참조용으로 표시됩니다. "훈련 데이터 집합", "검증 데이터 집합" 및 "테스트 데이터 집합" 옆의 상자에 각 집합에 포함할 경과 시간을 나타내는 값을 입력합니다.

JMP^{PRO} 폴드 수 설정

"검증 열 생성" 보고서의 이 섹션에서는 K 폴드 교차 검증의 폴드 수인 K 를 지정할 수 있습니다. K 의 기본값 및 최소값은 4이며, 최대값은 그룹화 열(지정된 경우)의 수준 수입니다. 수준이 4개 미만인 그룹화 열을 지정하면 오류 메시지가 표시됩니다.

JMP^{PRO} 옵션

"검증 열 생성" 보고서의 이 섹션에서는 다음과 같은 추가 옵션을 지정할 수 있습니다.

새 열 이름 검증 열 이름을 지정할 수 있습니다.

검증 열 유형 검증 열 유형을 "고정" 또는 "계산식"으로 지정할 수 있습니다.

- 고정 열 유형은 고정 값을 포함하는 열을 생성합니다.
- 계산식 열 유형은 계산식 열을 검증 열로 생성합니다.

난수 시드값 (계산식 열 유형을 지정한 경우에는 사용 불가능) 이후에 동일한 검증 열을 재현하기 위해 난수 시드값을 설정할 수 있습니다.

JMP^{PRO} 시작

"검증 열 생성" 보고서에서 **시작**을 클릭하면 검증 열이 데이터 테이블에 추가됩니다. 검증 열에는 층화, 그룹화 또는 절단점 변수를 제공하는 "노트" 특성이 있습니다.

JMP^{PRO} 검증 열 생성 플랫폼 옵션

"검증 열 생성"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

다음 옵션에 대한 자세한 내용은 JMP 사용의 에서 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.

JMP PRO 검증 열 생성 플랫폼의 추가 예

이 예에서는 두 개의 절단점 검증 열을 비교합니다. 첫 번째는 절단점 열만 사용하여 생성되고 두 번째는 절단점 열과 배치 ID 열을 모두 사용하여 생성됩니다. 데이터는 미국 전역의 16 개 기상 관측소에서 1 년 동안 수집한 주간 기상 데이터입니다. 모든 기상 관측소가 매주 (년 중)에 대한 주간 온도 측정값을 가지고 있는 것은 아닙니다. 각 기상 관측소에 대해 처음 60%의 관측값을 훈련에 사용하고, 다음 25% 관측값을 검증에 사용하고, 마지막 15% 관측값을 테스트에 사용하려고 합니다.

절단점을 사용하여 검증 열 생성

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Functional Data/Weekly Weather Data.jmp**를 엽니다.
2. **분석 > 예측 모델링 > 검증 열 생성**을 선택합니다.
3. **Week of Year**를 선택하고 **절단점 열**을 클릭합니다.
4. **확인**을 클릭합니다.
5. "절단점 결정 기준" 옆의 목록에서 **비율**을 선택합니다.
6. "훈련 데이터 집합", "검증 데이터 집합" 및 "테스트 데이터 집합" 옆의 상자에 각각 "0.60", "0.25", "0.15"를 입력합니다.
7. "새 열 이름" 옆의 상자에 "Cutpoint Validation"을 입력합니다.
8. **시작**을 클릭합니다.
"Cutpoint Validation"이라는 검증 열이 데이터 테이블에 추가됩니다.
9. **분석 > 테이블 생성**을 선택합니다.
10. ID를 클릭하고 "행에 대한 놓기 영역"으로 드래그합니다.
11. **Cutpoint Validation**을 클릭하고 N 위로 드래그합니다.
12. **행 %**를 클릭하고 셀 위로 드래그합니다.

그림 12.4 절단점 검증 열 비율

ID	Cutpoint Validation		
	훈련 행 %	검증 행 %	테스트 행 %
Bristol, TN	62.79%	20.93%	16.28%
Corpus Christi, TX	70.27%	18.92%	10.81%
Denver, CO	58.14%	23.26%	18.60%
Goodwell, OK	64.71%	17.65%	17.65%
Grand Canyon National Park, AZ	53.66%	26.83%	19.51%
Greenville, ME	60.47%	20.93%	18.60%
Harrisburg, PA	60.53%	26.32%	13.16%
Helena, MT	60.00%	25.00%	15.00%
Lynchburg, VA	71.79%	17.95%	10.26%
Madera, CA	67.50%	22.50%	10.00%
Miami Beach, FL	63.16%	26.32%	10.53%
Minneapolis/St. Paul, MN	57.50%	27.50%	15.00%
N. Myrtle Beach, SC	69.05%	16.67%	14.29%
Peason Ridge Range, LA	57.89%	28.95%	13.16%
Philip, SD	59.52%	23.81%	16.67%
Terra Haute, IN	47.22%	33.33%	19.44%

그림 12.4에서는 일부 기상 관측소에서 제공하는 훈련, 검증 및 테스트 데이터 집합 비율이 정확하지 않음을 보여 줍니다. 정확한 백분율을 얻으려면 배치 ID 열을 사용합니다.

절단점과 배치 ID 를 사용하여 검증 열 생성

1. 분석 > 예측 모델링 > 검증 열 생성을 선택합니다.
2. Week of Year 를 선택하고 절단점 열을 클릭합니다.
3. ID 를 선택하고 절단점 배치 ID 를 클릭합니다.
4. 확인을 클릭합니다.
5. "절단점 결정 기준" 옆의 목록에서 비율을 선택합니다.
6. "훈련 데이터 집합", "검증 데이터 집합" 및 "테스트 데이터 집합" 옆의 상자에 각각 "0.60", "0.25", "0.15" 를 입력합니다.
7. "새 열 이름" 옆의 상자에 "Cutpoint Batch Validation" 을 입력합니다.
8. 시작을 클릭합니다.
"Cutpoint Batch Validation" 이라는 검증 열이 데이터 테이블에 추가됩니다.
9. 분석 > 테이블 생성을 선택합니다.
10. ID 를 클릭하고 "행에 대한 놓기 영역" 으로 드래그합니다.
11. Cutpoint Batch Validation 을 클릭하고 N 위로 드래그합니다.
12. 행 % 를 클릭하고 셀 위로 드래그합니다.

그림 12.5 배치 ID 비율을 사용한 절단점 검증 열

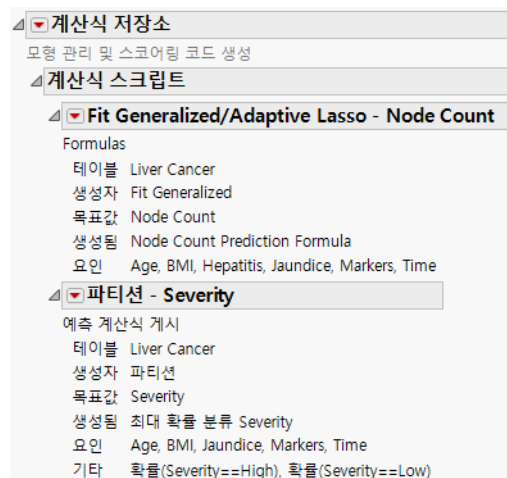
ID	Cutpoint Batch Validation		
	훈련 행 %	검증 행 %	테스트 행 %
Bristol, TN	60.47%	25.58%	13.95%
Corpus Christi, TX	59.46%	24.32%	16.22%
Denver, CO	60.47%	25.58%	13.95%
Goodwell, OK	58.82%	26.47%	14.71%
Grand Canyon National Park, AZ	60.98%	24.39%	14.63%
Greenville, ME	60.47%	25.58%	13.95%
Harrisburg, PA	60.53%	26.32%	13.16%
Helena, MT	60.00%	25.00%	15.00%
Lynchburg, VA	58.97%	25.64%	15.38%
Madera, CA	60.00%	25.00%	15.00%
Miami Beach, FL	60.53%	26.32%	13.16%
Minneapolis/St. Paul, MN	60.00%	25.00%	15.00%
N. Myrtle Beach, SC	59.52%	26.19%	14.29%
Peason Ridge Range, LA	60.53%	26.32%	13.16%
Philip, SD	59.52%	26.19%	14.29%
Terra Haute, IN	61.11%	25.00%	13.89%

그림 12.5에서는 절단점 배치 ID 열을 사용하면 각 기상 관측소의 훈련, 검증 및 테스트 데이터 집합에 대한 비율이 지정된 값에 훨씬 가깝다는 것을 보여 줍니다.

계산식 저장소 플랫폼은 JMP Pro에서만 사용할 수 있습니다.

계산식 저장소는 모형을 구성, 비교 및 프로파일링하기 위한 저장소입니다. 계산식 저장소에 게시된 모형에 대해 JMP 내부 또는 외부 배포를 위한 스코어링 코드를 생성할 수 있습니다. 모형 탐색 작업의 경우 계산식 저장소를 사용하여 후보 모형을 JMP 데이터 테이블 외부에 저장할 수 있습니다. 계산식 저장소에서 모형 프로파일러 및 모형 비교 플랫폼에 액세스할 수 있습니다. 나중에 사용하기 위해 선택한 모형을 현재 JMP 테이블에 저장하거나 스코어링할 새 데이터와 함께 JMP 테이블에 저장할 수 있습니다. C, Python, JavaScript, SAS 또는 SQL로 스코어링 코드를 생성하면 JMP 외부 환경에서 사용할 수 있습니다.

그림 13.1 계산식 저장소의 예



목차

계산식 저장소 플랫폼 개요.....	195
계산식 저장소의 예	195
계산식 저장소 플랫폼 시작.....	196
예측 계산식을 계산식 저장소에 게시하는 플랫폼.....	196
계산식 저장소 플랫폼 옵션.....	197
계산식 저장소 모형 옵션.....	199
계산식 저장소 플랫폼에서 스코어링 코드 생성.....	199

JMP PRO 계산식 저장소 플랫폼 개요

계산식 저장소는 배포할 모형을 구성, 비교, 프로파일링 및 스코어링하기 위한 저장소입니다. 모형은 예측 계산식입니다. 예측 계산식은 계산식 저장소에 열 스크립트로 저장됩니다. 예측 계산식을 JMP 테이블에 추가하여 데이터를 스코어링할 수 있습니다. 또한 계산식 저장소를 사용하여 예측 계산식에 대한 스코어링 코드를 생성하면 JMP 외부 환경에서 모형을 쉽게 배포할 수 있습니다.

계산식 저장소를 사용하면 다음 작업을 수행할 수 있습니다.

- 예측 계산식을 데이터 테이블이 아닌 다른 곳에 저장
- 여러 데이터 테이블의 예측 계산식을 공통 위치에 저장
- 예측 계산식 내에 중간 정리 계산식 저장
- 모형 비교
- 모형 프로파일링
- JMP 에서 새 데이터를 스코어링하기 위해 데이터 테이블에 예측 계산식 추가
- JMP 외부에 모형을 배포하기 위한 스코어링 코드 (C, Python, JavaScript, SAS 또는 SQL) 생성
- 계산식 저장소는 생성된 스코어링 코드에 중간 계산식을 포함합니다. 열 계산식이 있는 입력을 통해 생성된 모형은 해당 계산식을 스코어링 코드에 포함합니다. 예를 들어 모형에 재코딩된 변수가 있는 경우 재코딩 계산식이 스코어링 코드에 포함됩니다. 이렇게 하면 처리되지 않은 원시 데이터에 예측 계산식을 적용할 수 있습니다.

참고: 결측값 정보화는 지원되는 중간 계산식이 아닙니다.

JMP PRO 계산식 저장소의 예

이 예에서는 샘플 데이터 테이블의 기존 스크립트를 사용하여 모형을 생성한 후 모형의 예측 계산식을 계산식 저장소에 게시합니다. 데이터 테이블에는 연구에 참여한 환자의 간암 중증도에 대한 데이터가 포함되어 있습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Liver Cancer.jmp** 를 엽니다.
2. "Lasso Poisson, Validation Column" 스크립트 옆의 녹색 삼각형을 클릭합니다.
3. "검증 열을 사용한 Poisson 적응형 Lasso 회귀" 옆의 빨간색 삼각형을 클릭하고 **열 저장 > 예측 계산식 게시**를 선택합니다.

이 옵션을 선택하면 일반화 적합 모형에 대한 예측 계산식이 포함된 계산식 저장소가 열립니다.

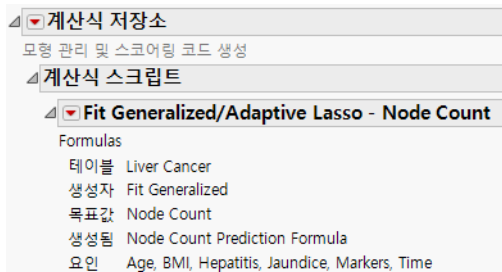
4. 예측 계산식을 데이터 테이블에 추가하려면 계산식 저장소의 "일반화 적합 - Node Count" 빨간색 삼각형에서 **스크립트 실행**을 선택합니다.

참고 : 3 단계에서 열 저장 > 예측 계산식 저장을 사용하면 3 단계와 4 단계의 결과를 얻을 수 있습니다. 그러나 이 경우에는 예측 계산식이 계산식 저장소에 포함되지 않습니다.

- JMP 외부에서 사용할 스코어링 코드를 생성하려면 "일반화 적합 - Node Count"의 빨간색 삼각형에서 코드 유형을 선택합니다. 코드가 포함된 스크립트 창이 나타납니다. 자세한 내용은 "계산식 저장소 플랫폼에서 스코어링 코드 생성"에서 확인하십시오.
- 계산식 저장소를 저장하려면 파일 > 저장을 선택합니다.

참고 : 계산식 저장소는 JMP 보고서 파일 (*.jrp)로 저장됩니다.

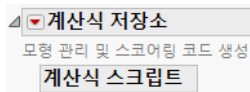
그림 13.2 일반화 모형이 있는 계산식 저장소



JMP PRO 계산식 저장소 플랫폼 시작

분석 > 예측 모델링 > 계산식 저장소를 선택하여 계산식 저장소를 시작합니다.

그림 13.3 시작 창의 빈 계산식 저장소



또는 열려 있는 계산식 저장소가 없는 경우 "게시" 명령을 선택하면 계산식 저장소가 열립니다.

JMP PRO 예측 계산식을 계산식 저장소에 게시하는 플랫폼

예측 계산식을 게시하고 스코어링 코드를 생성하는 플랫폼은 다음과 같습니다.

- 판별
- 최소 제곱 회귀
- 로지스틱 회귀
- 파티션
- Uplift

- K 최근접 이웃
- Naive Bayes
- 신경망
- 잠재 계층 분석
- 주성분
- 일반화 회귀
- PLS
- 가우스 과정

예측 계산식을 계산식 저장소에 게시하지 않는 플랫폼에서는 예측 계산식을 데이터 테이블에 저장할 수 있습니다. 데이터 테이블에서 **열에서 계산식 추가**를 선택하여 예측 계산식을 계산식 저장소에 추가합니다. 그러나 스코어링 코드는 이러한 모형에 대해 올바르게 작동하지 않을 수 있습니다. 스코어링 코드에 대한 자세한 내용은 "[계산식 저장소 플랫폼에서 스코어링 코드 생성](#)"에서 확인하십시오.

또한 계산식 편집기에서 직접 계산식을 게시할 수 있습니다. 계산식을 선택하고 마우스 오른쪽 버튼을 클릭한 후 **계산식 저장소에 게시**를 선택합니다.

JMP PRO 계산식 저장소 플랫폼 옵션

"계산식 저장소"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

열에서 계산식 추가 데이터 테이블 열의 기존 예측 계산식을 현재 계산식 저장소에 추가할 수 있습니다. 테이블에 계산식이 하나뿐인 경우 해당 계산식이 저장소에 추가됩니다. 열 계산식이 여러 개인 경우 대화상자에서 원하는 계산식을 선택할 수 있습니다. 데이터 테이블에서 계산식 열이 선택되면 대화상자에서 해당 열이 미리 선택됩니다.

스크립트 표시 현재 계산식 저장소에 있는 모든 계산식에 대한 스크립트가 포함된 새 "계산식 창"을 열거나, 열려 있는 "계산식 창"에 추가합니다.

스크립트 복사 현재 계산식 저장소의 모든 스크립트를 클립보드에 복사합니다.

계산식을 함수로 복사 현재 계산식 저장소에서 모형을 선택하여 JSL에서 사용할 함수로 복사할 수 있습니다. 선택한 모형은 클립보드에 복사됩니다. 함수는 열 대신 스칼라 변수에 대해 작동합니다. 선택한 각 모형에 대해 하나의 네임스페이스가 생성됩니다. 각 네임스페이스에는 모형 입력을 모수로 사용하고 모형 예측 변수를 반환하는 `predict` 함수 하나가 있습니다.

계산식을 변환으로 복사 현재 계산식 저장소에서 복사할 모형을 선택할 수 있습니다. 선택한 모형은 클립보드의 `Transform Columns()` 문 내에 복사됩니다.

스크립트 실행 현재 계산식 저장소의 모형을 JMP 데이터 테이블의 새 열에 저장할 수 있습니다. 중간 열 계산식이 있는 계산식은 중간 열을 JMP 데이터 테이블에 추가합니다.

참고 : 데이터 테이블이 두 개 이상 열려 있으면 "스크립트 실행"에 사용할 데이터 테이블을 지정할 수 있는 "테이블 선택" 창이 나타납니다.

C 코드 생성, Python 코드 생성, JavaScript 코드 생성, SAS 코드 생성 및 SQL 코드 생성 현재 계산식 저장소에서 코드 생성을 위한 모형을 선택할 수 있습니다. 선택한 모형에 대한 각 스코어링 코드 (C, Python, JavaScript, SAS DS2 또는 SQL 형식)가 포함된 새 스크립트 창이 나타납니다. 이 코드를 사용하면 선택한 환경 또는 프레임워크에서 모형을 쉽게 배포할 수 있습니다. 자세한 내용은 "[계산식 저장소 플랫폼에서 스코어링 코드 생성](#)"에서 확인하십시오.

모형 비교 현재 계산식 저장소에서 모형 비교 유틸리티를 사용하여 비교할 모형을 선택할 수 있습니다. 여러 테이블의 모형이 저장소에 저장된 경우 먼저 관심 테이블을 선택한 후 관심 모형을 선택합니다. Shift 키를 누른 채 "모형 비교" 옵션을 선택하면 새 창에서 모형 비교가 시작됩니다. 자세한 내용은 "[모형 비교](#)"에서 확인하십시오.

참고 : 데이터 테이블이 두 개 이상 열려 있으면 "모형 비교"에 사용할 데이터 테이블을 지정할 수 있는 "테이블 선택" 창이 나타납니다.

모형 비교 제거 모든 모형 비교 보고서를 현재 계산식 저장소에서 제거합니다.

프로파일러 현재 계산식 저장소에서 프로파일러를 사용하여 프로파일링할 모형을 선택할 수 있습니다. 여러 테이블의 모형이 저장소에 저장된 경우 먼저 관심 테이블을 선택한 후 관심 모형을 선택합니다. Shift 키를 누른 채 "프로파일러" 옵션을 선택하면 새 창에서 프로파일러가 시작됩니다. 자세한 내용은 **Profilers**에서 확인하십시오.

참고 : 데이터 테이블이 두 개 이상 열려 있으면 "프로파일러"에 사용할 데이터 테이블을 지정할 수 있는 "테이블 선택" 창이 나타납니다.

프로파일러 제거 모든 프로파일러를 현재 계산식 저장소에서 제거합니다.

다음 옵션에 대한 자세한 내용은 **JMP 사용**의 에서 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.

JMP PRO 계산식 저장소 모형 옵션

계산식 저장소에 저장된 각 예측 계산식에는 개별 옵션 메뉴가 있습니다. 메인 메뉴 옵션에 해당하는 표시, 복사, 실행 및 코드 생성 옵션 외에 다음과 같은 개별 메뉴 옵션이 포함되어 있습니다.

새 열 이름 바꾸기 모형 이름을 바꿀 수 있습니다. 스크립트를 실행하여 예측 계산식을 JMP 데이터 테이블에 추가하면 이 새 이름이 열 이름으로 적용됩니다. 또한 생성된 코드에도 이 이름이 포함됩니다.

제거 계산식 저장소에서 모형을 제거합니다. 이 명령은 실행 취소할 수 없습니다.

JMP PRO 계산식 저장소 플랫폼에서 스코어링 코드 생성

스코어링 코드를 생성하는 목적은 JMP에서 생성된 모형을 프로덕션 환경 또는 선택한 기타 프레임워크에서 쉽게 사용하기 위해서입니다. 많은 플랫폼에서 예측 계산식을 계산식 저장소에 게시하지만 모든 예측 계산식이 전체 코드를 생성하는 것은 아닙니다. 코드는 전체 코드이거나 코드 조각일 수 있으며, 구현을 위해 추가 프로그래밍이 필요한 지원되지 않는 함수를 포함할 수도 있습니다. 대상 언어로 추가 프로그래밍이 필요한 지원되지 않는 함수 호출을 나타내는 오류 메시지가 표시되는 경우가 종종 있습니다. 좀 더 복잡한 함수를 구현하기 위해 대상 언어별 라이브러리가 필요할 수도 있습니다. 이러한 라이브러리는 아래의 언어별 섹션에서 설명합니다. 예를 들어 다음과 같은 작업을 수행할 수 있습니다.

- 생성된 SAS 코드를 사용하여 모형을 SAS 모형 관리자에 배포합니다.
- 생성된 SQL 코드를 사용하여 데이터베이스 내 스코어링으로 ETL 프로세스를 확대합니다.
- 생성된 C 코드로 빌드된 응용 프로그램을 사용하여 데이터 변환 파이프라인에 대한 노드를 생성합니다.
- 생성된 Python 코드를 사용하여 실시간 스코어링 결과를 표시할 Jupyter 노트를 생성합니다.
- 생성된 JavaScript를 포함하는 웹 응용 프로그램으로 고객이 자신의 데이터를 직접 스코어링할 수 있습니다.

C, Python 및 JavaScript 언어의 경우 생성된 코드를 배포하거나 컴파일할 때 .h 파일 및 유틸리티 라이브러리와 같은 지원 코드를 포함해야 합니다. 이러한 파일은 JMP 설치 폴더 아래의 Scoring 폴더에 있습니다.

팁: 사용자 컴퓨터에서 설치 폴더 위치를 찾으려면 JSL 명령 `Get Path Variable("$ALL_HOME/Scoring")`을 사용하십시오.

C 코드 생성된 C 스코어링 코드는 라이브러리로 컴파일된 후 응용 프로그램에 연결되어야 합니다. 정적 연결 또는 동적 연결 방식을 사용할 수 있습니다. 컴파일 및 연결에 필요한 `jmp_lib.h`, `jmp_parms.h`, `jmp_score.h` 파일은 JMP 설치 폴더 아래의 Scoring/C 폴더에서 찾을 수 있습니다.

Python 코드 Python 스코어링 응용 프로그램을 실행하는 데 필요한 `jmp_score.py` 파일은 JMP 설치 폴더 아래의 **Scoring/Python** 폴더에서 찾을 수 있습니다. 모형을 배포할 때 `jmp_score.py`를 실행 가능한 스코어링 코드와 동일한 위치에 복사합니다.

생성된 코드를 호출하여 데이터를 스코어링하는 경우 입력 데이터가 JMP에서 생성된 코드와 호환되는 데이터 구조로 전달되어야 합니다. Python 스코어링 코드는 입력 및 출력 인수가 사전과 유사한 유형이어야 합니다. Spark, Dask 및 Pandas 프레임워크의 **DataFrame** 행 개체 또는 표준 Python "dict" 클래스와 같은 키를 사용하여 값에 액세스하고 쓸 수 있습니다.

선형 대수 연산자 또는 NumPy Python 라이브러리가 필요한 고급 연산자(예: **Vec Quadratic**, **Design Norm**)를 사용합니다.

JavaScript 코드 JavaScript 스코어링 응용 프로그램을 실행하는 데 필요한 `jmp_score.js` 파일은 **Scoring/JavaScript** 폴더에서 찾을 수 있습니다.

SAS 코드 PROC DS2 문으로 래핑한 생성된 코드 조각은 SAS In-Database Code Accelerator를 포함한 SAS 응용 프로그램에서 사용할 수 있습니다. 코드는 SAS 창에 추가되며 코드 앞의 주석 블록에 변수 이름 매핑 섹션을 포함합니다. 중간 열과 관련된 임시 변수 이름은 drop 명령을 통해 DATA 스텝 끝에서 삭제됩니다.

팁 : 범주형 반응에 대한 로지스틱 또는 신경망 모형과 같이 ifmax 호출이 있는 모형의 경우 임시 변수 선언을 **method run()** 문 앞으로 이동하십시오.

SQL 코드 select 문으로 래핑된 SQL 코드 조각은 대부분의 주 데이터베이스 서버에 대한 SQL 쿼리에 사용할 수 있습니다.

참고 : " 자리 표시자 " 또는 "ERROR"가 포함된 코드는 지원되지 않는 함수 호출을 나타냅니다.

14 장

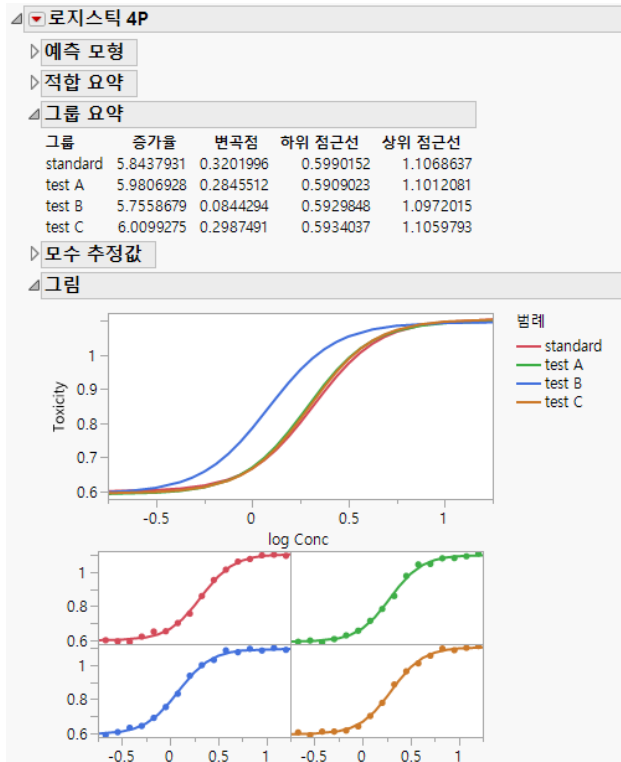
곡선 적합

기본 제공 비선형 모델을 데이터에 적합

곡선 적합 플랫폼을 사용하면 모수에 대해 비선형인 모델을 적합시킬 수 있습니다. 많은 경우, 특히 물리학과 생물학에서 잘 알려진 비선형 방정식은 변수 간의 관계를 설명합니다. 예를 들어 약물 생리학 실험에서는 약물에 대한 반응 강도가 어떻게 변하는지를 약물 농도 함수로 나타낼 수 있습니다. 흔히 시그모이드 곡선은 반응 강도를 약물 농도 함수로 정확하게 모델링합니다. 또한 시간 경과에 따른 인구 규모를 모델링할 수 있는 지수 성장 곡선도 예로 들 수 있습니다.

곡선 적합 플랫폼에서는 다항식, 로지스틱, 프로빗, Gompertz, 지수, 정상점, 약동학, 비율 및 용해 모형과 같은 사전 정의된 모형을 제공합니다. 미리 정의된 모형을 사용하면 모형 계산식을 생성하거나 모수 추정값의 시작 값을 지정할 필요가 없습니다. 시작 값을 지정하고 모형 계산식을 생성하려면 지정된 비선형 모델을 적합시킬 수 있는 비선형 플랫폼을 사용합니다. 자세한 내용은 "[비선형 회귀](#)"에서 확인하십시오.

그림 14.1 곡선 적합 플랫폼에서 비선형 적합의 예



목차

곡선 적합 플랫폼 개요.....	203
곡선 적합 플랫폼의 예.....	203
곡선 적합 플랫폼 시작.....	206
곡선 적합 보고서.....	208
모형 비교 보고서.....	208
모형 적합 보고서.....	209
곡선 적합 옵션.....	210
모형 없는 비교 보고서.....	212
모형 적합 옵션.....	213
병렬 검증.....	215
모수 추정값 비교.....	217
동등성 검증.....	218
곡선 적합 플랫폼의 추가 예.....	219
곡선 적합 플랫폼에 대한 통계 상세 정보.....	221
곡선 적합 모형에 대한 통계 상세 정보.....	222
용해 곡선 분석에 대한 통계 상세 정보.....	232

곡선 적합 플랫폼 개요

일부 모형은 모수에 대해 선형(예: 2차 또는 기타 다항식)이고, 다른 모형은 선형이 되도록 변환(예: x 의 로그 변환을 사용하는 경우)할 수 있습니다. 이러한 상황에서는 모형 적합 또는 X로 Y 적합 플랫폼이 더 적절합니다. 선형 모형 적합에 나오는 예에서는 산소 섭취량과 달리기 시간 사이의 유의한 선형 관계를 보여 줍니다. 모형 적합에 대한 자세한 내용은 선형 모형 적합에서 확인하십시오. X로 Y 적합에 대한 자세한 내용은 기본 분석에서 확인하십시오.

곡선 적합 플랫폼을 사용하면 모수에 대해 비선형인 모형을 적합시킬 수 있습니다. 이 장의 처음에 나오는 예에서는 비선형 관계 분석(농도 함수로 나타낸 약물 독성)을 보여 줍니다. 독성에 대한 농도 효과는 낮은 투여량에서 높은 투여량으로 변하므로 이 관계는 비선형입니다.

다음은 선형 및 비선형 함수에 대한 방정식 예입니다.

$$\text{선형 함수: } Y = \beta_0 + \beta_1 e^x$$

$$\text{비선형 함수: } Y = \beta_0 + \beta_1 e^{\beta_2 x}$$

곡선 적합 플랫폼에서는 다항식, 로지스틱, 프로빗, Gompertz, 지수, 정상점, 약동학, 비율 및 용해 모형과 같은 사전 정의된 모형을 제공합니다. 그룹화 변수를 지정하면 그룹화 변수의 각 수준에 대해 개별 모형 모수를 추정할 수 있습니다. 적합 모형과 추정 모수는 그룹화 변수의 전체 수준에서 비교할 수 있습니다. 모수 및 비모수 기법 둘 다에 대한 옵션과 함께 용해 프로파일을 비교하는 옵션 집합도 있습니다.

곡선 적합을 사용하면 예측 계산식을 생성하기 위한 모형을 만들 수도 있습니다. 그런 다음 비선형의 모수 상한과 하한을 설정합니다. 자세한 내용은 "[모수 한계 설정의 예](#)"에서 확인하십시오. 또한 JMP Pro에서는 곡선 적합 플랫폼 내에서 보조 변수 집합을 지정하고 일반화 회귀 모형을 적합시켜 이러한 변수가 반응에 어떤 영향을 미치는지 확인할 수 있습니다.

곡선 적합 플랫폼의 예

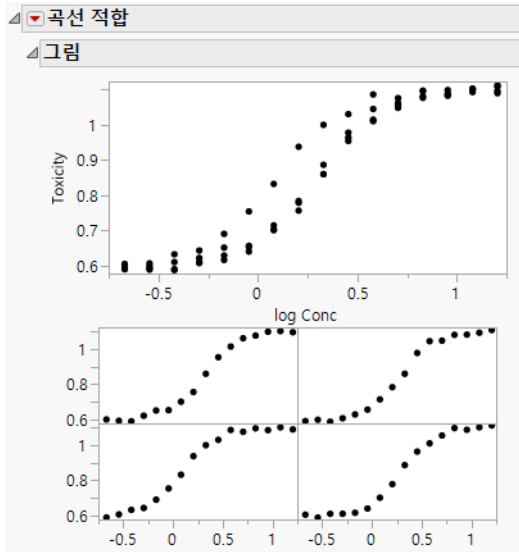
이 예에서는 약물 농도의 함수로 나타낸 독성에 대한 모형을 생성하는 방법을 보여 줍니다. standard 약물 제제가 있고 이를 세 개의 새로운 제제와 비교하려고 합니다.

각 약물의 특정 농도에서 생존 세포에 대한 비생존 세포의 독성 비율에 관심이 있습니다. 사전 조사를 통해 각 약물 제제의 16개 농도에 대한 독성 비율을 알고 있습니다. 비율이 높으면 독성이 더 많음을 나타내며 이는 약물 개발에 해로울 수 있습니다. 농도 값 범위를 줄이고 곡선의 차이를 더 쉽게 감지할 수 있도록 로그 농도가 계산되었습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Nonlinear Examples/Bioassay.jmp 를 엽니다.
2. **분석 > 전문 모델링 > 곡선 적합** 을 선택합니다.
3. Toxicity 를 Y, 반응 역할에 할당합니다.
4. log Conc 를 X, 회귀변수 역할에 할당합니다.

5. Formulation 을 그룹 역할에 할당합니다 .
6. 확인을 클릭합니다 .

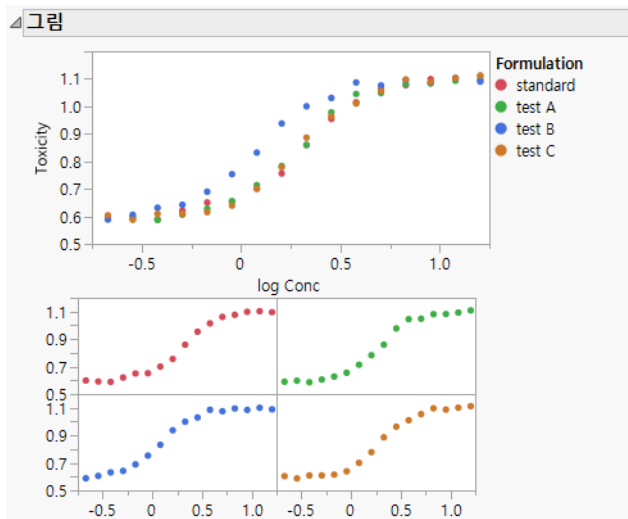
그림 14.2 초기 곡선 적합 보고서



" 그림 " 보고서에는 각 제제의 적합 모형에 대한 중첩 그림이 포함되어 있습니다 .

7. 각 약물 제제를 식별하는 범례를 표시하려면 그래프 중 하나를 마우스 오른쪽 버튼으로 클릭하고 **행 범례**를 선택합니다 . **Formulation** 열을 선택하고 **확인**을 클릭합니다 .

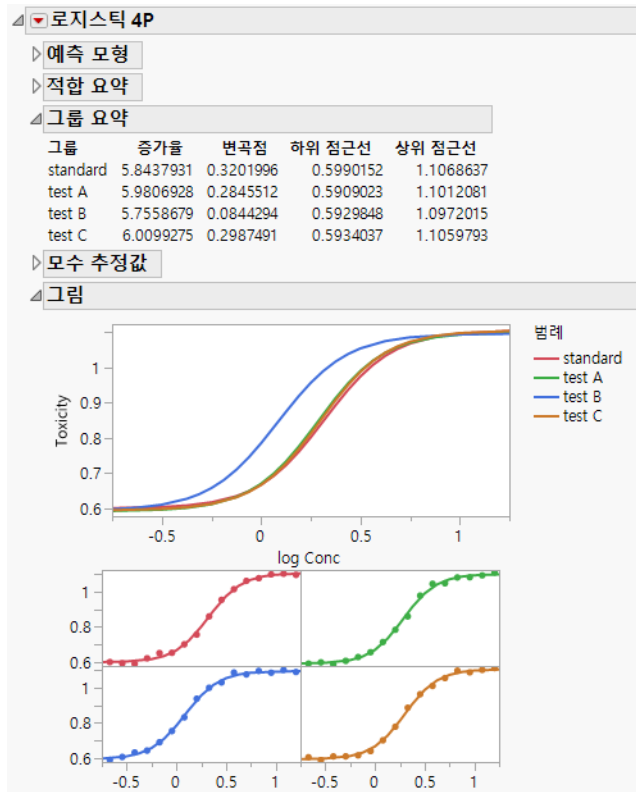
그림 14.3 그림 범례가 포함된 곡선 적합 보고서



곡선이 S 형태로 나타나므로 시그모이드 곡선이 적합할 수 있습니다. 표 14.1에서는 곡선 적합 플랫폼에서 제공하는 여러 유형의 모형에 대한 계산식과 그래픽 표현을 보여 줍니다.

- "곡선 적합"의 빨간색 삼각형을 클릭하고 **시그모이드 곡선 > 로지스틱 곡선 > 로지스틱 4P** 적합을 선택합니다.

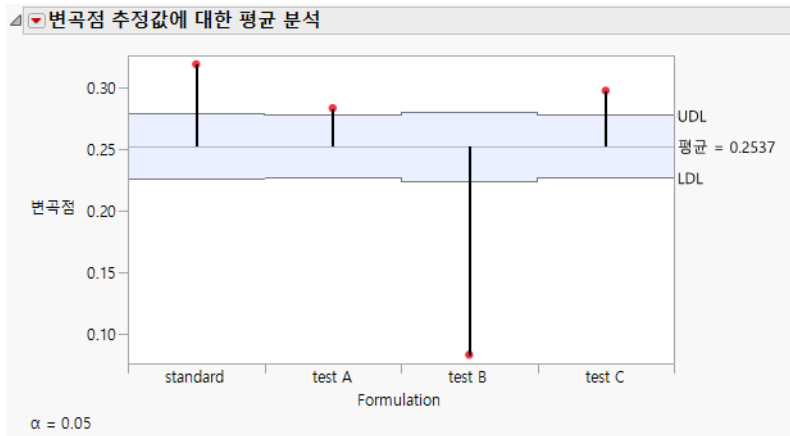
그림 14.4 로지스틱 4P 보고서



"로지스틱 4P" 보고서가 나타납니다. 각 약물 제제에 대한 별도의 그림도 있습니다. 적합 곡선 그림을 보면 test B 곡선이 다른 곡선에 비해 빠르게 상승하므로 B 제제가 다를 수 있음을 알 수 있습니다. 변곡점 모수로 인해 상승이 발생합니다.

- "로지스틱 4P"의 빨간색 삼각형을 클릭하고 **모수 추정값 비교**를 선택합니다.

그림 14.5 모수 비교 보고서



test B 제제에 대한 변곡점 모수가 평균 변곡점보다 유의하게 낮습니다. 이는 [그림 14.4](#)에 표시된 그림과 일치합니다. 약물 제제 B는 다른 제제보다 독성 비율이 더 높습니다.

곡선 적합 플랫폼 시작

곡선 적합 플랫폼을 시작하려면 **분석 > 전문 모델링 > 곡선 적합**을 선택합니다.

그림 14.6 곡선 적합 플랫폼 시작 창

읽인 데이터 형식 함수 열 함수 열

읽인 데이터 형식입니다.

열 선택

4개 열

- Concentration
- log Conc
- Formulation
- Toxicity

선택한 열 역할 지정

Y, 반응	필수 숫자
X, 회귀변수	필수 숫자
그룹	선택적
Z, 보조	선택적
가중치	선택적 숫자
빈도	선택적 숫자
기준	선택적

작업

확인 취소 제거 재호출 도움말

Detailed description: This is the 'Curve Fit Platform Start Window'. It shows a list of variables on the left: Concentration, log Conc, Formulation, and Toxicity. The 'Formulation' variable is highlighted in red. In the center, there is a table to assign roles to the selected variables. The roles assigned are: Y (Response) - Required Number, X (Regression Variable) - Required Number, Group - Optional, Z (Auxiliary) - Optional, Weight - Optional Number, Frequency - Optional Number, and Criterion - Optional. On the right, there are buttons for '확인' (OK), '취소' (Cancel), '제거' (Remove), '재호출' (Re-call), and '도움말' (Help).

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

데이터 형식 선택

곡선 적합 시작 창에는 다양한 유형의 데이터 형식 탭이 있습니다. 사용하는 데이터 형식에 따라 탭을 선택합니다.

쌓인 데이터 형식 데이터 테이블의 각 행이 단일 관측값에 해당하는 경우 선택합니다. 반응, 회귀변수 및 그룹 변수에 대해 각각 별도의 열이 있습니다.

함수 행 데이터 테이블의 각 행이 그룹 변수의 한 수준에 대한 전체 반응 함수에 해당하는 경우 선택합니다. 각 열은 회귀변수의 수준에 해당합니다.

팁 : 열에 FDE X 열 특성이 포함된 경우 이 열 특성은 x 값을 지정하는 데 사용됩니다. 열에 열 특성이 포함되어 있지 않으면 열 이름에 지정된 x 값의 수준이 "함수 행" 데이터 형식에 사용됩니다. 열 이름이 너무 복잡하면 문제가 있는 열을 나타내는 오류 메시지가 표시되고 x 값은 $x=1, 2, \dots, n$ 으로 지정됩니다. 여기서 n 은 관측값 수입니다.

함수 열 데이터 테이블의 각 열이 그룹 변수의 한 수준에 대한 전체 반응 함수에 해당하는 경우 선택합니다. 각 행은 회귀변수의 수준에 해당합니다.

열 역할 지정

시작 창에는 열 역할을 지정하기 위한 다음 옵션이 포함되어 있습니다.

Y, 반응 Y 변수를 지정합니다.

X, 회귀변수 ("쌓인 데이터 형식" 및 "함수 열" 탭에만 사용 가능) X 변수를 지정합니다.

그룹 ("쌓인 데이터 형식" 및 "함수 행" 탭에만 사용 가능) 그룹화 변수를 지정합니다. 적합 모형에는 그룹화 변수의 각 수준에 대한 개별 모수가 있습니다. 이를 통해 그룹화 변수의 전체 수준에서 적합 모형과 추정 모수를 비교할 수 있습니다.

Z, 보조 ("쌓인 데이터 형식" 및 "함수 행" 탭에만 사용 가능) 하나 이상의 보조 변수를 지정합니다. 보조 변수는 곡선 적합 플랫폼의 계산에 사용되지 않으며 이 변수를 포함해도 결과에 영향을 주지 않습니다. 보조 변수는 향후 곡선 적합 결과를 분석하는 데 사용할 수 있는 변수입니다. 보조 변수를 지정하면 "모수 테이블 생성" 옵션으로 생성된 테이블에 포함됩니다. JMP Pro 에서 하나 이상의 보조 변수와 그룹화 변수를 지정하면 적합 모형에 대해 "곡선 DOE 분석" 옵션을 사용할 수 있습니다. 자세한 내용은 "[곡선 DOE 분석](#)"에서 확인하십시오.

가중치 관측값의 가중치를 포함하는 변수를 지정합니다.

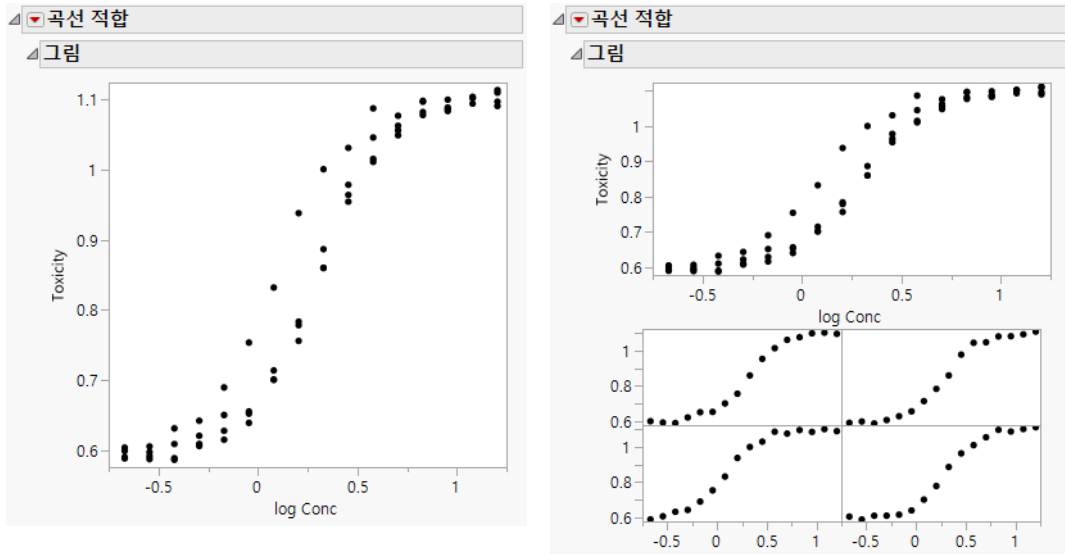
빈도 ("쌓인 데이터 형식" 탭에만 사용 가능) 관측값의 빈도를 포함하는 변수를 지정합니다.

기준 변수의 모든 수준에 대해 개별 분석을 수행할 변수를 지정합니다.

곡선 적합 보고서

처음에는 "곡선 적합" 보고서에 Y 대 X 그림만 포함됩니다. "그룹" 변수를 지정하거나 "함수 열" 탭을 사용하면 적합 모형의 각 그룹에 대한 중첩 및 개별 그림이 보고서에 포함됩니다.

그림 14.7 곡선 적합 보고서: 그룹화 변수 없음 (왼쪽), 그룹 변수 있음 (오른쪽)



모형을 적합시키면 적합 모형이 그림에 나타납니다(플랫폼 시작 창에서 그룹화 변수를 지정하지 않은 경우). 그림 위에 "모형 비교" 보고서가 나타나고 그림 아래에 모형 보고서가 나타납니다. 새 모형이 적합될 때마다 "모형 비교" 보고서에 추가되고 새 모형 보고서가 보고서 창에 추가됩니다. 자세한 내용은 "[모형 비교 보고서](#)" 및 "[모형 적합 보고서](#)"에서 확인하십시오.

참고: 데이터 테이블에서 제외된 모든 행은 곡선 적합 그림에서도 숨겨집니다.

모형 비교 보고서

곡선 적합 플랫폼의 "모형 비교" 보고서에는 여러 모형을 비교하는 데 사용되는 적합 통계량이 표시됩니다.

그림 14.8 모형 비교 보고서

곡선 적합									
모형 비교									
모형	AICc	AICc 가중치	.2 .4 .6 .8	BIC	SSE	MSE	RMSE	R ²	
로지스틱 4P	-372.3021	1		-348.9054	0.005325	0.0001109	0.0105327	0.9980584	
Gompertz 4P	-327.5372	1.903e-10		-304.1405	0.0107173	0.0002233	0.0149425	0.9960922	

AICc 두 개 이상의 모형을 비교하는 데 사용할 수 있는 추정된 통계 모형의 적합도를 나타내는 측도를 제공합니다. AICc 는 작은 표본에 맞게 AIC 를 조정한 것입니다. AICc 는 데이터 점 수가 모수 수보다 두 개 이상 많은 경우에만 계산할 수 있습니다. AICc 값이 가장 낮은 모형이 최적 모형이며 이 예에서는 로지스틱 4P 입니다. 자세한 내용은 [선형 모형 적합](#)에서 확인하십시오.

AICc 가중치 합계가 1 이 되는 정규화된 AICc 값을 제공합니다. AICc 가중치는 적합 모형 중 하나가 true 일 때 특정 모형이 true 모형일 확률로 해석할 수 있습니다. 따라서 AICc 가중치가 1 에 가장 가까운 모형이 최량 적합 모형입니다. 이 예에서는 로지스틱 4P 모형이 분명히 더 나은 적합입니다. AICc 가중치는 비결측 AICc 값만 사용해서 계산됩니다.

$$\text{AICcWeight} = \exp[-0.5(\text{AICc} - \min(\text{AICc}))] / \sum(\exp[-0.5(\text{AICc} - \min(\text{AICc}))])$$

여기서 $\min(\text{AICc})$ 는 적합 모형 중 가장 작은 AICc 값입니다. AICc 가중치 열은 내림차순으로 정렬됩니다.

BIC 여러 모형을 비교할 때 유용한 모형 적합 측도를 제공합니다 (가능도 함수를 기반으로 함). BIC 값이 낮은 모형이 더 나은 적합입니다. 자세한 내용은 [선형 모형 적합](#)에서 확인하십시오.

SSE 각 관측값과 예측값 간의 차이에 대한 제곱합입니다.

MSE 각 값의 오차 제곱 평균을 제공합니다.

RMSE 랜덤 오차의 표준편차를 추정하는 MSE 의 제곱근입니다.

R² 랜덤 오차가 아니라 모형에 기인할 수 있는 반응의 변동 비율을 추정합니다. R² 값이 1 에 가장 가까운 모형이 더 나은 적합입니다.

모형 비교 플랫폼에서는 잔차와 실제값 그림과 같은 추가 옵션을 제공합니다. 자세한 내용은 ["모형 비교"](#)에서 확인하십시오.

모형 적합 보고서

곡선 적합 플랫폼에서 적합되는 각 개별 모형에 대한 보고서가 생성됩니다. 각 모형 보고서에는 다음 섹션이 포함됩니다.

예측 모형 예측 계산식과 모수의 대수 형식을 표시합니다.

적합 요약 "모형 비교" 보고서와 동일한 적합 통계량을 표시합니다.

그룹 요약 (그룹화 변수가 있는 경우에만 표시됨) 그룹 모수 추정값을 표시합니다.

모수 추정값 모수 추정값, 표준 오차, 카이제곱 통계량, p 값 및 신뢰 구간을 표시합니다. 추정값의 상관관계수 및 공분산에 대한 별도의 테이블도 있습니다.

팁 : 정렬 기준으로 사용할 열을 클릭하여 "모수 추정값" 보고서 테이블을 정렬할 수 있습니다.

그림 적합 모형을 사용한 데이터 그림을 표시합니다 ([그림 14.7](#)). 그림은 플랫폼 시작 창에서 그룹화 변수를 선택한 경우에만 표시됩니다.

곡선 적합 옵션

"곡선 적합"의 빨간색 삼각형 메뉴에는 여러 범주의 기본 제공 모형이 포함되어 있습니다. 특정 데이터에 사용할 수 없는 모형은 회색 메뉴 옵션입니다. 개별 모형 상세 정보 및 데이터 요구 사항은 [표 14.1](#)에서 확인하십시오.

다항 1차 ~ 5차 다항식 모형 및 거듭제곱 모형을 적합시킵니다.

시그모이드 곡선 로지스틱, 프로빗, Gompertz 및 Weibull 모형을 적합시킵니다. 이러한 모형은 S 형태이며 상위 점근선과 하위 점근선이 모두 있습니다. 로지스틱 2P, 3P 및 4P 모형과 프로빗 2P 및 4P 모형은 대칭입니다. 로지스틱 5P, 프로빗 3P 및 두 개의 Gompertz 모형은 대칭이 아닙니다. 로지스틱 2P와 프로빗 2P는 반응이 0에서 1 사이인 경우에만 사용할 수 있습니다. Weibull 성장은 반응 값과 회귀변수 값이 모두 음수가 아닌 경우에만 사용할 수 있습니다. 시그모이드 곡선의 예로는 학습 곡선과 종양 성장 모델링이 있으며 둘 다 처음에는 증가하다가 차츰 감소합니다.

지수적 성장 및 감소 지수, 이중 지수, 기계적 성장, 하이브리드 지수 및 셀 성장 모형을 적합시킵니다. 지수 2P와 3P는 비슷하지만 3P 모형에는 점근선이 있습니다. 이중 지수 모형은 두 개의 개별 성장 및 감소 프로세스가 있다고 가정합니다. 기계적 성장 및 지수 3P 모형은 항상 증가(또는 감소)하지만 증가율(또는 감소율)이 느려서 모형에 점근선이 생깁니다. 지수적 성장 및 감소의 예로는 각각 바이러스 확산과 약물 반감기가 있습니다.

정상점 모형 가우스 정상점, ExGaussian 정상점(지수 수정 가우스 정상점) 및 로렌츠 정상점 모형을 적합시킵니다. 이러한 모형은 정상점까지 증가한 후 감소합니다. 가우스 정상점 모형은 가우스 PDF(확률 밀도 함수)의 척도화된 결과입니다. ExGaussian 정상점 모형은 가우스 정상점 모형과 비슷하지만 한쪽으로 치우친 형태일 수 있습니다. 로렌츠 정상점 모형은 Cauchy 분포(연속형 확률 분포)의 척도화된 결과입니다. 이러한 모형은 일부 화학 농도 시험 및 인공 신경망에 사용할 수 있습니다.

약동학 모형 1구획 경구 투여 모형, 2구획 정맥 투여 모형 및 이중 지수 4P 모형을 적합시킵니다. 이 옵션은 체내 약물 농도를 모델링하는 데 사용됩니다.

비율 방정식 Michaelis-Menten 모형 및 역 Michaelis-Menten 모형과 여러 1차 비율 및 2차 비율 모형을 적합시킵니다. Michaelis-Menten 모형은 효소 반응 속도를 기질 농도와 연관시키는 생화학적 운동 모형입니다. 1차 및 2차 비율 모형은 화학 반응을 모델링하는 데 유용합니다.

Antoine 방정식 적합 증기 압력을 온도 함수로 모델링하는 데 주로 사용되는 Antoine 모형을 적합시킵니다. Antoine 모형에는 수평 점근선과 수직 점근선이 모두 있습니다.

용해 곡선 분석 (반응 값과 회귀변수 값이 모두 음수가 아닌 경우에만 사용 가능) 용해 곡선 분석을 위한 옵션 하위 메뉴가 포함되어 있습니다. 용해 곡선은 시간 경과에 따른 정제 용해 속도를 측정합니다. 용해 곡선 분석은 새 정제의 용해 곡선을 표준 또는 대조 정제의 용해 곡선과 비교합니다. 각 새 정제에 대해 별도의 분석이 수행됩니다. "용해 곡선 분석" 메뉴에서는 용해 곡선 비교를 위한 모수 및 비모수 기법을 모두 제공합니다. 기본 제공 용해 곡선 모형 중 하나를 사용하는 경우 적합한 각 개별 모형에 대해 표준 모형 적합 보고서가 생성됩니다.

모형 없는 비교 용해 곡선을 비교하기 위한 비모수 기법을 제공합니다. "모형 없는 비교" 하위 메뉴에서 비모수 기법 중 하나를 사용하면 방법별 용해 곡선 보고서가 생성됩니다. 자세한 내용은 "모형 비교 보고서"에서 확인하십시오.

참고 : 모형 없는 비교 기법을 사용하려면 모두 곡선에 결측값이 없어야 합니다.

F1 분석 각 시점에서 대조 정제의 곡선과 시험 정제의 곡선 간 차이 비율을 측정하는 F1 차이 요인을 사용하여 분석을 수행합니다. 자세한 내용은 "곡선 적합 모형에 대한 통계 상세 정보"에서 확인하십시오. F1에 대한 신뢰 구간은 붓스트랩 절차를 사용하여 계산됩니다. 붓스트랩 절차에서는 BCa(편향 수정 및 가속) 백분위수 방법을 사용합니다. 자세한 내용은 Efron 연구 자료 (1981)에서 확인하십시오.

"F1 분석" 옵션을 선택하면 "용해 곡선 규격" 창이 나타납니다. 이 창에서는 대조 수준, 신뢰 구간을 계산하는 데 사용되는 유의 수준, 붓스트랩 표본 수 및 난수 시드값을 지정할 수 있습니다. 데이터가 쌓인 데이터 형식인 경우에는 대조군과 시험군 내의 개별 곡선을 식별하는 곡선 ID 열도 지정해야 합니다.

F2 분석 대조 정제의 곡선과 시험 정제의 곡선 간 용해 비율 유사성을 측정하는 F2 유사성 요인을 사용하여 분석을 수행합니다. 자세한 내용은 "곡선 적합 모형에 대한 통계 상세 정보"에서 확인하십시오. F2에 대한 신뢰 구간은 붓스트랩 절차를 사용하여 계산됩니다. 붓스트랩 절차에서는 BCa(편향 수정 및 가속) 백분위수 방법을 사용합니다. 자세한 내용은 Efron 연구 자료 (1981)에서 확인하십시오. F2의 하한이 50 보다 크면 시험 정제의 용해가 대조 정제의 용해와 유사하다는 결론을 내립니다 (Paixão et al. 2017 참조).

"F2 분석" 옵션을 선택하면 "용해 곡선 규격" 창이 나타납니다. 이 창에서는 대조 수준, 신뢰 구간을 계산하는 데 사용되는 유의 수준, 붓스트랩 표본 수 및 난수 시드값을 지정할 수 있습니다. 데이터가 쌓인 데이터 형식인 경우에는 대조군과 시험군 내의 개별 곡선을 식별하는 곡선 ID 열도 지정해야 합니다.

다변량 거리 대조 정제의 곡선과 시험 정제의 곡선 간 다변량 거리를 측정하는 Mahalanobis 거리 M 을 사용하여 분석을 수행합니다. 자세한 내용은 "곡선 적합 모형에 대한 통계 상세 정보"에서 확인하십시오. M 에 대한 신뢰 구간은 다변량 정규 가정을 사용하여 계산됩니다. M 의 상한이 최대 차이보다 작으면 시험 정제의 용해가 대조 정제의 용해와 유사하다는 결론을 내립니다 (Paixão et al. 2017 참조).

"다변량 거리" 옵션을 선택하면 "용해 곡선 규격" 창이 나타납니다. 이 창에서는 대조 수준 및 신뢰 구간을 계산하는 데 사용되는 유의 수준을 지정할 수 있습니다. 데이터가 쌓인 데이터 형식인 경우에는 대조군과 시험군 내의 개별 곡선을 식별하는 곡선 ID 열도 지정해야 합니다.

Higuchi 곡선 시차 성분이 있는 모형과 버스트 성분이 있는 모형을 포함하여 Higuchi 모형을 적합시킵니다.

Hixson-Crowell 곡선 시차 성분이 있는 모형을 포함하여 Hixson-Crowell 모형을 적합시킵니다.

Korsmeyer-Peppas 곡선 시차 성분이 있는 모형과 버스트 성분이 있는 모형을 포함하여 Korsmeyer-Peppas 모형을 적합시킵니다.

시그모이드 곡선 로지스틱, 프로빗 및 Weibull 모형을 적합시킵니다. 자세한 내용은 "[시그모이드 곡선](#)"에서 확인하십시오.

다음 옵션에 대한 자세한 내용은 **JMP 사용**의 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.

모형 없는 비교 보고서

곡선 적합 플랫폼에서 세 가지 모형 없는 비교 기법인 F1 분석, F2 분석 또는 다변량 거리 중 하나를 사용하면 모형 없는 비교 보고서가 생성됩니다. 보고서 제목은 방법 이름에 해당합니다. 각 모형 없는 비교 보고서에는 탭 형식의 곡선 비교 보고서가 포함되어 있습니다. 시험군이 여러 개인 경우 각 모형 없는 비교 보고서에는 요약 테이블이 포함됩니다. 또한 각 보고서 맨 위에는 대조군으로 사용된 그룹화 변수의 수준을 나타내는 노트가 있습니다.

요약 테이블

요약 테이블에는 대조군과 비교되는 각 시험군에 대한 행이 있습니다. 관측된 요인 값에 대한 열과 해당 방법(F1 분석, F2 분석 또는 다변량 거리)의 신뢰 하한 및 상한에 대한 열이 있습니다. "F2 분석" 및 "다변량 거리" 옵션에 대한 요약 테이블에는 해당 행의 시험 곡선이 대조 곡선과 유사한지 여부를 나타내는 결론 열도 포함됩니다.

비교

"비교" 섹션에는 대조군과 비교되는 각 시험군에 대한 탭이 있습니다. 각 탭에는 지정된 기법에 대한 그림, 값 테이블 및 결과 테이블이 포함됩니다. F1 분석 또는 F2 분석 기법을 사용하는 경우 붓스트랩 결과 그림도 표시됩니다.

그림에는 각각 "대조" 및 "시험" 라벨이 지정된 두 개의 평균 용해 곡선이 표시됩니다. 또한 대조군과 시험군의 각 시점에 대한 평균, 표준편차 및 CV(변동 계수)를 제공하는 요약 통계량 테

이블도 있습니다. 요약 통계량 테이블에서 값이 노란색으로 강조 표시되면 다음 조건 중 하나가 참입니다.

- 시험군에 평균 용해가 85% 보다 큰 시점이 두 개 이상 있습니다.
- 시험군에서 첫 번째 시점의 CV 값이 20% 보다 큼니다.
- 시험군에서 첫 번째 시점 이후에 CV 값이 10% 보다 큰 시점이 있습니다.

F1 분석 또는 F2 분석 기법을 사용하는 경우 결과 테이블은 붓스트랩 절차를 통해 가져옵니다. 테이블에는 관측된 F1 또는 F2 값, 붓스트랩 표본 수, 붓스트랩 평균, 붓스트랩 중앙값, BCa 신뢰 하한/상한, 왜도 및 첨도가 포함됩니다. F2 분석의 경우 결론 섹션도 있습니다. "붓스트랩 결과" 섹션에는 각 붓스트랩 표본에 대한 F1 또는 F2 값과 붓스트랩 표본의 정규 분위수 그림이 포함됩니다.

다변량 거리 기법을 사용하는 경우 테이블에는 관측된 Mahalanobis 거리 값, 계산된 신뢰 하한/상한 및 최대 차이가 포함됩니다. 최대 차이는 대조 정제와 시험 정제 간의 차이가 모든 시점에서 10%인 경우의 Mahalanobis 거리입니다. 자세한 내용은 "용해 곡선 분석에 대한 통계 상세 정보"에서 확인하십시오. M 의 상한이 최대 차이보다 작으면 시험 정제의 용해가 대조 정제의 용해와 유사하다는 결론을 내립니다.

모형 없는 비교 분석 옵션

"F1 분석", "F2 분석" 및 "다변량 거리" 보고서의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

곡선 DOE 분석 (시작 창에서 그룹화 변수와 하나 이상의 보조 변수를 지정한 경우에만 사용 가능. "함수 열" 형식을 사용하는 경우에는 사용 불가능) 해당 요인 (F1, F2 또는 다변량 거리) 을 보조 변수의 함수로 모델링하기 위해 모형 없는 비교 보고서 내에서 "일반화 회귀" 보고서를 시작합니다. 자세한 내용은 "곡선 DOE 분석"에서 확인하십시오.

적합 제거 "곡선 적합" 보고서에서 모형 없는 비교 보고서를 제거합니다.

모형 적합 옵션

곡선 적합 플랫폼의 각 모형 보고서에는 다음 옵션 중 일부 또는 전체가 포함된 빨간색 삼각형 메뉴가 있습니다.

병렬 검정 곡선이 X 축을 따라 이동할 때 모형이 비슷한지 여부를 판별하는 데 도움이 됩니다. 특정 상황에서는 그룹 간 추가 비교를 수행하기 전에 병렬 처리를 설정하는 것이 중요합니다. 플랫폼 시작 창에서 그룹 변수를 지정한 경우에만 이 옵션을 사용할 수 있습니다. 이 옵션은 고차 다항식을 제외한 선형 회귀 모형과 시그모이드 모형에 사용할 수 있습니다. 자세한 내용은 "병렬 검정"에서 확인하십시오.

곡선 아래 면적 적합 곡선 아래 면적을 제공합니다. 이 옵션은 1 구획, 2 구획, 가우스 정상점 및 로렌츠 정상점 모형에 사용할 수 있습니다. 이중 지수 4P 모형에도 이 옵션을 사용할 수 있지

만 모든 모수가 양수일 때만 가능합니다. 통합 범위는 모형의 유형에 따라 다르며 보고서에 지정됩니다.

플랫폼 시작 창에서 그룹화 변수를 지정한 경우 그룹 간 추정값 비교를 위해 평균 분석이 수행됩니다. 그룹에 대한 결과가 결정 한계를 초과하면 결과가 AUC의 전체 평균과 다르다고 간주됩니다.

반응 정상점까지의 시간 적합 곡선의 정상점에서 회귀변수 X 의 추정값을 표시합니다. 추정값의 표준 오차도 표시됩니다. 이 옵션은 셀 성장 4P 및 1 구획 모형에 사용할 수 있습니다.

반응 정상점 적합 곡선의 정상점에서 반응 Y 의 추정값을 표시합니다. 추정값의 표준 오차도 표시됩니다. 이 옵션은 셀 성장 4P 및 1 구획 모형에 사용할 수 있습니다.

모수 추정값 비교 그룹화 변수의 전체 수준에서 모수의 동일성 검정을 위한 분석을 제공합니다. 플랫폼 시작 창에서 그룹 변수를 지정한 경우에만 이 옵션을 사용할 수 있습니다. 자세한 내용은 "**모수 추정값 비교**"에서 확인하십시오.

동등성 검정 그룹화 변수의 전체 수준에서 모형의 동등성 검정을 위한 분석을 제공합니다. 플랫폼 시작 창에서 그룹 변수를 지정한 경우에만 이 옵션을 사용할 수 있습니다. 자세한 내용은 "**동등성 검정**"에서 확인하십시오.

JMP PRO 곡선 DOE 분석 (시작 창에서 그룹화 변수와 하나 이상의 보조 변수를 지정한 경우에만 사용 가능.) 함수 열 탭을 사용하는 경우에는 사용 불가능) 곡선 적합 플랫폼 내에서 "일반화 회귀" 보고서를 시작합니다. 일반화 회귀 모형은 보조 변수를 모형 효과로 사용하여 비선형 모형의 각 모수에 적합됩니다. 기본적으로 2차 요인 모형이 적합되고 "추정 방법"은 "최량 부분집합"입니다. 모형의 항 수가 20 개를 초과하거나 함수 수가 1000 개를 초과하면 "추정 방법"이 자동으로 "가지 치기 전진 선택"으로 전환됩니다. 데이터 테이블이 확증적 선별 설계에서 나온 경우 "추정 방법"이 자동으로 "2 단계 전진 선택"으로 전환됩니다. 아니면 원래 데이터 테이블에서 원하는 모형 적합을 정의하는 모형 스크립트를 지정할 수도 있습니다. 그러면 모형이 결합되어 반응 프로파일러를 회귀변수 및 보조 변수의 함수로 생성합니다. 그런 다음 CDOE 프로파일러를 사용하여 보조 변수가 반응에 어떤 영향을 주는지 탐색할 수 있습니다.

"곡선 DOE 분석" 보고서에는 다음과 같은 빨간색 삼각형 메뉴 옵션이 있습니다.

모형 모수에 대한 일반화 회귀 각 모형 모수에 대한 "일반화 회귀" 보고서를 표시하거나 숨깁니다. "일반화 회귀" 모형 보고서에 대한 자세한 내용은 선형 모형 적합에서 확인하십시오.

진단 그림 반응 변수에 대한 실제값 대 예측값 그림과 잔차 그림을 표시하거나 숨깁니다.

CDOE 프로파일러 보조 변수에 따라 반응이 어떻게 변하는지 탐색할 수 있는 CDOE 프로파일러를 표시하거나 숨깁니다. "CDOE 프로파일러"의 빨간색 삼각형 메뉴 옵션에 대한 자세한 내용은 Profilers에서 확인하십시오.

예측 계산식 저장 반응에 대한 예측 계산식을 데이터 테이블의 새 열에 저장합니다.

모수 테이블 생성 모수 추정값, 표준 오차 및 t 비를 데이터 테이블에 저장합니다. 플랫폼 시작 창에서 그룹 변수를 지정한 경우에만 이 옵션을 사용할 수 있습니다.

실제값 대 예측값 그림 세로 축에 실제 Y 값을 표시하고 가로 축에 예측 Y 값을 표시합니다.

잔차 대 예측값 그림 세로 축에 잔차를 표시하고 가로 축에 예측 Y 값을 표시합니다.

프로파일러 적합 예측 함수의 프로파일러를 표시하거나 숨깁니다. 도함수는 X 변수에 대한 예측 함수의 도함수입니다. 프로파일러에 대한 자세한 내용은 **Profilers** 에서 확인하십시오.

계산식 저장 다양한 계산식 열을 데이터 테이블에 저장하기 위한 옵션이 포함되어 있습니다. 시작 창에서 "함수 행" 또는 "함수 열" 탭 중 하나를 사용하는 경우 "쌓인 데이터 저장" 옵션만 사용할 수 있습니다.

예측 계산식 저장 예측 방정식을 저장합니다.

예측값의 표준 오차 저장 예측값의 표준 오차를 저장합니다.

모수 예측 계산식 저장 예측 방정식을 모수 형식으로 저장합니다. 이렇게 하면 사용자 비선형 플랫폼에서 적합 모형을 사용하려는 경우에 유용합니다.

잔차 계산식 저장 잔차를 저장합니다.

스튜던트화 잔차 계산식 저장 표준화 잔차를 추정된 표준편차로 나누는 스튜던트화 잔차 계산식을 저장합니다.

1 차 도함수 저장 X 변수에 대한 예측 함수의 도함수를 저장합니다.

1 차 도함수의 표준 오차 저장 1 차 도함수의 표준 오차 방정식을 저장합니다.

역추정 예측 계산식 저장 Y 에서 X 를 예측하기 위한 방정식을 저장합니다.

쌓인 데이터 저장 열을 새 데이터 테이블에 저장합니다. 데이터 테이블에는 쌓인 형식의 원래 데이터와 반응 예측값 열 및 잔차 열이 포함됩니다.

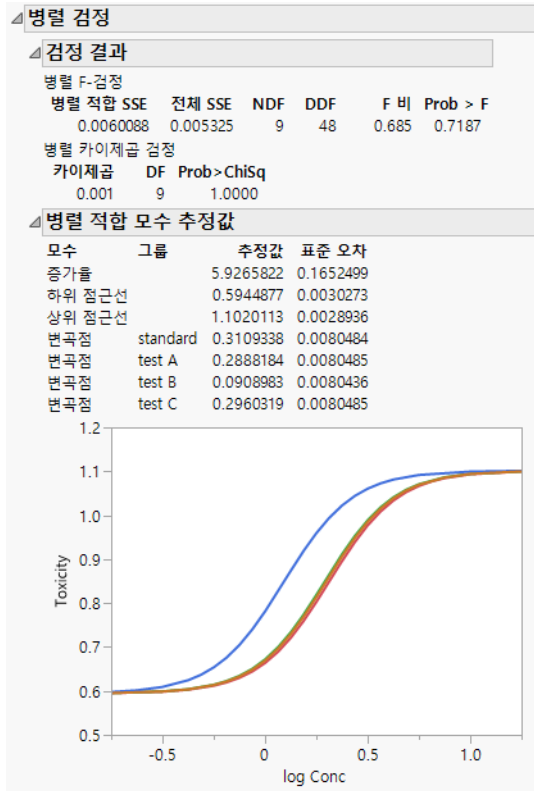
사용자 역추정 예측 특정 Y 값에 대한 X 값을 예측합니다. 역추정 예측에 대한 자세한 내용은 선형 모형 적합에서 확인하십시오.

적합 제거 모형 보고서, "모형 비교" 보고서의 항목 및 그림의 적합선을 제거합니다.

병렬 검정

곡선 적합 플랫폼 또는 곡선 적합 플랫폼의 모형 적합 빨간색 삼각형 메뉴에 있는 "병렬 검정" 옵션은 그룹 간의 적합 모형이 형상은 동일하지만 X 축을 따라 이동하는지 검정하기 위한 분석을 제공합니다(그림 14.9). Bioassay 예에서 약물 제제 B에 대한 곡선은 다른 세 곡선의 왼쪽으로 이동합니다. 그러나 곡선의 형상이 여전히 같은지(병렬) 또는 B 제제가 다른지는 알 수 없습니다. 병렬 검정을 통해 서로 다른 약물 제제의 형상이 비슷하고 가로 축을 따라 이동하는지 여부를 알 수 있습니다. 적합 모형의 빨간색 삼각형 메뉴에서 **병렬 검정**을 선택하여 보고서를 추가합니다.

그림 14.9 병렬 검정



보고서는 다음과 같은 결과를 제공합니다.

검정 결과 병렬 F-검정 및 카이제곱 검정 결과를 제공합니다. 이 검정에서는 전체 모형의 오차 제곱합 (전체 SSE) 과 축소 모형의 오차 제곱합 (적합 SSE) 을 비교합니다. 전체 모형은 각 그룹에 서로 다른 모수를 제공합니다. 축소 모형은 변곡점을 제외한 모든 모수를 그룹이 공유하도록 강제 적용합니다. p 값이 작으면 그룹 모형이 서로 유의하게 다르다는 것을 나타냅니다. 따라서 모수를 공유한 축소 모형은 충분하지 않으며 전체 모형을 사용해야 합니다. **그림 14.9** 에서 p 값이 0.05 보다 크므로 곡선 사이에 차이가 있다는 결론을 내릴 증거가 충분하지 않습니다. 이 경우에는 모수를 공유한 축소 모형이 적절합니다.

적합 통계량은 다음과 같이 계산됩니다.

F 비 $[(\text{적합 SSE} - \text{전체 SSE})/\text{NDF}]/[\text{전체 SSE}/\text{DDF}]$

카이제곱 적합 SSE - 전체 SSE

병렬 적합 모수 추정값 축소 모형의 모수 추정값을 제공합니다 (변곡점을 제외한 동일한 모수). 축소 모형의 적합 곡선 그림이 제공됩니다. 약물 제제 B의 변곡점이 다른 세 약물 제제의 변곡점보다 훨씬 낮습니다.

상대 잠재력 그룹화 변수의 각 수준에 대한 잠재력 및 상대 잠재력을 제공합니다. 잠재력은 $10^{(EC_{50})}$ 입니다. 여기서 EC_{50} 은 기준선과 최대값 사이의 중간 반응이 생성되는 농도입니다. 로지스틱 2P, 3P 및 4P 의 경우 잠재력은 $10^{(변곡점\ 모수)}$ 입니다. 상대 잠재력은 그룹화 변수의 한 수준에 대한 잠재력을 다른 수준의 잠재력으로 나눈 값입니다. 예를 들어 [그림 14.10](#) 에서 **standard** 에 대한 **test A** 의 상대 잠재력은 $Potency_{test\ A}/Potency_{standard}$ 입니다.

그림 14.10 그룹별 상대 잠재력

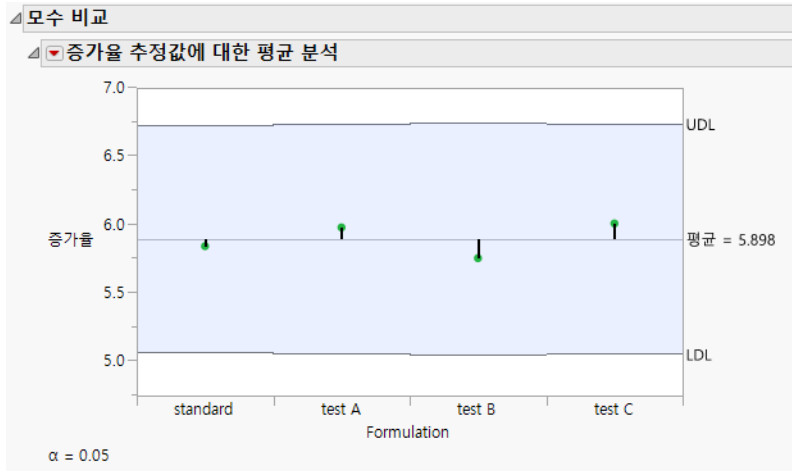
상대 잠재력			
상대 잠재력 대 standard			
그룹	잠재력	상대 잠재력	표준 오차
standard	2.0461329	1	0
test A	1.9445466	1.0522416	0.0246535
test B	1.2328161	1.6597227	0.0388878
test C	1.977115	1.0349084	0.0242474
상대 잠재력 대 test A			
그룹	잠재력	상대 잠재력	표준 오차
standard	2.0461329	0.950352	0.0222663
test A	1.9445466	1	0
test B	1.2328161	1.5773209	0.036957
test C	1.977115	0.9835273	0.0230435
상대 잠재력 대 test B			
그룹	잠재력	상대 잠재력	표준 오차
standard	2.0461329	0.6025103	0.014117
test A	1.9445466	0.6339864	0.0148545
test B	1.2328161	1	0
test C	1.977115	0.6235429	0.0146098
상대 잠재력 대 test C			
그룹	잠재력	상대 잠재력	표준 오차
standard	2.0461329	0.9662691	0.0226392
test A	1.9445466	1.0167486	0.0238219
test B	1.2328161	1.6037388	0.037576
test C	1.977115	1	0

[그림 14.10](#) 의 **상대 잠재력 대 standard** 패널에서 약물 제제 A 와 C 의 상대 잠재력은 거의 1 입니다. 이는 해당 제제의 잠재력이 **standard** 제제의 잠재력과 유사하다는 것을 나타냅니다. 약물 제제 B 의 잠재력은 **standard** 의 잠재력보다 낮습니다. 이는 농도 함수로 나타내는 약물 제제 B 의 독성이 **standard** 보다 빠르게 증가한다는 의미입니다.

병렬 검정에서 곡선이 평행하므로 상대 잠재력을 계산할 수 있습니다. 상대 잠재력을 기반으로 B 제제가 다른 약물 제제보다 더 강력하다는 결론을 내릴 수 있습니다. 이전 조사 결과와 함께 볼 때 약물 제제 B의 독성이 더 강한 것 같습니다.

모수 추정값 비교

곡선 적합 플랫폼의 "모수 추정값 비교" 옵션은 그룹화 변수의 전체 수준에서 모수의 동일성 검정 결과를 제공합니다. 각 모수에 대해 해당 모수가 전체 평균과 같은지 여부를 검정하는 ANOM(평균 분석) 보고서가 있습니다. 모수에 대한 결과가 결정 한계를 초과하면 해당 모수가 전체 평균과 다릅니다. [그림 14.11](#)에서는 증가율 추정값에 대한 ANOM 보고서를 보여 줍니다. 적합 모형의 빨간색 삼각형 메뉴에서 **모수 추정값 비교**를 선택하여 보고서를 추가합니다.

그림 14.11 증가율 추정값에 대한 모수 비교

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

유의 수준 설정 검정의 유의 수준을 설정합니다.

요약 보고서 표시 모수 추정값, 결정 한계 및 모수의 한계 초과 여부를 포함하는 보고서를 표시하거나 숨깁니다.

표시 옵션 결정 한계, 음영 및 중심선을 표시하거나 숨기기 위한 옵션이 포함되어 있습니다. 점 모양을 변경하는 옵션도 포함되어 있습니다.

평균 분석 보고서에 대한 자세한 내용은 기본 분석에서 확인하십시오.

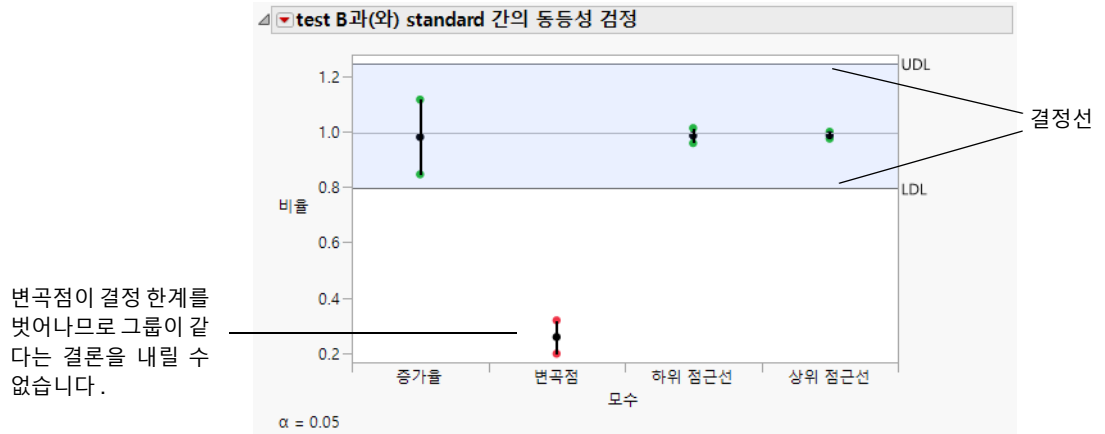
동등성 검정

곡선 적합 플랫폼의 "동등성 검정" 옵션은 그룹화 변수의 전체 수준에서 모형의 동등성 검정을 위한 분석을 제공합니다(그림 14.12). 옵션을 선택한 후에는 다른 모든 수준과 대조하여 검정할 그룹화 변수 수준을 지정합니다. 모든 수준과 선택한 수준을 비교하는 보고서가 있습니다. 적합 모형의 빨간색 삼각형 메뉴에서 **동등성 검정**을 선택하여 보고서를 추가합니다.

모수의 동일성 검정은 모수의 비율을 분석하여 수행됩니다. 기본 결정선은 0.8 및 1.25 비율 값에 배치되어 25% 차이를 나타냅니다.

모든 신뢰 구간이 결정선 내에 있으면 두 그룹이 실제로 동일합니다. 단일 구간이 선을 벗어나면(그림 14.12) 그룹이 같다는 결론을 내릴 수 없습니다. 약물 제제 B의 변곡점이 standard보다 낮으며, 이는 이전 결과와 일치합니다.

그림 14.12 동등성 검정



"동등성"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

유의 수준 설정 검정의 유의 수준을 설정합니다. 기본값은 0.05입니다.

결정선 설정 비율에 대한 결정 한계를 변경합니다. 기본값은 0.8 과 1.25 로 설정되어 25% 차이를 나타냅니다.

요약 보고서 표시 모수 추정값, 결정 한계 및 모수의 한계 초과 여부를 포함하는 보고서를 표시하거나 숨깁니다.

표시 옵션 결정 한계, 음영 및 중심선을 표시하거나 숨기기 위한 옵션이 포함되어 있습니다. 점 모양을 변경하는 옵션도 포함되어 있습니다. 추가 형식 지정 옵션을 보려면 그래프를 마우스 오른쪽 버튼으로 클릭하고 **사용자 정의**를 선택합니다.

곡선 적합 플랫폼의 추가 예

"용해 곡선 분석" 옵션을 사용하여 용해 프로파일을 기준으로 대조 정제와 가장 유사한 정제를 생성하는 요인 설정을 판별합니다. 각 그룹에 6개의 개별 정제가 포함된 16개의 시험 정제 그룹이 있습니다. 각 시험 정제 그룹은 네 가지 요인의 서로 다른 조합으로 구성됩니다. 대조 정제의 요인 설정은 알 수 없습니다. 먼저 시험 정제의 용해 곡선을 대조 정제의 용해 곡선과 비교합니다. 그런 다음 "곡선 DOE 분석" 옵션을 사용하여 적절한 요인 설정을 판별합니다.

용해 곡선 분석

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Dissolution DoE.jmp 를 엽니다.
2. **분석 > 전문 모델링 > 곡선 적합** 을 선택합니다.
3. "함수 행" 탭을 클릭합니다.
4. Dissolution60 ~ Dissolution360 을 선택하고 **Y, 반응**을 클릭합니다.
5. Batch 를 선택하고 **그룹**을 클릭합니다.

6. Polymer A, Polymer B, Total Polymer 및 Compression Force 를 선택하고 **Z**, 보조를 클릭합니다.
7. **확인**을 클릭합니다.
8. "곡선 적합"의 빨간색 삼각형을 클릭하고 **용해 곡선 분석 > 모형 없는 비교 > F2 분석**을 선택합니다.
9. "용해 곡선 규격" 창에서 **Batch**의 대조 수준을 R01로 지정합니다.
10. (선택 사항)"용해 곡선 규격" 창에서 "난수 시드값" 옆에 "1234"를 입력합니다.
11. **확인**을 클릭합니다.

그림 14.13 F2 분석 보고서

F2 분석				
참조: R01				
그룹	F2	90% BCa 하한	90% BCa 상한	결론
A01	38.98	37.067	41.073	유사하지 않음
A02	76.189	70.283	83.053	유사함
A03	40.263	38.118	42.75	유사하지 않음
A04	60.558	56.102	65.004	유사함
A05	37.199	35.593	38.613	유사하지 않음
A06	36.14	34.365	38.095	유사하지 않음
A07	70.32	64.452	77.655	유사함
A08	42.429	40.117	44.955	유사하지 않음
A09	42.943	39.548	46.749	유사하지 않음
A10	50.25	47.025	54.547	유사하지 않음
A11	33.246	31.524	35.227	유사하지 않음
A12	29.53	28.316	30.677	유사하지 않음
A13	84.864	79.669	93.739	유사함
A14	33.312	32.29	34.607	유사하지 않음
A15	37.005	35.454	38.685	유사하지 않음
A16	28.922	27.894	29.979	유사하지 않음

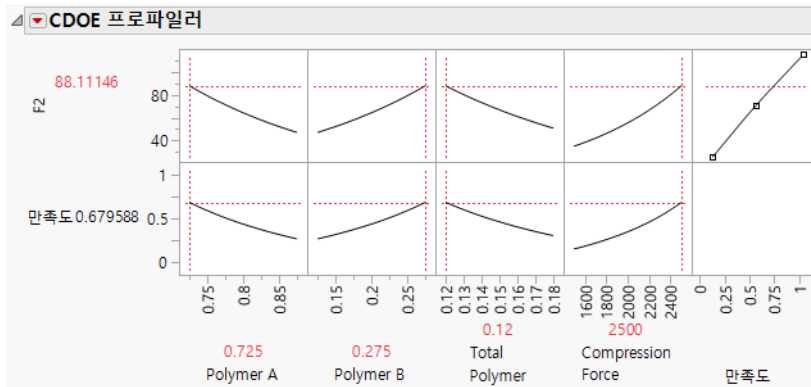
F2 유사성 요인에 따르면 4개의 정제 (A02, A04, A07, A13)가 대조 정제와 유사한 용해 프로파일을 갖습니다. 해당 정제는 모두 F2에 대한 90% 신뢰 하한이 50보다 큼니다. "비교" 섹션에서 각각 다른 탭을 선택하여 각 시험군과 대조군 간의 개별 비교를 볼 수 있습니다.

F2에 대한 곡선 DOE 분석

F2 분석을 통해 대조 정제와 유사한 정제를 판별할 수 있지만 왜 유사한지 이유는 확인할 수 없습니다. "곡선 DOE 분석" 옵션은 F2를 시작 창에 지정된 4개의 보조 변수에 대한 함수로 모델링하는 일반화 회귀 분석을 수행합니다.

1. "F2 분석"의 빨간색 삼각형을 클릭하고 **곡선 DOE 분석**을 선택합니다.
F2 값이 클수록 대조 정제와 더 유사하다는 것을 나타내므로 F2를 최대화하여 최적의 요인 설정을 판별할 수 있습니다.
2. "CDOE 프로파일러"의 빨간색 삼각형을 클릭하고 **최적화 및 만족도 > 만족도 함수**를 선택합니다.
3. "CDOE 프로파일러"의 빨간색 삼각형을 클릭하고 **최적화 및 만족도 > 만족도 최대화**를 선택합니다.

그림 14.14 F2 에 대한 곡선 DOE 분석



다음 요인 설정은 용해 프로파일을 기준으로 대조 정제와 가장 유사한 정제를 생성합니다.

Polymer A - 0.725

Polymer B - 0.275

Total Polymer - 0.12

Compression Force - 2500

곡선 적합 플랫폼에 대한 통계 상세 정보

이 섹션에는 곡선 적합 플랫폼에 대한 통계 상세 정보가 포함되어 있습니다.

- “곡선 적합 모형에 대한 통계 상세 정보”
- “용해 곡선 분석에 대한 통계 상세 정보”

곡선 적합 모형에 대한 통계 상세 정보

표 14.1에는 "곡선 적합"의 빨간색 삼각형 메뉴에서 사용할 수 있는 모형에 대한 계산식이 요약되어 있습니다.

표 14.1 곡선 적합 모형 계산식

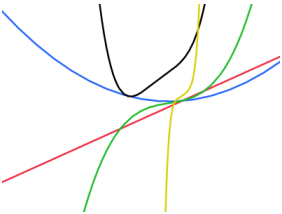
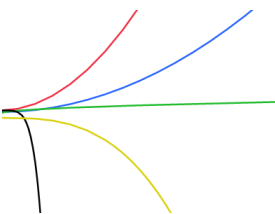
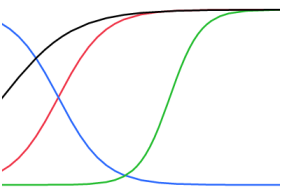
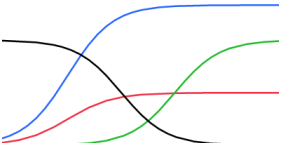
모형	계산식
다항 	$\beta_0 + \sum_{i=1}^k \beta_i x^i$ 여기서 k는 다항식의 차수입니다. 이러한 모형은 모형 적합 플랫폼과 X로 Y 적합 플랫폼을 사용하여 적합시킬 수도 있습니다.
거듭제곱 모형 	$a + bx^c$ a = 절편 b = 기울기 c = 거듭제곱 반응 값과 x 값이 모두 음수가 아닌 경우에만 사용할 수 있습니다.
로지스틱 2P 	$\frac{1}{1 + \exp(-a(x - b))}$ a = 증가율 b = 변곡점 모든 반응 값이 0에서 1 사이인 경우에만 사용할 수 있습니다.
로지스틱 3P 	$\frac{c}{1 + \exp(-a(x - b))}$ a = 증가율 b = 변곡점 c = 점근선

표 14.1 곡선 적합 모형 계산식 (계속)

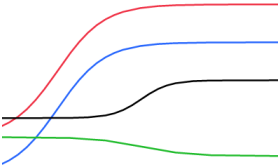
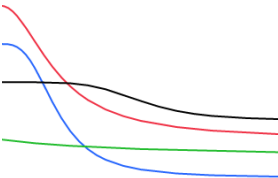
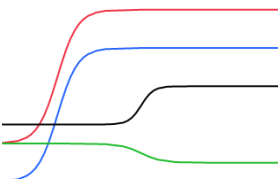
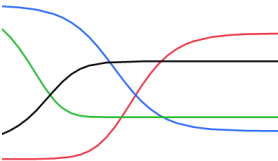
모형	계산식
로지스틱 4P 	$c + \frac{d - c}{1 + \exp(-a(x - b))}$ a = 증가율 b = 변곡점 c = 하위 점근선 d = 상위 점근선
로지스틱 4P Rodbard 	$c + \frac{d - c}{1 + (x/b)^a}$ a = 증가율 b = 변곡점 c = 하위 점근선 d = 상위 점근선 x 값이 양수인 경우에만 사용할 수 있습니다.
로지스틱 4P Hill 	$c + \frac{d - c}{1 + 10^{(-a(x - b))}}$ a = 증가율 b = 변곡점 c = 하위 점근선 d = 상위 점근선
로지스틱 5P 	$c + \frac{d - c}{(1 + \exp(-a(x - b)))^f}$ a = 증가율 b = 변곡점 c = 점근선 1 d = 점근선 2 f = 검정력

표 14.1 곡선 적합 모형 계산식 (계속)

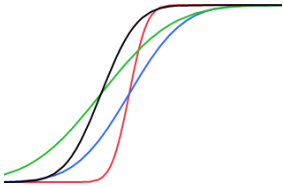
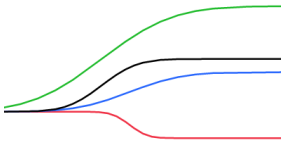
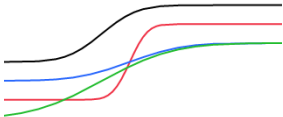
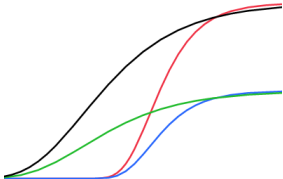
모형	계산식
프로빗 2P 	$\Phi(a \cdot (x - b))$ a = 증가율 b = 변곡점 Φ = 정규 분포 CDF 모든 반응 값이 0에서 1 사이인 경우에만 사용할 수 있습니다.
프로빗 3P 	$c \cdot \Phi(a \cdot (x - b))$ a = 증가율 b = 변곡점 c = 점근선 Φ = 정규 분포 CDF 또한 0에 점근선이 있습니다.
프로빗 4P 	$c + (d - c) \cdot \Phi(a \cdot (x - b))$ a = 증가율 b = 변곡점 c = 하위 점근선 d = 상위 점근선 Φ = 정규 분포 CDF
Gompertz 3P 	$a \exp(-\exp(-b(x - c)))$ a = 점근선 b = 증가율 c = 변곡점

표 14.1 곡선 적합 모형 계산식 (계속)

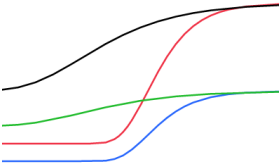
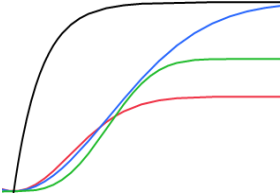
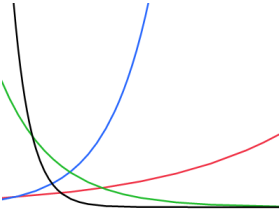
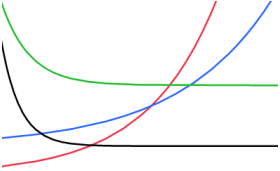
모형	계산식
Gompertz 4P 	$a + (b - a)\exp(-\exp(-c(x - d)))$ a = 하위 점근선 b = 상위 점근선 c = 증가율 d = 변곡점
Weibull 성장 	$a(1 - \exp(-(x/c)^b))$ a = 상위 점근선 b = 증가율 c = 변곡점 반응 값과 x 값이 모두 음수가 아닌 경우에만 사용할 수 있습니다.
지수 2P 	$a\exp(bx)$ a = 척도 b = 증가율
지수 3P 	$a + b\exp(cx)$ a = 점근선 b = 척도 c = 증가율

표 14.1 곡선 적합 모형 계산식 (계속)

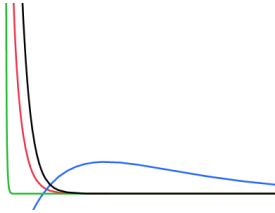
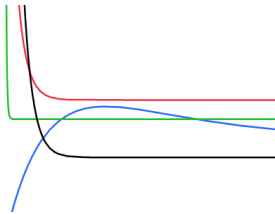
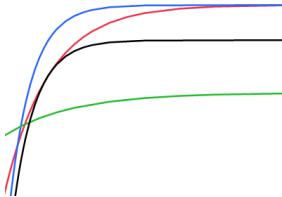
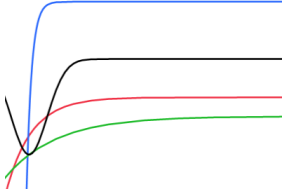
모형	계산식
이중 지수 4P	$a \exp(-bx) + c \exp(-dx)$
	$a = \text{척도 1}$ $b = \text{감소율 1}$ $c = \text{척도 2}$ $d = \text{감소율 2}$ 반응 값이 양수인 경우에만 사용할 수 있습니다.
이중 지수 5P	$a + b \exp(-cx) + d \exp(-fx)$
	$a = \text{점근선}$ $b = \text{척도 1}$ $c = \text{감소율 1}$ $d = \text{척도 2}$ $f = \text{감소율 2}$
기계적 성장	$a(1 - b \exp(-cx))$
	$a = \text{점근선}$ $b = \text{척도}$ $c = \text{증가율}$
하이브리드 지수	$ab - a(b - 1) \exp(-(cx)^d)$
	$a = \text{최대값}$ $b = \text{척도}$ $c = \text{비율}$ $d = \text{거듭제곱}$ 반응 값과 x 값이 모두 음수가 아닌 경우에만 사용할 수 있습니다.

표 14.1 곡선 적합 모형 계산식 (계속)

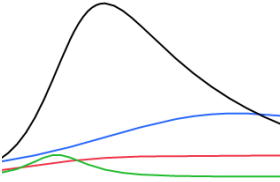
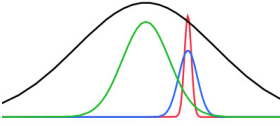
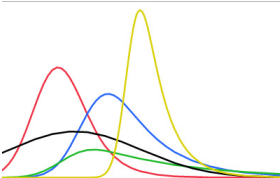
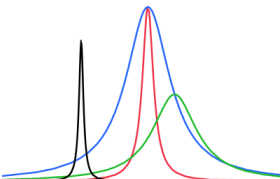
모형	계산식
셀 성장 4P 	$\frac{ab}{(a-b)\exp(-cx) + b\exp(dx)}$ a = 사멸 비율 d가 0인 경우 정상점 값 b = 시간 0에서 반응 c = 셀 분화 비율 d = 셀 사멸 비율 반응 값이 양수이고 x 값이 음수가 아닌 경우에만 사용할 수 있습니다.
가우스 정상점 	$a\exp\left(-\frac{1}{2}\left(\frac{x-b}{c}\right)^2\right)$ a = 정상점 값 b = 임계점 c = 증가율
ExGaussian 정상점 (지수 수정 가우스 정상점) 	$ad \cdot \exp\left(0.5d(2b + dc^2 - 2x) + \log\left(\phi\left(\frac{x-b-dc}{c}\right)\right)\right)$ a = AUC(곡선 아래 면적) b = 위치 c = 척도 d = 람다 ϕ = 표준 정규 pdf
로렌츠 정상점 	$\frac{ab^2}{(x-c)^2 + b^2}$ a = 정상점 값 b = 증가율 c = 임계점

표 14.1 곡선 적합 모형 계산식 (계속)

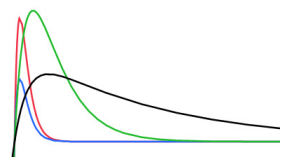
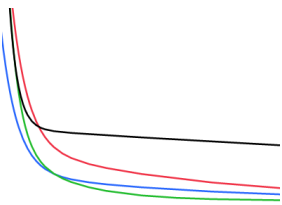
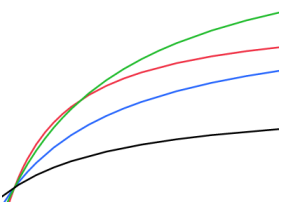
모형	계산식
1 구획 경구 투여 정제 → 신 장 에 서 처 리 $\xrightarrow{b}$ 배출 	$\frac{abc}{c-b}(\exp(-bx) - \exp(-cx))$ a = 곡선 아래 면적 b = 제거율 c = 흡수율 반응 값과 x 값이 모두 양수인 경우에만 사용할 수 있습니다.
2 구획 정맥 투여 샷 (shot) → 혈액 $\xrightarrow{b}$ 조직 혈액 $\xleftarrow{c}$ 조직 혈액 $\xrightarrow{d}$ ↓ 	$\frac{a}{\alpha - \beta}((\alpha - b)\exp(-\alpha x) - (\beta - b)\exp(-\beta x))$ $\alpha = \frac{1}{2}(b + c + d + \sqrt{(b + c + d)^2 - 4bd})$ $\beta = \frac{1}{2}(b + c + d - \sqrt{(b + c + d)^2 - 4bd})$ a = 초기 농도 b = 유입 속도 c = 유출 속도 d = 제거율 반응 값과 x 값이 모두 양수인 경우에만 사용할 수 있습니다.
Michaelis Menten 	$\frac{ax}{b + x}$ a = 최대 속도 b = Michaelis 상수 반응 값과 x 값이 모두 양수인 경우에만 사용할 수 있습니다.

표 14.1 곡선 적합 모형 계산식 (계속)

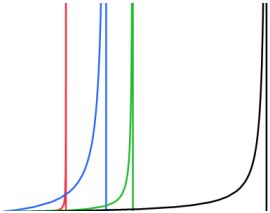
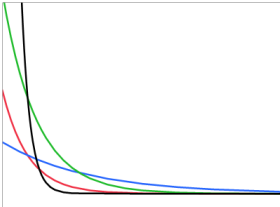
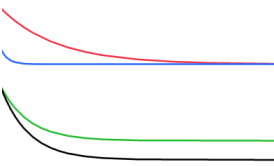
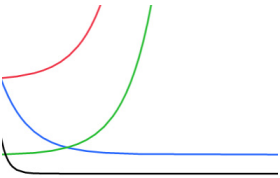
모형	계산식
역 Michaelis-Menten 	$\frac{bx}{a-x}$ a = 최대 속도 b = Michaelis 상수 반응 값과 x 값이 모두 양수인 경우에만 사용할 수 있습니다.
1차 비율 	$a \exp(-bx)$ a = 초기 값 b = 비율 상수 x 값이 음수가 아닌 경우에만 사용할 수 있습니다.
한계를 사용한 1차 	$b + (a - b) \exp(-cx)$ a = 초기 값 b = 값 제한 c = 비율 상수 x 값이 음수가 아닌 경우에만 사용할 수 있습니다.
평형을 사용한 1차 	$\frac{a(c + b \exp(-(b - c)x))}{(b + c)}$ a = 초기 값 b = 전진 비율 c = 후진 비율 x 값이 음수가 아닌 경우에만 사용할 수 있습니다.

표 14.1 곡선 적합 모형 계산식 (계속)

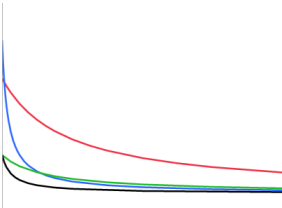
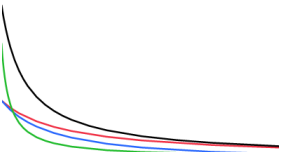
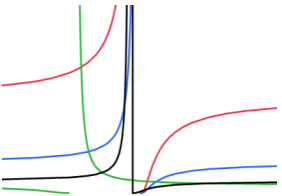
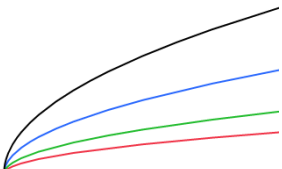
모형	계산식
2차 	$\frac{1}{(bx + 1/a)}$ a = 초기 값 b = 비율 상수 x 값이 음수가 아닌 경우에만 사용할 수 있습니다.
두 성분을 사용한 2차 	$\frac{ba}{(b + a)\exp(bcx) - a}$ a = 초기 값 b = 오프셋 c = 비율 상수 모든 x 값이 음수가 아닌 경우에만 사용할 수 있습니다.
Antoine 방정식 	$10^{a - \frac{b}{c + x}}$ a = 성분별 상수 b = 성분별 상수 c = 성분별 상수 반응 값이 0보다 큰 경우에만 사용할 수 있습니다. 세로 점근선은 10^a에 있습니다. 가로 점근선은 $-c$에 있습니다.
Higuchi 	$a\sqrt{x}$ a = 릴리스 상수 모든 x 및 y 값이 음수가 아닌 경우에만 사용할 수 있습니다.

표 14.1 곡선 적합 모형 계산식 (계속)

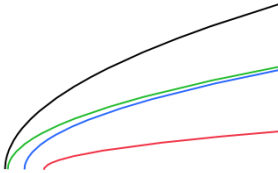
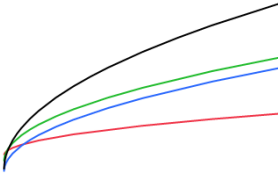
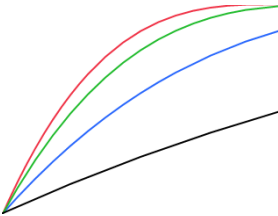
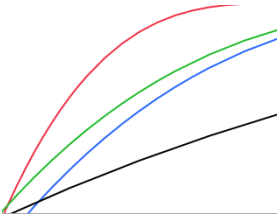
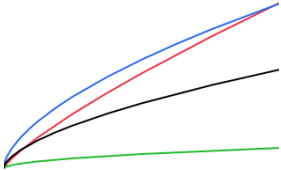
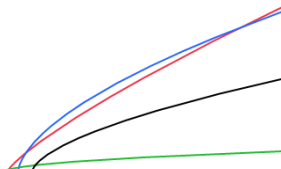
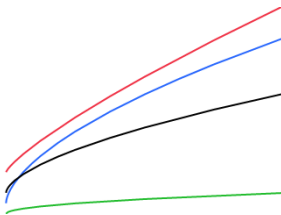
모형	계산식
시차를 사용한 Higuchi 	$a\sqrt{x-b}$ a = 릴리스 상수 b = 시차 모든 x 및 y 값이 음수가 아닌 경우에만 사용할 수 있습니다.
버스트를 사용한 Higuchi 	$b + a\sqrt{x}$ a = 릴리스 상수 b = 버스트 모든 x 및 y 값이 음수가 아닌 경우에만 사용할 수 있습니다.
Hixson-Crowell 	$100(1 - (1 - ax)^3)$ a = 릴리스 상수 모든 x 및 y 값이 음수가 아닌 경우에만 사용할 수 있습니다.
시차를 사용한 Hixson-Crowell 	$100(1 - (1 - a(x-b))^3)$ a = 릴리스 상수 모든 x 및 y 값이 음수가 아닌 경우에만 사용할 수 있습니다.

표 14.1 곡선 적합 모형 계산식 (계속)

모형	계산식
Korsmeyer-Peppas 	ax^b a = 릴리스 상수 b = 확산 지수 모든 x 및 y 값이 음수가 아닌 경우에만 사용할 수 있습니다.
시차를 사용한 Korsmeyer-Peppas 	$a(x-c)^b$ a = 릴리스 상수 b = 확산 지수 c = 시차 모든 x 및 y 값이 음수가 아닌 경우에만 사용할 수 있습니다.
버스트를 사용한 Korsmeyer-Peppas 	$c+ax^b$ a = 릴리스 상수 b = 확산 지수 c = 버스트 모든 x 및 y 값이 음수가 아닌 경우에만 사용할 수 있습니다.

용해 곡선 분석에 대한 통계 상세 정보

이 섹션에는 곡선 적합 플랫폼의 모형 없는 용해 분석에 사용되는 요인을 계산하기 위한 계산식이 포함되어 있습니다. 이 계산식에서는 다음과 같은 표기를 사용합니다.

- μ_{Ri} 는 t_i 시점에서 대조 곡선 배치에 대한 용해 값의 평균입니다.
- μ_{Ti} 는 t_i 시점에서 시험 곡선 배치에 대한 용해 값의 평균입니다.
- $i = 1, ..., N$ 은 시점 수입니다.

F1 차이 요인

F1 차이 요인은 각 시점에서 대조 정제의 곡선과 시험 정제의 곡선 간 차이 비율을 측정합니다. F1 은 다음 계산식을 사용하여 계산됩니다.

$$F1 = 100 \times \left(\frac{\sum_{i=1}^N |(\mu_{Ri} - \mu_{Ti})|}{\sum_{i=1}^N \mu_{Ri}} \right)$$

F2 유사성 요인

F2 유사성 요인은 대조 정제의 곡선과 시험 정제의 곡선 간 용해 비율 유사성을 측정합니다. F2 는 다음 계산식을 사용하여 계산됩니다 (Paixão et al. 2017 참조).

$$F2 = 50 \log \left(\frac{100}{\sqrt{1 + \frac{D^2}{p}}} \right)_{10}$$

다음은 각 요소에 대한 설명입니다.

$$D^2 = \sum_{i=1}^N (\mu_{Ri} - \mu_{Ti})^2$$

p - 시점 수

Mahalanobis 거리

Mahalanobis 거리 M 은 대조 정제의 곡선과 시험 정제의 곡선 간 다변량 거리를 측정합니다. M 은 다음 계산식을 사용하여 계산됩니다 (Paixão et al. 2017 참조).

$$M = \sqrt{(\mathbf{R} - \mathbf{T})^T (S_{pooled})^{-1} (\mathbf{R} - \mathbf{T})}$$

다음은 각 요소에 대한 설명입니다.

$$\mathbf{R} = [\mu_{R1}, \dots, \mu_{RN}]^T$$

$$\mathbf{T} = [\mu_{T1}, \dots, \mu_{TN}]^T$$

$$S_{pooled} = (\mathbf{SR} + \mathbf{ST})/2$$

$\mathbf{SR}$ = 대조 곡선 값의 공분산 행렬

ST = 시험 곡선 값의 공분산 행렬

10% 최대 차이는 다음 계산식을 사용하여 계산됩니다 .

$$MaxDiff\ 10\% = \sqrt{m^T (S_{pooled})^{-1} m}$$

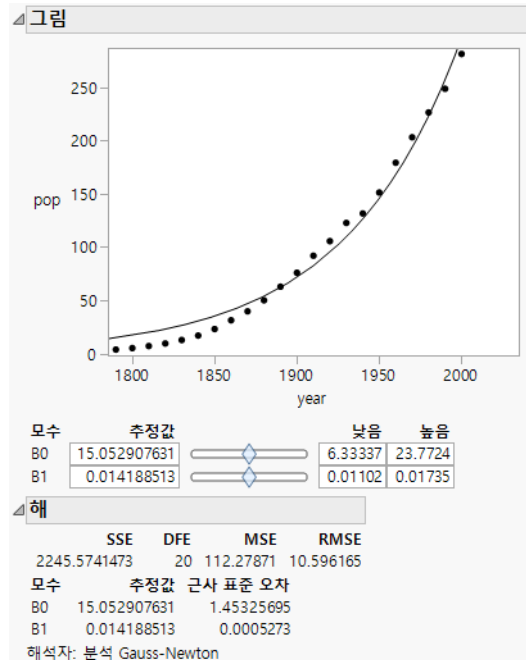
여기서 m 은 10의 Nx1 벡터입니다.

비선형 회귀

데이터에 사용자 비선형 모형 적합

비선형 플랫폼은 모수에 대해 비선형인 모형에 사용됩니다. 이 장에서는 모형 계산식 및 추정할 모수가 포함된 사용자 비선형 모형에 대해 중점적으로 다룹니다. 기본 최소 제곱 손실 함수 또는 사용자 손실 함수를 사용하여 모형을 적합시킵니다. 이 플랫폼은 전체 관측값에서 손실 함수의 합을 최소화합니다.

그림 15.1 사용자 비선형 적합의 예



비선형 플랫폼에서는 다항식, 로지스틱, Gompertz, 지수, 정상점 및 약동학 모형과 같은 사전 정의된 모형도 제공합니다. 자세한 내용은 "[곡선 적합](#)"에서 확인하십시오.

참고: 일부 모형은 모수에 대해 선형 (예: 2 차 또는 기타 다항식) 이거나, 선형이 되도록 변환 (예: x 의 로그 변환을 사용하는 경우) 할 수 있습니다. 이러한 상황에서는 모형 적합 또는 X 로 Y 적합 플랫폼이 더 적절합니다. 이 플랫폼에 대한 자세한 내용은 선형 모형 적합 및 기본 분석에서 확인하십시오.

목차

비선형 플랫폼의 예	237
비선형 플랫폼 시작	239
비선형 적합 보고서	240
비선형 플랫폼 옵션	244
모형 라이브러리를 사용하여 계산식 생성	248
비선형 모형 라이브러리 사용자 정의	250
비선형 플랫폼의 추가 예	251
최대 가능도의 예	252
이항 오차가 있는 프로빗 모형의 예	255
Poisson 손실 함수의 예	255
모수 한계 설정의 예	257
왼쪽 중도절단 데이터 분석의 예	260
Weibull 손실 함수 적합의 예	261
단순 생존 분포 적합의 예	264
비선형 플랫폼에 대한 통계 상세 정보	266
프로파일 가능도 신뢰 한계에 대한 통계 상세 정보	266
사용자 손실 함수에 대한 통계 상세 정보	267
도함수에 대한 통계 상세 정보	268
유효 비선형 모델링에 대한 통계 상세 정보	270

비선형 플랫폼의 예

이 예에서는 모형에 대한 계산식 열을 생성한 후 비선형 플랫폼으로 모형을 적합시켜 사용자 모형을 적합시킵니다. 이 방법을 사용하면 기본 제공 모형을 적합시키는 것보다 몇 단계가 더 필요하지만 비선형 모형을 적합시킬 수 있습니다. 또한 사용자 손실 함수를 제공하고 적합 프로세스에 대한 몇 가지 다른 옵션을 지정할 수도 있습니다.

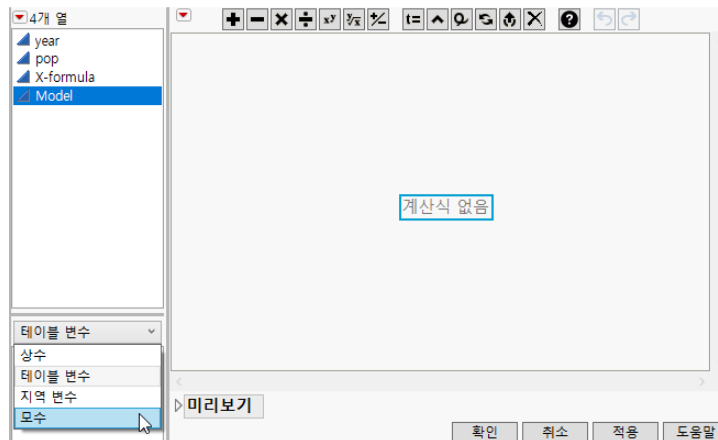
팁 : 사용자 모형이 여러 데이터 테이블에 자주 사용되는 모형인 경우 모형 라이브러리에 모형을 추가할 수 있습니다. 자세한 내용은 "[비선형 모형 라이브러리 사용자 정의](#)"에서 확인하십시오.

계산식 열 생성

사용자 모형을 적합시키려면 먼저 초기 모수 추정값으로 모형 열을 생성해야 합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Nonlinear Examples/US Population.jmp 를 엽니다.
2. 모형이라는 새 열을 생성합니다.
3. 모형 열을 마우스 오른쪽 버튼으로 클릭하고 **열 특성 > 계산식**을 선택합니다.
계산식 편집기가 나타납니다.
4. 열 목록 아래의 목록에서 **모수**를 선택합니다.

그림 15.2 모수 선택



5. **새 모수**를 선택합니다.
6. " 이름 " 에 "B0" 을 입력합니다.
7. " 값 " 에 "3.9" 를 입력합니다. 이 값이 모수의 초기 추정값입니다.
8. **확인**을 클릭합니다.
9. **새 모수**를 선택합니다.

10. " 이름 " 에 "B1" 을 입력하고 " 값 " 에 "0.022" 를 입력합니다 .
11. **확인**을 클릭합니다 .
12. 계산식 편집기 함수 , year 열 및 모수를 사용하여 모형 계산식을 입력합니다 .

그림 15.3 완료된 모형 계산식

$$B0 \cdot \text{Exp} \left(B1 \cdot (year - 1790) \right)$$

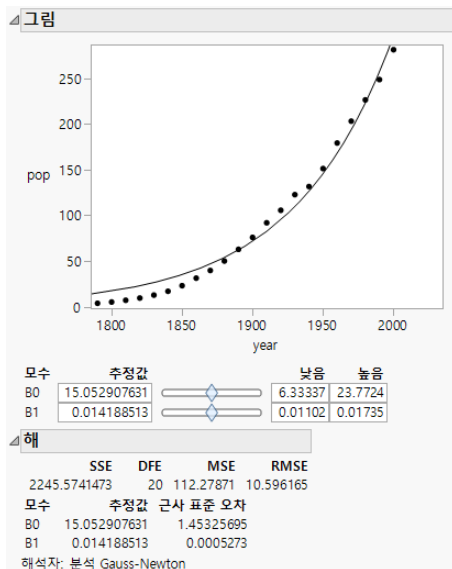
팁 : Exp 명령을 찾으려면 " 초월함수 " 옆의 회색 삼각형을 클릭합니다 .

13. **확인**을 클릭합니다 .

비선형 모형 적합

1. 분석 > 전문 모델링 > 비선형을 선택합니다 .
2. 모형을 **X, 예측변수 계산식** 역할에 할당합니다 .
3. pop 를 **Y, 반응** 역할에 할당합니다 .
4. **확인**을 클릭합니다 .
5. " 제어판 " 에서 **시작**을 클릭하여 모형을 적합시킵니다 .

그림 15.4 그림 및 해 보고서



해 보고서에 최종 모수 추정값이 다른 적합 통계량과 함께 표시됩니다. 적합 모형은 그림에 표시됩니다.

그룹화 변수가 있는 모형의 모수

계산식 편집기에서 모수를 추가할 때 **범주로 확장**, **열 선택** 체크박스에 주목하십시오. 이 옵션은 한 번에 여러 모수(예: 범주형 변수의 각 수준에 대해 하나씩)를 추가하는 데 사용됩니다. 이 옵션을 선택하면 열을 선택할 수 있는 대화상자가 나타납니다. 선택 후 "모수" 목록에 이름이 **D_열** 이름인 새 모수가 나타납니다. 여기서 **D**는 모수에 지정한 이름입니다. 이 모수를 계산식에 사용하면 그룹화 변수의 각 수준에 대한 개별 모수를 포함하는 Match 표현식이 삽입됩니다.

비선형 플랫폼 시작

비선형 플랫폼을 시작하려면 **분석 > 전문 모델링 > 비선형**을 선택합니다.

그림 15.5 비선형 플랫폼 시작 창

최소 제공 또는 사용자 손실 함수를 사용하여 비선형 모형을 적합시킵니다.

열 선택

7개 열

- Y
- X
- Model Y
- Loss
- Model2 Y
- Loss2
- Loss with No Model Column

모형 라이브러리

선택한 열 역할 지정

Y, 반응	선택적 숫자
X, 예측변수 계산식	Model Y
그룹	선택적
가중치	선택적 숫자
빈도	선택적 숫자
손실	Loss
기준	선택적

X 예측 변수 열에 모수가 포함된 계산식이 있거나 X 예측 변수 열이 기본 제공 모형과 함께 사용할 독립 변수입니다.

사용자 계산식 적합 옵션

예측 변수
재설정
Parameter({b0 = 0, b1 = 0}, b0 + b1 * X)

손실
재설정
if(:Y == 1,
-Log(1 / (1 + Exp(-:Model Y))),
-Log(1 - 1 / (1 + Exp(-:Model Y)))
)

☐ 수치 미분만
☐ 중간 계산식 확장

작업

확인

취소

제거

재로출

도움말

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

비선형 플랫폼 시작 창에는 다음과 같은 기능이 있습니다.

Y, 반응 Y 변수를 선택합니다.

X, 예측변수 계산식 X 변수를 선택하거나, 모수를 사용한 모형 계산식이 포함된 열을 선택합니다.

참고 : 모수를 사용한 모형 계산식이 포함되지 않은 열을 선택하고 사용자 계산식도 지정하지 않으면 곡선 적합 플랫폼이 시작됩니다.

참고 : 할당이 없는 모수 목록 항목이 포함된 계산식 열을 선택하면 JMP 에서 자동으로 해당 모수를 0 으로 설정하고 로그에 메시지를 표시합니다.

그룹 그룹화 변수를 지정합니다 . 적합 모형에는 그룹화 변수의 각 수준에 대한 개별 모수가 있습니다 . 이를 통해 그룹화 변수의 전체 수준에서 적합 모형과 추정 모수를 비교할 수 있습니다 .

참고 : 계산식에 각 그룹 수준에 대한 개별 모수가 없으면 JMP 에서 그룹 변수가 무시됩니다.

가중치 관측값의 가중치를 포함하는 변수를 지정합니다 .

빈도 관측값의 빈도를 나타내는 변수를 지정합니다 .

손실 손실 함수를 제공하는 계산식 열을 지정합니다 .

기준 변수의 모든 수준에 대해 개별 분석을 수행할 변수를 지정합니다 .

모형 라이브러리 계산식 열 생성을 위해 초기값을 선택하는 데 도움이 되는 " 모형 라이브러리 " 도구를 실행합니다 . 자세한 내용은 " [모형 라이브러리를 사용하여 계산식 생성](#) " 에서 확인하십시오 .

사용자 계산식 적합 옵션 모수를 사용한 모형 계산식이 포함되지 않은 열을 "X, 예측변수 계산식 " 역할에 지정하는 경우 이 옵션을 사용하여 사용자 계산식을 생성할 수 있습니다 . " 예측 변수 " 필드에서 모수, 모수 값, 선택한 X 변수를 사용하는 사용자 계산식을 지정합니다 . 그런 다음 **재설정** 버튼을 클릭하여 플랫폼을 시작합니다 .

수치 미분만 수치 미분만 사용합니다 . 이 옵션은 모형이 정리되지 않아 해석 도함수를 사용할 수 없는 경우에 유용합니다 . 또한 어려운 상황에서도 수렴에 도달하는 데 도움이 될 수 있습니다 . 이 옵션은 "X, 예측변수 계산식 " 역할에 계산식 열이 제공되는 경우에만 사용됩니다 .

중간 계산식 확장 모형에 대한 성분 열이 자체 계산식이 있는 열인 경우 해당 열이 다른 열을 참조하면 JMP 에서 내부 계산식을 대체하도록 지정합니다 . 성분 열이 확장되지 않도록 하려면 **기타** 열 특성을 사용하여 이름이 " 계산식 확장 " 이고 값이 0 인 특성을 할당합니다 . 이 옵션은 "X, 예측변수 계산식 " 역할에 계산식 열이 제공되는 경우에만 사용됩니다 .

비선형 적합 보고서

초기 " 비선형 적합 " 보고서에는 " 제어판 " 과 " 그림 " 이 포함됩니다 ([그림 15.6](#) 참조) .

제어판 적합 프로세스를 제어하기 위한 옵션을 제공합니다 .

시작 적합 프로세스를 시작합니다 .

중지 적합 프로세스를 중지합니다 .

단계 적합 프로세스를 한 번에 하나씩 순서대로 반복합니다 .

재설정 계산식에 대해 편집 가능한 값을 재설정하고, 반복 값을 재설정 후 이러한 새 값으로 SSE 를 계산합니다 .

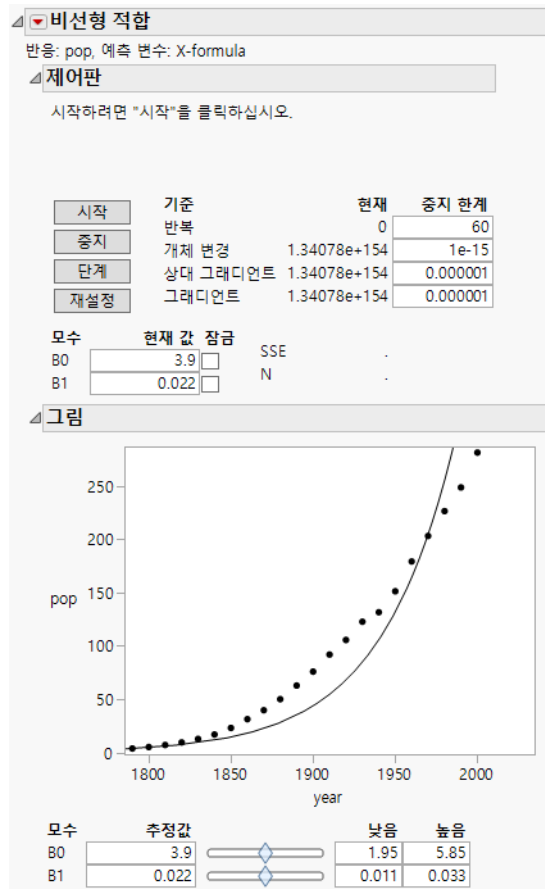
기준 적합 프로세스의 반복 측도를 표시합니다.

현재 각 기준의 현재 값을 표시합니다.

중지 한계 "기준" 아래에 나열된 측도에 대한 한계를 설정합니다.

그림 X 변수가 하나뿐인 모형의 X 및 Y 변수 그림을 표시합니다. 현재 값에 기반한 모형이 그림에 표시됩니다. 모수의 현재 값을 변경하려면 그림 아래의 슬라이더 또는 편집 상자를 사용합니다.

그림 15.6 초기 비선형 적합 보고서



시작을 클릭하여 모형을 적합시키면 다음과 같은 추가 항목이 보고서에 포함됩니다 ([그림 15.7](#) 참조).

추정값 저장 현재 모수 값을 계산식 열의 모수에 저장합니다.

신뢰 한계 모든 모수에 대한 신뢰 구간을 계산합니다. 이 구간은 프로파일 가능도 신뢰 구간이며 "해" 보고서에 표시됩니다. 신뢰 한계 계산에는 각 모수의 각 한계에 대한 새 반복 집합이 포함되며 반복을 통해 한계를 찾지 못하는 경우가 종종 있습니다. **알파 편집 및 수렴 기준** 옵션

선은 신뢰 구간 계산에 사용됩니다. **CL에 대한 목표 SSE**에 대한 자세한 내용은 "[프로파일 가능도 신뢰 한계에 대한 통계 상세 정보](#)"에서 확인하십시오.

해 모수 추정값 및 기타 통계량을 표시합니다.

SSE 잔차 오차 제곱합을 표시합니다. SSE는 최소화하는 것이 목표입니다. 사용자 손실 함수가 지정된 경우 이 값은 손실 함수의 합입니다.

DFE 오차의 자유도, 즉 사용된 관측값 수에서 적합된 모수의 수를 뺀 값을 표시합니다.

MSE 평균 제곱 오차를 표시합니다. 이 값은 잔차 오차 분산의 추정값, 즉 SSE를 DFE로 나눈 값입니다.

RMSE 잔차 오차의 표준편차, 즉 MSE의 제곱근을 추정합니다.

R² (모형에 절편이 있고 잔차 합이 0에 가까운 경우에만 사용 가능) R² 값을 표시합니다.

모수 적합 계산식에서 모수에 지정한 이름을 나열합니다.

추정값 생성된 모수 추정값을 나열합니다. 비선형 회귀의 경우 모든 것이 작동하는 것처럼 보여도 이 추정값에 문제가 있을 수 있습니다.

근사 표준 오차 선형 회귀와 유사하게 계산되는 근사 표준 오차를 나열합니다. 이 값은 RMSE와 도함수 교차곱 역행렬 대각값의 제곱근과의 곱에 의해 생성됩니다.

참고: 관측값 수가 적합 모형의 모수 수보다 작거나 같으면 근사 표준 오차가 보고되지 않습니다.

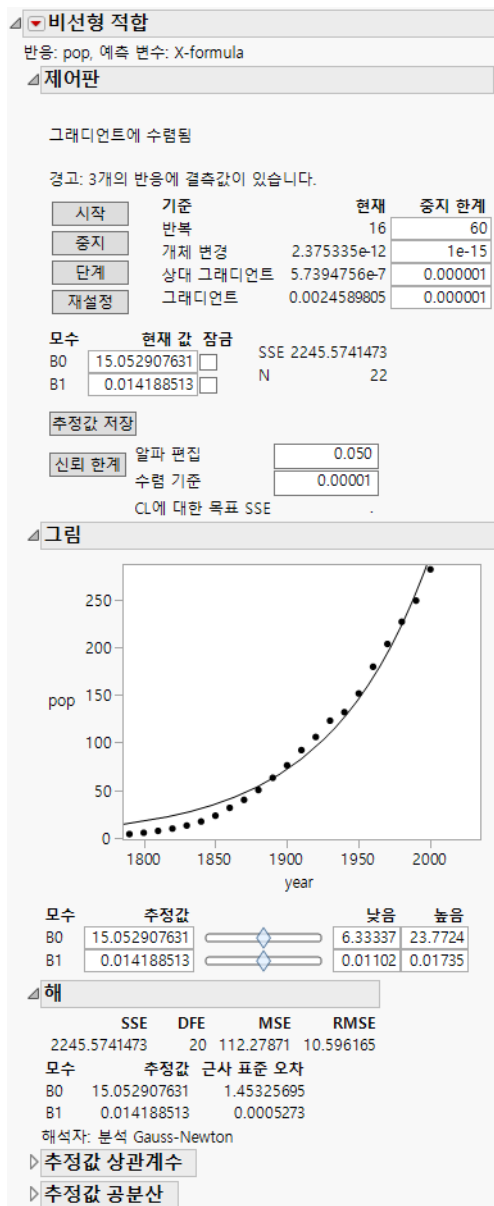
CL 하한 및 CL 상한 모수에 대한 신뢰 한계를 표시합니다. "제어판"에서 **신뢰 한계**를 클릭하지 않으면 이 값은 결측입니다. 신뢰 구간에 대한 자세한 내용은 "[프로파일 가능도 신뢰 한계에 대한 통계 상세 정보](#)"에서 확인하십시오.

제외된 데이터 제외된 행에 대한 적합 통계량을 보여 주는 보고서를 표시합니다. 이 옵션은 모형 적합에 사용되지 않은 관측값에 대해 모형을 검증하는 경우에 유용합니다. 이 기능을 "해 저장" 옵션과 함께 사용하여 제외 항목을 변경하고 다른 제외 항목을 반영하는 새 보고서를 얻을 수 있습니다.

추정값 상관계수 모수 추정값 간의 상관을 표시합니다.

추정값 공분산 모수 추정값 간의 공분산을 표시합니다.

그림 15.7 적합 모형 보고서



비선형 플랫폼 옵션

"비선형 적합"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

모수 경계 모수의 경계를 설정합니다. 이 옵션을 선택하면 "제어판"에 편집 가능한 상자가 나타납니다. 필드를 비워 두면 경계가 없는 모수임을 나타냅니다.

그림 X 변수가 하나뿐인 모형의 X 및 Y 변수 그림을 표시하거나 숨깁니다. 그림에 표시된 모형은 모수의 현재 값을 기반으로 합니다. 모수의 현재 값을 변경하려면 그림 아래의 슬라이더 또는 편집 상자를 사용합니다. 시작할 때 그룹 변수를 지정하면 각 그룹에 대한 곡선이 표시됩니다.

반복 옵션 적합 알고리즘에 대한 옵션을 지정합니다.

반복 로그 반복 테이블을 표시하거나 숨깁니다. 이 옵션을 선택하면 테이블에 후속 반복이 기록됩니다.

수치 미분만 모형이 정리되지 않아 해석 도함수를 사용할 수 없는 경우에 유용합니다. 또한 어려운 상황에서도 수렴에 도달하는 데 도움이 될 수 있습니다.

중간 계산식 확장 수식에 대한 성분 열이 자체 계산식이 있는 열인 경우 해당 열이 다른 열을 참조하면 JMP에서 내부 계산식을 대체하도록 지정합니다. 성분 열이 확장되지 않도록하려면 **기타** 열 특성을 사용하여 이름이 "계산식 확장"이고 값이 0인 특성을 할당합니다.

Newton Gauss-Newton(정규 최소 제곱의 경우) 또는 Newton-Raphson(손실 함수가 있는 모형의 경우)이 최적화 방법인지 여부를 지정합니다.

QuasiNewton SR1 QuasiNewton SR1을 최적화 방법으로 지정합니다.

QuasiNewton BFGS QuasiNewton BFGS를 최적화 방법으로 지정합니다.

현재 추정값 채택 추정값이 수렴되지 않은 경우에도 JMP에서 현재 추정값을 사용하여 해 보고서 생성하도록 지정합니다.

도함수 표시 비선형 계산식의 도함수를 JMP 로그에 표시합니다. 도함수에 대한 기술 정보는 "[도함수에 대한 통계 상세 정보](#)"에서 확인하십시오.

언스레드 주 계산 스레드에서 반복을 실행합니다. 대부분의 경우 JMP는 별도의 계산 스레드에서 계산을 수행합니다. 이렇게 하면 비선형 계산 중에 다른 작업을 수행하는 동안 JMP의 응답성이 향상됩니다. 그러나 주 스레드에서 실행해야 하는 일부 고립 상황(예: 디스플레이 루틴을 호출하는 파생 효과가 있는 모형)이 있으므로 이 옵션을 설정해야 합니다.

프로파일러 반응 표면을 보기 위한 다양한 프로파일러를 제공합니다.

프로파일러 예측 프로파일러를 표시합니다. 프로파일러를 사용하면 각 x 변수에서 차례대로 표면의 수직 단면을 볼 수 있으며 요인의 최적 값을 찾을 수 있습니다.

등고선 프로파일러 등고선 프로파일러를 표시합니다. 등고선 프로파일러를 사용하면 2차원 등고선과 3차원 그물 그림을 볼 수 있습니다.

표면 프로파일러 3 차원 표면 그림을 생성합니다. 이 옵션은 X 변수가 두 개 이상 있는 모형에만 사용할 수 있습니다.

모수 프로파일러 예측 프로파일러를 표시하고 모수의 함수로 나타낸 SSE 또는 손실을 프로파일링합니다.

모수 등고선 프로파일러 등고선 프로파일러를 표시하고 모수의 함수로 나타낸 SSE 또는 손실 등고선을 그림니다.

모수 표면 프로파일러 3 차원 표면 그림을 생성하고 모수의 함수로 나타낸 SSE 또는 손실을 프로파일링합니다. 이 옵션은 모수가 두 개 이상 있는 모형에만 사용할 수 있습니다.

SSE 격자 해 추정값 주변의 값 격자를 생성하고 각 값의 오차 제곱합을 계산합니다. 해 추정값의 SSE 가 최소값이어야 합니다. 이 옵션을 선택하면 다음과 같은 기능이 포함된 **출력에 대한 격자 지정** 보고서가 표시됩니다.

모수 모형의 모수를 나열합니다.

최소값 격자 계산에 사용된 최소 모수 값을 표시합니다. 기본적으로 최소값은 해 추정값에서 근사 표준 오차의 2.5 배를 뺀 값입니다.

최대값 격자 계산에 사용된 최대 모수 값을 표시합니다. 기본적으로 최대값은 해 추정값에서 근사 표준 오차의 2.5 배를 더한 값입니다.

점의 수 각 모수에 대해 생성할 점의 수를 제공합니다. 새 격자 테이블의 총 점 수를 계산하려면 모든 "점의 수" 값을 곱합니다. 처음에 "점의 수" 값은 처음 두 모수에 대해 11 이고 나머지 모수에 대해 3 입니다. 새 값을 지정하는 경우 홀수 값을 사용하여 격자 테이블에 해 추정값이 포함되도록 해야 합니다. 모수에 대해 "점의 수" 를 0 으로 설정하면 격자 테이블에 해 추정값만 기록됩니다.

시작을 클릭하면 JMP 에서 새 테이블에 점 격자를 생성합니다. 강조 표시된 행은 테이블에 해가 있는 경우 해 추정값 행을 나타냅니다.

원래 모수로 되돌리기 플랫폼을 원래 모수 값 (계산식 열 모수에 지정된 값) 으로 재설정합니다.

해 저장 현재 모수 추정값과 요약 통계량을 포함하는 "저장된 모형" 보고서를 생성합니다. 여러 모형의 결과를 저장하고 비교할 수 있습니다. 이 기능은 서로 다른 모수 제한에 기반한 모형 또는 다른 옵션을 사용하여 적합된 모형을 비교하려는 경우에 유용합니다. 특정 모형에 대한 라디오 버튼을 클릭하면 "그림" 에 해당 모형이 표시되고 "제어판" 에 모수 추정값이 표시됩니다.

사용자 추정값 모수 함수의 추정값을 제공합니다. 사용자는 모수만 포함된 표현식을 제공합니다. 그러면 JMP 에서 현재 모수 추정값을 사용하여 표현식을 계산하고 1 차 Taylor 급수 근사를 사용하여 표현식의 표준 오차도 계산합니다.

사용자 추정 프로파일러 사용자 표현식에 대한 프로파일러를 생성할 수 있습니다. 이 유형의 프로파일러는 지정한 비선형 모형에 유용합니다. 모수와 하나 이상의 요인이 포함된 표현식을 제공합니다. 표현식에 변환을 적용하는 옵션도 있습니다. "확인" 을 클릭한 후 요인의 초기 값을 지정합니다. 표현식의 프로파일러는 전체 요인 범위에 대한 현재 모수 추정값을 기반으로 합니다. 지정된 초기값에 프로파일러가 설정됩니다. "사용자 추정 프로파일러" 옵션을

사용하여 여러 개의 사용자 프로파일을 추가할 수 있습니다. 각 사용자 프로파일러의 빨간색 삼각형 메뉴에는 프로파일러 그래프 및 요인 설정에 대한 옵션이 포함되어 있습니다. 자세한 내용은 **Profilers** 에서 확인하십시오. 프로파일러를 제거하려면 해당하는 빨간색 삼각형을 클릭하고 "프로파일러 제거"를 선택합니다. 자세한 내용은 "[사용자 프로파일러 생성](#)"에서 확인하십시오.

참고 : 기본적으로 표현식 및 요인에 대한 프로파일러 축 범위가 좁을 수 있습니다. 축을 두 번 클릭하고 최소값과 최대값을 새로 지정하여 범위를 조정할 수 있습니다.

사용자 역추정 예측 주어진 Y 값에 대한 X 값을 추정합니다. 또한 추정된 X 의 표준 오차를 계산합니다. JMP에서 모형을 반전할 수 있어야 합니다. 표준 오차는 반전 표현식을 사용하는 1차 Taylor 급수 근사를 기반으로 합니다. 신뢰 구간은 t -분위수를 표준 오차와 함께 사용하며 Wald 구간입니다.

예측 표현식 표시 보고서 맨 위에 예측 모형 또는 손실 함수를 표시합니다.

예측 신뢰 한계 저장 모형 예측에 대한 점근적 신뢰 한계를 저장합니다. 이 값은 주어진 X 값에서 평균 Y 에 대한 신뢰 구간입니다.

개별 신뢰 한계 저장 개별 예측에 대한 점근적 신뢰 한계를 저장합니다. 이 값은 주어진 X 값에서 개별 Y 값에 대한 신뢰 구간입니다.

계산식 저장 모형 결과를 데이터 테이블 옆에 저장하기 위한 옵션을 제공합니다.

예측 계산식 저장 현재 모수 추정값을 사용하여 예측 계산식을 저장합니다.

예측값의 표준 오차 저장 모형 예측에 대한 표준 오차를 저장합니다. 이 값은 주어진 X 에 대해 평균 Y 를 예측하기 위한 표준 오차입니다. 계산식은 $\text{Sqrt}(\text{VecQuadratic}(\text{matrix1}, \text{vector1}))$ 형식입니다. 여기서 **matrix1**은 모수 추정값과 관련된 공분산 행렬이고 **vector1**은 각 모수에 대한 모형의 편도함수 합성입니다.

개별값의 표준 오차 저장 개별 예측에 대한 표준 오차를 저장합니다. 이 값은 주어진 X 값에 대해 개별 Y 값을 예측하기 위한 표준 오차입니다. 계산식은 $\text{Sqrt}(\text{VecQuadratic}(\text{matrix1}, \text{vector1}) + \text{mse})$ 형식입니다. 여기서 **matrix1**은 모수 추정값과 관련된 공분산 행렬, **vector1**은 각 모수에 대한 모형의 편도함수 합성, **mse**는 오차 분산 추정값입니다.

잔차 계산식 저장 잔차를 계산하기 위한 계산식을 저장합니다.

예측 신뢰 한계 계산식 저장 모형 예측에 대한 신뢰 구간을 계산하는 계산식을 저장합니다. 이 값은 주어진 X 에 대한 평균 Y 의 신뢰 구간입니다.

개별 신뢰 한계 계산식 저장 개별 예측에 대한 신뢰 구간을 계산하는 계산식을 저장합니다. 이 값은 주어진 X 에 대한 개별 Y 의 신뢰 구간입니다.

역추정 예측 계산식 저장 모형의 역추정, 역추정 예측의 표준 오차 및 개별 역추정 예측의 표준 오차에 대한 계산식을 저장합니다.

지정된 풀이 계산식 저장 간단한 사례의 경우 "역추정 예측 계산식 저장"과 동일합니다. 그러나 이 명령을 사용하면 계산식이 여러 변수의 함수가 될 수 있으며 표현식을 대체할 수 있습니다. 이 기능은 계산식에서 한 번만 나오는 가역 연산자 및 함수를 쉽게 풀이하는 데만 사용할 수 있습니다.

이 명령을 선택하면 해석할 변수를 선택할 수 있는 대화상자가 나타납니다. 결과 테이블의 열 이름을 편집할 수도 있습니다. 또한 이 대화상자에서 이름을 값으로 대체할 수도 있습니다. 이러한 경우 해당 값에 대한 계산식이 풀이됩니다.

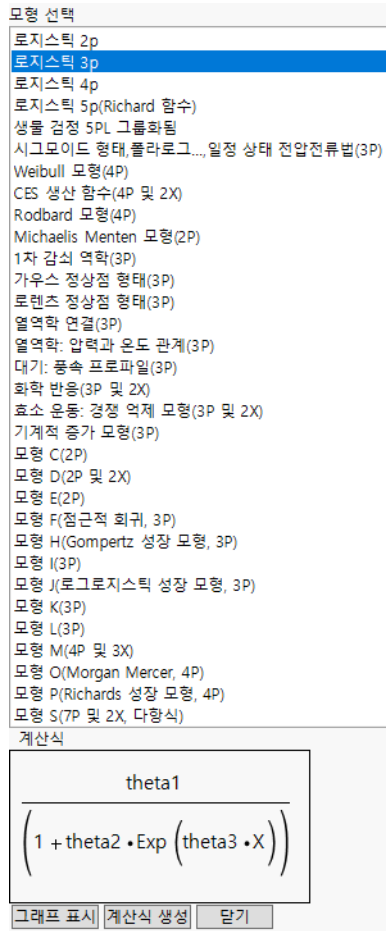
참고 : 표준 오차, 신뢰 구간 및 가설 검정은 최소 제곱 추정이 수행되거나, 최대 가능도 추정이 적절한 음의 로그 가능도와 함께 사용되는 경우에만 올바릅니다.

추정값을 테이블에 저장 모수 추정값을 새 데이터 테이블에 저장합니다.

모형 라이브러리를 사용하여 계산식 생성

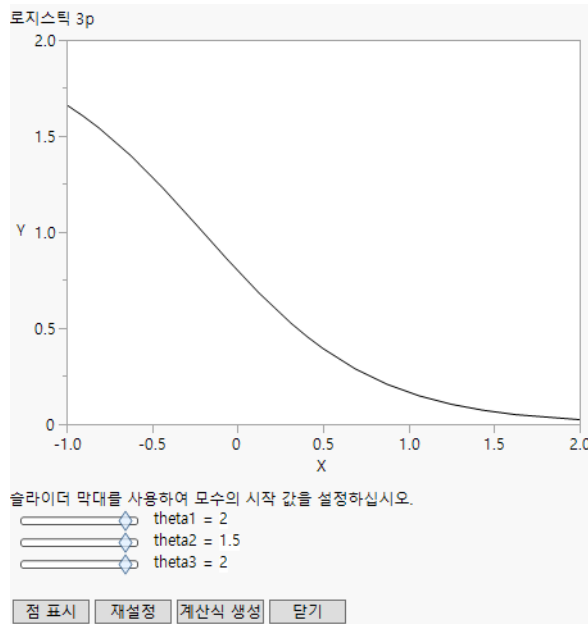
비선형 플랫폼의 모형 라이브러리를 사용하여 모수 및 초기값이 포함된 계산식 열을 생성할 수 있습니다. 비선형 시작 창에서 **모형 라이브러리** 버튼을 클릭하여 라이브러리를 엽니다. 목록에서 모형을 선택하면 **계산식** 상자에 해당 모형의 계산식이 표시됩니다.

그림 15.8 비선형 모형 라이브러리 대화상자



그래프 표시를 클릭하면 1모수 모형의 2D 이론적 곡선과 2모수 모형의 3D 표면 그림이 표시됩니다. 설명(X) 변수가 세 개 이상 있는 모형에는 그래프를 사용할 수 없습니다. 그래프 창에서 슬라이더를 사용하거나, 값을 직접 클릭한 후 입력하여 모수의 기본 초기값을 변경합니다.

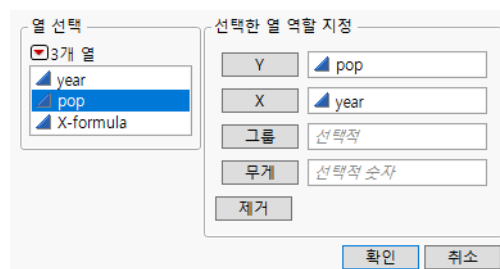
그림 15.9 모형 라이브러리의 그래프 예



재설정 버튼은 모수의 초기값을 기본값으로 다시 설정합니다.

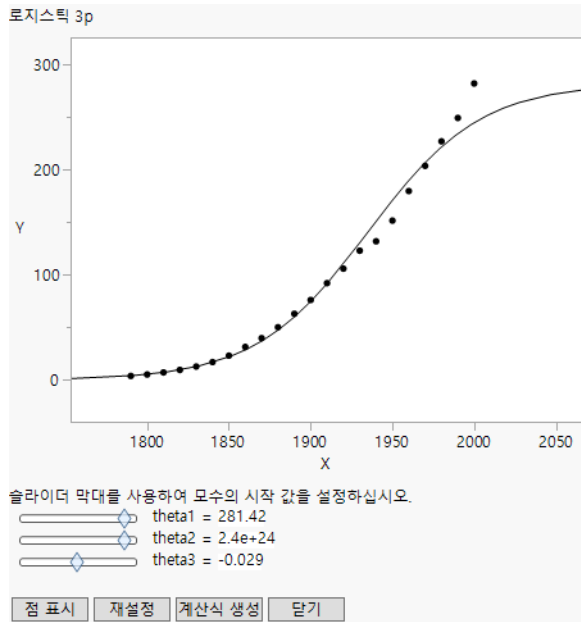
실제 데이터 점을 그림에 중첩하려면 **점 표시**를 클릭합니다. X 및 Y 역할과 선택적 그룹 역할에 열을 할당하라는 대화상자가 열립니다([그림 15.10](#) 참조). 그룹 역할을 사용하면 범주형 변수의 모든 수준에 모형을 적합시킬 수 있습니다. 여기에 그룹 역할을 지정하는 경우 플랫폼 시작 창의 그룹 열도 지정합니다.

그림 15.10 역할 선택



대부분의 모형에서 시작 값은 상수입니다. 점을 표시하면 모수 값을 조정하여 모형이 모수의 여러 값에 따라 얼마나 잘 적합되는지 확인할 수 있습니다. 미국 인구 예의 경우 [그림 15.11](#)에 점이 표시되어 있습니다.

그림 15.11 점 표시



점 표시를 사용한 후 **계산식 생성**을 클릭하면 모형 라이브러리에서 선택한 모형을 따라 명명된 새 데이터 테이블 열이 생성됩니다. 이 열에는 최신 모수 시작 값의 함수로 나타낸 계산식이 있습니다.

참고 : **그래프 표시** 또는 **점 표시** 버튼을 사용하기 전에 **계산식 생성**을 클릭하면 X 및 Y 역할과 선택적 그룹 역할을 제공해야 합니다 ([그림 15.10](#)). 그런 다음 원할 경우 모수 시작 값을 조정할 수 있도록 그림으로 돌아갑니다. 이때 **계산식 생성**을 다시 클릭하여 새 열을 생성합니다.

데이터 테이블에 계산식이 생성되면 비선형 시작 대화상자에서 새 열을 **X, 예측변수 계산식**으로 할당하여 분석을 계속합니다.

비선형 모형 라이브러리 사용자 정의

비선형 플랫폼의 모형 라이브러리는 JMP(Windows)가 포함된 폴더 또는 Application Package(macOS)의 Resources/Builtins 폴더에 기본 제공된 NonlinLib.jsl 스크립트에 의해 생성됩니다. 이 스크립트를 수정하여 비선형 라이브러리 스크립트를 사용자 정의할 수 있습니다.

참고 : 파일을 직접 편집하지 못할 수도 있습니다. 이 경우 스크립트를 열고 원하는 대로 편집한 후 컴퓨터의 다른 위치에 저장합니다. 이때 동일한 파일 이름을 사용하십시오. 그런 다음 Resources/Builtins 폴더의 기존 NonlinLib.jsl 스크립트를 삭제하고 새 NonlinLib.jsl 파일을 Resources/Builtins 폴더로 이동합니다.

모형을 추가하려면 `Listofmodellist#` 목록에 세 줄을 추가해야 합니다. 이 세 줄은 사실 다음과 같은 세 부분으로 구성된 목록입니다.

- 모형 (따옴표로 묶은 문자열)
- 모형 계산식 (표현식)
- 모형 척도

예를 들어 다음과 같은 형식의 "Simple Exponential Growth" 라는 모형을 추가하려는 경우를 가정해 보겠습니다.

$$y = b_1 e^{kx}$$

다음 줄을 `NonlinLib.jsl` 스크립트에 추가합니다.

```
{//Simple Exponential Growth  
  "Simple Exponential Growth",  
  Expr(Parameter({b1=2, k=0.5}, b1*exp(k * :X))),  
  lowx = -1; highx = 2; lowy = 0; highy = 2},
```

다음 사항에 유의하십시오.

- 첫 번째 줄은 여는 대괄호 (목록 시작) 와 선택적 주석입니다. 두 번째 줄은 모형 라이브러리 창에 표시되는 문자열입니다.
- `lowx`, `highx`, `logy` 및 `highy` 값은 이론적 그래프의 초기 창을 지정합니다.
- 위의 예에서는 쉼표가 마지막 문자로 사용됩니다. `Listofmodellist#` 목록의 마지막 항목에는 쉼표를 생략할 수 있습니다.
- 모형에 세 개 이상의 모수가 사용되는 경우 마지막 줄 (그래프 한계 포함) 을 따옴표로 묶은 "String Not Available" 문자열로 바꿉니다.

모형을 삭제하려면 `Listofmodellist#` 목록에서 해당하는 세 줄 목록을 삭제합니다.

비선형 플랫폼의 추가 예

이 섹션에는 비선형 플랫폼을 사용하는 예가 포함되어 있습니다.

- " 최대 가능도의 예 "
- " 이항 오차가 있는 프로빗 모형의 예 "
- "Poisson 손실 함수의 예 "
- " 모수 한계 설정의 예 "
- " 왼쪽 중도절단 데이터 분석의 예 "
- "Weibull 손실 함수 적합의 예 "
- " 단순 생존 분포 적합의 예 "

최대 가능도의 예

이 예에서는 비선형 플랫폼을 사용하여 손실 함수를 최소화합니다. 손실 함수는 음의 로그 가능도 함수이므로 최대 가능도 추정값을 생성합니다.

모형 추정

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Nonlinear Examples/Logistic w Loss.jmp 를 엽니다 .
2. **분석 > 전문 모델링 > 비선형**을 선택합니다 .
3. Model Y 를 **X, 예측변수 계산식** 역할에 할당합니다 .
이 열은 선형 모형을 포함합니다 . 시작 창의 " 예측 변수 " 섹션에 모형이 표시됩니다 .
4. Loss 를 **손실** 역할에 할당하여 사용자 손실 함수를 정의합니다 .

그림 15.12 비선형 시작 창

최소 제곱 또는 사용자 손실 함수를 사용하여 비선형 모형을 적합시킵니다.

The screenshot shows the Nonlinear Start Window with the following sections:

- 열 선택 (Column Selection):** A list of 7 columns: Y, X, Model Y, Loss, Model2 Y, Loss2, and Loss with No Model Column. The 'Loss' column is selected.
- 선택한 열 역할 지정 (Selected Column Role Assignment):** A table for assigning roles to columns:

역할 (Role)	선택한 열 (Selected Column)
Y, 반응 (Y, Response)	Model Y
X, 예측변수 계산식 (X, Predictor Formula)	Model Y
그룹 (Group)	선택적 (Optional)
가중치 (Weight)	선택적 숫자 (Optional Number)
빈도 (Frequency)	선택적 숫자 (Optional Number)
손실 (Loss)	Loss
기준 (Criteria)	선택적 (Optional)
- 작업 (Actions):** Buttons for '확인' (OK), '취소' (Cancel), '제거' (Remove), '재호출' (Reopen), and '도움말' (Help).
- 모형 라이브러리 (Model Library):** A button to open the model library.
- 사용자 계산식 적합 옵션 (User Formula Fit Options):**
 - 예측 변수 (Predictor Variable):** Parameter({b0 = 0, b1 = 0}, b0 + b1 * X)
 - 손실 (Loss):**

```

if( :Y == 1,
  -Log( 1 / (1 + Exp( -:Model Y )) ),
  -Log( 1 - 1 / (1 + Exp( -:Model Y )) )
)

```
 - ☐ 수치 미분만 (Only Numerical Differentiation)
 - ☐ 중간 계산식 확장 (Expand Intermediate Formulas)

5. **확인**을 클릭합니다 .

그림 15.13 비선형 적합 제어판

비선형 적합

예측 변수: Model Y, 손실: Loss

제어판

시작하려면 "시작"을 클릭하십시오.
결측값 증가를 방지하기 위해 파라미터가 0인 것을 작은 값으로 재설정합니다.

	기준	현재	중지 한계
시작	반복	0	60
중지	개제 변경	1.34078e+154	1e-15
단계	상대 그래디언트	1.34078e+154	0.000001
재설정	그래디언트	1.34078e+154	0.000001

☒ 손실이 음의 로그 가능도임

모수	현재 값	잠금	손실
b0	1e-100	☐	.
b1	1e-100	☐	N

6. " 제어판 " 에서 **시작**을 클릭합니다 .

그림 15.14 해 보고서

해				
손실	DFE	평균 손실	제곱근 평균 손실	
7.8070527905	18	0.4337252		0.6585781
모수	추정값	근사 표준 오차		
b0	-5.511451052	2.43728165		
b1	0.0347553478	0.01595111		
해석자: 분석 NR				

" 해 " 보고서의 " 손실 " 값은 모수 추정값에서 계산된 음의 로그 가능도입니다 .

사용자 프로파일러 생성

사용자 손실 함수를 사용하여 모형이 적합되었으므로 표준 프로파일러를 사용할 수 없습니다. 사용자 프로파일러를 사용하여 모형 예측을 탐색합니다.

1. " 비선형 적합 " 의 빨간색 삼각형을 클릭하고 **사용자 추정 프로파일러**를 선택합니다 .
2. 모형을 사용자 추정 표현식으로 입력합니다 . 이렇게 하려면 " 표현식 입력 " 창에 다음 표현식을 입력합니다 .

$$1/(1 + \text{Exp}(b0 + b1*x))$$

3. " 변환 " 목록에서 " 로짓 " 을 선택합니다 .

그림 15.15 완료된 표현식 입력 창

4. **확인**을 클릭합니다 .

5. 초기값에 "100" 을 입력합니다 .

6. **확인**을 클릭합니다 .

기본적으로 프로파일러는 반응 및 요인의 범위가 좁기 때문에 해석하기가 어렵습니다. 축 설정을 조정하여 의미 있는 범위를 사용할 수 있습니다.

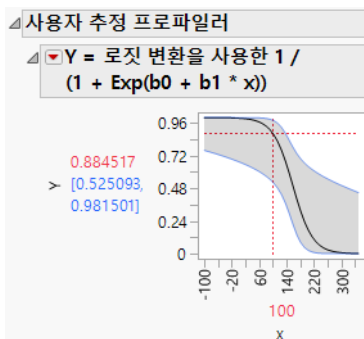
7. 세로 축을 두 번 클릭합니다 ." 척도 " 개요에서 0 을 최소값으로 지정하고 1 을 최대값으로 지정합니다 .

8. **확인**을 클릭합니다 .

9. 가로 축을 두 번 클릭합니다 ." 척도 " 개요에서 -100 을 최소값으로 지정하고 400 을 최대값으로 지정합니다 .

10. **확인**을 클릭합니다 .

그림 15.16 사용자 추정 프로파일러



표시된 사용자 프로파일러는 사용자 표현식으로 제목이 지정되고 전체 x 변수 값에 대한 반응 값을 표시합니다 . 95% 신뢰 구간도 프로파일러에 표시됩니다 .

이항 오차가 있는 프로빗 모형의 예

비선형 플랫폼에서 수치 미분을 사용하여 이항 오차가 있는 프로빗 모형에 대한 최대 가능도 추정값을 구합니다. 이 예의 데이터 테이블에는 가열 및 균열 처리 시간을 다르게 한 후 준비 상태를 테스트한 주괴 수가 포함되어 있습니다. 준비된 평균 주괴 수는 테스트한 수와 주어진 가열 및 균열 처리 시간에서 주괴 사용 준비가 완료될 확률의 곱입니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Ingots2.jmp** 를 엽니다.

2. **분석 > 전문 모델링 > 비선형**을 선택합니다.

3. **P** 를 **X**, **예측변수 계산식** 역할에 할당합니다.

프로빗 모형을 사용하여 **P** 열에 다음과 같은 모형 계산식이 포함됩니다.

Normal Distribution(**b0+b1*Heat+b2*Soak**)

Normal Distribution 함수에 대한 인수는 처리의 선형 모형입니다.

4. **Loss** 를 **손실** 역할에 할당합니다.

이항 오차를 지정하기 위해 **Loss** 열에 다음 계산식을 사용하는 손실 함수가 포함됩니다.

$-(Nready * \text{Log}(p) + (Ntotal - Nready) * \text{Log}(1 - p))$

5. **수치 미분만** 옵션을 선택합니다.

6. **확인**을 클릭합니다.

7. **시작**을 클릭합니다.

이 플랫폼에서는 [그림 15.17](#)에 표시된 모수 추정값을 얻기 위해 수치 SR1 방법을 사용했습니다.

그림 15.17 Ingots2 데이터의 해

해				
손실	DFE	평균 손실	제공된 평균 손실	
47.479945327	16	2.9674966	1.7226423	
모수	추정값	근사 표준 오차		
b0	-2.893415255	0.51357485		
b1	0.0399554549	0.01202873		
b2	0.0362537778	0.15060404		
해석자: 수치 SR1				

Poisson 손실 함수의 예

비선형 플랫폼을 사용하여 Poisson 모형을 개수 데이터에 적합시킵니다. Poisson 분포에 다음과 같은 파라미터화가 사용됩니다.

$$P(Y = n) = \frac{e^{-\mu} \mu^n}{n!}, n = 0, 1, 2, \dots$$

여기서 μ 는 단일 모수이거나, 여러 모수가 있는 선형 모형일 수 있습니다. 많은 교재와 논문에서 모형을 변환하고 반복 재가중 최소 제곱을 사용하여 적합시킬 수 있는 방법을 보여 줍니다.

(Nelder and Wedderburn 1972 연구 자료). 그러나 JMP에서는 모형을 직접 적합시키는 것이 더 간단합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Ship Damage.jmp 를 엽니다 .
2. **분석 > 전문 모델링 > 비선형**을 선택합니다 .
3. model 을 **X, 예측변수 계산식** 역할에 할당합니다 .
4. Poisson 을 **손실** 역할에 할당합니다 .
5. **확인**을 클릭합니다 .
6. b0 의 **현재 값** (초기값) 을 1 로 설정하고 다른 모수는 0 으로 설정합니다 .

그림 15.18 새 모수 입력

제어판

시작하려면 "시작"을 클릭하십시오.

	기준	현재	중지 한계
시작	반복	0	60
중지	개제 변경	1.34078e+154	1e-15
단계	상대 그래디언트	1.34078e+154	0.000001
재설정	그래디언트	1.34078e+154	0.000001

☒ 손실이 음의 로그 가능도임

모수	현재 값	잠금	손실
b0	1	☐	.
B	0	☐	N
C	0	☐	
D	0	☐	
E	0	☐	
Yr 65	0	☐	
Yr 70	0	☐	
Yr 75	0	☐	
Used 75	0	☐	

7. **시작**을 클릭합니다 .
8. **신뢰 한계** 버튼을 클릭합니다 .

"해" 보고서가 나타납니다. 결과에는 모수 추정값, 신뢰 구간 및 기타 요약 통계량이 포함됩니다.

그림 15.19 Poisson 손실 예의 해 테이블

해					
	손실	DFE	평균 손실	제곱근 평균 손실	
	68.281234087	25	2.7312494	1.6526492	
모수	추정값	근사 표준 오차	CL 하한	CL 상한	
b0	-6.405914771	0.21744445	-6.8430512	-5.9896833	
B	-0.543353982	0.17758996	-0.881379	-0.183552	
C	-0.687418388	0.32904722	-1.3764541	-0.0745299	
D	-0.075979835	0.2905789	-0.6715296	0.47523936	
E	0.325581683	0.23587934	-0.143451	0.78520431	
Yr 65	0.6971496668	0.1496412	0.40752821	0.99512758	
Yr 70	0.8184263427	0.16977314	0.48728046	1.15369087	
Yr 75	0.4534435316	0.23317069	-0.0123098	0.90388895	
Used 75	0.3844880718	0.11827198	0.15341509	0.61742312	
해석자: 분석 NR					

참고 : 표준 오차, 신뢰 구간 및 가설 검정은 최소 제곱 추정이 수행되거나, 최대 가능도 추정이 적절한 음의 로그 가능도와 함께 사용되는 경우에만 올바릅니다.

모수 한계 설정의 예

이 예에서는 곡선 적합에서 예측 계산식을 저장한 후 비선형에서 모수 한계를 설정하는 방법을 보여 줍니다. 곡선 적합 분석법을 사용하면 모형을 적합시킨 후 비선형 플랫폼의 전체 분석법에서 예측 방정식을 사용할 수 있습니다. 이 방법은 더 많은 단계와 사용자 입력이 필요하지만 비선형 모형을 적합시킬 수 있습니다.

1. "곡선 적합 플랫폼의 예"의 1 단계 ~ 8 단계를 완료하여 모형을 적합시킵니다.
2. "로지스틱 4P"의 빨간색 삼각형을 클릭하고 **계산식 저장 > 모수 예측 계산식 저장**을 선택합니다.

Toxicity 예측 변수라는 새 열이 데이터 테이블에 나타납니다.

3. **분석 > 전문 모델링 > 비선형**을 선택합니다.
4. Toxicity를 **Y, 반응** 역할에 할당합니다.
5. Toxicity 예측 변수를 **X, 예측변수 계산식** 역할에 할당합니다.
6. Formulation을 **그룹** 역할에 할당합니다.
7. **확인**을 클릭합니다.

"비선형 적합" 창이 나타납니다. "제어판"에 모수 값과 잠금 옵션이 표시됩니다. 각 모수 앞에 나열된 문자는 곡선 적합 기능에서 예측 모형의 변수에 해당합니다.

그림 15.20 비선형 적합 제어판

비선형 적합
반응: Toxicity, 예측 변수: Toxicity 예측 변수, 그룹: Formulation

제어판
시작하려면 "시작"을 클릭하십시오.

	시작	기준	현재	중지 한계
반복			0	60
개제 변경			1.34078e+154	1e-15
단계			상대 그래디언트 1.34078e+154	0.000001
재설정			그래디언트 1.34078e+154	0.000001

모수	현재 값	잠금	SSE	
a_standard	5.8437931223	☐	N	.
b_standard	0.320199595	☐		.
c_standard	0.5990151869	☐		
d_standard	1.1068636768	☐		
a_test A	5.9806928009	☐		
b_test A	0.284551222	☐		
c_test A	0.5909023298	☐		
d_test A	1.101208093	☐		
a_test B	5.7558678529	☐		
b_test B	0.0844294427	☐		
c_test B	0.592984828	☐		
d_test B	1.0972015134	☐		
a_test C	6.0099274621	☐		
b_test C	0.298749146	☐		
c_test C	0.5934036815	☐		
d_test C	1.1059793	☐		

팁 : 사전 정보에서 값을 알고 있는 경우 모수를 잠글 수 있습니다.

8. "비선형 적합"의 빨간색 삼각형을 클릭하고 **모수 경계**를 선택합니다.

모수 상한 및 하한을 설정하기 위한 옵션이 모수 옆에 나타납니다.

9. [그림 15.21](#)에 표시된 대로 모수의 하한을 설정합니다. 사전 경험을 통해 약물의 최대 독성이 1.1 이상이라는 것을 알고 있습니다.

그림 15.21 모수 경계 설정

제어판

시작하려면 "시작"을 클릭하십시오.

시작	기준	현재	중지 한계
중지	반복	0	60
단계	개제 변경	1.34078e+154	1e-15
재설정	상대 그래디언트	1.34078e+154	0.000001
	그래디언트	1.34078e+154	0.000001

모수	현재 값	잠금	하한	상한	SSE
a_standard	5.8437931223	☐	-	-	N
b_standard	0.320199595	☐	-	-	
c_standard	0.5990151869	☐	-	-	
d_standard	1.1068636768	☐	1.1	-	
a_test A	5.9806928009	☐	-	-	
b_test A	0.284551222	☐	-	-	
c_test A	0.5909023298	☐	-	-	
d_test A	1.101208093	☐	1.1	-	
a_test B	5.7558678529	☐	-	-	
b_test B	0.0844294427	☐	-	-	
c_test B	0.592984828	☐	-	-	
d_test B	1.0972015134	☐	1.1	-	
a_test C	6.0099274621	☐	-	-	
b_test C	0.298749146	☐	-	-	
c_test C	0.5934036815	☐	-	-	
d_test C	1.1059793	☐	1.1	-	

10. **시작**을 클릭합니다.

해 보고서에 최종 모수 추정값이 다른 적합 통계량과 함께 표시됩니다. 적합 모형은 그림에 표시됩니다.

그림 15.22 비선형 적합 그림 및 모수 추정값

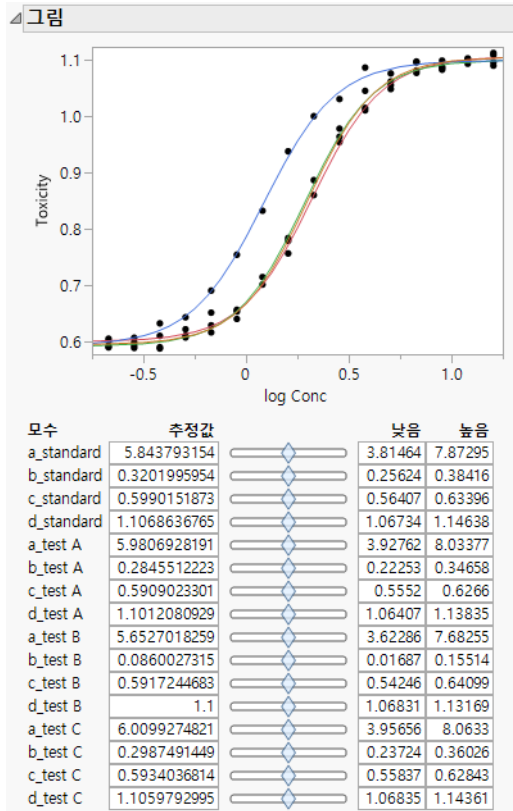


그림 아래의 옵션을 사용하여 모수 한계 및 추정값을 조정할 수 있습니다.

왼쪽 중도절단 데이터 분석의 예

비선형 플랫폼을 사용하여 왼쪽 중도절단 데이터를 분석할 수 있습니다. 이 예에서 왼쪽 중도절단 데이터의 경우 관측값에 대해 알려진 최소 시간을 나타내므로 0이 중도절단된 값입니다.

참고 : Tobit 모형은 경제학에서 반응이 양수 또는 0 이어야 하는 경우에 많이 사용됩니다. 이때 0 은 중도절단된 점을 나타냅니다. 이 모형은 0 에서 중도절단되는 정규 분포를 가정합니다. 관측값이 0 이면 왼쪽 중도절단으로 간주됩니다.

Tobit2.jmp 데이터 테이블의 경우 다음 사항에 유의하십시오.

- 반응 변수는 제품의 내구성에 대한 측도이며 0 보다 작을 수 없습니다 (durable, 0 에서 왼쪽 중도절단됨).
- age 와 liquidity 는 독립 변수입니다.

- 테이블에 모형과 Tobit 손실 함수도 포함되어 있습니다. 잔차 형식의 모형은 $\text{durable} - (b_0 + b_1 * \text{age} + b_2 * \text{liquidity})$ 입니다.

팁 : Tobit Loss 와 연결된 계산식을 보려면 열을 마우스 오른쪽 버튼으로 클릭하고 **계산식**을 선택합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Reliability/Tobit2.jmp 를 엽니다 .
2. **분석 > 전문 모델링 > 비선형**을 선택합니다 .
3. Model 을 선택하고 **X, 예측변수 계산식**을 클릭합니다 .
4. Tobit Loss 를 선택하고 **손실**을 클릭합니다 .
5. **확인**을 클릭합니다 .
6. **시작**을 클릭합니다 .
7. **신뢰 한계**를 클릭합니다 .

그림 15.23에 신뢰 한계와 함께 모수 추정값이 나와 있습니다.

그림 15.23 해 보고서

해					
	손실	DFE	평균 손실	제곱근 평균 손실	
	28.92596097	16	1.8078726	1.3445715	
모수	추정값	근사 표준 오차	CL 하한	CL 상한	
b0	15.277120063	16.0327167	-22.059913	54.7012259	
b1	-0.134007501	0.21893135	-0.7365041	0.32841861	
b2	-0.045135584	0.05826851	-0.1858906	0.09330941	
Sigma	5.5693492202	1.72814365	3.31613117	11.5291865	
해석자: 분석 NR					

Weibull 손실 함수 적합의 예

이 예에서 모형은 비선형 플랫폼의 Weibull, 로그 정규 및 지수 분포를 사용하여 생존 시간에 적합됩니다. 모형 적합에는 두 가지 효과만 있는 단순 생존 모형, 모든 효과가 있는 더 복잡한 모형, 전체 모형에 포함할 이산 효과 Cell Type에 대한 가변수 생성이 포함됩니다.

비선형 모형 적합은 모형 모수에 제공하는 초기값의 영향을 받는 경우가 많습니다. 이 예에서 적절한 초기값을 찾는 한 가지 방법은 먼저 비선형 플랫폼을 사용하여 선형 모형만 적합시키는 것입니다. 모형이 수렴되면 모수의 해 값이 비선형 모형의 초기 모수 값이 됩니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 VA Lung Cancer.jmp 를 엽니다 .

첫 번째 모형과 모든 손실 함수는 이미 데이터 테이블에 계산식으로 생성되어 있습니다 . Model 열의 계산식은 다음과 같습니다 .

$\text{Log}(\text{Time}) - (b_0 + b_1 * \text{Age} + b_2 * \text{Diag Time})$

2. **분석 > 전문 모델링 > 비선형**을 선택합니다 .
3. Model 을 선택하고 **X, 예측변수 계산식**을 클릭합니다 .
4. **확인**을 클릭합니다 .

5. **시작**을 클릭합니다.

그림 15.24 비선형 적합 제어판의 초기 모수 값

비선형 적합

예측 변수: Model

제어판

그래디언트에 수렴됨

	시작	기준	현재	중지 한계
반복			1	60
개제 변경			98.176266643	1e-15
상대 그래디언트			1.869125e-14	0.000001
그래디언트			2.403991e-13	0.000001

모수

모수	현재 값	잠금
b0	3.9398815293	☐
b1	0.0045053136	☐
b2	-0.012448139	☐

SSE 237.06333652
N 137

추정값 저장

신뢰 한계

알파 편집	수렴 기준
0.050	0.00001

CL에 대한 목표 SSE

해

	SSE	DFE	MSE	RMSE
	237.06333652	134	1.7691294	1.3300862

모수

모수	추정값	근사 표준 오차
b0	3.9398815293	0.65130758
b1	0.0045053136	0.01082544
b2	-0.012448139	0.0107535

해석자: 분석 Gauss-Newton

추정값 상관계수

추정값 공분산

보고서에서는 이 모형에 대한 최소 제곱 모수 추정값을 계산합니다.

6. **추정값 저장**을 클릭합니다.

열 계산식의 모수 추정값은 이 초기 비선형 적합 프로세스에 의해 추정된 값으로 설정됩니다.

Weibull 열에는 Weibull 계산식이 포함됩니다. 자세한 내용은 **Reliability and Survival Methods**에서 확인하십시오.

적합 프로세스를 계속하려면 다음을 수행하십시오.

7. **분석 > 전문 모델링 > 비선형**을 다시 선택합니다.

8. **Model**을 선택하고 **X, 예측변수 계산식**을 클릭합니다.

9. **Weibull loss**를 선택하고 **손실**을 클릭합니다.

10. **확인**을 클릭합니다.

그림 15.25 초기 비선형 적합 제어판

비선형 적합
예측 변수: Model, 손실: Weibull loss

제어판
시작하려면 "시작"을 클릭하십시오.

	기준	현재	중지 한계
시작	반복	0	60
중지	개제 변경	1.34078e+154	1e-15
단계	상대 그래디언트	1.34078e+154	0.000001
재설정	그래디언트	1.34078e+154	0.000001

☒ 손실이 음의 로그 가능도임

모수	현재 값	잠금	손실
b0	3.9398815293	☐	N
b1	0.0045053136	☐	N
b2	-0.012448139	☐	
sigma	1	☐	

이제 손실 함수에 **sigma** 라는 모수가 추가되었습니다. 이것은 분수의 분모에 있으므로 시작 값 1 이 **sigma** 에 적절합니다. 기본값 이외의 손실 함수를 사용하는 경우 "제어판"의 **손실이 음의 로그 가능도임** 상자가 기본적으로 선택됩니다.

11. **시작**을 클릭합니다.

그림 15.26 비선형 적합

비선형 적합
예측 변수: Model, 손실: Weibull loss

제어판
그래디언트에 수렴됨

	기준	현재	중지 한계
시작	반복	13	60
중지	개제 변경	4.145281e-10	1e-15
단계	상대 그래디언트	1.6644587e-9	0.000001
재설정	그래디언트	3.3666278e-8	0.000001

☒ 손실이 음의 로그 가능도임

모수	현재 값	잠금	손실
b0	5.6130465522	☐	227.61384948
b1	-0.012429709	☐	N
b2	-0.01091692	☐	137
sigma	1.1599937998	☐	

추정값 저장

신뢰 한계	알파 편집	수렴 기준
	0.050	
		0.00001

CL에 대한 목표 SSE

적합 프로세스는 Weibull 가능도 함수의 음의 로그를 최대화하여 모수를 추정합니다.

12. (선택 사항) **신뢰 한계**를 클릭하여 "해" 테이블의 모수에 대한 95% 신뢰 하한 및 상한을 표시합니다.

그림 15.27 해 보고서

해					
손실	DFE	평균 손실	제곱근 평균 손실		
227.61384948	133	1.7113823	1.3081981		
모수	추정값	근사 표준 오차	CL 하한	CL 상한	
b0	5.6130465522	0.67171997	4.31169667	6.96484024	
b1	-0.012429709	0.01114031	-0.0346969	0.00930037	
b2	-0.01091692	0.010399	-0.0289695	0.01220355	
sigma	1.1599937998	0.07823465	1.0209347	1.33040517	
해석자: 분석 NR					
▷ 추정값 상관계수					
▷ 추정값 공분산					

참고 : 신뢰 한계는 표준 점근적 신뢰 구간이 아니라 프로파일 가능도 신뢰 구간이므로 계산하는 데 시간이 걸릴 수 있습니다.

사전 정의된 지수 및 로그 정규 손실 함수를 사용하여 모형을 실행할 수도 있습니다. 다른 모형을 적합시키기 전에 모수 추정값을 최소 제곱 추정값으로 재설정합니다. 그렇지 않으면 모수 추정값이 수렴되지 않을 수 있습니다. 모수 추정값을 재설정하려면 다음을 수행하십시오.

13. (선택 사항) "비선형 적합"의 빨간색 삼각형을 클릭하고 **원래 모수로 되돌리기**를 선택합니다.

단순 생존 분포 적합의 예

이 섹션에는 비선형 플랫폼에서 단순 생존 분포를 적합시키는 예가 포함되어 있습니다.

지수, Weibull 및 극단값 손실 함수

비선형 플랫폼을 사용하여 지수, Weibull 및 극단값 손실 함수를 고장 시간 데이터에 적합시킬 수 있습니다. 이 예의 데이터는 총 30만 시간의 서비스 시간이 누적된 70개 디젤 팬을 대상으로 한 연구에서 가져왔습니다. 팬은 서로 다른 여러 시간에 서비스되었습니다. 반응은 팬의 고장 시간 또는 중도절단 시 실행 시간입니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Reliability/Fan.jmp 를 엽니다.

팁 : 손실 함수에 대한 계산식을 보려면 Fan.jmp 데이터 테이블에서 Exponential, Weibull 및 Extreme value 열을 마우스 오른쪽 버튼으로 클릭하고 **계산식**을 선택합니다.

2. **분석 > 전문 모델링 > 비선형**을 선택합니다.
3. Exponential 을 선택하고 **손실**을 클릭합니다.
4. **확인**을 클릭합니다.
5. 손실이 음의 로그 가능도임 체크박스가 선택되어 있는지 확인합니다.
6. **시작**을 클릭합니다.
7. **신뢰 한계**를 클릭합니다.
8. Exponential 대신 Weibull 과 Extreme value 를 선택하여 이 단계를 반복합니다.

그림 15.28 비선형 적합 결과

비선형 적합					
손실: Exponential					
제어판					
해					
손실	DFE	평균 손실	제공된 평균 손실		
135.17722247	69	1.9590902	1.399675		
모수	추정값	근사 표준 오차	CL 하한	CL 상한	
sigma	28703.333333	8285.93861	17114.2092	53600.0963	
해석자: 분석 NR					
추정값 상관계수					
추정값 공분산					

비선형 적합					
손실: Weibull					
제어판					
해					
손실	DFE	평균 손실	제공된 평균 손실		
135.15271994	68	1.98754	1.4098014		
모수	추정값	근사 표준 오차	CL 하한	CL 상한	
Alpha	26296.817205	12251.3882	13629.9746	106086.918	
Beta	1.0584462448	0.26825073	0.60609377	1.65792545	
해석자: 분석 NR					
추정값 상관계수					
추정값 공분산					

비선형 적합					
손실: Extreme value					
제어판					
해					
손실	DFE	평균 손실	제공된 평균 손실		
42.247996552	68	0.6212941	0.7882221		
모수	추정값	근사 표준 오차	CL 하한	CL 상한	
lambda	3.2694489763	0.46588966	2.61236736	4.66416689	
delta	0.9447814453	0.23944403	0.60315891	1.6501854	
해석자: 분석 NR					
추정값 상관계수					
추정값 공분산					

로그 정규 손실 함수

로그 정규 분포는 데이터 범위가 e 의 여러 거듭제곱인 경우에 유용합니다. 로그 정규 손실 함수는 모수의 시작 값에 매우 민감할 수 있습니다. 로그 정규 분포는 정규 분포와 비슷하므로 Time 의 자연 로그인 새 변수를 생성하고 분포를 사용하여 이 열의 평균 및 표준편차를 찾을 수 있습니다. 그런 다음 해당 값을 비선형 플랫폼의 시작 값으로 사용합니다. 이 예에서는 Time 의 자연 로그 평균이 4.72 이고 표준편차가 0.35 입니다.

1. 도움말 > 샘플 데이터 폴더를 선택하고 Reliability/Locomotive.jmp 를 엽니다.

팁 : 손실 함수에 대한 계산식을 보려면 `Locomotive.jmp` 데이터 테이블에서 `logNormal` 열을 마우스 오른쪽 버튼으로 클릭하고 **계산식**을 선택합니다.

2. **분석 > 전문 모델링 > 비선형**을 선택합니다.
3. `logNormal` 을 선택하고 **손실**을 클릭합니다.
4. **확인**을 클릭합니다.
5. " 평균 " 옆의 상자에 "4.72" 를 입력합니다.
6. " 시그마 " 옆의 상자에 "0.35" 를 입력합니다.
7. **시작**을 클릭합니다.
8. **신뢰 한계**를 클릭합니다.

그림 15.29 해 보고서

해					
손실		DfE	평균 손실	제곱근 평균 손실	
237.09354534		94	2.5222718	1.5881662	
모수	추정값	근사 표준 오차	CL 하한	CL 상한	
Mu	5.1169247193	0.10416697	4.93735853	5.35966436	
sigma	0.7054940302	0.0932136	0.55446171	0.93294904	
해석자: 수치 NR					

로그 정규 모수의 최대 가능도 추정값은 평균의 경우 5.11692이고 시그마의 경우 0.7055입니다. 로그 정규 분포의 중앙값에 대한 해당 추정값은 5.11692의 역대수($e^{5.11692}$)로, 약 167입니다. 이는 기관차 엔진의 일반적인 수명을 나타냅니다.

비선형 플랫폼에 대한 통계 상세 정보

이 섹션에는 비선형 플랫폼에 대한 통계 상세 정보가 포함되어 있습니다.

- "프로파일 가능도 신뢰 한계에 대한 통계 상세 정보"
- "사용자 손실 함수에 대한 통계 상세 정보"
- "도함수에 대한 통계 상세 정보"
- "유효 비선형 모델링에 대한 통계 상세 정보"

프로파일 가능도 신뢰 한계에 대한 통계 상세 정보

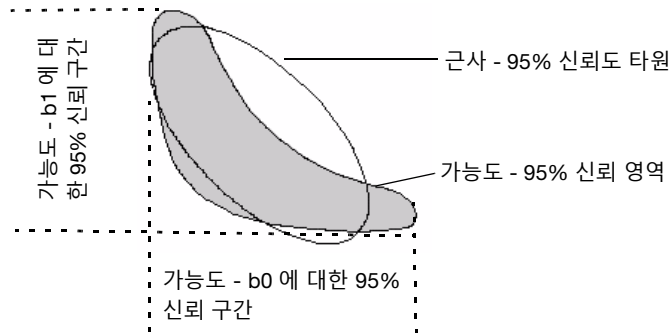
비선형 플랫폼에서 모수의 신뢰 상한 및 하한은 각 모수를 다른 모수에 대해 최소화한 후 해당 모수의 값을 검색하는 것을 기반으로 합니다. 검색을 통해 해의 최소 SSE보다 특정 값만큼 큰 SSE를 생성하는 값을 찾습니다. 이 차이의 목표는 F 분포를 기반으로 합니다. 이러한 구간을 가능도 신뢰 구간 또는 프로파일 가능도 신뢰 구간이라고도 합니다(Bates and Watts 1988, Ratkowsky 1990).

프로파일 신뢰 한계는 모두 목표 SSE로 시작합니다. 이 값은 F -검정이 주어진 유의 수준에서 해 SSE와 유의하게 다르다고 간주하는 오차 제곱합(또는 손실 함수의 합)입니다. 손실 함수가 음의 로그 가능도로 지정되면 F 분위수 대신 카이제곱 분위수가 사용됩니다. 각 모수의 신뢰 상한은 SSE가 목표 SSE에 도달할 때까지 모수 값이 증가합니다. 모수 값이 상승하므로 다른 모든 모수는 프로파일링된 모수의 변경에 따라 최소 제곱 추정값으로 조정됩니다. 개념상 이것은 내포된 반복의 복합 집합입니다. 내부적으로는 Johnston과 DeLong이 개발한 1회 반복 작업을 통해 이를 수행하는 방법이 있습니다. 자세한 내용은 SAS/ETS User's Guide(SAS Institute Inc. 2020)에서 확인하십시오.

그림 15.30에서는 음영 영역 내부에 최소 제곱(또는 최소 손실) 해가 있는 목표 SSE 또는 음의 가능도의 등고선을 보여 줍니다.

- 점근적 표준 오차는 타원체가 있는 영역에 가깝고 극단(수평 및 수직 접선)의 모수 값을 사용하는 신뢰 구간을 생성합니다.
- 프로파일 신뢰 한계는 근사 타원체 대신 실제 영역의 극단에서 모수 값을 찾습니다.

그림 15.30 모수에 대한 신뢰 한계 다이어그램



가능도 신뢰 구간은 근사 표준 오차에서 계산된 신뢰 구간보다 신뢰도가 높습니다. 특정 한계를 찾을 수 없으면 다음 한계에 대한 계산이 시작됩니다. 수렴이 어려운 경우 다음을 시도해 보십시오.

- 더 큰 알파 값을 사용하면 구간이 짧아지고 동작이 개선될 수 있습니다.
- 신뢰 한계 기준을 완화합니다.

사용자 손실 함수에 대한 통계 상세 정보

비선형 플랫폼에서는 기본 잔차 제곱합 이외의 함수를 최소화하거나 최대화할 수 있습니다. 이 섹션에서는 이를 수행하는 방식을 수리적으로 보여 줍니다.

$f(\beta)$ 를 모형으로 가정해 보겠습니다. 그런 다음 비선형 플랫폼에서 다음과 같이 정의된 손실 함수의 합을 최소화하려고 시도합니다.

$$L = \sum_{i=1}^n \rho(f(\beta))$$

각 행에 대한 손실 함수 $\rho(\bullet)$ 은 데이터 테이블에 있는 다른 변수의 함수일 수 있습니다. 여기에는 0이 아닌 1차 및 2차 도함수가 있어야 합니다. 기본 $\rho(\bullet)$ 함수인 잔차 제곱은 다음과 같습니다.

$$\rho(f(\beta)) = (y - f(\beta))^2$$

사용자 손실 함수가 있는 모형을 지정하려면 데이터 테이블에 변수를 생성하고 손실 함수를 구성합니다. 비선형 플랫폼을 시작한 후 손실 함수가 포함된 열을 손실 변수로 선택합니다.

비선형 최소화 계산식은 모형에 대해 $\rho(\bullet)$ 의 처음 두 도함수를 가져오고, 다음과 같이 그래디언트 및 근사 헤시안 행렬을 구성하여 작동합니다.

$$\begin{aligned} L &= \sum_{i=1}^n \rho(f(\beta)) \\ \frac{\partial L}{\partial \beta_j} &= \sum_{i=1}^n \frac{\partial \rho(f(\beta))}{\partial f} \frac{\partial f}{\partial \beta_j} \\ \frac{\partial^2 L}{\partial \beta_j \partial \beta_k} &= \sum_{i=1}^n \left[\frac{\partial^2 \rho(f(\beta))}{(\partial f)^2} \frac{\partial f}{\partial \beta_j} \frac{\partial f}{\partial \beta_k} + \frac{\partial \rho(f(\beta))}{\partial f} \frac{\partial^2 f}{\partial \beta_k \partial \beta_j} \right] \end{aligned}$$

$f(\bullet)$ 가 모수에 대해 선형인 경우 마지막 방정식의 두 번째 항은 0 입니다. 그렇지 않은 경우에도 첫 번째 항에 비해 합계가 작을 수 있으며 다음을 사용합니다.

$$\frac{\partial^2 L}{\partial \beta_j \partial \beta_k} \approx \sum_{i=1}^n \frac{\partial^2 \rho(f(\beta))}{(\partial f)^2} \frac{\partial f}{\partial \beta_j} \frac{\partial f}{\partial \beta_k}$$

잔차 합이 작으므로 ρ 이 잔차 제곱이면 두 번째 항이 작을 수 있습니다. 절편 항이 있는 경우 항은 0 입니다. 최소 제곱의 경우 이 항은 Gauss-Newton 과 Newton-Raphson 을 구별하는 항입니다.

참고: 표준 오차, 신뢰 구간 및 가설 검정은 최소 제곱 추정치 수행되거나, 최대 가능도 추정이 적절한 음의 로그 가능도와 함께 사용되는 경우에만 올바릅니다.

도함수에 대한 통계 상세 정보

비선형 플랫폼에서는 가장 일반적인 연산을 사용하는 계산식에 대해 기호 도함수를 사용합니다. 이 섹션에서는 도함수 표현식의 결과 유형을 보여 줍니다.

Negative Exponential.jmp 비선형 샘플 데이터 예제를 열면 Nonlinear 열의 실제 계산식이 다음과 같습니다.

`Parameter({b0=0.5, b1=0.5}, b0*(1-Exp(-b1*X)))`

계산식 편집기를 사용하는 경우 계산식의 Parameter 블록이 숨겨집니다. 계산식은 이러한 방식으로 열에 저장되고 비선형 시작 대화상자에 나타납니다. 두 모수 **b0**과 **b1**은 초기값이 지정되고 적합 계산식에 사용됩니다.

비선형 플랫폼에서는 계산식의 별도 복사본을 만들고 이를 편집하여 표현식에서 모수를 추출합니다. 그런 다음 모수에 대한 참조를 해당 모수가 추정된 위치에 매핑합니다. 비선형에서는 모수에 대해 예측 계산식의 해석 도함수를 사용합니다. **도함수 표시** 명령을 사용하면 다음과 같은 결과 계산식이 로그에 나열됩니다.

예측 모형

`b0 * First(T#1=1-(T#2=Exp(-b1*X)), T#3=-(-1*T#2*X))`

모수에 대한 모형의 도함수

`{T#1, T#3*b0}`

도함수 기능은 다음과 같이 작동합니다.

- 하위 표현식을 반복해서 계산하지 않도록 하기 위해 예측 모형에는 도함수 계산에 필요한 하위 표현식 값을 저장하는 할당이 스레드됩니다. **T#1**, **T#2** 등과 같은 이름에 할당이 수행됩니다.
- 예측 모형은 평가된 추가 하위 표현식이 필요한 경우 **First** 함수를 사용합니다. 이 함수는 첫 번째 인수 표현식의 값을 반환하고 다른 인수도 평가합니다. 이 경우 도함수에 대한 추가 할당이 필요합니다.
- 도함수 테이블은 그 자체가 표현식 목록이며, 각 모수에 대해 하나의 표현식을 적합시킵니다. 예를 들어 **b0**에 대한 모형의 도함수는 **T#1**이며, 예측 모형의 스레드는 $1 - (\text{Exp}(-b1 \cdot X))$ 입니다. **b1**에 대한 도함수는 **T#3*b0**이며, 위의 할당을 대체할 경우 $-(-1 \cdot \text{Exp}(-b1 \cdot X) \cdot X) \cdot b0$ 입니다. 최적화가 많이 수행되지만 연산이 항상 최적으로 결합되는 것은 아닙니다. 이중 부정을 제거하지 않는 **T#3**의 표현식으로 이를 확인할 수 있습니다.

손실 함수를 지정하면 계산식 편집기에서 모수(있는 경우)에 대한 도함수를 사용합니다. 그리고 모형에 대해 1차 도함수와 2차 도함수를 사용합니다(있는 경우).

도함수 메커니즘이 함수의 해석 도함수를 사용하는 방법을 모르면 **NumDeriv** 함수를 사용하여 수치 미분을 수행합니다. 이 경우 플랫폼에서는 인수의 델타 변경과 관련하여 함수의 변경을 평가하는 데 사용되는 델타를 표시합니다. 적절한 수치 미분을 얻으려면 다른 델타 설정으로 실험해야 할 수도 있습니다.

팁

주어진 모형을 나타내는 방법에는 항상 여러 가지가 있으며, 어떤 방법은 다른 형식보다 훨씬 우수한 성능을 보입니다. Ratkowsky 연구 자료(1990)에서 이러한 대체 형식을 다룹니다.

계산식의 여러 위치에서 반복되는 하위 표현식이 있는 경우 임시 변수에 할당하는 것이 좋습니다. 그런 다음 나중에 계산식에서 이 변수를 참조합니다. 예를 들어 위의 모형 계산식 중 하나는 다음과 같습니다.

```
If(Y==0, Log(1/(1+Exp(model))), Log(1 - 1/(1 + Exp(model))));
```

표현식을 인수 분해한 후 지역 변수에 할당하여 계산식을 단순화할 수 있습니다.

```
temp=1/(1+Exp(model));
If(Y==0, Log(temp), Log(1-temp));
```

도함수 기능은 할당과 조건 전체에서 도함수를 추적할 수 있습니다.

유효 비선형 모델링에 대한 통계 상세 정보

비선형 플랫폼을 사용하는 경우 다항식을 중심화하는 것이 좋습니다.

모수에 대해 선형인 완전 다항식 항이 있는 경우 항상 다항식을 중심화하는 것이 좋습니다. 이렇게 하면 최적화를 위한 수치 표면의 조건이 개선됩니다. 다음과 같은 표현식을 예로 들어보겠습니다.

$$a_1 + b_1x + c_1x^2$$

다음과 같이 변환해야 합니다.

$$a_2 + b_2(x - \bar{x}) + c_2(x - \bar{x})^2$$

두 모형은 모수 변환을 제외하고는 동일하지만 모형이 비선형이면 두 번째 모형을 훨씬 쉽게 적합시킬 수 있습니다.

모수 변환은 쉽게 풀이할 수 있습니다.

$$a_1 = a_2 - b_2\bar{x} + c_2\bar{x}^2$$

$$b_1 = b_2 - 2c_2\bar{x}$$

$$c_1 = c_2$$

반복 수가 계속 최대값까지 진행되면 최대 반복 수를 늘리거나 수렴 기준 중 하나를 완화합니다.

모든 문제에 대해 잘 맞는 총괄 최적화 방법은 실제로 없습니다. JMP에는 비선형 플랫폼에서 해를 구할 수 있는 문제의 범위를 확대할 수 있도록 **Newton**, **QuasiNewton BFGS**, **QuasiNewton SR1** 및 **수치 미분만** 등의 옵션이 있습니다.

기본 설정이 특정 문제에 대한 해로 수렴할 수 없는 경우 이 설정의 다양한 조합을 사용하여 수렴 승산을 높입니다.

일부 모형은 모수의 시작 값에 매우 민감합니다. 새 시작 값으로 작업하는 것이 효과적일 때가 많습니다. 시작 값을 편집하고 **재설정**을 클릭하여 결과를 확인합니다. 그럼도 유용합니다. 슬라이

더를 사용하여 더 잘 적합되도록 곡선을 시각적으로 수정합니다. 모수 프로파일러가 도움을 줄 수 있지만 데이터 집합이 작은 경우 외에는 너무 느릴 수 있습니다.

16 장

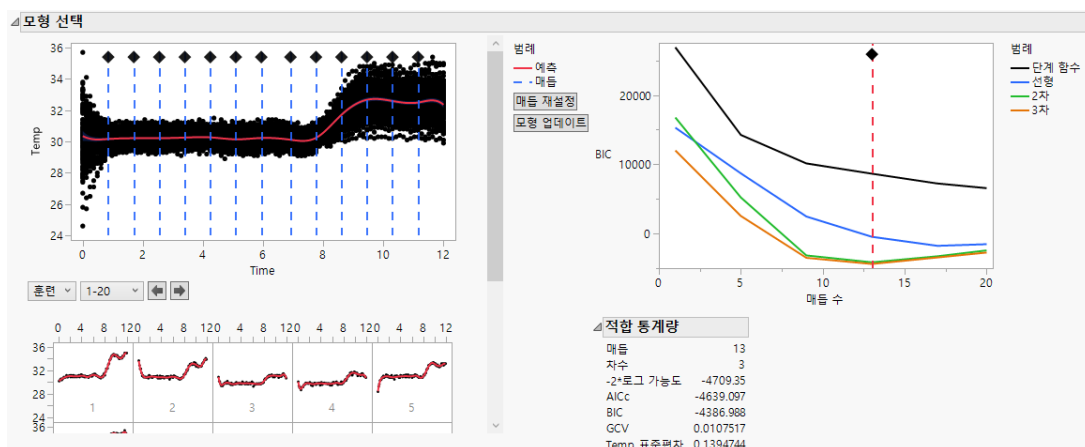
JMP^{PRO} 함수 데이터 탐색기 함수 데이터 분석 수형

함수 데이터 탐색기 플랫폼은 JMP Pro에서만 사용할 수 있습니다.

FDE(함수 데이터 탐색기)는 함수, 신호, 스펙트럼 또는 계열 데이터를 위한 플랫폼입니다. 이 플랫폼은 탐색적 데이터 분석 도구 또는 차원 축소 기법으로 사용할 수 있습니다. 차원 축소의 경우 FDE 플랫폼은 함수 데이터를 다른 JMP 플랫폼에서 분석할 수 있는 형식으로 변환합니다.

출력 데이터에 대한 몇 가지 유형의 변환 및 입력 데이터에 대한 정렬을 포함하여 데이터 전처리 도구를 FDE 플랫폼에서 사용할 수 있습니다. B-스플라인, P-스플라인, Fourier 또는 소파동 기저 모형을 데이터에 적합시켜 함수 모형이 생성됩니다. 모형이 적합되면 함수 모형에 대해 함수 PCA(함수 주성분 분석)가 자동으로 수행됩니다. 함수 PCA의 결과(예: 함수 주성분 스코어)는 특징 추출 또는 차원 축소를 위해 별도의 데이터 테이블에 저장됩니다. 기저 함수 모형을 먼저 적합시키지 않고 데이터에 대해 함수 주성분 분석을 직접 수행할 수도 있습니다.

FDE 플랫폼에는 DOE(실험 설계) 옵션도 포함되어 있습니다. DOE 옵션 중 하나를 사용하면 시작 창에 지정된 보조 변수를 사용하여 함수 데이터의 기능을 모델링할 수 있습니다. 이렇게 하면 모형 적합을 사용하여 보조 변수에 따라 함수 반응이 어떻게 변하는지 확인할 수 있습니다. FDE에서 B-스플라인 모형 선택



목차

함수 데이터 탐색기 플랫폼 개요	275
함수 데이터 탐색기의 예	275
함수 데이터 탐색기 플랫폼 시작	279
함수 데이터 탐색기 보고서	282
모형 보고서	283
함수 데이터 탐색기 플랫폼 옵션	288
함수 데이터 탐색기 그룹 옵션	290
데이터 처리 옵션	291
모형 보고서 옵션	295
함수 데이터 탐색기 플랫폼의 추가 예	297
여러 함수 공정의 예	297
함수 DOE 분석의 예	301
함수 데이터 탐색기 플랫폼에 대한 통계 상세 정보	303
함수 모형 적합에 대한 통계 상세 정보	304
함수 요약 보고서에 대한 통계 상세 정보	304

JMP PRO 함수 데이터 탐색기 플랫폼 개요

함수 데이터는 측정값 집합이 곡선 또는 이미지를 구성하는 연속 도메인에 대해 기록된 데이터로 정의할 수 있습니다. 대개 도메인은 시간이고 측정값 집합은 ID 변수로 정의됩니다. 도메인의 특정 지점 t 에서 ID 수준 i 에 대한 함수 데이터 관측값을 $f_i(t)$ 로 표기합니다. 함수 데이터 탐색기 플랫폼을 사용하면 함수 데이터를 탐색하고 분석할 수 있습니다.

함수 데이터의 형식은 조밀하거나 희박할 수 있습니다. 조밀한 함수 데이터는 관측값이 ID 변수의 모든 수준에 대해 균등한 간격의 동일한 점 격자 위에 있을 때 발생합니다. 희박한 함수 데이터는 관측값이 도메인 전체에서 균등하지 않은 간격으로 배치되어 ID 수준의 관측값 수가 다를 때 발생합니다. 함수 데이터 탐색기 플랫폼에서는 두 가지 형태의 함수 데이터를 모두 처리할 수 있습니다.

함수 데이터를 여러 가지 방법으로 표현할 수 있지만 일반적으로 다음과 같은 두 경우로 분류할 수 있습니다.

- 관심 반응 $f(t)$ 는 함수 형태를 가집니다.
- 함수 형태를 갖는 하나 이상의 공변량 $f(t)'$ 가 있습니다. 이를 함수 또는 신호 공정이라고도 합니다.

함수 데이터 탐색기 플랫폼은 모든 유형의 함수 데이터에 대한 탐색 도구로 유용합니다. 그러나 이 플랫폼의 강점은 많은 함수 공정(스칼라 반응과 연관될 수 있음)을 사용하고 이후 모델링에 사용할 주요 특징을 추출하는 것입니다. 이 작업을 수행하려면 먼저 함수 모형을 데이터에 적합시킵니다. 그런 다음 이 함수 모형에 대해 함수 PCA(함수 주성분 분석)가 수행됩니다. FPC(함수 주성분) 스코어와 같은 함수 PCA의 결과는 저장된 후 다른 모델링 플랫폼(예: 모형 적합 플랫폼의 일반화 회귀 분석법)에서 특징 추출 및 분석에 사용됩니다. 또는 FDE 플랫폼 내에서 보조 변수 집합을 지정하고 일반화 회귀 모형을 적합시켜 이러한 변수가 반응에 어떤 영향을 미치는지 확인할 수 있습니다.

함수 데이터 분석에 대한 자세한 내용은 Ramsay and Silverman(2005) 연구 자료에서 확인하십시오.

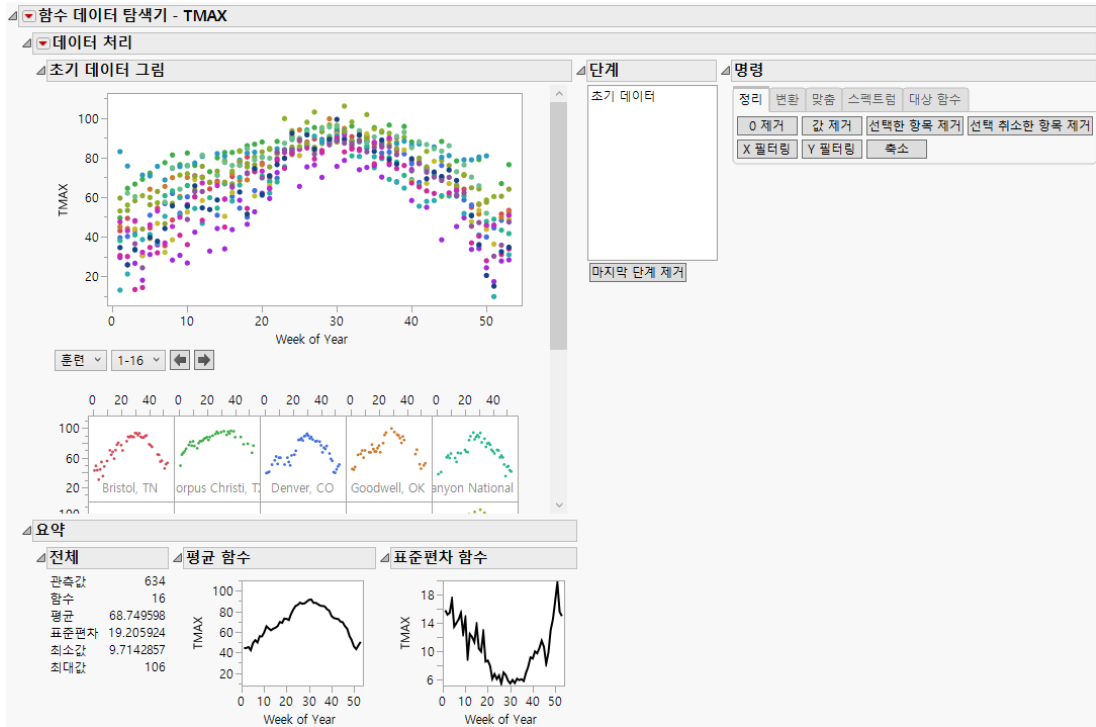
JMP PRO 함수 데이터 탐색기의 예

이 예에서는 미국 전역의 16개 기상 관측소에서 수집한 주간 기상 데이터를 분석합니다. 모든 기상 관측소가 매주(년 중)에 대한 주간 온도 측정값을 가지고 있는 것은 아니므로 이것은 희소 함수 데이터의 예입니다. 일별 온도는 주간 평균으로 요약됩니다. 데이터 테이블의 "Weather Station Locations" 스크립트를 실행하여 위치 지도를 표시합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Functional Data/Weekly Weather Data.jmp**를 엽니다.
2. **분석 > 전문 모델링 > 함수 데이터 탐색기**를 선택합니다.
3. **TMAX**를 선택하고 **Y, 출력**을 클릭합니다.
4. **Week of Year**를 선택하고 **X, 입력**을 클릭합니다.

5. ID 를 선택하고 ID, 함수를 클릭합니다 .
6. 확인을 클릭합니다 .

그림 16.1 초기 함수 데이터 탐색기 보고서



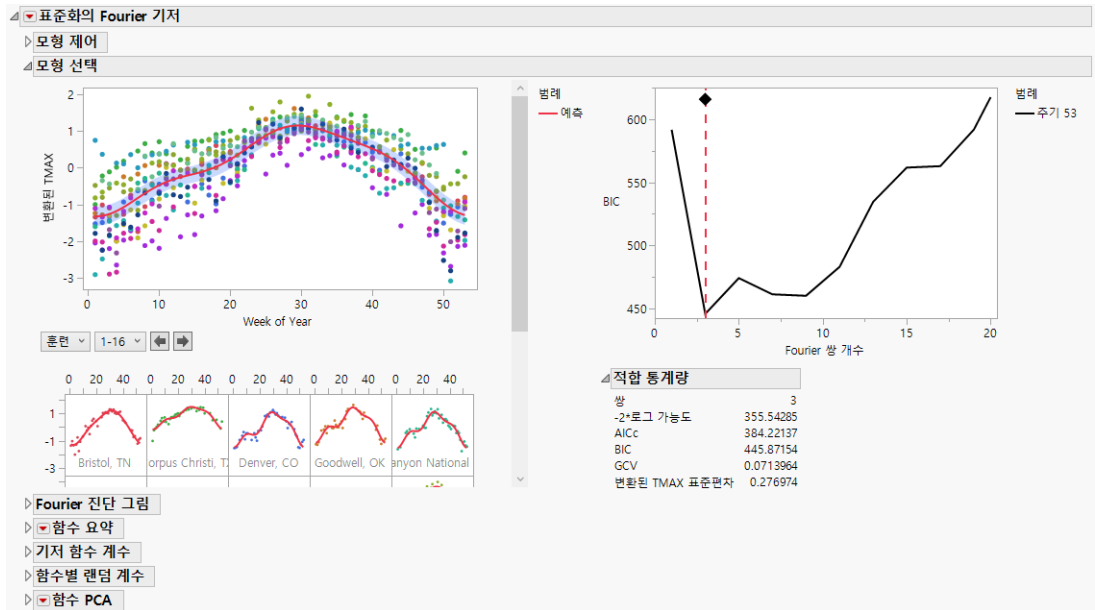
초기 "함수 데이터 탐색기" 보고서에는 원시 데이터 그림, 요약 통계량, 데이터의 함수 평균 및 함수 표준편차 요약 그림이 포함되어 있습니다. 데이터 처리 옵션이 포함된 "명령" 섹션도 있습니다. 데이터 처리 옵션은 "데이터 처리"의 빨간색 삼각형 메뉴에서도 액세스할 수 있습니다. 대개 모델링하기 전에 출력 데이터를 표준화하는 것이 좋습니다.

7. "명령" 섹션에서 "변환" 탭을 클릭합니다 .
8. **표준화** 버튼을 클릭합니다 .

지정된 변환을 기반으로 데이터 그림과 요약 통계량이 업데이트됩니다. "단계" 목록에 "표준화"가 추가됩니다.

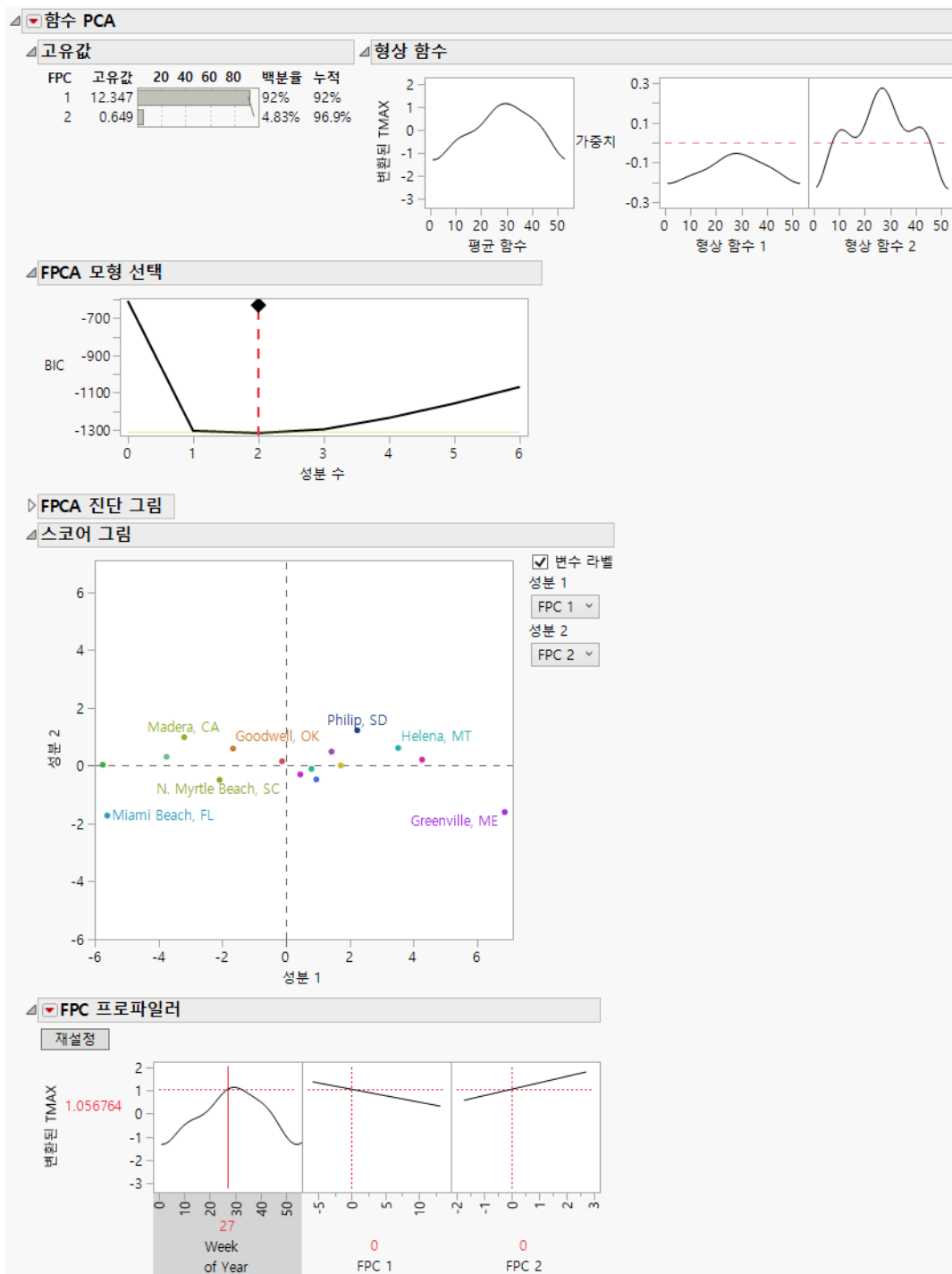
9. "함수 데이터 탐색기"의 빨간색 삼각형을 클릭하고 **모형 > Fourier 기저**를 선택합니다 .

그림 16.2 Fourier 기저 모형 보고서



"Fourier 기저 " 보고서에는 선택한 모형에 대한 정보를 제공하는 여러 보고서가 포함되어 있습니다 . " 모형 선택 " 보고서에는 BIC 적합 기준에 따라 최량 적합 모형이 표시됩니다 . 기상 데이터의 경우 선택된 Fourier 기저 모형의 주기는 53 이고 기저 함수 쌍은 3 개입니다 . 모형에 대한 적합 통계량과 계수도 제공됩니다 . 아래로 스크롤하여 " 함수 PCA " 보고서를 봅니다 .

그림 16.3 함수 PCA 보고서



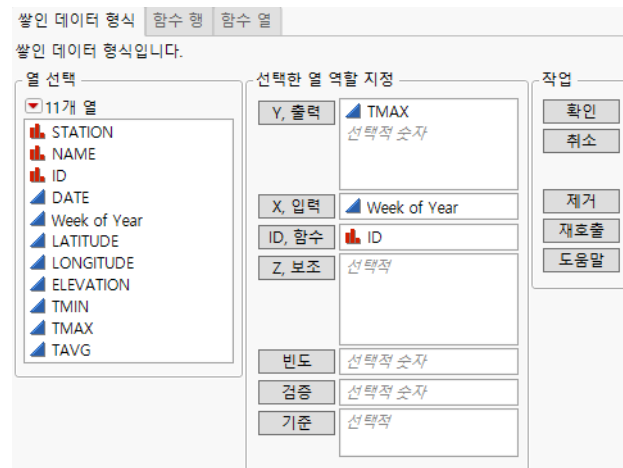
" 함수 PCA " 보고서에서는 처음 두 고유값이 데이터 변동의 거의 97% 를 설명한다는 것을 보여 줍니다 . " 모형 선택 " 그래프에서 빨간색 파선을 클릭한 후 3 FPC 로 드래그하면 처음 세 고유값이 데이터 변동의 99% 를 설명한다는 것을 확인할 수 있습니다 . 그러나 첫 번째 고유값만으로 92% 를 설명합니다 . " 스코어 그림 " 을 사용하여 다른 함수의 이상치인 개별 함수를 감지할 수 있습니다 . " 스코어 그림 " 에서 "Miami Beach, FL" 과 "Greenville, ME" 위치를 제외한 대부분의 위치가 함께 모여 있습니다 . 개별 함수 그림이 나올 때까지 위로 스크롤합니다 . Miami Beach 위치의 함수는 더 편평하므로 다른 위치에 비해 온도 변동이 적다는 것을 나타냅니다 . Greenville 위치의 함수는 최대값이 더 낮으므로 다른 위치에 비해 온도가 일관되게 낮음을 나타냅니다 .

팁 : 이상치를 더 잘 식별하려면 " 스코어 그림 " 보고서에서 " 변수 라벨 " 옵션을 선택 취소하십시오 .

JMP PRO 함수 데이터 탐색기 플랫폼 시작

분석 > 전문 모델링 > 함수 데이터 탐색기를 선택하여 함수 데이터 탐색기 플랫폼을 시작합니다 .

그림 16.4 함수 데이터 탐색기 시작 창



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

데이터 형식 선택

함수 데이터 탐색기 시작 창에는 다양한 유형의 데이터 형식 탭이 있습니다 . 사용하는 데이터 형식에 따라 탭을 선택합니다 .

쌓인 데이터 형식 데이터 테이블의 각 행이 단일 관측값에 해당하는 경우 선택합니다 . 출력 , 입력 및 ID 변수에 대한 개별 열이 있습니다 .

참고 : 데이터 형식 중 " 쌓인 데이터 형식 " 만 함수 공정을 여러 개 지정할 수 있습니다. " 쌓인 데이터 형식 " 탭의 **Y, 출력**에 둘 이상의 열을 할당하면 각 Y 변수가 개별적으로 분석됩니다. " 그룹 적합 " 보고서에는 Y 변수에 대한 개별 보고서가 포함됩니다.

함수 행 데이터 테이블의 각 행이 ID 변수의 한 수준에 대한 전체 출력 함수에 해당하는 경우 선택합니다. 각 열은 입력 변수의 수준에 해당합니다.

팁 : 열에 FDE X 열 특성이 포함된 경우 이 열 특성은 입력 변수의 값을 지정하는 데 사용됩니다. 열에 열 특성이 포함되어 있지 않으면 열 이름에 지정된 입력 변수의 수준이 " 함수 행 " 데이터 형식에 사용됩니다. 열 이름이 너무 복잡하면 문제가 있는 열을 나타내는 오류 메시지가 표시되고 값은 $x = 1, 2, \dots, n$ 으로 지정됩니다. 여기서 n 은 함수의 관측값 수입니다. 예를 들어 t20 2 및 x30v20 은 문제가 있는 열 이름입니다.

함수 열 데이터 테이블의 각 열이 ID 변수의 한 수준에 대한 전체 출력 함수에 해당하는 경우 선택합니다. 각 행은 입력 변수의 수준에 해당합니다.

열 역할 지정

시작 창에는 열 역할을 지정하기 위한 다음 옵션이 포함되어 있습니다.

Y, 출력 함수 공정 $f(t)$ 를 할당합니다. ID 변수의 각 수준에 대해 두 개 이상의 관측된 출력 값이 있어야 합니다.

참고 : 관측된 출력 값이 두 개 미만인 함수는 분석에서 제거됩니다.

X, 입력 (" 쌓인 데이터 형식 " 및 " 함수 열 " 탭에만 사용 가능) 입력 변수 t 를 할당합니다. " X, 입력 " 에 변수를 지정하지 않으면 행 번호가 사용됩니다. 행 번호를 사용하면 입력 도메인에서 관측값의 간격이 균등하다고 가정합니다.

ID, 함수 (" 쌓인 데이터 형식 " 및 " 함수 행 " 탭에만 사용 가능) 각 함수에 ID 변수를 할당합니다. 쌓인 데이터의 경우 ID 변수가 할당되지 않으면 모든 관측값을 하나의 함수에서 가져온다고 가정합니다.

Z, 보조 (" 쌓인 데이터 형식 " 및 " 함수 행 " 탭에만 사용 가능) 하나 이상의 보조 변수를 할당합니다. 보조 변수는 함수 데이터 탐색기 플랫폼의 계산에 사용되지 않으며 이 변수를 포함해도 결과에 영향을 주지 않습니다. 보조 변수는 향후 함수 데이터 탐색기의 결과를 분석하는 데 사용할 수 있는 변수입니다. 보조 변수를 지정하면 " 데이터 저장 " 및 " 요약 저장 " 옵션으로 생성되는 테이블의 **Supplementary** 열 그룹에 포함됩니다. 이러한 열에는 원래 데이터 테이블에 지정된 열 특성이 그대로 유지됩니다. 적합 모형에 " 함수 DOE 분석 " 옵션도 사용할 수 있습니다. 자세한 내용은 " **함수 DOE 분석** " 에서 확인하십시오.

빈도 (" 쌓인 데이터 형식 " 탭에만 사용 가능) 분석의 각 행에 대한 빈도를 나타내는 숫자 값이 포함된 열을 할당합니다. 빈도 열의 효과는 데이터 테이블을 확장하여 정수 빈도가 k 인 행을 k 개의 동일한 행으로 확장하는 것입니다.

검증 검증 데이터 집합을 정의하는 선택적 숫자 열을 할당합니다. 이 열에는 구분되는 값이 두 개만 포함되어야 합니다. 더 작은 값이 훈련 데이터 집합을 정의하고 더 큰 값이 검증 데이터

터 집합을 정의합니다. 값이 세 개 이상인 경우 가장 작은 값이 훈련 데이터 집합을 정의하고 다른 모든 값은 검증 데이터 집합을 정의합니다. FDE 플랫폼에서는 검증 열을 사용하여 모델을 훈련하고 평가합니다. 검증에 대한 자세한 내용은 "[JMP 모델링의 검증](#)"에서 확인하십시오.

참고 : "검증" 옵션을 사용하면 각 함수의 관측값 표본이 아니라 전체 함수를 홀드아웃할 수 있습니다. 따라서 ID 값이 동일한 모든 관측값이 테스트 또는 검증으로 분류되어야 합니다. ID 값이 동일한 관측값이 두 데이터 집합에 모두 포함될 수 없습니다. 이 유형의 검증 열에 대한 자세한 내용은 "[그룹화 검증 열](#)"에서 확인하십시오.

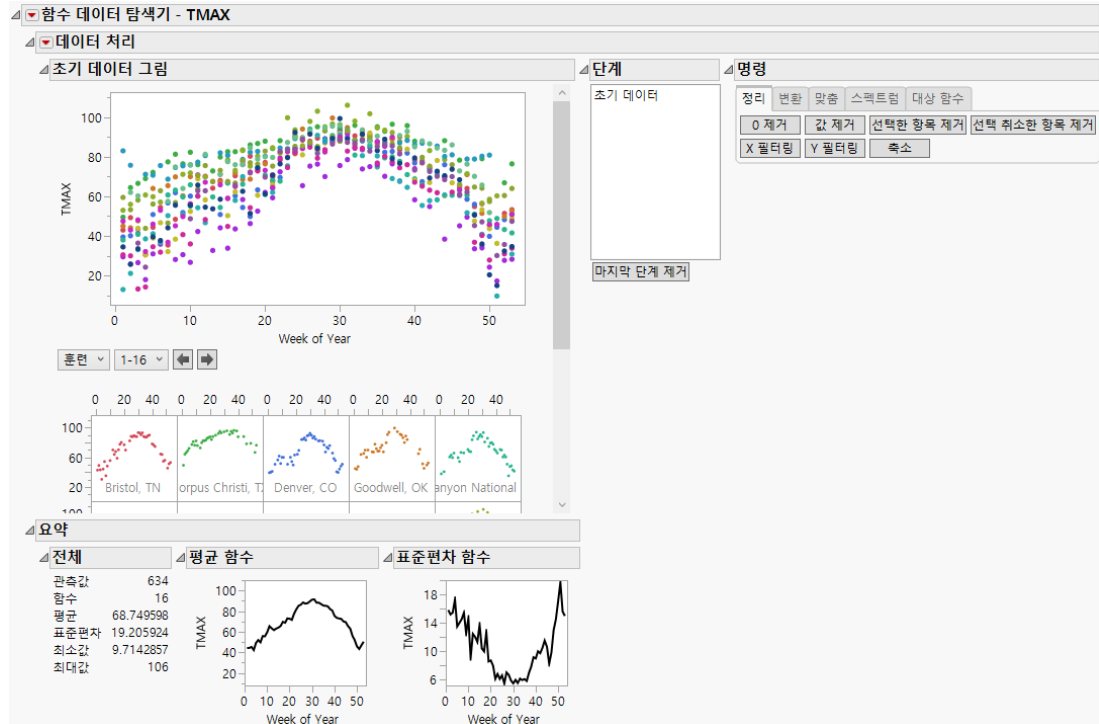
주의 : "열 선택" 목록에서 아무 열도 선택하지 않은 상태로 "검증" 버튼을 클릭하고 데이터 테이블에 검증 열을 추가하면 그룹화 검증 열이 되지 않습니다. 대신 검증 열 생성 플랫폼을 사용합니다. 자세한 내용은 "[그룹화 검증 열](#)"에서 확인하십시오.

기준 변수의 각 수준에 대한 개별 분석으로 구성된 보고서를 생성하는 열을 할당합니다. 기준 변수가 둘 이상 할당되면 기준 변수의 가능한 각 수준 조합에 대해 개별 분석이 생성됩니다.

JMP PRO 함수 데이터 탐색기 보고서

초기 " 함수 데이터 탐색기 " 보고서에는 " 데이터 처리 " 보고서와 " 요약 " 보고서가 포함됩니다.

그림 16.5 초기 함수 데이터 탐색기 보고서



"데이터 처리" 보고서에는 모든 관측값의 초기 데이터 그림과 ID 변수 수준에 해당하는 개별 데이터 그림 격자가 포함되어 있습니다. 격자에 표시되는 최대 개별 그림 수는 20개입니다. 격자 위에는 표시할 개별 그림을 대화식으로 선택할 수 있는 드롭다운 메뉴와 화살표가 있습니다. 검증 데이터 집합 또는 대상 함수를 지정한 경우 첫 번째 드롭다운 메뉴를 사용하여 훈련 데이터 집합, 검증 데이터 집합 또는 대상 함수 중에서 표시되는 개별 그림의 출처를 선택할 수 있습니다. 함수가 20개를 초과하는 경우 두 번째 드롭다운 메뉴와 화살표를 사용하여 다양한 그림 그룹을 표시할 수 있습니다. 모든 그래프에는 입력 변수의 값에 대한 함수 반응 데이터가 표시됩니다. "데이터 처리" 보고서의 "명령" 섹션에 데이터 처리 버튼도 포함되어 있습니다.

"요약" 보고서에는 관측값 수, 함수 수, 전체 평균, 표준편차, 최소값 및 최대값을 포함한 전체 요약 통계량 테이블이 포함되어 있습니다. 함수 평균 및 함수 표준편차 그림도 있습니다. 그림에 표시된 함수 요약 통계량은 입력 변수의 각 고유 값에서 계산됩니다.

"데이터 처리" 및 "요약" 보고서는 모형을 적합시키거나 직접 함수 주성분 분석을 수행하기 전에 데이터를 전처리하는 데 유용합니다. "데이터 처리" 보고서에 있는 데이터 정리, 변환, 맞춤 및 스펙트럼 버튼은 "데이터 처리"의 빨간색 삼각형 메뉴에 있는 옵션에 해당합니다. 자세한 내용

은 "데이터 처리 옵션"에서 확인하십시오. 전처리 단계를 수행할 때마다 그래프와 요약 결과가 자동으로 업데이트됩니다. 여러 전처리 단계를 순서대로 수행할 수 있습니다. 각 단계가 "데이터 처리" 보고서의 "단계" 목록에 추가됩니다. 가장 최근 단계를 제거하려면 "단계" 개요 아래의 **마지막 단계 제거** 버튼을 클릭합니다.

JMP PRO 모형 보고서

"함수 데이터 탐색기"의 빨간색 삼각형 메뉴에 있는 "모형" 하위 메뉴 옵션을 사용하여 모형을 데이터에 적합시킬 수 있습니다. 사용 가능한 모형은 "모형"에서 확인하십시오. 다른 유형의 모형을 데이터에 적합시킬 때마다 모형 보고서가 나타납니다. 각 모형 보고서에는 다음 보고서가 포함됩니다.

- "모형 제어"
- "모형 선택"
- "진단 그림"
- "함수 요약"
- "계수"
- "함수별 랜덤 계수"
- "함수 PCA"

JMP PRO 모형 제어

"모형 제어" 보고서를 사용하면 "모형 선택" 보고서에서 비교할 모형의 모수를 정의할 수 있습니다. "모형 제어" 보고서의 모양은 적합되는 모형 유형에 따라 달라집니다. 소파동 모형에는 모형 제어를 사용할 수 없습니다.

B-스플라인 및 P-스플라인 모형 제어

B-스플라인 또는 P-스플라인 모형이 적합되면 다음 모수를 지정할 수 있습니다.

매듭 수 각 스플라인의 매듭 수를 추가 또는 제거하거나 범위를 지정합니다. 매듭 수는 0이 아닌 정수여야 합니다.

참고: B-스플라인 모형에 허용되는 최대 매듭 수는 함수당 최대 관측값 수 또는 고유 입력 수보다 한 개 적습니다. P-스플라인 모형에 허용되는 최대 매듭 수는 고유 입력 수보다 두 개 적습니다. 최대값보다 큰 값을 지정하면 경고 메시지가 나타납니다.

스플라인 차수 "모형 선택" 보고서에서 스플라인 차수 적합을 추가하거나 제거합니다.

Fourier 기저 모형 제어

Fourier 기저 모형이 적합되면 다음 모수를 지정할 수 있습니다.

Fourier 쌍 개수 비교할 Fourier 쌍 개수를 추가 또는 제거하거나 범위를 지정합니다.

주기 함수의 주기를 변경합니다.

모형 제어를 지정한 후 **시작**을 클릭하면 "모형 선택" 보고서에 업데이트된 모형이 표시됩니다.

팁 : 모형을 적합시키기 전에 모형 제어를 지정하려면 Shift 키를 누른 채 "함수 데이터 탐색기"의 빨간색 삼각형을 클릭한 후 원하는 모형을 선택합니다. 자세한 내용은 "[모형](#)"에서 확인하십시오.

JMP PRO 모형 선택

"모형 선택" 보고서에는 전체 예측 그림, 개별 예측 그림 격자, 해 경로 그림 및 테이블이 포함되어 있습니다. 개별 예측 그림 격자는 "데이터 처리" 보고서의 개별 그림 격자와 레이아웃 및 컨트롤이 동일합니다. 한 번에 최대 20개의 그림이 표시됩니다. 개별 예측 그림의 다양한 그룹을 볼 수 있는 드롭다운 메뉴와 화살표가 있습니다.

예측 그림에는 원시 데이터 및 현재 모형의 예측 곡선이 표시됩니다. 검증 데이터 집합이 있는 경우 검증 데이터 집합에 포함된 함수에 대해서는 예측 곡선이 표시되지 않습니다. 전체 예측 그림의 곡선은 평균 예측 곡선입니다. 개별 예측 그림의 곡선은 각 특정 함수에 대한 예측 곡선입니다. B-스플라인 모형의 경우 전체 예측 그림에 매듭 위치도 표시됩니다. 파란색 슬라이더 막대를 다른 위치로 드래그하여 매듭 위치를 변경할 수 있습니다. 새 매듭 위치에 따라 모형 보고서를 업데이트하려면 **모형 업데이트** 버튼을 클릭합니다. 매듭을 기본 위치로 재설정하려면 **매듭 재설정** 버튼을 클릭합니다.

소파동 모형의 경우 "계수" 탭도 있습니다. 이 탭에는 전체 계수 그림과 개별 계수 그림의 격자가 포함됩니다. 계수 그림은 각 해상도 내에서 계수의 상대 중요도를 나타내는 파선으로 구성됩니다. 그림에서 선의 위치는 입력 변수(가로 축)와 해상도 번호(세로 축)에 의해 결정됩니다. 이러한 숫자는 "소파동 계수" 테이블의 열 이름에 있습니다. 개별 그림의 경우 그림의 각 선 색상은 해당 계수의 부호에 의해 결정됩니다. 파란색은 양수 계수를 나타내고 빨간색은 음수 계수를 나타냅니다. 그림의 각 선 길이는 계수에 의해 결정됩니다. 길이는 각 해상도 내에서 가장 큰 절대값 계수로 척도화된 해당 계수입니다. 따라서 최대 계수의 선 길이가 가장 길고 계수가 작을수록 선이 짧아집니다. 평균 그림의 경우 계수는 각 함수에 대해 평균화됩니다. 각 선의 색상과 길이는 평균화 및 척도화된 계수에 의해 결정됩니다.

참고 : 계수 그림은 최량 적합 소파동 모형에만 사용할 수 있습니다.

해 경로 그림 및 연결된 테이블의 모양은 모형 유형에 따라 다릅니다.

B-스플라인 및 P-스플라인 모형

해 경로 그림에서는 정의된 매듭 수에 대해 표시된 모형 선택 기준을 보여 줍니다. 각 스플라인 차수에 대한 별도의 해 경로가 있습니다. BIC(Bayes 정보 기준)이 기본 적합 기준입니다. 모형의 빨간색 삼각형 옵션인 **모형 선택**을 사용하여 선택 기준을 변경할 수 있습니다. 현재 해는 해 경로 그림에서 수직 점선으로 지정됩니다. 기본적으로 선택된 스플라인 차수와 매듭 수는 모형 선택

기준 값이 가장 작은 모형에 해당합니다. 현재 모형 선택을 변경하려면 세로선 맨 위의 슬라이더를 드래그하거나, 해 경로 그림 또는 범례에서 특정 스플라인을 클릭합니다. 이렇게 하면 "모형 선택" 보고서의 예측 그림과 다른 모든 보고서의 정보가 자동으로 업데이트됩니다.

해 경로 그림 아래의 테이블은 현재 해 모형에 대한 정보가 포함된 적합 통계량 테이블입니다. 여기에는 매듭 수, 스플라인 차수, $-2 \times$ 로그 가능도, AICc/BIC/GCV 모형 적합 기준 값 및 반응 표준편차 값이 표시됩니다. 반응 표준편차는 적합 모형의 잔차 시그마로 정의됩니다. P-스플라인 모형을 선택한 경우 벌점 모수 λ (람다)도 표시됩니다.

Fourier 모형

해 경로 그림에서는 정의된 주기 동안 Fourier 쌍 개수에 대해 표시된 모형 선택 기준을 보여 줍니다. BIC(Bayes 정보 기준)가 기본 적합 기준입니다. 모형의 빨간색 삼각형 옵션인 **모형 선택**을 사용하여 선택 기준을 변경할 수 있습니다. 현재 해는 해 경로 그림에서 수직 점선으로 지정됩니다. 기본적으로 모형 선택 기준 값이 가장 작은 모형에 해당하는 Fourier 쌍 개수 위치에 슬라이더가 있습니다. 수직 점선 맨 위의 슬라이더를 드래그하여 현재 모형의 Fourier 쌍 개수를 변경할 수 있습니다. 슬라이더를 드래그하면 "모형 선택" 보고서의 예측 그림과 다른 모든 보고서의 정보가 자동으로 업데이트됩니다.

해 경로 그림 아래의 테이블은 현재 해 모형에 대한 정보가 포함된 적합 통계량 테이블입니다. 여기에는 Fourier 쌍 개수, $-2 \times$ 로그 가능도, AICc/BIC/GCV 모형 적합 기준 값 및 반응 표준편차 값이 표시됩니다. 반응 표준편차는 적합 모형의 잔차 시그마로 정의됩니다.

소파동 모형

해 경로 그림에서는 소파동 모형을 정의하는 모형 번호에 대해 표시된 모형 선택 기준을 보여 줍니다. BIC(Bayes 정보 기준)가 기본 적합 기준입니다. 모형의 빨간색 삼각형 옵션인 **모형 선택**을 사용하여 선택 기준을 변경할 수 있습니다. 현재 모형 선택은 해 경로 그림에서 수직 점선으로 지정됩니다. 기본적으로 모형 선택 기준 값이 가장 작은 모형이 선택됩니다.

해 경로 그림 아래의 테이블에는 모형 번호, 해당 모형 이름 및 현재 모형 선택이 표시됩니다. 이 테이블은 모형 선택 기준에 따라 정렬되며 최량 적합 모형이 맨 위에 있습니다. 수직 점선 맨 위의 슬라이더를 드래그하거나 테이블에서 직접 모형을 선택하여 다른 소파동 모형을 선택합니다. 다른 모형을 선택하면 "모형 선택" 보고서의 예측 그림과 다른 모든 보고서의 정보가 자동으로 업데이트됩니다.

JMP PRO 진단 그림

"진단 그림" 보고서에는 "실제값 대 예측값 그림"과 "잔차 대 예측값 그림"이 포함되어 있습니다. 이러한 그림은 현재 모형이 데이터에 얼마나 적합한지 평가하는 데 도움이 됩니다. "진단 그림" 보고서는 기본적으로 닫혀 있습니다.

JMP[®] PRO 함수 요약

ID 변수의 각 수준에 대한 함수 PCA의 요약을 표시합니다. 데이터 변동을 1% 넘게 설명하는 고유값과 관련된 함수 주성분이 기본적으로 표시됩니다. 평균, 표준편차, 중앙값, 최소값, 최대값, 통합 차이, RISE(제곱근 적분 제곱 오차) 및 RIFS(제곱근 적분 함수 제곱)도 표시됩니다. 통합 차이 및 RISE 요약 값은 ID 특정 함수가 전체 평균 함수와 얼마나 다른지 판단하는 데 사용됩니다. RIFS 요약 값은 최적 곡선 적합을 위해 사용됩니다. 자세한 내용은 "[함수 요약 보고서에 대한 통계 상세 정보](#)"에서 확인하십시오.

함수 요약 옵션

"함수 요약"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

함수 요약 사용자 정의 "함수 요약" 보고서에 표시되는 FPC 수와 요약 통계량을 선택할 수 있는 창을 표시합니다. 표시할 FPC 수가 지정되면 "함수 PCA" 보고서도 업데이트됩니다. 또한 "요약 저장" 옵션으로 생성되는 데이터 테이블에 각 함수에 대한 그래프를 포함할지 여부를 결정하는 "그래프 저장" 체크박스도 있습니다.

참고 : B-스플라인, P-스플라인 또는 Fourier 모형에 지정할 수 있는 최대 FPC 수는 10 개입니다. 소파동 모형 또는 직접 함수 PCA의 경우 지정할 수 있는 최대 FPC 수는 20 개입니다.

팁 : 함수 공정이 여러 개인 경우 Ctrl 키를 클릭한 채 "함수 요약 사용자 정의"를 선택하면 동일한 요약 값을 표시하도록 모든 "함수 요약" 보고서를 사용자 정의할 수 있습니다.

요약 저장 "함수 요약" 보고서에 지정된 요약 통계량을 새 데이터 테이블에 저장합니다. 새 데이터 테이블의 이름에 모형 적합이 설명되어 있습니다. 이 데이터 테이블에는 고유 함수, 평균 함수, 예측 함수 및 조건부 예측 함수에 대한 계산식 열이 포함됩니다. "함수 요약 사용자 정의" 창에서 "그래프 저장" 옵션을 선택한 경우 각 함수에 대해 지정된 모형 적합 및 원시 데이터의 그래프 이미지를 포함하는 열도 있습니다. 데이터 테이블에는 예측 계산식 및 조건부 예측 계산식에 대한 예측 프로파일러를 시작하는 프로파일러 스크립트가 있습니다. 이러한 계산식은 입력 변수, ID 변수 및 고유 함수의 함수입니다.

팁 : 소파동 모형의 경우 "요약 저장" 옵션을 사용하여 생성된 데이터 테이블에 소파동 계수도 저장됩니다.

관리도 빌더 관리도 빌더를 사용하여 함수 주성분을 분석합니다. "함수 요약" 테이블의 각 함수 주성분에 대해 개별 차트 및 "한계 요약" 테이블이 생성됩니다. 자세한 내용은 품질 및 공정 방법에서 확인하십시오.

JMP[®] PRO 계수

지정된 모형에 대한 추정 계수를 표시합니다.

기저 함수 모형의 경우 추정된 기저 함수 계수와 해당 표준편차가 표시됩니다. 이러한 계수는 ID 변수의 모든 수준에서 공통이며 혼합 모형 프레임워크의 고정 추정값입니다. 계수에 대한 표준 오차 및 신뢰 구간을 보려면 테이블을 마우스 오른쪽 버튼으로 클릭하고 **열**을 선택합니다.

소과동 모형의 경우 ID 변수의 각 수준에 대해 추정된 소과동 계수가 표시됩니다. 부소과동에 대한 열이 있고 나머지 계수에 대한 일련의 열이 있습니다. 각 계수 열은 입력 도메인에서 해당 해상도 및 위치로 식별됩니다. 계수 테이블에는 임계값을 사용하여 찾은 소과동 계수의 최소 표현이 포함되어 있습니다(Donoho, 1995 참조).

JMP PRO 함수별 랜덤 계수

각 기저 함수 및 함수 공정 조합에 대해 추정된 랜덤 계수를 표시합니다. 이 값은 ID 변수의 각 수준에 대해 고유하며 혼합 모형 프레임워크의 임의 효과 추정값입니다. 소과동 모형의 경우 "함수별 랜덤 계수" 보고서가 표시되지 않습니다.

JMP PRO 함수 PCA

적합된 함수 모형에 대해 함수 PCA(함수 주성분 분석)가 수행됩니다. "함수 PCA" 보고서에는 각 FPC(함수 주성분)에 해당하는 고유값이 가장 큰 것부터 순서대로 나열됩니다. 각 FPC에 의해 설명되는 변동 백분율과 누적 백분율이 나열되고 막대 차트에 표시됩니다. 평균 함수의 그래프 및 각 형상 함수에 대한 그래프가 있습니다. 이것은 고유 함수의 값입니다.

"함수 PCA" 보고서에서 모형 선택을 수행하여 선택한 함수 주성분 수를 다시 정의할 수 있습니다. BIC(Bayes 정보 기준) 대 FPC 수를 보여 주는 해 경로 그림이 있습니다. 현재 FPC 수는 해 경로 그림에서 수직 점선으로 지정됩니다. FPC 수가 다른 모형의 적합이 비슷할 수 있습니다. 따라서 해 경로 그림에서는 BIC 통계량 값의 구간을 나타내는 영역을 제공합니다. 녹색 영역과 노란색 영역이 있습니다. 녹색 영역에는 최소 BIC에서 최소 BIC + 4까지 구간에 해당하는 값이 포함되고, 노란색 영역에는 최소 BIC + 4에서 최소 BIC + 10까지 구간에 해당하는 값이 포함됩니다. 기본적으로 녹색 영역 내에서 FPC 수가 가장 적은 모형이 선택됩니다. 세로선 맨 위의 슬라이더를 드래그하여 FPC 수를 변경할 수 있습니다. 슬라이더를 드래그하면 "함수 PCA" 보고서의 다른 정보가 자동으로 업데이트됩니다.

참고: 전체 y 축 척도에 비해 좁은 영역은 그림에서 보기 어려울 수 있습니다. y 축을 확대하면 영역이 더 보기 좋게 표시됩니다.

직접 함수 PCA가 수행되면 전체 예측 그림과 개별 예측 그림 격자도 표시됩니다. 개별 예측 그림 격자는 "데이터 처리" 보고서의 개별 그림 격자와 레이아웃 및 컨트롤이 동일합니다. 한 번에 최대 20개의 그림이 표시되며, 개별 예측 그림의 다양한 그룹을 볼 수 있는 드롭다운 메뉴와 화살표가 있습니다. FPC 수를 업데이트하면 예측 그림도 자동으로 업데이트됩니다.

예측 그림에는 원시 데이터 및 현재 모형에 해당하는 예측 곡선이 표시됩니다. 검증 데이터 집합이 있는 경우 검증 데이터 집합에 포함된 함수에 대해서는 예측 곡선이 표시되지 않습니다. 전체 예측 그림의 곡선은 지정된 FPC 수에서 평균 예측 곡선입니다. 개별 예측 그림의 곡선은 지정된 FPC 수에서 각 특정 함수에 대한 예측 곡선입니다.

참고: 단일 함수만 모델링된 경우 " 함수 PCA" 보고서가 표시되지 않습니다. 그렇지 않은 경우 JMP 에서 함수 PCA 를 수행할 수 없으면 " 함수 PCA" 보고서에 오류 메시지가 나타납니다.

" 함수 PCA" 의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

진단 그림 " 실제값 대 예측값 그림 " 과 " 잔차 대 예측값 그림 " 을 표시하거나 숨깁니다. 이러한 그림은 선택된 함수 주성분 수를 고려하여 모형이 데이터에 얼마나 적합한지 평가하는 데 도움이 됩니다.

스코어 그림 FPC 스코어의 스코어 그림을 표시하거나 숨깁니다. " 성분 선택 " 아래의 목록을 사용하여 스코어 그림의 각 축에 표시할 FPC 를 지정할 수 있습니다. FPC 가 하나뿐인 경우 $y=x$ 선에 FPC 스코어가 표시되고 성분 변경 목록은 표시되지 않습니다. 스코어 그림은 이상치를 감지하는 데 유용합니다. FPC 스코어의 경우 스코어 그림을 사용하여 ID 변수에서 이상치 함수가 있는 수준을 감지할 수 있습니다. 스코어 그림에서 점을 선택하면 FPC 프로파일러가 해당 함수의 스코어로 설정됩니다.

팁: 스코어 그림의 점을 커서로 가리키면 ID 변수의 해당 수준에 대한 적합 곡선 그림이 표시됩니다.

FPC 프로파일러 FPC 스코어의 프로파일러를 표시하거나 숨깁니다. "FPC 프로파일러"에는 입력 변수에 대한 열과 각 FPC 스코어에 대한 열이 포함됩니다. 대상 함수를 지정하면 프로파일러 그래프 위에 추가 버튼 옵션이 있습니다. 대상 함수를 최적화하고 대상 프로파일러를 표시하거나 숨길 수 있습니다. " 대상 프로파일러 표시 "를 선택하면 두 개의 프로파일러가 보고서에 추가됩니다. 하나는 대상 함수와의 차이를 측정하고 다른 하나는 대상 함수와의 적분 오차를 측정합니다. "FPC 프로파일러"의 빨간색 삼각형 메뉴 옵션에 대한 자세한 내용은 **Profilers**에서 확인하십시오.

팁: 프로파일러에서 모든 FPC 스코어를 0으로 재설정하려면 "재설정" 버튼을 사용하십시오.

FPC 수 사용자 정의 " 함수 PCA"에 표시할 FPC 스코어 수를 지정합니다. 이 옵션에서 FPC 스코어 수를 지정하면 " 함수 요약" 보고서도 업데이트됩니다. 평균 모형을 보려면 FPC 수를 0으로 설정합니다.

JMP PRO 함수 데이터 탐색기 플랫폼 옵션

" 함수 데이터 탐색기 "의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

요약 다음과 같은 함수 요약 통계량 옵션을 제공하는 하위 메뉴입니다.

평균 그림 " 요약 " 보고서에 함수 평균 그림을 표시하거나 숨깁니다. 기본적으로 설정되어 있습니다.

표준편차 그림 " 요약 " 보고서에 함수 표준편차 그림을 표시하거나 숨깁니다. 기본적으로 설정되어 있습니다.

중앙값 그림 " 요약 " 보고서에 함수 중앙값 그림을 표시하거나 숨깁니다 .

모형 다음과 같은 모형 옵션을 제공하는 하위 메뉴입니다 .

B- 스플라인 기저 스플라인 (B- 스플라인) 모형을 데이터에 적합시킵니다 . 비주기 데이터의 경우 B- 스플라인 모형을 사용합니다 .

P- 스플라인 별점 기저 스플라인 (P- 스플라인) 모형을 데이터에 적합시킵니다 .

Fourier 기저 Fourier 기저 모형을 데이터에 적합시킵니다 . 주기 데이터의 경우 Fourier 기저 모형을 사용합니다 . 주기 모형은 함수의 시작 위치와 완료 위치가 같다고 가정합니다 . 자세한 내용은 "[Fourier 기저 모형](#) "에서 확인하십시오 .

소파동 여러 소파동 모형을 데이터에 적합시킵니다 . 소파동 모형은 정상점이 많이 포함된 데이터에 유용한 기저 함수 모형의 한 유형입니다 . 이 옵션을 사용하려면 데이터가 균등한 간격의 격자에 있어야 합니다 . 데이터 간격이 균등하지 않으면 소파동 루틴 전에 자동으로 격자가 생성됩니다 .

"소파동" 옵션을 선택하면 Haar, Daubechies, Symlet, Coiflet 및 배직교 계열의 모형에서 여러 소파동 모형이 적합됩니다 . 이것은 모두 모수에 따라 정상점의 형태와 유형이 다른 유연한 함수입니다 . 이러한 모형의 모형 적합 결과는 "모형 선택 " 보고서의 테이블에 표시되며 최량 적합 모형이 맨 위에 오도록 정렬됩니다 . 자세한 내용은 "[소파동 모형](#) "에서 확인하십시오 .

소파동에 대한 자세한 내용은 Nason 연구 자료 (2008)에서 확인하십시오 .

참고 : 고유 입력 값이 세 개 미만인 경우 Fourier 기저 모형이나 P- 스플라인 모형을 데이터에 적합시킬 수 없으며 경고 메시지가 나타납니다 .

모형 제어 모형을 적합시키기 전에 "모형 제어 " 패널을 열 수 있는 하위 메뉴를 표시합니다 . 자세한 내용은 "[모형 제어](#) "에서 확인하십시오 . 또한 이 옵션을 사용하면 모형을 적합시키기 전에 "모형 선택 "의 빨간색 삼각형 메뉴 옵션을 통해 모형 선택 기준을 변경할 수 있습니다 .

직접 함수 PCA (단일 함수만 있는 경우에는 사용 불가능) 기저 함수 모형을 먼저 적합시키지 않고 데이터에 대해 함수 주성분 분석을 직접 수행합니다 . 이렇게 하면 특히 대규모 데이터 집합의 경우 계산 시간이 감소합니다 . 직접 함수 PCA 는 다음과 같이 구현됩니다 .

1. 입력 데이터를 0 에서 1 사이로 맞추고 관측값을 입력 값의 공통 격자에 보간합니다 .
2. 데이터에 대해 함수 주성분 분석을 수행합니다 .
3. 각 격자점에 매듭이 있는 P- 스플라인 모형을 사용하여 첫 번째 고유 함수를 평활화합니다 .
4. 데이터에서 첫 번째 평활 고유 함수를 제거하고 , 많은 양의 데이터 변동이 설명될 때까지 [2 단계 ~ 4 단계](#)를 반복합니다 .

직접 함수 PCA 를 수행하면 " 함수 PCA " 보고서가 표시됩니다 . 자세한 내용은 "[함수 PCA](#) "에서 확인하십시오 .

데이터 저장 처리된 데이터를 새 데이터 테이블에 저장합니다 . 처리된 데이터는 쌍인 데이터 형식으로 저장됩니다 .

다음 옵션에 대한 자세한 내용은 **JMP 사용**의 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.

JMP PRO 함수 데이터 탐색기 그룹 옵션

함수 데이터 탐색기 그룹 옵션은 시작 창의 "쌓인 데이터 형식" 탭에서 함수 공정을 여러 개 지정한 경우에만 사용할 수 있습니다.

요약 모든 함수 공정에 대한 함수 요약 통계량 그림을 표시하거나 숨기는 옵션 하위 메뉴입니다. 자세한 내용은 "**요약**"에서 확인하십시오.

데이터 처리 모든 함수 공정에 적용할 수 있는 데이터 처리 옵션 하위 메뉴입니다. 자세한 내용은 "**데이터 처리 옵션**"에서 확인하십시오.

모형 모든 함수 공정에 적용시킬 수 있는 모형 옵션 하위 메뉴입니다. 자세한 내용은 "**모형**"에서 확인하십시오.

데이터 저장 모든 함수 공정에 대해 처리된 데이터를 새 데이터 테이블에 저장합니다. 처리된 데이터는 쌓인 데이터 형식으로 저장됩니다.

요약 저장 각 "함수 요약" 보고서에 지정된 요약 통계량을 새 데이터 테이블에 저장합니다. 새 데이터 테이블의 이름에 모형 적합이 설명되어 있습니다. 이 데이터 테이블에는 각 Y 변수에 대한 고유 함수, 평균 함수, 예측 함수 및 조건부 예측 함수의 계산식 열이 포함됩니다. 데이터 테이블에는 각 Y 변수에 대해 예측 계산식 및 조건부 예측 계산식에 대한 예측 프로파일러를 시작하는 프로파일러 스크립트가 있습니다. 이러한 계산식은 입력 변수, ID 변수 및 고유 함수의 함수입니다.

행으로 배열 창에서 가로로 표시되는 보고서 수를 지정합니다.

다음 옵션에 대한 자세한 내용은 **JMP 사용**의 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.

JMP PRO 데이터 처리 옵션

함수 데이터 탐색기 플랫폼의 "데이터 처리" 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

정리 다음과 같은 데이터 정리 옵션이 포함된 하위 메뉴입니다.

0 제거 값이 0 인 관측값을 제거합니다. 데이터에 0 이 없으면 0 값이 없음을 나타내는 경고가 표시됩니다.

값 제거 데이터에서 제거할 값을 지정할 수 있는 규격 창을 표시합니다.

선택한 항목 제거 데이터 테이블에서 선택한 행에 해당하는 관측값을 제거합니다.

선택 취소한 항목 제거 데이터 테이블에서 선택하지 않은 행에 해당하는 관측값을 제거합니다.

주의: "선택한 항목 제거" 및 "선택 취소한 항목 제거"는 행 번호를 제거합니다. "자동 재계산" 이 활성화된 경우 이러한 옵션을 사용하기 전에 행을 추가하거나 삭제해야 합니다.

X 필터링 지정된 구간을 벗어나는 X 값을 제거합니다. "X 필터링" 옵션을 선택하면 "다음 값 미만" 및 "다음 값 초과" 값을 지정해야 합니다. 지정된 구간을 벗어나는 X 값은 분석에 사용되지 않습니다.

Y 필터링 지정된 구간을 벗어나는 Y 값을 제거합니다. "Y 필터링" 옵션을 선택하면 "다음 값 미만" 및 "다음 값 초과" 값을 지정해야 합니다. 지정된 구간을 벗어나는 Y 값은 분석에 사용되지 않습니다.

축소 다음 기법 중 하나를 사용하여 X 값에 대한 데이터를 줄입니다.

- "격자" 탭을 사용하여 관측값을 값의 공통 격자에 보간합니다. 격자 크기를 지정할 수 있습니다. 기본적으로 격자 크기는 가장 긴 함수의 값 개수입니다. 이는 최대 허용 격자 크기이기도 합니다.

- "계급" 탭을 사용하여 고유 X 값에서 균등한 간격으로 지정된 수의 계급을 생성합니다. 각 함수 (또는 "ID, 함수" 변수의 수준)에 대해 계급 내의 관측값을 평균화하여 해당 계급 수준에 대한 Y 값을 생성합니다.
- "줄임" 탭을 사용하여 X 값에 대한 모든 N 개 관측값을 제거합니다. 여기서 N 은 지정된 줄임 비율에 의해 결정됩니다. 이 작업은 각 함수 (또는 "ID, 함수" 변수의 수준)에 대해 수행됩니다. 줄임 비율의 기본값은 2 이며, 이 값은 각 함수에서 관측값의 절반을 제거합니다.

참고 : "제거" 옵션을 사용하면 지정된 관측값이 분석 및 모델링 보고서에서 제외되지만 데이터 테이블에서 해당 관측값이 변경되지는 않습니다.

변환 다음과 같은 데이터 변환 옵션이 포함된 하위 메뉴입니다.

중심화 결과를 가운데에 맞춥니다.

표준화 평균 0, 분산 1 이 되도록 데이터를 중심에 맞추고 척도를 조정하여 결과를 표준화합니다.

범위 0 ~ 1 0 ~ 1 범위에 들어가도록 결과의 척도를 조정합니다.

제곱근 결과의 제곱근을 계산하여 데이터를 변환합니다. 출력 값은 음수가 아니어야 합니다.

제곱 결과의 제곱을 계산하여 데이터를 변환합니다.

로그 결과의 자연 로그를 계산하여 데이터를 변환합니다.

지수 결과의 지수 함수를 계산하여 데이터를 변환합니다.

부정 결과에 음수를 취해 데이터를 변환합니다.

로짓 결과의 로짓 함수를 계산하여 데이터를 변환합니다. 출력 값은 0 에서 1 사이여야 합니다.

로그 X 입력의 자연 로그를 계산하여 데이터를 변환합니다.

맞춤 다음과 같은 입력 데이터 맞춤 옵션이 포함된 하위 메뉴입니다.

행 맞춤 입력 값을 행 번호로 바꿉니다.

최대 맞춤 각 ID 수준에 대해 관측된 최대 출력 값을 사용하여 함수를 정렬합니다. 관측된 최대 출력 값과 관련된 입력 값은 각 ID 수준에 대해 0 으로 설정되고, 다른 입력 값은 관측된 최대값과 0 사이의 차이에 따라 위 또는 아래로 이동합니다.

최소 맞춤 각 ID 수준에 대해 관측된 최소 출력 값을 사용하여 함수를 정렬합니다. 관측된 최소 출력 값과 관련된 입력 값은 각 ID 수준에 대해 0 으로 설정되고, 다른 입력 값은 관측된 최소값과 0 사이의 차이에 따라 위 또는 아래로 이동합니다.

0 ~ 1 정렬 입력 값의 범위가 0 ~ 1 이 되도록 출력 함수를 정렬합니다.

팁 : "0 ~ 1 정렬" 은 P-스플라인 모형을 적합시킬 때 특히 유용합니다.

함수별 정렬 각 함수가 입력 값의 전체 최소값에서 시작하여 입력 값의 전체 최대값에서 끝나도록 출력 함수를 정렬합니다.

DTW(동적 시간 뒤틀림) (함수가 두 개 이상인 경우에만 사용 가능) DTW(동적 시간 뒤틀림)를 사용하여 출력 함수를 정렬합니다. DTW는 둘 이상의 함수를 서로 맞추기 위해 최적의 뒤틀림을 찾는 함수 맞춤 기법입니다. DTW 옵션을 선택하면 "대조 함수 선택" 창이 나타납니다. 이를 사용하여 대조 함수를 선택합니다. 대조 함수는 나머지 함수를 맞춤 때 사용할 기준 함수입니다.

대조 함수를 선택하고 "확인"을 클릭하면 나머지 쿼리 함수 목록과 함께 뒤틀림 함수 그림이 표시됩니다. 뒤틀림 함수 그림에서 대조 함수는 y 축에 있고 선택한 쿼리 함수는 x 축에 있습니다. 빨간색 대각선 ($y=x$) 과의 편차는 맞춤을 개선하기 위해 쿼리 함수의 입력에 뒤틀림이 적용되었음을 나타냅니다.

스펙트럼 스펙트럼 데이터에 유용한 다음 옵션이 포함된 하위 메뉴입니다.

SNV 표준 정규 변량 방법을 데이터에 적용합니다. 이 방법은 평균 0, 표준편차 1 이 되도록 각 개별 함수 (ID 변수의 수준) 를 중심에 맞추고 척도를 조정하여 출력을 표준화합니다.

MSC 승법 산포 수정을 데이터에 적용합니다. 각 개별 함수 (ID 변수의 수준) 에 대해 단순 선형 회귀가 적합됩니다. 여기서 반응은 함수의 출력 값이고 회귀변수는 평균 함수의 출력 값입니다. 원래 출력 값 y_{it} 는 다음 방정식을 사용하여 새 값 y_{it}^* 로 대체됩니다.

$$y_{it}^* = \frac{y_{it} - \bar{y}_t}{b_i}$$

여기서 b_i 는 함수 i 에 대한 단순 선형 회귀에서 얻은 기울기입니다. 자세한 내용은 Geladi et al 연구 자료 (1985) 에서 확인하십시오.

Savitzky-Golay Savitzky-Golay 방법을 사용할 수 있는 옵션을 제공합니다. 자세한 내용은 Savitzky and Golay 연구 자료 (1964) 에서 확인하십시오.

참고 :Savitzky-Golay 방법과 관련된 모든 옵션을 사용하려면 입력 데이터가 균등한 간격의 격자에 있어야 하고 적어도 하나의 함수에 7개 이상의 데이터 점이 포함되어야 합니다. 데이터가 균등한 간격의 격자에 있지 않은 경우 Savitzky-Golay 옵션을 선택하면 균등한 간격의 격자에 데이터가 자동으로 배치됩니다.

필터 Savitzky-Golay 필터를 데이터에 적용합니다. 이 방법은 로컬 다항식을 도메인 전체의 여러 점 집합에 적합시킵니다. 최소 제곱을 사용하여 다항식이 적합되고 각 적합의 점 수는 대역폭에 의해 결정됩니다. 이 옵션을 선택하면 0, 1, 2 차 다항식과 최대 10 개의 대역폭에 대해 여러 적합이 생성됩니다. 각 함수의 최량 적합 모형은 AIC 를 기준으로 선택됩니다. 다항식 차수와 대역폭은 함수마다 다를 수 있습니다.

1 차 도함수 2 차 또는 3 차 다항식만 사용하여 Savitzky-Golay 필터를 데이터에 적용한 후 1 차 도함수를 사용합니다. 필터가 다항식을 적합시키므로 도함수가 분석적으로 계산됩니다.

2 차 도함수 3 차 다항식만 사용하여 Savitzky-Golay 필터를 데이터에 적용한 후 2 차 도함수를 사용합니다. 필터가 다항식을 적합시키므로 도함수가 분석적으로 계산됩니다.

기준 수정 각 개별 함수에서 기준 함수를 뺍니다. 기준 수정은 제거할 추세 또는 기준을 알고 있을 때 사용됩니다. 예를 들어 데이터 측정 방식의 아티팩트가 원인일 수 있습니다. 일반적으로 정보는 데이터의 정상점에 있으므로 이러한 영역은 기준 모형에 포함되지 않습니다.

이 옵션을 선택하면 기준 수정 창이 표시됩니다. 이 창에는 데이터를 표시하는 선택 그림과 기준 모형을 지정하는 옵션 집합이 포함되어 있습니다. 기준 수정 창에는 다음 옵션이 있습니다.

기준 모형 기준 함수에 대한 모형 유형을 지정합니다. 선형, 2 차, 3 차, 2 모수 지수 또는 3 모수 지수 모형을 지정할 수 있습니다.

수정 영역 기준 함수를 빼는 영역을 지정합니다. 전체 함수 영역에서 기준을 빼거나, 기준 모형을 생성하는 데 사용된 영역에서만 기준을 뺄 수 있습니다.

기준 영역 선택 그림에 파란색 세로선 한 쌍을 추가하거나 제거합니다. 선은 처음에서로 포개져 있습니다. 이 선을 이동하여 기준 모형에 포함하지 않을 데이터 영역을 지정할 수 있습니다. 파란색 선 쌍 사이에 있는 데이터 영역은 기준 모형에 포함되지 않습니다.

앵커 지점 선택 그림에 빨간색 세로선을 추가하거나 제거합니다. 이 선은 기준 모형에 적용되는 데이터 점을 지정합니다.

대상 함수 (함수가 두 개 이상인 경우에만 사용 가능) 대상 함수를 로드할 수 있습니다.

대상 로드 대상 함수를 지정할 수 있는 창을 표시합니다. 대상 함수는 모든 함수가 대상 함수(대조 함수 또는 황금 곡선이라고도 함)처럼 보이도록 곡선 매칭에 사용됩니다.

대상 함수를 지정하는 경우 해당 함수의 데이터는 모형 적합에 사용되지 않습니다. 대상 함수를 지정하면 FPC 프로파일러에 다른 옵션이 추가됩니다. 자세한 내용은 "[FPC 프로파일러](#)"에서 확인하십시오.

참고 : 다른 전처리 단계를 수행하기 전에 대상 함수를 로드해야 합니다.

DTW(동적 시간 뒤틀림) 옵션

뒤틀림 함수 그림 뒤틀림 함수 그림을 표시하거나 숨깁니다. 기본적으로 설정되어 있습니다.

거리 행렬 저장 거리 행렬을 별도의 데이터 테이블에 저장합니다. 거리 행렬은 함수를 군집화하는 데 유용할 수 있습니다. 거리 행렬 데이터 테이블에는 "계층적 군집화" 스크립트가 포함되어 있습니다.

뒤틀림 함수 저장 뒤틀림 함수를 별도의 데이터 테이블에 저장합니다. 데이터 테이블의 각 행에는 DTW 조정 입력 변수, 원래 입력 변수 및 ID 변수가 포함됩니다.

JMP PRO 모형 보고서 옵션

함수 데이터 탐색기 플랫폼에서 데이터에 모형을 적합시키면 모형 보고서가 나타납니다. 모형 보고서의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

모형 선택 모형 선택 기준으로 사용할 옵션이 포함된 하위 메뉴를 표시합니다. AICc, BIC 및 GCV 중에서 선택할 수 있습니다. 자세한 내용은 **선형 모형 적합**에서 확인하십시오.

그림 기반 (소파동 또는 직접 함수 PCA 에는 사용 불가능) 모든 기저 함수 그림을 하나의 그래프에 표시하거나 숨깁니다.

진단 그림 (소파동 또는 직접 함수 PCA 에는 사용 불가능) "진단 그림" 보고서를 표시하거나 숨깁니다. 자세한 내용은 "**진단 그림**"에서 확인하십시오.

함수 요약 "함수 요약" 보고서를 표시하거나 숨깁니다. 자세한 내용은 "**함수 요약**"에서 확인하십시오.

기저 함수 계수 (직접 함수 PCA 에는 사용 불가능) 해당 모형 적합에 대한 계수 보고서를 표시하거나 숨깁니다. 자세한 내용은 "**계수**"에서 확인하십시오.

함수별 랜덤 계수 (소파동 또는 직접 함수 PCA 에는 사용 불가능) "함수별 랜덤 계수" 보고서를 표시하거나 숨깁니다. 자세한 내용은 "**함수별 랜덤 계수**"에서 확인하십시오.

함수 PCA "함수 PCA" 보고서를 표시하거나 숨깁니다. 자세한 내용은 "**함수 PCA**"에서 확인하십시오.

소파동 DOE 분석 (시작 창에서 하나 이상의 보조 변수를 지정하고 소파동 모형이 데이터에 적합된 경우에만 사용 가능) FDE 플랫폼 내에서 "일반화 회귀" 보고서를 시작합니다. 보조 변수를 모형 효과로 사용하여 일반화 회귀 모형이 소파동 계수에 적합됩니다. 소파동 계수가 많을 수 있으므로 가장 큰 영향을 미치는 일부 계수만 적합됩니다. 이러한 계수 부분집합은 각 소파동 계수의 에너지에 의해 결정됩니다. 소파동 계수의 에너지는 다음과 같이 계산됩니다.

$$E_{j,k} = \sum_{i=1}^N d_{j,k,i}^2$$

여기서 j = 해상도, k = 입력 도메인에서의 위치, i = 함수, $d = (j,k,i)$ 에 대한 소파동 계수입니다. 총 에너지는 계수에 대한 모든 에너지 값의 합입니다.

$$E = \sum_j \sum_k E_{j,k}$$

에너지 값은 "소파동 DOE 분석" 보고서에 표시되는 "에너지" 테이블에 사용됩니다. "에너지" 테이블에는 다음 열이 포함되어 있습니다.

계수 해상도와 입력 도메인 내 위치로 식별되는 소파동 계수입니다.

에너지 해당 소파동 계수에 대한 상대 에너지입니다. 상대 에너지는 모든 소파동 계수에 대해 최대 에너지로 표준화된 에너지 값입니다. 이 값은 $E_{j,k}/\max_{j,k}(E_{j,k})$ 로 계산됩니다.

백분율 총 에너지 대비 해당 소파동 계수의 에너지 백분율입니다. 이 값은 해당 계수가 모든 계수의 합에 비해 얼마나 큰지 설명합니다.

누적 총 에너지의 누적 백분율입니다.

총 에너지의 90% 이상을 누적하여 설명하는 계수가 "에너지" 테이블에 표시됩니다.

"에너지" 테이블의 부소파동 및 각 소파동 계수에 대해 일반화 회귀 모형이 적합됩니다. 기본적으로 2 차 요인 모형이 적합되고 "추정 방법"은 "최량 부분집합"입니다. 범주형 2 차 효과는 포함되지 않습니다. 모형의 항 수가 20 개를 초과하거나 함수 수가 1000 개를 초과하면 "추정 방법"이 자동으로 "가지 치기 전진 선택"으로 전환됩니다. 아니면 원래 데이터 테이블에서 원하는 모형 적합을 정의하는 모형 스크립트를 지정할 수도 있습니다.

보조 변수를 사용하여 개별 소파동 계수를 모델링하면 보조 변수 값이 특정 입력 값 주위의 정상점 위치에 어떤 영향을 주는지 확인할 수 있습니다. 이는 각 소파동 계수가 입력 도메인의 특정 위치와 직접적인 관련이 있기 때문입니다. 그런 다음 FDOE 프로파일러를 사용하여 보조 변수가 반응에 어떤 영향을 주는지 탐색할 수 있습니다.

팁 : "소파동 DOE 분석" 옵션이 로드되지 않으면 데이터에 대해 표준화 전처리 단계를 사용하고, 중심화 및 척도화된 데이터에서 소파동 모형을 다시 적합시킨 다음, 소파동 DOE 분석을 시작해 보십시오.

"소파동 DOE 분석" 보고서에는 다음과 같은 빨간색 삼각형 메뉴 옵션이 있습니다.

소파동 계수에 대한 일반화 회귀 각 소파동 계수에 대한 "일반화 회귀" 보고서를 표시하거나 숨깁니다. "일반화 회귀" 모형 보고서에 대한 자세한 내용은 **선형 모형 적합**에서 확인하십시오.

FDOE 프로파일러 보조 변수에 따라 반응이 어떻게 변하는지 탐색할 수 있는 FDOE 프로파일러를 표시하거나 숨깁니다. "FDOE 프로파일러"의 빨간색 삼각형 메뉴 옵션에 대한 자세한 내용은 **Profilers**에서 확인하십시오.

참고 : 소파동 DOE 분석 보고서의 FDOE 프로파일러는 대상 함수를 지원하지 않습니다. 대상 함수를 검토하려면 "함수 DOE 분석" 옵션을 대신 사용하십시오.

열 저장 "예측 계산식 저장" 및 "잔차 계산식 저장" 옵션이 있는 목록을 표시합니다. 이러한 옵션은 해당 계산식을 데이터 테이블의 새 열에 저장합니다.

함수 DOE 분석 (시작 창에서 하나 이상의 보조 변수를 지정한 경우에만 사용 가능) FDE 플랫폼 내에서 "일반화 회귀" 보고서를 시작합니다. 일반화 회귀 모형은 보조 변수를 모형 효과로 사용하여 각 FPC 스코어 함수에 적합됩니다. 기본적으로 2 차 요인 모형이 적합되고 "추정 방법"은 "최량 부분집합"입니다. 범주형 2 차 효과는 포함되지 않습니다. 모형의 항 수가 20 개를 초과하거나 함수 수가 1000 개를 초과하면 "추정 방법"이 자동으로 "가지 치기 전진 선택"으로 전환됩니다. 아니면 원래 데이터 테이블에서 원하는 모형 적합을 정의하는 모형 스크립트를 지정할 수도 있습니다. 보조 변수를 사용하여 FPC 스코어를 모델링하면 모형 적합을 사용하여 보조 변수에 따라 반응이 어떻게 변하는지 확인할 수 있습니다. 보조 변수가 반응에 어떤 영향을 주는지 탐색하려면 FDOE 프로파일러를 사용합니다.

" 함수 DOE 분석 " 보고서에는 다음과 같은 빨간색 삼각형 메뉴 옵션이 있습니다 .

FPC 스코어에 대한 일반화 회귀 각 FPC 스코어에 대한 " 일반화 회귀 " 보고서를 표시하거나 숨깁니다 . " 일반화 회귀 " 모형 보고서에 대한 자세한 내용은 선행 모형 적합에서 확인하십시오 .

진단 그림 반응 변수에 대한 실제값 대 예측값 그림과 잔차 그림을 표시하거나 숨깁니다 .

FDOE 프로파일러 보조 변수에 따라 반응이 어떻게 변하는지 탐색할 수 있는 FDOE 프로파일러를 표시하거나 숨깁니다 . 대상 함수를 지정하면 프로파일러 그래프 위에 추가 버튼 옵션이 있습니다 . 대상 함수를 최적화하고 대상 프로파일러를 표시하거나 숨길 수 있습니다 . " 대상 프로파일러 표시 " 를 선택하면 두 개의 프로파일러가 보고서에 추가됩니다 . 하나는 대상 함수와의 차이를 측정하고 다른 하나는 대상 함수와의 적분 오차를 측정합니다 . " FDOE 프로파일러 " 의 빨간색 삼각형 메뉴 옵션에 대한 자세한 내용은 **Profilers** 에서 확인하십시오 .

열 저장 " 예측 계산식 저장 " 및 " 잔차 계산식 저장 " 옵션이 있는 목록을 표시합니다 . 이러한 옵션은 해당 계산식을 데이터 테이블의 새 열에 저장합니다 .

데이터 저장 모델링된 데이터를 새 데이터 테이블에 저장합니다 . 모델링된 데이터는 쌓인 데이터 형식으로 저장됩니다 .

적합 제거 지정된 적합에 대한 모형 보고서를 제거합니다 .

JMP PRO 함수 데이터 탐색기 플랫폼의 추가 예

이 섹션에는 함수 데이터 탐색기 플랫폼을 사용하는 예가 포함되어 있습니다 .

- " 여러 함수 공정의 예 "
- " 함수 DOE 분석의 예 "

JMP PRO 여러 함수 공정의 예

이 예에서는 함수 데이터 탐색기 플랫폼을 사용하여 모형을 데이터에 적합시키고 함수 주성분을 새 데이터 테이블에 저장합니다 . 그런 다음 모형 적합 플랫폼의 일반화 회귀 분석법으로 함수 주성분을 분석합니다 .

효소 생산을 분석하려고 합니다 . Yield 는 유전자 변형 효모에 의해 생산된 효소의 양입니다 . 12 시간 동안 동일한 시간 간격으로 수집된 배치당 100 개의 공정 측정값이 있습니다 .

함수 모형 적합

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Functional Data/Fermentation Process.jmp** 를 엽니다 .
2. **분석 > 전문 모델링 > 함수 데이터 탐색기**를 선택합니다 .
3. " 쌓인 데이터 형식 " 탭에서 **Ethanol ~ pH** 를 선택하고 **Y, 출력**을 클릭합니다 .

4. Time 을 선택하고 **X, 입력**을 클릭합니다.
5. BatchID 를 선택하고 **ID, 함수**를 클릭합니다.
6. **확인**을 클릭합니다.
7. " 함수 데이터 탐색기 그룹 " 의 빨간색 삼각형을 클릭하고 **데이터 처리 > 맞춤 > 0 ~ 1 정렬**을 선택합니다. 이렇게 하면 각 " 함수 데이터 탐색기 " 보고서에서 입력 변수의 범위가 0 에서 1 사이로 맞춰집니다.
8. " 함수 데이터 탐색기 그룹 " 의 빨간색 삼각형을 클릭하고 **모형 > B- 스플라인**을 선택합니다. 이렇게 하면 각 함수 공정에 B- 스플라인 모형이 적합됩니다.

그림 16.6 Ethanol 에 대한 함수 데이터 탐색기 보고서

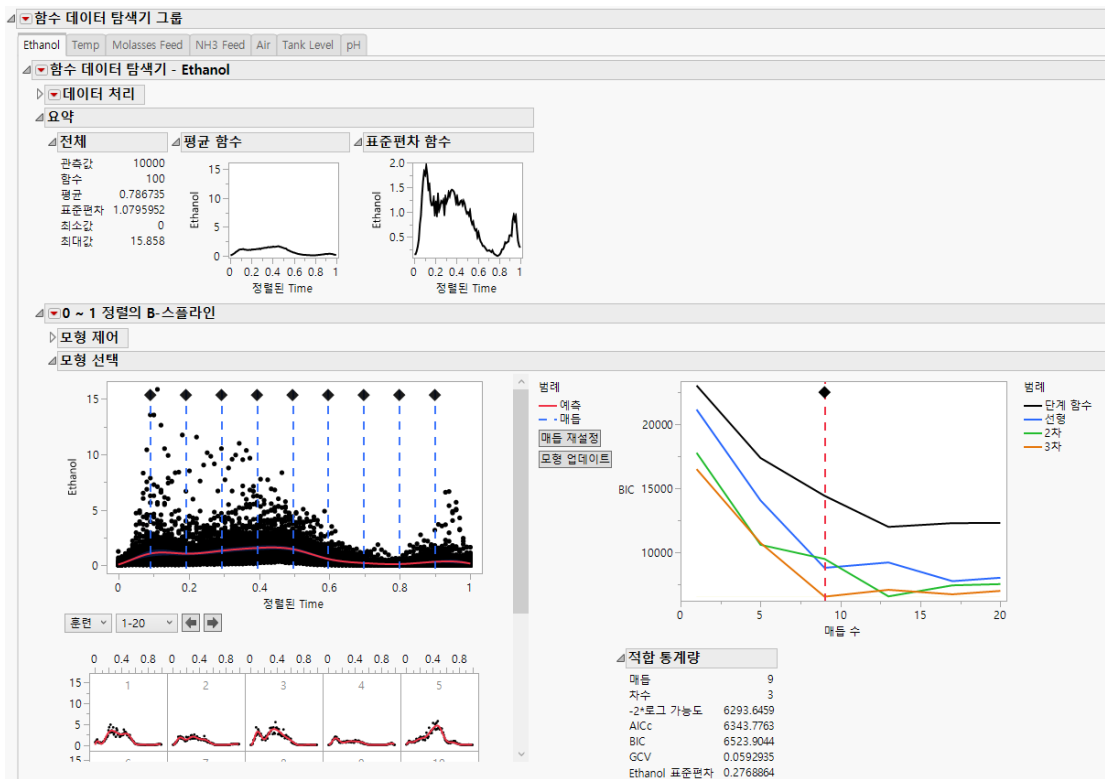


그림 16.7 Ethanol 에 대한 함수 요약 보고서

함수 요약

BatchID	FPC 1	FPC 2	FPC 3	FPC 4	FPC 5	FPC 6	FPC 7	FPC 8	FPC 9	FPC 10	평균	표준편차	중앙값	최소값	최대값	통합 차이	제곱근 적분	제곱근 적분
1	0.6845185	0.4272826	0.3090167	-0.174881	-0.261418	0.2534972	0.1329424	0.0932066	-0.0227	0.0295575	1.2856117	1.2574705	0.7225838	0.0275282	3.4532315	-0.297809	0.9637933	1.7832112
2	0.2003363	-0.162552	0.1458985	-0.006107	0.0059254	0.0042297	0.0488981	0.090565	-0.044089	-0.007963	0.848664	0.71297	0.7468215	0.0333247	2.0280261	-0.736315	0.317081	1.1007059
3	1.0570202	0.1670821	-0.025524	0.1554196	0.1599784	-0.213975	0.0257032	0.1420075	-0.077034	0.0378548	1.5221477	1.3465031	1.4533318	-0.018659	3.8349448	-0.060525	1.1267728	2.0232175
4	-0.225892	-0.342518	0.0456513	0.0469127	0.095793	-0.009734	0.0396208	0.0254217	-0.006387	0.0042403	0.5248517	0.4519468	0.4568174	0.0261908	1.6643826	-1.061051	0.4292202	0.6897988
5	0.7066471	0.8878144	-0.416058	0.1123721	0.0512395	0.1779	0.0420387	-0.099458	0.0015278	-0.002762	1.4912892	1.4352359	0.828356	0.068398	4.6865036	-0.091374	1.2326783	2.0610547
6	-0.164355	0.0402334	0.0698129	0.0188387	-0.004275	-0.031643	-0.01363	-0.001999	0.0493236	0.0057484	0.648615	0.4651278	0.6306233	0.0425897	1.5070566	-0.936894	0.1937517	0.7960802
7	0.4840975	0.0604706	-0.02545	0.138502	-0.044549	-0.079275	-0.003466	-0.047315	-0.096165	-0.073317	1.188724	0.8264775	1.3331523	0.0450854	2.337651	-0.395091	0.5320304	1.4384966
8	0.0598675	0.3232751	-0.194786	0.1468215	0.0999753	-0.021161	-0.007182	-0.06898	0.0521077	-0.059809	0.8913351	0.7786373	0.7196766	0.0029737	2.5753376	-0.69338	0.4349227	1.1820858
9	-0.35775	-0.244627	0.0950901	0.0453786	0.0302305	0.027442	0.0319439	0.0218378	0.024305	-0.022845	0.4330103	0.3329415	0.4001359	0.0494026	1.0654932	-1.153327	0.4507855	0.5436156
10	0.1030905	0.3791323	-0.32985	-0.309402	0.1014623	0.1254908	-0.011579	-0.009238	-0.031353	-0.041127	1.0564703	0.84232	0.8304595	0.068398	3.0158387	-0.528012	0.622775	1.343811

기저 함수 계수

모수	평균 추정값	표준편차 추정값
정렬된 Time[0]	0.068398	0
정렬된 Time[0.04]	0.3628335	0.6072559
정렬된 Time[0.1]	1.448524	2.3982214
정렬된 Time[0.19]	0.8753212	0.9861143
정렬된 Time[0.29]	1.3486495	1.4351881
정렬된 Time[0.39]	1.5859873	1.4953555
정렬된 Time[0.49]	1.69221	1.2326668
정렬된 Time[0.6]	0.396788	0.7335251
정렬된 Time[0.7]	0.2061623	0.2751322
정렬된 Time[0.8]	0.0235344	0
정렬된 Time[0.89]	0.3734323	0.6151888
정렬된 Time[0.95]	0.3903473	1.3136877
정렬된 Time[1]	0.1561339	0.1818355

그림 16.6 및 그림 16.7에서는 함수 공정 변수 중 하나인 Ethanol 에 대한 모형 보고서를 보여 줍니다. 전체 보고서를 스크롤하여 각 공정 변수에 대한 모형 적합을 확인합니다. 그런 다음 " 함수 요약 " 보고서의 FPC 를 분석에 사용합니다.

FPC 저장 및 Yield 결과 연결

1. Ctrl 키를 누른 채 " 함수 요약 " 의 빨간색 삼각형을 클릭하고 **함수 요약 사용자 정의**를 선택합니다.
 2. **표시할 FPC 수 입력** 옆의 상자에 "3" 을 입력합니다.
 3. **모든 요약 선택 취소** 상자를 클릭합니다.
 4. **확인**을 클릭합니다.
 5. " 함수 데이터 탐색기 그룹 " 의 빨간색 삼각형을 클릭하고 **요약 저장**을 선택합니다.
 6. **도움말 > 샘플 데이터 폴더**를 선택하고 Functional Data/Fermentation Process Batch Yield Results.jsp 를 엽니다.
 7. 함수 데이터 탐색기 모형 요약 .jmp 데이터 테이블에서 BatchID 를 마우스 오른쪽 버튼으로 클릭하고 **연결 ID** 를 선택 취소합니다.
 8. 함수 데이터 탐색기 모형 요약 .jmp 데이터 테이블에서 BatchID 를 마우스 오른쪽 버튼으로 클릭하고 **연결 참조 > Fermentation Process Batch Yield Results.jsp** 를 선택합니다.
- 이렇게 하면 yield 데이터 테이블과 요약 데이터 테이블이 결합됩니다.

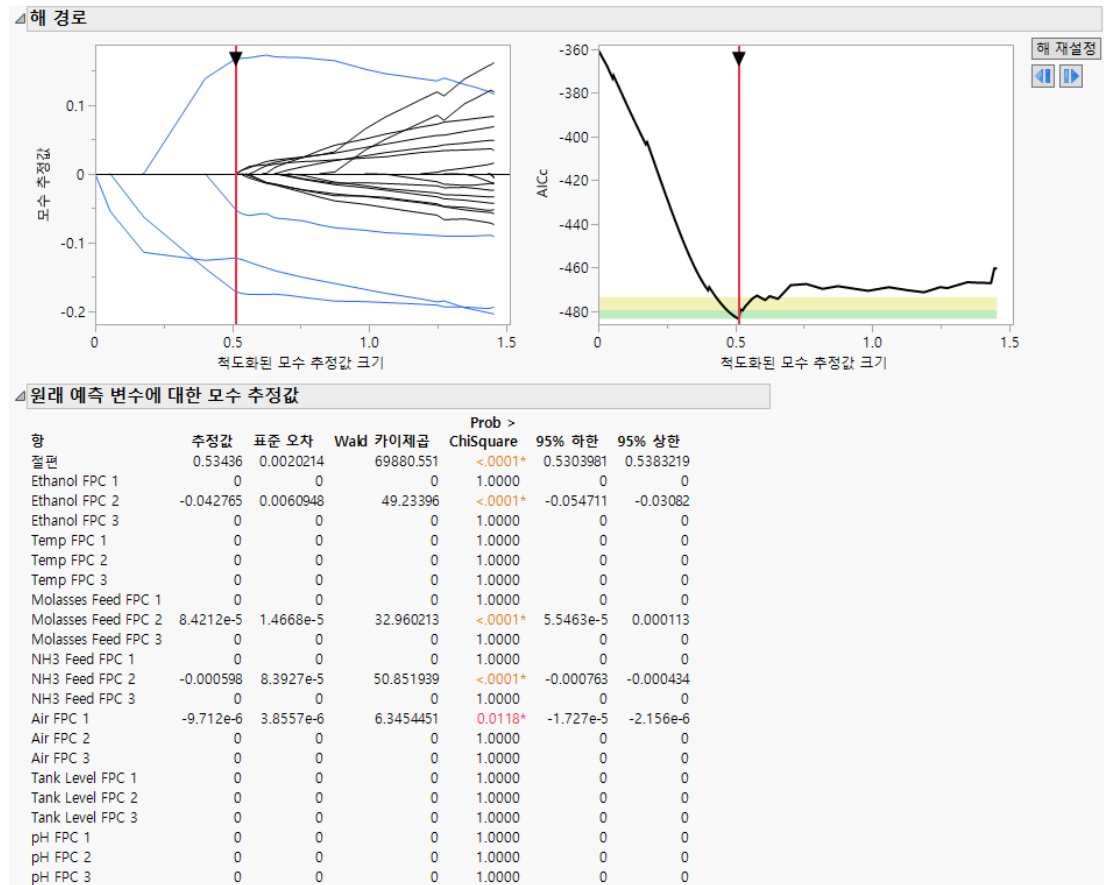
일반화 회귀 모형 적합

모형 적합 플랫폼의 일반화 회귀 분석법을 사용하여 함수 공정 변수가 Yield 에 미치는 영향을 확인합니다.

1. 함수 데이터 탐색기 모형 요약 .jmp 데이터 테이블에서 **분석 > 모형 적합**을 선택합니다.

2. "BatchID 에 의해 Fermentation Process Batch Yield Results 로 참조됨 " 옆의 삼각형을 클릭합니다.
3. Yield[BatchID] 를 선택하고 **Y** 를 클릭합니다.
4. Time 및 BatchID 를 제외한 나머지 열을 선택하고 **추가**를 클릭합니다.
5. " 분석법 " 을 " 일반화 회귀 " 로 변경합니다.
6. **실행**을 클릭합니다.
7. **적응형** 상자를 선택합니다.
8. **시작**을 클릭합니다.

그림 16.8 Batch Yield 에 대한 일반화 회귀 보고서



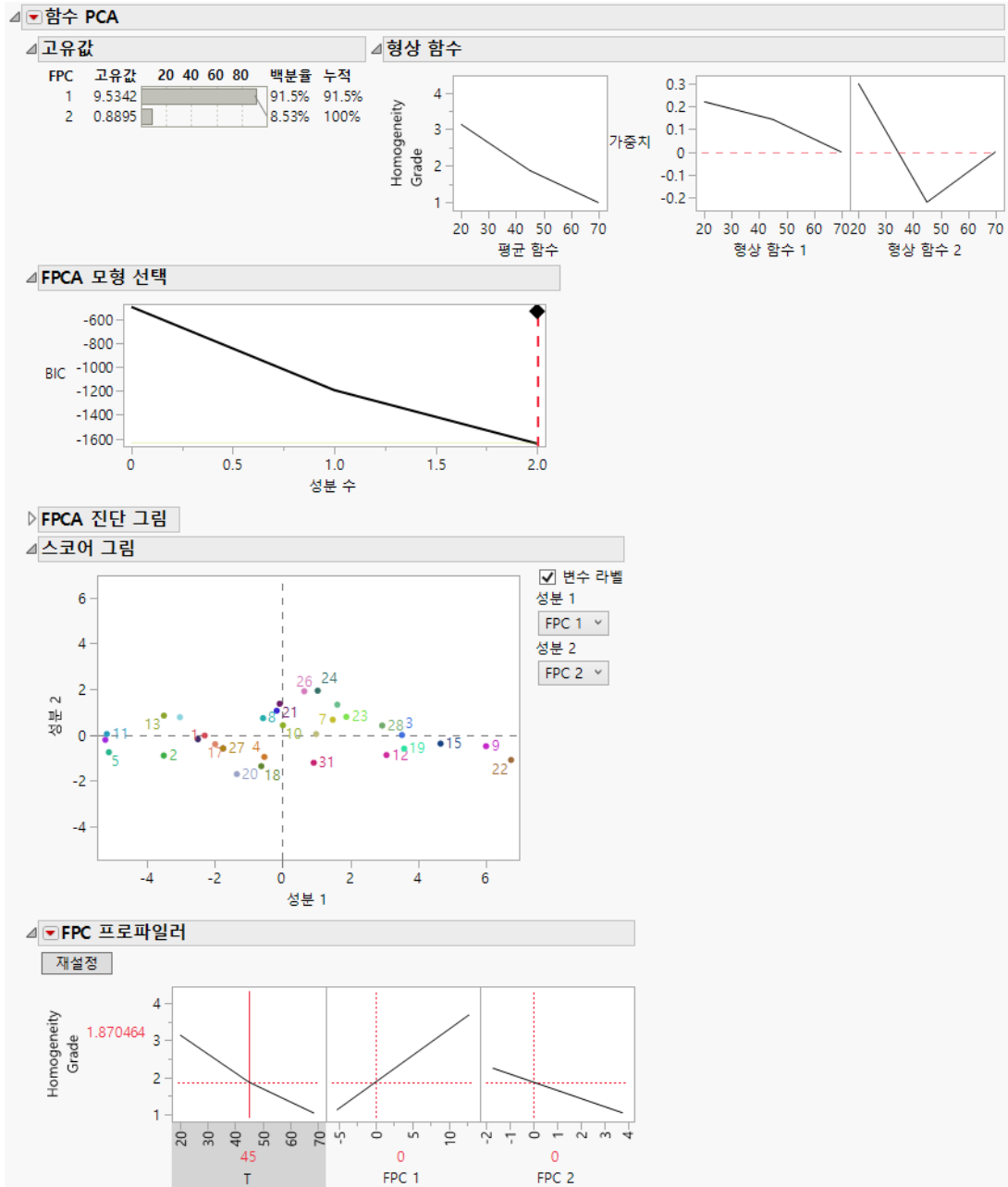
"일반화 회귀" 보고서에 따르면 Ethanol, Molasses Feed, NH3 Feed 및 Air의 특정 성분이 Yield에 유의한 영향을 줍니다. 모형의 R^2 은 0.73225입니다. 먼저 FDE를 사용하여 함수 공정에 대해 차원 축소를 수행하면 변수의 수를 크게 줄이는 동시에 적절한 예측 모형을 생성할 수 있습니다.

함수 DOE 분석의 예

이 예에서는 함수 데이터 탐색기를 사용하여 성분 조합이 다양한 온도에서 화합물의 동질성 등급에 어떤 영향을 미치는지 확인합니다. 화합물을 구성하는 세 가지 성분은 **Active**, **Water** 및 **Solvent** 라는 라벨이 지정된 기본 성분입니다. 32 개의 성분 조합이 있으며 이러한 조합은 **Formulation** 열의 값으로 표시됩니다. 각 배합 (**Formulation**) 에 대해 섭씨 20 도에서 70 도까지 5 도 간격으로 **Homogeneity Grade** 가 측정됩니다. 서로 다른 온도에 대한 이러한 측정값을 그림에 표시하여 각 배합에 대한 곡선을 만들 수 있습니다. 함수 데이터 탐색기 플랫폼을 사용하여 **Active**, **Water** 및 **Solvent** 값에 따라 곡선이 어떻게 변하는지 확인합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Functional Data/Formulation For Homogeneity DOE.jmp** 를 엽니다.
2. **분석 > 전문 모델링 > 함수 데이터 탐색기**를 선택합니다.
3. " 쌓인 데이터 형식 " 탭에서 **Homogeneity Grade** 를 선택하고 **Y, 출력**을 클릭합니다.
4. **T** 를 선택하고 **X, 입력**을 클릭합니다.
5. **Formulation** 을 선택하고 **ID, 함수**를 클릭합니다.
6. **Solvent**, **Active**, **Water** 를 선택하고 **Z, 보조**를 클릭합니다.
이러한 변수는 초기 함수 데이터 분석에 사용되지 않습니다. 그러나 여기서 보조 변수로 지정하면 나중에 모형을 적합시킬 때 사용할 수 있습니다.
7. **확인**을 클릭합니다.
8. " 함수 데이터 탐색기 " 의 빨간색 삼각형을 클릭하고 **모형 > B- 스플라인**을 선택합니다.
최량 적합 모형은 매듭이 하나인 선형 스플라인 모형입니다.

그림 16.9 Homogeneity Grade 에 대한 함수 PCA



" 함수 PCA" 보고서에서는 데이터의 변동을 설명하려면 두 개의 함수 주성분이 필요하다는 것을 보여 줍니다. "FPC 프로파일러"에서는 서로 다른 온도에서 FPC가 Homogeneity Grade

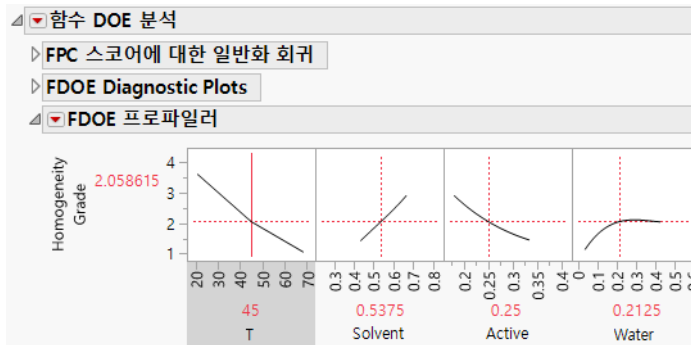
에 미치는 영향을 보여 줍니다. 그러나 실제 성분이 Homogeneity Grade 에 어떤 영향을 미치는지 알 수 있는 방법은 없습니다.

9. " 초기 데이터의 B- 스플라인 " 옆에 있는 빨간색 삼각형을 클릭하고 **함수 DOE 분석**을 선택합니다.

" 함수 DOE 분석 " 옵션은 세 개의 보조 변수를 예측 변수로 사용하여 일반화 회귀 모델을 각 주성분 스코어에 적합시킵니다. 데이터 테이블에 "Model" 스크립트가 포함되어 있으므로 적합한 모델은 스크립트에 정의된 모델을 기반으로 합니다.

참고 : 데이터 테이블에 "Model" 스크립트가 포함되어 있지 않으면 2 차 요인 모델이 적합됩니다.

그림 16.10 Homogeneity Grade 에 대한 FDOE 프로파일러



"FDOE 프로파일러" 를 사용하면 실제로 관심 있는 예측 변수에 따라 Homogeneity Grade 가 어떻게 변하는지 탐색할 수 있습니다. 전반적으로 Active 양이 증가하면 Homogeneity Grade 가 감소하는 것으로 보입니다.

참고 : 이 예에서는 Active, Solvent 및 Water 값의 합이 1 이 되도록 제한됩니다. 이 제약 조건은 프로파일러에 반영됩니다.

JMP PRO 함수 데이터 탐색기 플랫폼에 대한 통계 상세 정보

이 섹션에는 함수 데이터 탐색기 플랫폼에 대한 통계 상세 정보가 포함되어 있습니다.

- " 함수 모델 적합에 대한 통계 상세 정보 "
- " 함수 요약 보고서에 대한 통계 상세 정보 "

함수 모형 적합에 대한 통계 상세 정보

함수 데이터 탐색기 플랫폼의 모든 모형 적합은 기저 함수 확장에 의존합니다. 기저 함수는 독립 함수의 집합입니다. 함수 $f(t)$ 는 ϕ_k 로 표시되는 K 개 기저 함수의 선형 조합을 사용하여 근사화할 수 있습니다. 일반적으로 함수는 다음과 같이 근사화됩니다.

$$\hat{f}(t) = \sum_{k=1}^K c_k \phi_k(t)$$

여기서 c_k 모수는 기저 계수입니다. 평활 정도는 기저 함수의 수 K 에 의해 결정됩니다. 기저 함수 확장에 대한 자세한 내용은 Ramsay and Silverman(2005) 연구 자료에서 확인하십시오.

JMP PRO Fourier 기저 모형

Fourier 기저를 정의하는 $k=1, \dots, K$ 개 함수는 $\phi_0=1$, $\phi_{2k-1}=\sin(r\omega t)$ 및 $\phi_{2k}=\cos(r\omega t)$ 가 되도록 정의됩니다. 그러면 근사 함수가 다음과 같이 정의됩니다.

$$\hat{f}_i(t) = c_{i0} + c_{i1} \sin \omega t + c_{i2} \cos \omega t + c_{i3} \sin 2\omega t + c_{i4} \cos 2\omega t + \dots$$

이는 $A=2\pi/\omega$ 로 정의되는 주기 A 를 갖는 주기 함수에 근사한 값을 산출합니다. 계수는 고정 기저 함수 계수와 함수별 랜덤 계수의 조합입니다. 각 c 는 다음과 같이 정의됩니다.

$$c_{ik} = \beta_k + \alpha_{ik}$$

여기서 β_k 는 기저 함수 k 의 고정 계수이고, α_{ik} 는 특정 함수 공정 i 에 대한 기저 함수 k 의 랜덤 계수입니다. Fourier 기저 모형의 경우 β_k 및 α_{ik} 모수의 추정값은 각각 "기저 함수 계수" 테이블과 "함수별 랜덤 계수" 테이블에서 확인할 수 있습니다.

Fourier 기저 모형에는 하나의 절편 항과 Fourier 쌍이라고 하는 동일한 수의 사인 항 및 코사인 항이 있습니다. 따라서 K 는 항상 홀수입니다. 예를 들어 $K=7$ 이면 절편 항 하나와 세 개의 Fourier 쌍이 있습니다.

함수 요약 보고서에 대한 통계 상세 정보

함수 데이터 탐색기 플랫폼의 "함수 요약" 보고서에는 통합 차이, RISE(제곱근 적분 제곱 오차) 및 RIFS(제곱근 적분 함수 제곱) 요약 통계량이 포함됩니다. 이 섹션에서는 이러한 요약 통계량을 정의합니다. 이 섹션에 나오는 모든 방정식은 입력 값이 0~1 범위에 맞춰져 있다고 가정합니다. 모든 함수에 대해 다음 표기가 사용됩니다.

$\hat{f}_i(x)$ 입력 값 x 에서 곡선 i 에 대한 추정 함수

$\hat{m}(x)$ 입력 값 x 에서의 추정 평균 함수

통합 차이

통합 차이는 ID 특정 곡선의 점과 전체 평균 곡선의 점에 대한 평균 차이입니다. 이 값을 사용하면 ID 특정 곡선이 대체로 평균 곡선 위에 있는지 아니면 아래에 있는지 판별할 수 있습니다. 곡선 i 에 대한 통합 차이는 다음과 같이 정의됩니다.

$$ID_i = \int_0^1 (\hat{f}_i(x) - \hat{m}(x)) dx$$

RISE

RISE(제곱근 적분 제곱 오차)는 ID 특정 곡선과 전체 평균 곡선 사이의 평균 거리입니다. 이 값을 사용하면 전체 평균 곡선과 가장 유사하거나 가장 다른 ID 특정 곡선을 판별할 수 있습니다. 곡선 i 에 대한 RISE 요약 값은 다음과 같이 정의됩니다.

$$RISE_i = \sqrt{\int_0^1 (\hat{f}_i(x) - \hat{m}(x))^2 dx}$$

RIFS

RIFS(제곱근 적분 함수 제곱)는 최적 곡선 시나리오를 분석하기 위한 요약 값입니다. 최적 곡선을 미리 빼는 경우 가장 작은 RIFS 값을 사용하여 최적 곡선에 가장 가까운 함수를 판별할 수 있습니다. 곡선 i 에 대한 RIFS 요약 값은 다음과 같이 정의됩니다.

$$RIFS_i = \sqrt{\int_0^1 \hat{f}_i(x)^2 dx}$$

17 장

가우스 과정

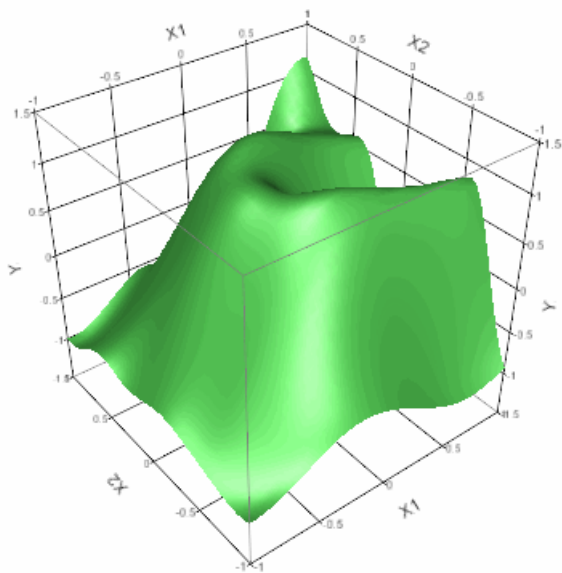
평활 모형을 사용하여 데이터 적합

가우스 과정 플랫폼을 사용하여 연속형 반응과 하나 이상의 예측 변수 간의 관계를 모델링할 수 있습니다. 이 유형의 모형은 유한 요소 코드의 출력과 같은 컴퓨터 시뮬레이션 실험에서 흔히 볼 수 있으며 데이터를 완벽하게 보간하는 경우가 많습니다. 가우스 과정은 오차 항이 없는 모형을 처리할 수 있습니다. 이러한 모형은 입력 값이 동일하면 항상 동일한 출력 값이 생성됩니다.

가우스 과정 플랫폼에서는 공간 상관 모형을 데이터에 적합시킵니다. 독립 변수의 값이 멀어질수록 두 관측값 사이의 반응 상관관계가 감소합니다.

이 플랫폼을 사용하는 한 가지 목적은 추가 분석 및 최적화에 사용할 수 있는 예측 계산식을 얻는 것입니다.

그림 17.1 가우스 과정 예측 표면 예제



목차

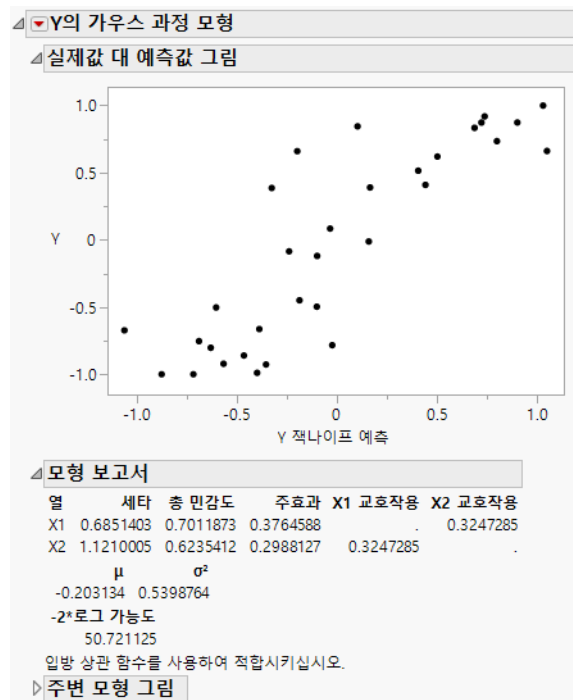
가우스 과정 플랫폼의 예.....	309
가우스 과정 플랫폼 시작.....	311
가우스 과정 보고서.....	312
실제값 대 예측값 그림.....	312
모형 보고서.....	312
주변 모형 그림.....	313
가우스 과정 플랫폼 옵션.....	313
가우스 과정 플랫폼의 추가 예.....	314
가우스 과정 모형의 예.....	314
범주형 예측 변수가 있는 가우스 과정 모형의 예.....	316
가우스 과정 플랫폼에 대한 통계 상세 정보.....	318
연속형 예측 변수가 있는 모형에 대한 통계 상세 정보.....	318
범주형 예측 변수가 있는 모형에 대한 통계 상세 정보.....	319
분산 계산식 파라미터화에 대한 통계 상세 정보.....	320
모형 적합에 대한 통계 상세 정보.....	320

가우스 과정 플랫폼의 예

이 예에서는 가우스 과정 플랫폼을 사용하여 반응에 대한 두 변수의 탐색적 검정력을 찾습니다. 반응에 대한 결정적 방정식과 함께 두 변수의 공간 채움 설계에서 가져온 데이터를 사용합니다. 열 계산식에서 반응에 대한 방정식을 볼 수 있습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 2D Gaussian Process Example.jmp 를 엽니다.
2. **분석 > 전문 모델링 > 가우스 과정**을 선택합니다.
3. X1 과 X2 를 선택하고 **X** 를 클릭합니다.
4. Y 를 선택하고 **Y** 를 클릭합니다.
5. **상관 유형 > 입방**을 선택합니다.
6. **JMP PRO 빠른 GASP** 를 선택 취소합니다.
7. **확인**을 클릭합니다.

그림 17.2 가우스 과정 보고서

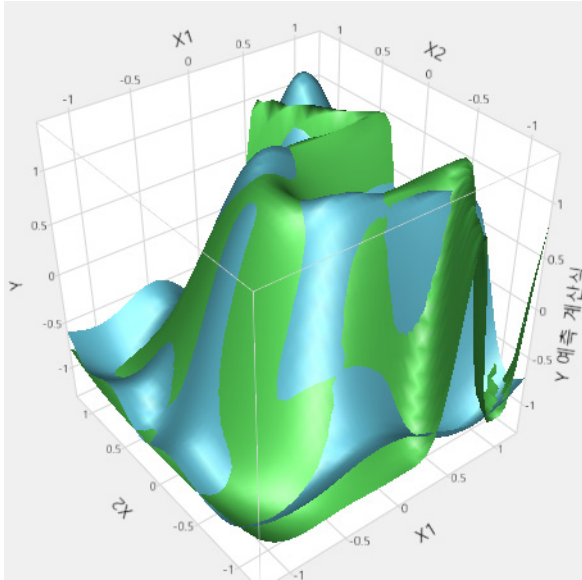


참고: 최소화 루틴의 다른 시작점, 상관 유형 선택 및 너겟 모수 포함으로 인해 추정된 모수가 다를 수 있습니다.

이제 적합된 표면을 원래 표면과 비교하여 시각화합니다.

8. "Y의 가우스 과정 모형" 옆의 빨간색 삼각형을 클릭하고 **예측 계산식 저장**을 선택합니다.
9. **그래프 > 표면 그림**을 선택합니다.
10. $X_1 \sim Y$ 예측 계산식을 선택하고 **열**을 클릭합니다.
11. **확인**을 클릭합니다.
12. "표면" 열에서 "Y 예측 계산식"에 대해 **양쪽**을 선택합니다.

그림 17.3 실제 Y와 예측 Y에 대한 3D 표면 그림



두 표면이 비슷합니다. 반응 Y에 대한 X_1 및 X_2 의 영향을 시각화할 수 있습니다. 그림을 회전하여 다른 각도에서 볼 수 있습니다. 반응에 대한 요인의 영향을 이해하는 데 사용할 수 있는 또 다른 도구에는 주변 그림이 있습니다.

가우스 과정 플랫폼 시작

분석 > 전문 모델링 > 가우스 과정을 선택하여 가우스 과정 플랫폼을 시작합니다.

그림 17.4 가우스 과정 시작 창

연속형 반응 변수와 하나 이상의 연속형 예측 변수 간의 관계를 보간을 사용한 스플라인으로 모델링합니다.

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

Y 분석할 연속형 열을 할당합니다.

X 설명 변수로 사용할 열을 할당합니다. "빠른 GASP" 옵션이 지정된 경우 JMP Pro 에서 범주형 변수를 사용할 수 있습니다.

너겟 모수 추정 추정 절차에 능형 회귀 모수를 도입합니다. 능형 회귀 모수는 반응에 잡음 또는 랜덤성이 있고 예측 모형으로 완벽한 보간 대신 잡음을 평활화하려는 경우에 유용합니다.

JMP PRO 빠른 GASP 빠른 GASP 알고리즘을 사용하기 위한 옵션입니다. 빠른 GASP 는 계산 속도를 높이기 위해 가우스 과정 모형을 작은 조각 (블록) 으로 분할합니다. 블록을 사용하면 여러 CPU 와 병렬 처리를 사용할 수 있습니다.

참고 : 관측값이 2,500 개가 넘으면 빠른 GASP 알고리즘이 필요합니다.

빠른 GASP 에 대한 자세한 내용은 Parker(2015) 연구 자료에서 확인하십시오.

상관 유형 모형의 상관 구조를 선택합니다. 이 플랫폼에서는 공간 상관 모형을 데이터에 적합시킵니다. 여기서는 독립 변수의 값이 멀어질수록 두 관측값 사이의 반응 상관이 감소합니다.

가우스 두 점 사이의 상관을 점 사이 거리에 관계없이 항상 0 이 아닌 것으로 제한합니다.

입방 충분히 떨어져 있는 두 점 사이의 상관이 0 이 되도록 허용합니다. 이 방법은 3 차 스플라인을 일반화한 것입니다.

JMP PRO 빠른 GASP 알고리즘은 3 차 상관 함수를 지원하지 않습니다.

최소 세타 값 적합 모형에 사용할 최소 세타 값을 설정합니다. 기본값은 0 입니다. 세타 값은 정규 회귀 모형의 기울기 모수와 유사합니다. 세타 값이 작으면 변수가 예측값에 거의 영향을 미치지 않음을 나타냅니다.

JMP PRO 블록 크기 빠른 GASP 알고리즘에 사용되는 각 계산 블록의 관측값 수입입니다. 블록당 25 개 이상의 관측값이 있어야 하고 데이터 집합의 최대 행 수는 2,500 개입니다.

가우스 과정 보고서

가우스 과정 보고서에는 실제값 대 예측값 그림, 모형 보고서 및 주변 모형 그림이 표시됩니다.

- "실제값 대 예측값 그림"
- "모형 보고서"
- "주변 모형 그림"

실제값 대 예측값 그림

"가우스 과정" 보고서의 "실제값 대 예측값 그림"에서는 Y 축에 실제 Y 값이 표시되고 X 축에 잭나이프 예측값이 표시됩니다. 적합도의 한 가지 측도는 점 배치가 그림의 대각선($Y = X$)을 얼마나 잘 따르는지 살펴보는 것입니다.

잭나이프 값은 제외된 각 Y에 대해 모형이 관련 행으로 재적합되지 않는다는 점에서 정확한 잭나이프 값이 아닙니다. 오히려 관련된 각 Y에 대한 예측 모형에서 행이 제외되지만 상관 모수는 행의 기여도를 유지합니다. 데이터를 완벽하게 보간하는 가우스 과정의 경우 이 잭나이프 절차는 입력과 동일하지 않은 예측을 제공합니다.

모형 보고서

가우스 과정 플랫폼의 "모형 보고서"에서는 모형 모수 추정값에 대한 함수 ANOVA 테이블을 보여 줍니다. 구체적으로 말하면 함수 기반 방법을 사용하여 변동을 계산하는 분산 분석 테이블입니다.

세타 가우스 과정 모형 모수 추정값입니다. 자세한 내용은 "[가우스 과정 플랫폼에 대한 통계 상세 정보](#)"에서 확인하십시오.

총 민감도 각 요인에 대한 주효과 및 모든 교호작용 항의 합입니다. 이 값은 한 요인과 해당 요인의 모든 이원 교호작용이 반응 변수에 미치는 영향을 나타내는 측도입니다.

총 변동은 전체 실험 공간에 대한 통합 변동입니다.

주효과 각 요인의 함수 주효과는 해당 요인 단독에 의한 통합 총 변동입니다. 주효과는 모형의 각 요인에 대한 함수 효과와 총 변동의 비율입니다.

교호작용 함수 교호작용 효과는 주효과와 비슷한 방식으로 계산됩니다.

JMP PRO **범주형 입력** 모형에 범주형 요인이 포함된 경우 각 범주형 요인에 대한 상관 행렬이 제공됩니다. 비대각 항목은 가우스 과정 모형 모수 추정값에 해당합니다. 자세한 내용은 "**범주형 예측 변수가 있는 모형에 대한 통계 상세 정보**"에서 확인하십시오.

평균과 시그마² 평균 및 분산 모형 모수입니다.

너겟 추정된 너겟 값입니다. 가우스 과정 시작 창에서 **너겟 모수 추정**을 선택한 경우 너겟 값이 보고됩니다. JMP에서 특이 공분산 행렬을 방지하기 위해 너겟 모수를 추가한 경우에도 너겟 값이 보고됩니다.

-2* 로그 가능성도 음의 로그 가능성도 함수의 두 배에 대한 추정값입니다. 자세한 내용은 **선형 모형 적합**에서 확인하십시오.

주변 모형 그림

가우스 과정 보고서에서 모형의 각 요인에 대한 주변 그림이 나타납니다. 각 그림마다 하나를 제외한 모든 요인은 데이터에서 다른 요인의 범위에 대해 균등한 분포를 사용하여 통합됩니다. 나머지 관심 요인에 대한 주변 반응 예측은 파란색으로 표시됩니다. 주변 모형이 데이터를 얼마나 잘 적합시키는지 나타내기 위해 관측된 반응 값에 대한 데이터 점이 포함됩니다. 요인 수가 적은 모형의 경우 이러한 그림을 진단 그림으로 사용할 수 있습니다.

가우스 과정 플랫폼 옵션

"가우스 과정"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

프로파일러 표준 프로파일러를 엽니다. 자세한 내용은 **Profilers**에서 확인하십시오.

등고선 프로파일러 등고선 프로파일러를 엽니다. 자세한 내용은 **Profilers**에서 확인하십시오.

표면 프로파일러 표면 프로파일러를 엽니다. 자세한 내용은 **Profilers**에서 확인하십시오.

예측 계산식 저장 활성 데이터 테이블에 새 예측 계산식 열을 생성합니다.

분산 계산식 저장 활성 데이터 테이블에 새 분산 계산식 열을 생성합니다.

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 "**계산식 저장소**"에서 확인하십시오.

JMP PRO **분산 계산식 게시** 분산 계산식을 생성하여 계산식 저장소 플랫폼에 계산식 열 스크립트로 저장합니다. "계산식 저장소" 보고서가 열려 있지 않으면 "계산식 저장소" 보고서가 생성됩니다. 자세한 내용은 "**계산식 저장소**"에서 확인하십시오.

잭나이프 예측값 저장 잭나이프 예측값을 활성 데이터 테이블에 저장합니다. 이 값은 실제값 대 예측값 그림의 X 축 값입니다.

다음 옵션에 대한 자세한 내용은 **JMP 사용**의에서 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.

가우스 과정 플랫폼의 추가 예

이 섹션에는 가우스 과정 플랫폼을 사용하는 예가 포함되어 있습니다.

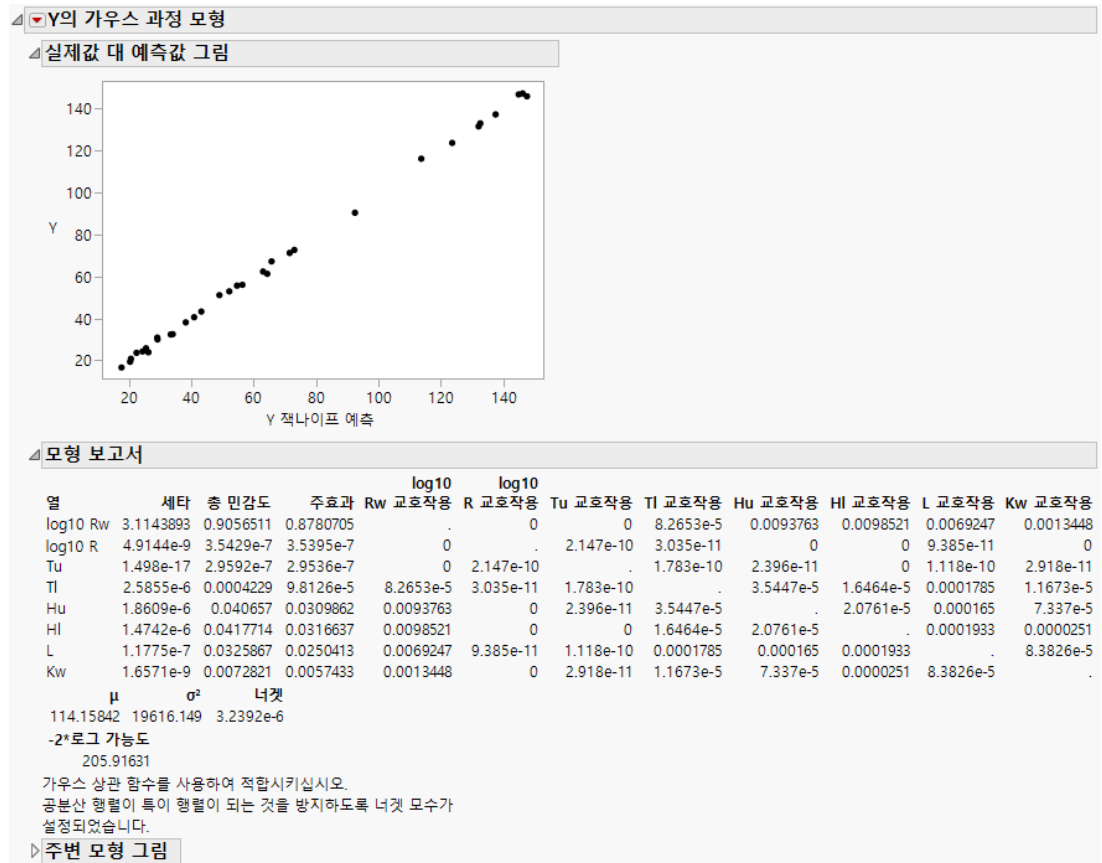
- "가우스 과정 모형의 예"
- "범주형 예측 변수가 있는 가우스 과정 모형의 예"

가우스 과정 모형의 예

지정된 공학 모형이 주어지면 가우스 과정 모형을 사용하여 모형에 포함된 요인이 반응에 미치는 영향을 이해할 수 있습니다. 이 데이터는 지표면에서 두 개의 대수층을 통과하여 뚫은 보어홀을 통한 물의 흐름을 보여 줍니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Design Experiment/Borehole Latin Hypercube.jmp 를 엽니다.
2. **분석 > 전문 모델링 > 가우스 과정**을 선택합니다.
3. $\log_{10} R_w \sim K_w$ 를 선택하고 **X** 를 클릭합니다.
4. **Y** 를 선택하고 **Y** 를 클릭합니다.
5.  JMP Pro 에서 분석을 더 빠르게 실행하려면 "빠른 GASP" 를 선택된 상태로 둡니다.
6. **확인**을 클릭합니다.

그림 17.5 Borehole Latin Hypercube 보고서

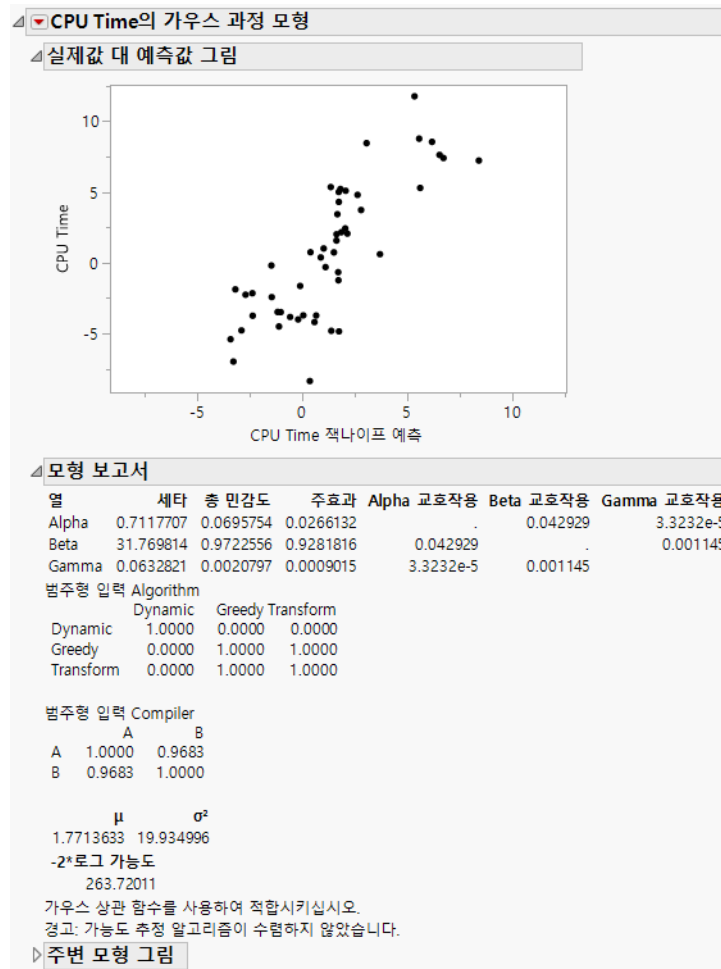


실제값 대 예측값 그림의 데이터가 $Y=X$ 선을 따라 분포되어 있으므로 가우스 과정 예측 모형이 실제 함수와 유사한 근사값을 산출한다는 것을 나타냅니다. "모형 보고서"에서 첫 번째 요인 $\log_{10} R_w$ 의 총 민감도가 가장 높음을 확인할 수 있습니다. $\log_{10} R_w$ 의 추정된 총 민감도는 반응의 변동을 90% 넘게 설명합니다. 세타 값이 작은 요인은 예측 계산식에 거의 또는 전혀 영향을 미치지 않습니다. 프로파일러를 사용하여 민감도를 시각화할 수 있습니다.

- "Y의 가우스 과정 모형"의 빨간색 삼각형을 클릭하고 **프로파일러**를 선택합니다.
반응 Y에 "반응 한계" 열 특성이 있으므로 만족도 함수가 자동으로 나타납니다.
- "예측 프로파일러"의 빨간색 삼각형을 클릭하고 **최적화 및 만족도 > 만족도 최대화**를 선택합니다.

참고: 범주형 요인이 포함된 모형에는 "빠른 GASP" 옵션을 사용해야 합니다. 자세한 내용은 "범주형 예측 변수가 있는 모형에 대한 통계 상세 정보"에서 확인하십시오.

그림 17.7 Algorithm Data 보고서



" 실제값 대 예측값 그림 "에서는 실제 CPU 시간과 예측 CPU 시간 사이의 강한 상관관계를 보여 줍니다. 이는 가우스 과정 예측 모형이 실제 함수와 유사한 근사값을 산출한다는 것을 나타냅니다. " 모형 보고서 "에서 Beta 예측 변수의 총 민감도가 가장 높습니다. 이는 연속형 예측 변수 중 Beta가 반응의 변동을 가장 많이 설명한다는 것을 나타냅니다. 각 범주형 예측 변수, 즉 Algorithm과 Compiler에 대한 별도의 " 범주형 입력 " 행렬이 있습니다. 이러한 행렬은 상관 행렬이며 각 범주형 예측 변수에 대한 수준 사이의 상관관계를 나타냅니다. 행렬의 비대각은 τ 모수입니다.

가우스 과정 플랫폼에 대한 통계 상세 정보

이 섹션에는 가우스 과정 플랫폼에 대한 통계 상세 정보가 포함되어 있습니다.

- "연속형 예측 변수가 있는 모형에 대한 통계 상세 정보"
- "범주형 예측 변수가 있는 모형에 대한 통계 상세 정보"
- "분산 계산식 파라미터화에 대한 통계 상세 정보"
- "모형 적합에 대한 통계 상세 정보"

연속형 예측 변수가 있는 모형에 대한 통계 상세 정보

가우스 과정 모형에 연속형 예측 변수만 포함된 경우 가우스 과정 플랫폼에서는 "가우스"와 "입방"이라는 두 개의 상관 구조를 구현합니다.

가우스 상관 구조는 맥이 2 인 곱 지수 상관 함수를 추정 모형으로 사용합니다. 이 모형은 Y 가 평균 μ 와 공분산 행렬 $\sigma^2 \mathbf{R}$ 의 정규 분포를 따른다고 가정합니다. $\mathbf{R}$ 행렬의 요소는 다음과 같이 정의됩니다.

$$r_{ij} = \exp \left(- \sum_{k=1}^K \theta_k (x_{ik} - x_{jk})^2 \right)$$

다음은 각 요소에 대한 설명입니다.

K = 연속형 예측 변수의 수

θ_k = k 번째 예측 변수의 세타 모수

x_{ik} = 개체 i 에 대한 k 번째 예측 변수의 값

x_{jk} = 개체 j 에 대한 k 번째 예측 변수의 값

입방 상관 구조도 Y 가 평균 μ 와 공분산 행렬 $\sigma^2 \mathbf{R}$ 의 정규 분포를 따른다고 가정합니다. $\mathbf{R}$ 행렬은 다음 요소로 구성됩니다.

$$r_{ij} = \prod_k \rho(d; \theta_k)$$

다음은 각 요소에 대한 설명입니다.

$$d = x_{ik} - x_{jk}$$

$$\rho(d;\theta) = \begin{cases} 1 - 6(d\theta)^2 + 6(|d|\theta)^3, & |d| \leq \frac{1}{2\theta} \\ 2(1 - |d|\theta)^3, & \frac{1}{2\theta} < |d| \leq \frac{1}{\theta} \\ 0, & \frac{1}{\theta} < |d| \end{cases}$$

자세한 내용은 [Santer\(2003\)](#) 연구 자료에서 확인하십시오. 입방 상관 구조에 사용되는 세타 모수는 문헌에 자주 사용되는 모수의 역수입니다. 세타가 모형에 영향을 미치지 않을 때 역수가 사용되어 rho의 값이 무한대가 아니라 0이 됩니다.

JMP PRO 범주형 예측 변수가 있는 모형에 대한 통계 상세 정보

가우스 과정 모형에 범주형 예측 변수가 포함된 경우 가우스 상관 구조가 사용됩니다. **R** 행렬의 요소는 다음과 같이 정의됩니다.

$$r_{ij} = \left(\prod_{p=1}^P \tau_{p_{ij}} \right) \exp \left(- \sum_{k=1}^K \theta_k (x_{ik} - x_{jk})^2 \right)$$

다음은 각 요소에 대한 설명입니다.

K = 연속형 예측 변수의 수

P = 범주형 예측 변수의 수

θ_k = k 번째 연속형 예측 변수의 세타 모수

x_{ik} = 개체 i 에 대한 k 번째 연속형 예측 변수의 값

x_{jk} = 개체 j 에 대한 k 번째 연속형 예측 변수의 값

$\tau_{p_{ij}}$ = 개체 i 에 대한 예측 변수 p 의 관측된 수준과 개체 j 에 대한 예측 변수 p 의 관측된 수준 사이의 상관

범주형 변수의 각 수준 조합에 대한 τ 모수가 있습니다. 여기서 τ_{ij} 는 개체 i 와 개체 j 의 관측된 수준에 의해 구성된 고유 조합에 해당합니다. 따라서 공분산 요소 r_{ij} 는 i 번째 관측값과 j 번째 관측값에서 얻은 범주형 예측 변수의 수준 조합에 따라 달라집니다. 자세한 내용은 [Qian et al. \(2012\)](#) 연구 자료에서 확인하십시오.

분산 계산식 파라미터화에 대한 통계 상세 정보

가우스 과정 플랫폼의 저장된 분산 계산식은 모형에 범주형 예측 변수가 포함된 경우를 제외하고 이전에 정의된 **R** 파라미터화를 사용합니다. 가우스 과정 모형에 범주형 예측 변수가 포함된 경우에는 저장된 분산 계산식에 다음과 같은 **R** 파라미터화가 사용됩니다.

$$r_{ij} = \exp \left(- \sum_{k=1}^K \theta_k (x_{ik} - x_{jk})^2 - \sum_{p=1}^P \phi_{p_{ij}} \right)$$

여기서 $\phi_{p_{ij}} = -\ln(\tau_{p_{ij}})$ 및 다른 모든 변수는 이전에 정의된 대로입니다.

모형 적합에 대한 통계 상세 정보

가우스 과정 모형의 모형 모수는 최대 가능도를 통해 적합됩니다. 적합된 모수는 플랫폼 보고서에 제공됩니다. 모수는 다음과 같습니다.

- μ = 가우스 과정 평균
- σ^2 = 가우스 과정 분산
- 세타는 **R** 정의에서 θ_k 값에 해당합니다.
- 범주형 입력 상관 행렬의 비대각은 **R** 정의에서 $\tau_{p_{ij}}$ 값에 해당합니다.

참고 : JMP 에서 분산 행렬을 반전할 수 있도록 능형 회귀 모수를 추가한 경우 보고서에 **공분산 행렬이 특이 행렬이 되는 것을 방지하도록 너겟 모수가 설정되었습니다.** 라는 노트가 포함됩니다.

시계열 분석

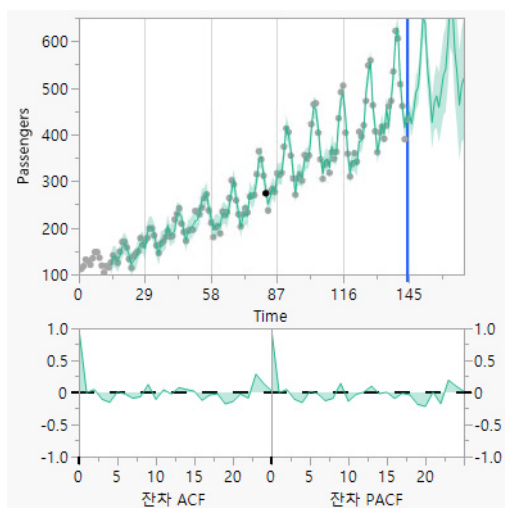
시계열 모형 및 전이 함수 적합

시계열 플랫폼을 사용하면 단변량 시계열을 탐색, 분석 및 예측할 수 있습니다. 시계열은 일정 간격의 일련의 기간 동안 얻은 관측값 집합입니다. 일반적으로 시간상 서로 가까운 관측값은 상관되어 있습니다. 시계열 방법론은 관측값 간의 이러한 종속성을 활용하여 미래에 시계열이 어떤 형태가 될지 더 잘 예측할 수 있습니다.

시계열 데이터의 공통 특성에는 계절성, 추세 및 자기상관이 포함됩니다. 시계열 플랫폼에서는 이러한 특성을 처리할 수 있는 옵션을 제공합니다. 변동도, 자기상관 그림, 편자기상관 그림 및 스펙트럼 밀도 그림과 같은 그래프를 사용하여 시계열을 설명하고 예측하는 데 적절한 모형 유형을 식별할 수 있습니다. 또한 이 플랫폼에는 데이터의 계절 또는 일반적 추세를 제거하여 분석을 단순화할 수 있는 몇 가지 분해 방법이 있습니다. 아니면 계절성과 장기 추세를 하나의 모형에 모두 포함할 수 있는 더 정교한 ARIMA 모형 및 상태 공간 평활 모형을 적합시킬 수 있습니다. Box-Cox 변환을 수행하고 변환된 계열을 분석 및 모델링할 수도 있습니다.

모형의 예측 성능을 평가하는 몇 가지 방법이 있습니다. "홀드백에 대한 예측" 기능은 시계열을 모형 생성을 위한 훈련 부분과 예측 성능 평가를 위한 홀드백 부분으로 분할합니다.

그림 18.1 예측 그림



목차

시계열 플랫폼 개요	323
시계열 플랫폼의 예	323
시계열 플랫폼 시작	326
시계열 보고서	327
시계열 그래프	328
시계열 기본 진단 차트	328
시계열 플랫폼 옵션	330
시계열 진단	330
차분 및 분해	331
시계열 모형	333
추가 옵션	340
추가 시계열 보고서	341
차분 보고서	341
분해 보고서	342
모형 비교 보고서	344
모형 보고서	346
상태 공간 평활 보고서	350
전이 함수 모형 보고서	352
스펙트럼 밀도 보고서	353
시계열 플랫폼의 추가 예	354
시간 ID 열 생성의 예	354
Box-Cox 변환의 예	359
홀드백 집합을 사용한 예	361
시계열 플랫폼에 대한 통계 상세 정보	363
스펙트럼 밀도에 대한 통계 상세 정보	363
X-11 분해에 대한 통계 상세 정보	364
지수 평활 모형에 대한 통계 상세 정보	364
ARIMA 모형에 대한 통계 상세 정보	368
전이 함수에 대한 통계 상세 정보	369

시계열 플랫폼 개요

시계열은 일정 간격의 일련의 기간 동안 얻은 관측값 집합입니다($y_1, y_2, \dots, y_N$). 예를 들어 분기별 판매 보고서, 월별 평균 온도, 흑점 개수 등이 시계열 데이터에 포함됩니다. 시계열 플랫폼을 사용하면 이러한 유형의 데이터에서 발견되는 패턴과 추세를 탐색할 수 있습니다. 그런 다음 이러한 패턴과 추세를 사용하여 미래의 결과를 예측할 수 있습니다.

시계열 데이터의 공통 특성에는 계절성, 추세 및 자기상관이 포함됩니다. **계절성**은 알려진 기간 동안 나타나는 패턴을 말합니다. 예를 들어 월별로 수집되는 데이터는 모든 데이터 수집 연도의 여름철에 유사한 형태를 보일 수 있습니다. **추세**는 시간에 따른 값의 점진적 증가 또는 감소와 같이 계열의 장기적인 움직임을 말합니다. **자기상관**은 계열의 각 점이 계열의 이전 값과 상관되는 정도를 나타냅니다.

시계열 플랫폼에는 다양한 모형과 예측 방법이 있습니다. 그러나 일부 방법은 추세 또는 계절성을 처리할 수 없습니다. 적절한 모형을 선택하려면 계열에 어떤 특성이 있는지 확인하는 것이 중요합니다. 시계열 플랫폼에서는 시계열의 진화를 설명하고 예측하는 데 적절한 모형 유형을 식별할 수 있는 변동도, 자기상관 그림, 편자기상관 그림 및 스펙트럼 밀도 그림과 같은 그래프를 제공합니다. 또한 이 플랫폼에는 데이터의 계절 또는 일반적 추세를 제거하여 분석을 탐색하고 단순화할 수 있는 몇 가지 차분 및 분해 방법이 있습니다. Box-Cox 변환을 보고 데이터에 적용할 수도 있습니다.

아니면 계절성과 장기 추세를 포함할 수 있는 더 정교한 모형을 적합시킬 수 있습니다. 플랫폼에서 이 기능을 가진 모형 중 하나는 고급 지수 평활 모형인 **Winter**의 가법 방법입니다. 또한 **ARIMA**(AutoRegressive Integrated Moving Average) 모형과 상태 공간 평활 모형을 적합시킬 수 있습니다. 이 두 가지 유형의 모형은 통계적으로 가장 복잡하지만 유연성도 가장 높습니다. 고급 지수 평활, **ARIMA** 및 상태 공간 평활 모형은 해석하기 어렵지만 예측에 탁월한 도구입니다.

시계열 플랫폼에서는 입력 계열과 함께 제공될 경우 전이 함수 모형도 적합시킬 수 있습니다.

시계열 플랫폼의 예

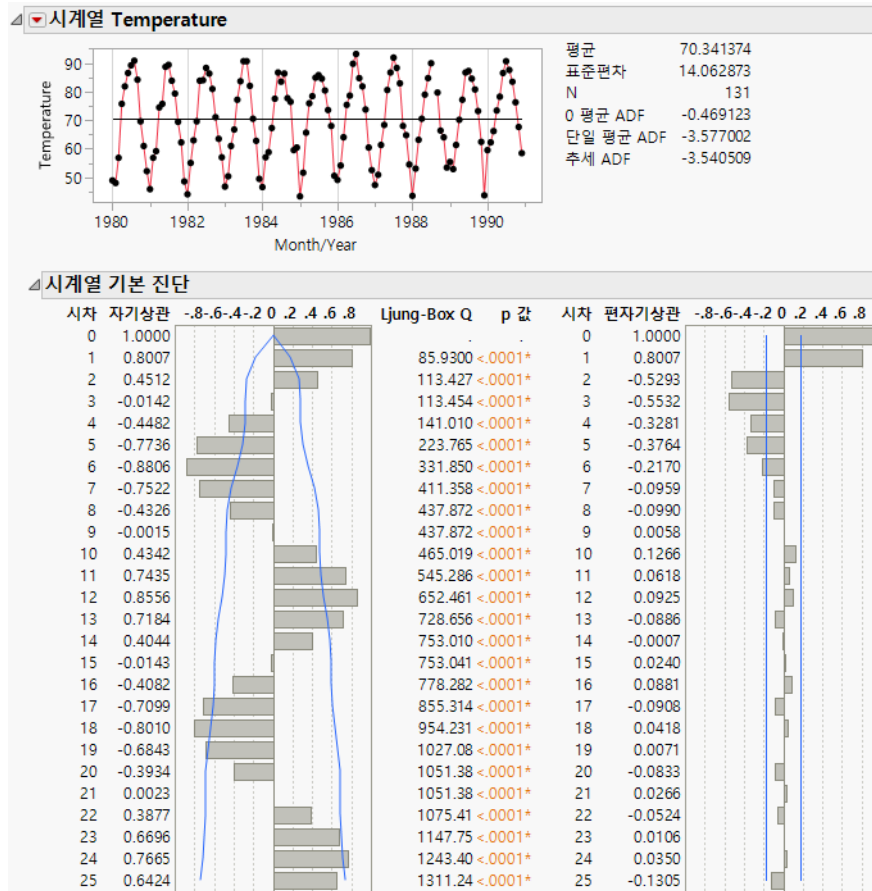
시계열 플랫폼을 사용하여 일련의 최대 월별 온도를 검토하고 향후 2 년간의 최대 월별 온도를 예측합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Time Series/Raleigh Temps.jmp 를 엽니다.
2. **분석 > 전문 모델링 > 시계열**을 선택합니다.
3. **Temperature** 를 선택하고 **Y, 시계열**을 클릭합니다.
4. **Month/Year** 를 선택하고 **X, 시간 ID** 를 클릭합니다.
5. **예측 주기** 옆의 상자에 "24" 를 입력합니다.

이 값은 데이터에 적합된 모형을 사용하여 예측되는 미래 기간의 수입니다. 다음 2 년, 즉 24 개월 동안의 월별 온도를 예측하려고 합니다.

6. **확인**을 클릭합니다.

그림 18.2 Raleigh Temps.jmp 에 대한 시계열 분석 보고서



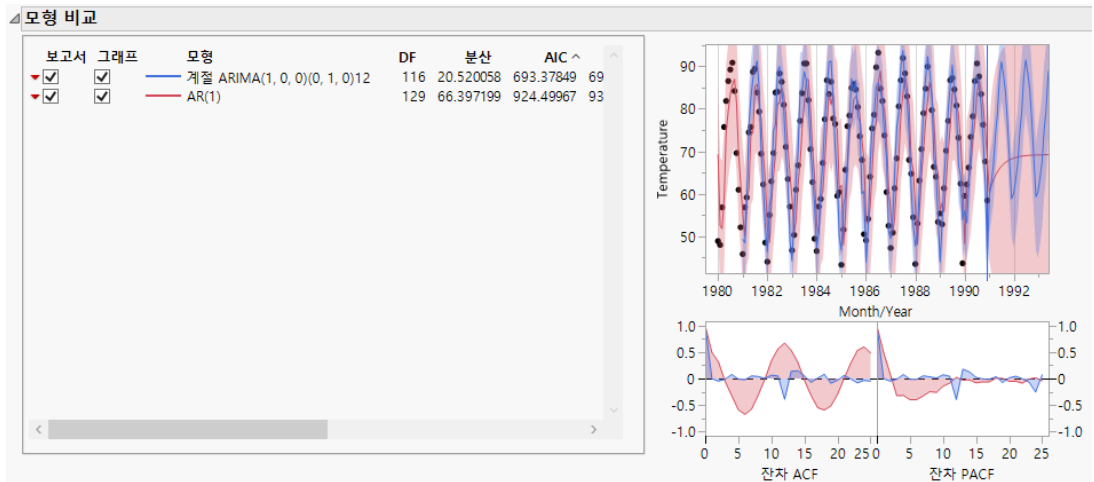
시계열 그래프에서는 계절이 순환적임을 보여 줍니다. 이 순환 성분은 자기상관 차트에서도 분명하게 확인할 수 있습니다. 1 시차의 점은 양의 상관관계가 있습니다 (" 자기상관 " 값 = 0.8007). 점이 멀어질수록 음의 상관관계가 된 후 다시 양의 상관관계를 보이고 패턴이 반복 됩니다. 시계열 그래프와 자기상관 차트는 시계열에 계절성 특성이 있다는 증거를 제공합니다.

7. " 시계열 " 의 빨간색 삼각형을 클릭하고 **ARIMA** 를 선택합니다.
8. 계절에 자기상관 증거가 보였으므로 " p , 자기회귀 차수 " 를 "1" 로 설정합니다.
9. **추정** 을 클릭합니다.
10. " 시계열 " 의 빨간색 삼각형을 클릭하고 **계절 ARIMA** 를 선택합니다.
11. 계절에 자기상관 증거가 보였으므로 "ARIMA" 상자에서 " p , 자기회귀 차수 " 를 "1" 로 설정합니다.
12. 계절에 계절성 증거가 보였으므로 "계절 ARIMA" 상자에서 " D , 차분 차수 " 를 "1" 로 설정합니다.

13. **추정**을 클릭합니다.

14. 모형 비교 테이블에서 두 모형에 대한 "그래프" 아래의 상자를 선택합니다.

그림 18.3 Raleigh Temps.jmp 에 대한 모형 비교 테이블



모형 비교 테이블은 AIC 통계량에 따라 내림차순으로 정렬됩니다. 즉, 최량 적합 모형이 보고서 맨 위에 나타납니다. 계절 ARIMA 모형의 AIC 값 (693.4) 이 정규 ARIMA 모형의 AIC 값 (924.5) 보다 훨씬 작습니다. 그래프에서는 ARIMA 모형이 관측된 점을 비교적 잘 예측하지만 잔차가 계절 ARIMA 모형의 잔차보다 크다는 것을 보여 줍니다. 또한 계절 ARIMA 모형의 예측 구간이 더 좁으므로 미래 관측값에 대해 보다 현실적인 예측을 제공합니다. 계절 이 계절 성분에 대한 증거를 보여 주었으므로 이러한 결과가 타당합니다.

시계열 플랫폼 시작

분석 > 전문 모델링 > 시계열을 선택하여 시계열 플랫폼을 시작합니다.

그림 18.4 시계열 시작 창

일정 간격의 시점에서 얻은 일련의 관측값을 모델링합니다.

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

시계열 플랫폼 시작 창에는 다음 옵션이 포함되어 있습니다.

Y, 시계열 하나 이상의 열을 시계열 변수로 할당합니다. Y 축에 표시됩니다.

입력 목록 하나 이상의 열을 입력 계열 변수로 할당합니다. "입력 시계열 패널"에 표시되고 전이 함수 모형에 사용됩니다. 입력 계열 변수는 시계열 또는 표시자로 숫자여야 합니다.

X, 시간 ID 시간 축(X 축)에 라벨을 지정하기 위한 하나의 변수를 할당합니다. "시간 ID"에 변수를 지정하지 않으면 행 번호가 대신 사용됩니다.

참고 : "X, 시간 ID" 변수를 사용하는 경우 "시간 주기" 열 특성을 사용하여 시간 주기를 지정할 수 있습니다. "매년", "매분기", "매월", "매주", "매일", "매시간", "매분" 및 "매초" 중에서 선택할 수 있습니다. 이렇게 하면 JMP에서 예측 값을 그림에 표시할 때 데이터 간격을 결정하는 데 도움이 됩니다. 주기를 지정하지 않으면 데이터가 일정한 간격의 수치형 데이터로 간주됩니다.

주의 : X, 시간 ID에 할당된 변수의 관측값은 간격이 일정하다고 가정합니다. 그러나 시계열 플랫폼에서는 타임스탬프가 증가하는지 여부만 확인합니다. 관측값의 간격이 일정하지 여부는 확인하지 않습니다.

기준 변수의 각 수준에 대한 개별 분석으로 구성된 보고서를 생성하는 열을 할당합니다. 기준 변수가 둘 이상 할당되면 기준 변수의 가능한 각 수준 조합에 대해 개별 분석이 생성됩니다.

참고 : " 기준 " 변수를 사용하는 경우 " 기준 " 변수의 각 수준에 대한 관측값 수에 따라 자기 상관 시차 수를 변경해야 할 수도 있습니다 . 시차 수는 1 보다 크고 수준당 관측값 수보다 작아야 합니다 .

자기상관 시차 자기상관 및 부분 상관을 계산하는 데 사용할 시차 수를 지정합니다 . 이 값은 이러한 상관을 계산하는 데 사용된 점 사이의 최대 주기 수입니다 . 1 보다 크고 행 수보다 작아야 합니다 . 기본 시차 수는 25 입니다 .

팁 : 최대 시차 수에 일반적으로 사용되는 규칙은 $n/4$ 이며 , 여기서 n 은 관측값 수입니다 .

예측 주기 데이터에 적합된 각 모형을 사용하여 예측되는 관측값 수를 지정합니다 . 기본 예측 수는 25 입니다 .

- " 홀드백에 대한 예측 " 옵션을 선택하는 경우 " 예측 주기 " 값은 모형 적합에 포함되지 않은 계열 끝의 관측값 수를 지정합니다 . 그러면 적합 모형을 사용하여 이러한 값을 예측합니다 .
- " 홀드백에 대한 예측 " 옵션을 선택하지 않는 경우 " 예측 주기 " 값은 계열이 끝난 후 예측되는 미래 관측값 수를 지정합니다 .

홀드백에 대한 예측 예측이 미래 관측값에 대해 수행되는지 아니면 홀드백 관측값에 대해 수행되는지 결정합니다 . " 홀드백에 대한 예측 " 옵션을 선택하면 " 예측 주기 " 옵션에 지정된 수에 따라 결정된 홀드백 집합에서 예측이 수행됩니다 .

Box-Cox 변환 사용 "Box-Cox 에 대한 람다" 옵션에 지정된 람다로 Box-Cox 변환을 사용하여 원래 데이터를 변환합니다 . 이 옵션을 선택하면 변환된 데이터에서 시계열 보고서의 모든 분석이 수행됩니다 . 그러나 모든 예측은 다시 변환되어 원래 척도로 보고됩니다 .

Box-Cox 에 대한 람다 ("Box-Cox 변환 사용 " 을 선택한 경우에만 사용 가능) 원래 데이터의 Box-Cox 변환에 사용되는 람다 모수를 지정합니다 .

참고 : 시계열 플랫폼은 시계열의 순서와 간격을 유지하기 위해 데이터 테이블의 제외된 행을 결측값으로 처리합니다 .

시계열 보고서

초기 시계열 보고서에는 시계열 그래프 , 시계열 변수에 대한 요약 통계량 및 검정 , 기본 진단 보고서가 표시됩니다 . 시작 창에서 " 입력 목록 " 에 열을 지정한 경우 " 전이 함수 분석 " 및 " 입력 시계열 패널 " 보고서가 표시됩니다 . " 전이 함수 분석 " 보고서와 " 입력 시계열 패널 " 보고서에는 시계열 보고서와 동일한 초기 정보가 포함됩니다 .

- “시계열 그래프”
- “시계열 기본 진단 차트”

시계열 그래프

시계열 그래프에는 각 시계열 대 시간 ID 가 표시됩니다. 시간 ID 를 지정하지 않으면 행 번호가 대신 사용됩니다. 홀드백 집합이 지정된 경우 훈련 데이터와 홀드백 데이터를 구분하는 세로 참조선이 그래프에 있습니다.

이 플랫폼에서는 ADF(Augmented Dickey-Fuller) 검정을 사용하여 몇 가지 안정성 검정도 수행합니다. 다음 검정 및 요약 통계량이 시계열 그래프 옆에 표시됩니다.

Box-Cox 에 대한 람다 (Box-Cox 변환을 지정한 경우에만 표시됨) Box-Cox 변환에 사용되는 람다 값입니다.

평균 표본 평균입니다.

표준편차 표본 표준편차입니다.

N 계열 길이입니다.

0 평균 ADF 0 평균을 사용하여 확률 보행과 대조하는 검정으로, 다음과 같이 정의됩니다.

$$x_t = \phi x_{t-1} + e_t$$

단일 평균 ADF 0 이 아닌 평균을 사용하여 확률 보행과 대조하는 검정으로, 다음과 같이 정의됩니다.

$$x_t - \mu = \phi(x_{t-1} - \mu) + e_t$$

추세 ADF 0 이 아닌 평균과 선형 추세를 사용하여 확률 보행과 대조하는 검정으로, 다음과 같이 정의됩니다.

$$x_t - \mu - \beta t = \phi[x_{t-1} - \mu - \beta(t-1)] + e_t$$

시계열 기본 진단 차트

"시계열 기본 진단" 차트에 표시되는 정보는 시계열 보고서의 빨간색 삼각형 메뉴 옵션에 따라 다릅니다. 진단 차트에서 정보를 표시하거나 숨기는 빨간색 삼각형 메뉴 옵션은 "자기상관", "편자기상관", "변동도" 및 "AR 계수"입니다. 기본적으로 "자기상관" 및 "편자기상관"이 표시됩니다.

자기상관 차트

"자기상관" 옵션은 "시계열 기본 진단" 차트에서 다음 열을 표시하거나 숨깁니다.

시차 점 사이의 주기 수입니다.

참고 : 시차 수는 기본적으로 0 에서 시작합니다. 시차 1 로 시작하는 상관을 계산하려면 그래프를 생성하기 전에 JMP 환경 설정을 수정합니다. **파일 > 환경 설정 > 플랫폼 > 시계열**을 선택한 후 **ACF 및 PACF 에 시차 0 표시 안 함**을 선택하십시오.

자기상관 k 번째 시차에 대한 자기상관이며 다음과 같이 계산됩니다.

$$r_k = \frac{c_k}{c_0} \text{ 여기서, } c_k = \frac{1}{N} \sum_{t=k+1}^N (y_t - \bar{y})(y_{t-k} - \bar{y})$$

및 $\bar{y}$ 는 시계열에서 비결측 점 N 개의 평균입니다. 정의에 따라 1 차 자기상관 (시차 0) 의 길이는 항상 1 입니다.

막대는 자기상관을 그래픽으로 나타냅니다. 파란색 곡선은 다음과 같이 계산되는 큰 시차 표준 오차의 두 배 (± 2 표준 오차) 를 나타냅니다.

$$SE_k = \sqrt{\frac{1}{N} \left(1 + 2 \sum_{i=1}^{k-1} r_i^2 \right)}$$

Ljung-Box Q 자기상관 그룹이 0 과 유의하게 다른지 여부를 검정하거나 모형의 잔차를 백색잡음과 구별할 수 있는지 검정하는 데 사용됩니다. Q 는 검정 통계량입니다.

p 값 Ljung-Box 검정의 p 값입니다.

편자기상관 차트

" 편자기상관 " 옵션은 " 시계열 기본 진단 " 차트에서 다음 열을 표시하거나 숨깁니다.

시차 점 사이의 주기 수입니다.

편자기상관 k 번째 시차에 대한 편자기상관입니다.

막대는 편자기상관을 그래픽으로 나타냅니다. 파란색 선은 근사 95% 예측 한계에 대한 ± 2 표준 오차를 나타내며, 여기서 표준 오차는 다음과 같이 계산됩니다.

$$SE_k = \frac{1}{\sqrt{n}}, \text{ 모든 } k \text{ 에 대해}$$

변동도 차트

" 변동도 " 옵션은 " 시계열 기본 진단 " 차트에서 다음 열을 표시하거나 숨깁니다.

시차 점 사이의 주기 수입니다.

변동도 변동도는 k 시차가 있는 점의 차이 분산을 측정하여 1 시차가 있는 점의 차이 분산과 비교합니다. 변동도는 다음과 같이 자기상관에서 계산됩니다.

$$V_k = \frac{1 - r_{k+1}}{1 - r_1}$$

여기서 r_k 는 시차 k 에서의 자기상관입니다.

AR 계수 차트

"AR 계수" 옵션은 "시계열 기본 진단" 차트에서 다음 열을 표시하거나 숨깁니다.

시차 점 사이의 주기 수입니다.

AR 계수 계수는 고차의 순수 자기회귀 모형을 적합시켜 얻을 수 있는 것과 비슷합니다.

시계열 플랫폼 옵션

"시계열"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

- “시계열 진단”
- “차분 및 분해”
- “시계열 모형”
- “추가 옵션”

시계열 진단

"시계열"의 빨간색 삼각형 메뉴에는 시계열 데이터를 탐색하기 위한 다음 진단 옵션이 포함되어 있습니다.

그래프 시계열 그림 모양을 제어하는 옵션이 포함된 하위 메뉴를 표시합니다.

시계열 그래프 시계열 그래프를 표시하거나 숨깁니다.

점 표시 시계열 그래프에 점을 표시하거나 숨깁니다.

연결선 시계열 그래프의 점을 연결하는 선을 표시하거나 숨깁니다.

평균 선 시계열 그래프에서 시계열의 평균을 나타내는 가로선을 표시하거나 숨깁니다.

자기상관 "시계열 기본 진단" 차트에서 자기상관 그림을 표시하거나 숨깁니다. 자기상관 그래프는 주어진 시간 간격(시차)에 대해 시계열에 있는 모든 점 쌍 사이의 상관을 설명합니다. 자세한 내용은 "[자기상관 차트](#)"에서 확인하십시오.

팁 : 표본의 자기상관 그래프를 **표본 자기상관 함수**라고도 합니다.

편자기상관 "시계열 기본 진단" 차트에서 편자기상관 그림을 표시하거나 숨깁니다. 편자기상관 그래프는 주어진 시간 간격(시차)에 대해 시계열에 있는 모든 점 쌍 사이의 부분 상관을 설명합니다. 자세한 내용은 "[편자기상관 차트](#)"에서 확인하십시오.

팁 : 자기상관 및 편자기상관 그래프는 시계열이 정상(시간 경과에 따라 고정 평균 및 표준 편차를 갖는다는 의미)인지 여부를 확인하고 시계열 적합에 적절한 모형을 결정하는 데 도움이 될 수 있습니다.

변동도 "시계열 기본 진단" 차트에서 변동도 그래프를 표시하거나 숨깁니다. 자세한 내용은 "[변동도 차트](#)"에서 확인하십시오.

AR 계수 "시계열 기본 진단" 차트에서 AR(자기회귀) 계수의 최소 제곱 추정값 그래프를 표시하거나 숨깁니다. 자세한 내용은 "[AR 계수 차트](#)"에서 확인하십시오.

스펙트럼 밀도 스펙트럼 밀도 그래프를 주기와 빈도의 함수로 표시하거나 숨깁니다. 스펙트럼 밀도 옵션은 데이터에 대한 두 가지 검정 결과를 제공하는 "백색잡음 검정" 보고서도 표시합니다. 자세한 내용은 "[스펙트럼 밀도 보고서](#)" 및 "[스펙트럼 밀도에 대한 통계 상세 정보](#)"에서 확인하십시오.

차분 및 분해

"시계열"의 빨간색 삼각형 메뉴에는 모델링할 데이터를 준비하기 위한 다음 옵션이 포함되어 있습니다.

차분 "차분 규격" 창을 표시합니다 ([그림 18.5](#)). 이 창에서 시계열에 적용할 차분 연산을 지정할 수 있습니다. 시계열의 값을 차분하면 비정상 계열을 정상 계열로 변환할 수 있습니다. 차분 계열은 다음 방정식에 의해 제공됩니다.

$$w_t = (1 - \mathbf{B})^d (1 - \mathbf{B}^s)^D y_t$$

여기서 t 는 시간 인덱스이고, $\mathbf{B}$ 는 $\mathbf{B}y_t = y_{t-1}$ 로 정의된 후방 연산자입니다.

참고 : 많은 시계열은 추세 또는 계절성이 있는 시계열과 같은 고정 평균을 보이지 않습니다. 이러한 비정상 계열은 ARMA 모형과 같이 정상 시계열을 가정하는 시계열 모형에 의한 설명에 적합하지 않습니다. 추세 및 / 또는 계절성을 제거하면 정상 차분 계열이 생성되고 안정성을 가정한 모형을 사용하여 계열을 설명할 수 있습니다.

그림 18.5 차분 규격 창

"차분 규격" 창을 사용하여 비계절 차분 차수 (d), 계절 차분 차수 (D) 및 기간별 관측값 수 (s)를 지정할 수 있습니다. 차분 차수의 값을 0으로 선택하면 해당 유형의 차분이 없다는 것을 나타냅니다. 차분 연산을 지정하고 **추정**을 클릭할 때마다 보고서 창에 새 "차분" 보고서가 표시됩니다. 자세한 내용은 "[시계열 플랫폼의 추가 예](#)"에서 확인하십시오.

분해 (시작 창에서 "홀드백에 대한 예측" 또는 "Box-Cox 변환 사용"을 선택한 경우에는 사용 불가능) 분해 방법 하위 메뉴를 표시합니다. 시계열 데이터 분해는 시계열에서 선형 추세와

계절 주기를 분리하고 제거합니다. 이렇게 하면 모형 추정에 도움이 될 수 있습니다. 세 가지 분해 옵션이 제공됩니다.

선형 추세 제거 선형 회귀 모형을 사용하여 시계열의 선형 추세를 추정하고 데이터에서 선형 추세를 제거합니다. 추세제거 계열에 대한 시계열 보고서가 선형 추세 정보와 함께 보고서 창에 추가됩니다. 자세한 내용은 "[시계열 보고서](#)" 및 "[선형 추세 보고서](#)"에서 확인하십시오.

순환 제거 단일 코사인파를 사용하여 시계열의 순환 성분을 추정한 후 데이터에서 순환 성분을 제거합니다. "순환 제거" 옵션을 선택하면 "순환 정의" 대화상자가 나타납니다. 이 대화상자 창에서 순환별 단위 수를 지정하고 데이터에서 상수를 뺄지 여부를 나타낼 수 있습니다. 순환제거 계열에 대한 시계열 보고서가 순환 정보와 함께 보고서 창에 추가됩니다. 자세한 내용은 "[시계열 보고서](#)" 및 "[순환 보고서](#)"에서 확인하십시오.

X11 미국 인구조사국(US Bureau of the Census)에서 개발한 X-11 방법을 사용하여 추세 및 계절 효과를 제거합니다 (Shiskin et. al. 1967 연구 자료 참조). X-11 방법에 대한 자세한 내용은 "[X-11 분해에 대한 통계 상세 정보](#)"에서 확인하십시오. 이 옵션을 선택하면 "분해 유형 선택" 대화상자가 나타납니다. 이 대화상자 창에서 승법 또는 가법 X-11 조정을 지정할 수 있습니다. **확인**을 클릭하면 X11 보고서가 보고서 창에 추가됩니다. 자세한 내용은 "[X11 보고서](#)"에서 확인하십시오.

X11 옵션은 월별 또는 분기별 데이터에만 사용할 수 있습니다. X, 시간 ID 열에 공백이나 결측값 없이 월 또는 분기별로 일정한 간격의 숫자 값이 포함되어야 합니다. 이러한 요구 사항을 충족하지 않는 시간 열에 대해 X11을 요청하면 JMP에서 오류가 반환됩니다. X11 분석에 적절한 시간 열의 예는 "[적절한 시간 ID 열 생성](#)"에서 확인하십시오.

참고 : **선형 추세 제거** 또는 **순환 제거** 옵션을 선택하면 추세제거 또는 순환제거 데이터가 포함된 열이 데이터 테이블에 추가됩니다. 이 열이 데이터 테이블에 이미 있는 경우 옵션을 선택하면 자동으로 기존 열을 덮어씁니다.

팁 : 일반적으로 선형 추세를 제거한 후 12 개월 주기와 같은 긴 주기를 제거하여 분해를 시작합니다. 그런 다음 6 개월 주기와 같은 짧은 주기를 제거할 수 있습니다.

시차 그림 표시 시간 t 의 관측값이 Y축에 있고 시간 $t + p$ 의 관측값이 X축에 있는 그림을 표시하거나 숨깁니다. $+/- p$ 는 시차라고 합니다. 이 그림은 시간 t 의 관측값이 시간 $t + p$ 의 다른 관측값과 어떤 관계가 있는지 확인하는 데 유용합니다. 그러나 그림에 구조가 있으면 시간에 따라 관측값 사이에 어떤 관계가 있음을 나타냅니다. 구조를 식별하면 시계열 모형을 생성할 때 도움이 됩니다.

Box-Cox 변환 그림 표시 (시작 창에서 "Box-Cox 변환 사용 " 을 지정한 경우에는 사용 불가능)
"Box-Cox 변환 그림 " 을 표시하거나 숨깁니다 . Box-Cox 변환은 람다 모수 (λ) 를 기반으로 하는 멱 변환이며 다음 계산식을 사용합니다 .

$$Y_{\lambda} = \begin{cases} \frac{y^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(y) & \text{단, } \lambda = 0 \end{cases}$$

"Box-Cox 변환 그림 "에서는 변환된 시계열 그래프와 람다 모수 (λ) 설정 옵션을 보여 줍니다 .
" 람다 " 상자에는 모수의 현재 값이 표시됩니다 . " 람다 " 슬라이더를 사용하여 모수 값을 변경할 수도 있으며 , 그러면 그림이 자동으로 업데이트됩니다 . " 낮음 " 및 " 높음 " 상자는 슬라이더 범위를 설정합니다 .

팁 : "Box-Cox 변환 그림 " 을 사용하여 다양한 람다 값에 대해 변환된 시계열을 확인할 수 있습니다 . 값을 결정 한 후에는 분석을 다시 시작하고 **Box-Cox 변환 사용**을 선택한 후 "Box-Cox 에 대한 람다 " 옆에 원하는 람다 값을 입력합니다 . 이렇게 하면 변환된 계열에 대해 시계열 플랫폼이 실행됩니다 .

교차 상관 (" 전이 함수 분석 " 의 빨간색 삼각형 메뉴에서만 사용 가능) 보고서에 교차 상관 그림을 표시하거나 숨깁니다 . 그림의 길이는 자기상관 그림 길이의 두 배 또는 $2 \times \text{ACF}$ 길이 + 1 입니다 . 이 그림에는 출력 계열 대 모든 입력 계열 그림이 숫자 형식과 그래픽 형식으로 포함됩니다 . 파란색 선은 2 표준 오차를 나타냅니다 .

참고 : 교차 상관 그림의 경우 귀무가설 하의 표준 오차가 $1/\sqrt{n-k}$ 로 계산됩니다 . 여기서 n 은 비결측 데이터 값의 수이고 k 는 자기상관 시차 수입니다 .

사전백색화 (" 입력 계열 " 의 빨간색 삼각형 메뉴에서만 사용 가능) 사전백색화 차수를 설정할 수 있는 " 사전백색화 규격 " 창을 표시합니다 . 사전백색화는 전이 함수 모형을 식별하는 데 사용되는 기법입니다 . 이 기법은 잔차가 백색잡음과 동일하도록 ARIMA 모형을 입력 계열에 적합시키는 방법을 사용합니다 . 그런 다음 동일한 모형을 사용하여 출력 계열을 추정합니다 . 필터링된 입력 계열과 필터링된 출력 계열 사이의 교차 상관을 사용하여 전이 함수 모형에 적절한 시차를 결정할 수 있습니다 . 사전백색화에 대한 자세한 내용은 Box et al. (1994) 연구 자료에서 확인하십시오 .

시계열 모형

"시계열"의 빨간색 삼각형 메뉴에는 다음 유형의 모형을 적합시키기 위한 옵션이 포함되어 있습니다 .

ARIMA 적합시킬 ARIMA 모형을 지정할 수 있는 "ARIMA 규격 " 창을 표시합니다 . ARIMA 모형은 과거 값과 일련의 오차 (**확률 충격** 또는 **개변**이라고도 함) 의 선형 조합을 통해 시계열의 미래 값을 예측합니다 . ARIMA 모형은 시계열에 대해 지정된 ARIMA 모형의 최대 가능성도 적합을 수행합니다 . 자세한 내용은 "ARIMA 모형 " 에서 확인하십시오 .

참고 : 일반적으로 ARIMA 모형은 $ARIMA(p,d,q)$ 로 표시됩니다. p, d 또는 q 가 0 이면 해당 문자가 삭제됩니다. 예를 들어 p 와 d 가 0 이면 모형은 $MA(q)$ 로 표시되는 단순 이동 평균 모형이 됩니다.

그림 18.6 ARIMA 규격 창

The image shows a software window titled "ARIMA 모형 지정" (ARIMA Model Specification). Inside, there is a section labeled "ARIMA" containing three input fields: "p. 자기회귀 차수" (p. Autoregressive order) with a value of 0, "d. 차분 차수" (d. Differencing order) with a value of 0, and "q. 이동 평균 차수" (q. Moving average order) with a value of 0. Below these fields is a dropdown menu for "예측 구간" (Forecast interval) set to "0.95". There are two checked checkboxes: "절편" (Intercept) and "제약 적합" (Constrained fit). At the bottom of the window are three buttons: "추정" (Estimate), "취소" (Cancel), and "도움말" (Help).

p, 자기회귀 차수 다항식 $\phi(B)$ 연산자의 차수 p 입니다.

d, 차분 차수 차분 연산자의 차수 d 입니다.

q, 이동 평균 차수 차분 연산자 $\theta(B)$ 의 차수 q 입니다.

예측 구간 예측 구간에 대해 0 에서 1 사이의 예측 수준을 설정할 수 있습니다.

절편 절편 항 μ 를 모형에 포함할지 여부를 결정합니다.

제약 적합 이 옵션을 선택하면 적합 절차에서 자기회귀 모수가 항상 안정 영역 내에 있고 이동 평균 모수가 가역 영역 내에 있도록 제한합니다.

팁 : 적합기가 최적 값을 찾기 어렵거나 적합 속도를 높이려면 " 제약 적합 " 옵션을 선택 취소하십시오. " 모형 요약 " 테이블을 사용하여 결과 적합 모형이 안정적이고 가역적인지 확인할 수 있습니다.

모형을 지정하고 **추정**을 클릭하면 모형 보고서가 보고서 창에 추가됩니다. 자세한 내용은 "[추가 시계열 보고서](#)" 에서 확인하십시오.

계절 ARIMA 적합시킬 계절 ARIMA 모형을 지정할 수 있는 " 계절 ARIMA 규격 " 창을 표시합니다. 이 창에는 "ARIMA 규격" 창과 동일한 요소가 있지만 계절 요소 규격도 포함됩니다. " 기간별 관측값 수 " 옵션을 사용하면 s 로 표시되는 기간별 관측값 수를 지정할 수 있습니다. 계절 ARIMA 모형에 대한 자세한 내용은 "[계절 ARIMA 모형](#)" 에서 확인하십시오.

참고 : 계절 ARIMA 모형은 " 계절 $ARIMA(p,d,q)(P,D,Q)s^s$ " 로 표시됩니다.

모형을 지정하고 **추정**을 클릭하면 모형 보고서가 보고서 창에 추가됩니다. 자세한 내용은 "[추가 시계열 보고서](#)" 에서 확인하십시오.

평활 모형 평활 모형 하위 메뉴를 표시합니다. 평활 모형을 선택하면 규격 창이 나타납니다. 자세한 내용은 "[단순 평활 평균 지정 창](#)" 및 "[평활 모형 창](#)" 에서 확인하십시오. 지정된 각 모

형에 대한 평활 모형 보고서가 보고서 창에 나타납니다. 자세한 내용은 "모형 보고서"에서 확인하십시오. 평활 모형은 모형에 의한 시계열의 진화를 나타냅니다.

$$y_t = \mu_t + \beta_t t + s(t) + a_t$$

다음은 각 요소에 대한 설명입니다.

μ_t = 시간에 따라 변하는 평균 항

β_t = 시간에 따라 변하는 기울기 항

$s(t)$ = 시간에 따라 변하는 계절 항 중 하나

a_t = 확률 충격

일반 평활 모형 방정식에 대한 자세한 내용은 "지수 평활 모형에 대한 통계 상세 정보"에서 확인하십시오. 다음과 같은 평활 모형을 사용할 수 있습니다.

단순 이동 평균 (시작 창에서 "홀드백에 대한 예측" 또는 "Box-Cox 변환 사용"을 선택한 경우에는 사용 불가능) 평활 창에서 정의된 인접한 여러 점의 평균을 사용하여 값을 추정하는 모형입니다. "단순 평활 평균 지정" 창에서 평활 창의 특성을 지정할 수 있습니다. 평활 창 옵션을 지정하면 "단순 이동 평균" 보고서가 표시됩니다. 기본적으로 이 보고서는 시간 창에서 연속 관측값의 평균과 동일한 값이 표시된 그림을 생성합니다. 여러 단순 이동 평균 모형을 추가하고 동일한 그림에 표시할 수 있습니다. 자세한 내용은 "단순 평활 평균 지정 창"에서 확인하십시오.

단순 지수 평활 수준 성분이 있는 모형입니다. 자세한 내용은 "단순 지수 평활"에서 확인하십시오.

이중 지수 평활 수준 성분과 추세 성분이 있는 모형입니다. 이것은 선형 지수 평활의 특수한 경우입니다. 자세한 내용은 "이중 (Brown) 지수 평활"에서 확인하십시오.

선형 지수 평활 수준 성분과 추세 성분이 있는 모형입니다. 자세한 내용은 "선형 (Holt) 지수 평활"에서 확인하십시오.

진폭감소 추세 선형 지수 평활 수준 성분과 진폭감소 추세 성분이 있는 모형입니다. 이 모형은 선형 추세보다 더 복잡한 추세를 나타내는 계열에 적절합니다. 자세한 내용은 "진폭감소 추세 선형 지수 평활"에서 확인하십시오.

계절 지수 평활 수준 성분과 계절 성분이 있는 모형입니다. 자세한 내용은 "계절 지수 평활"에서 확인하십시오.

Winters 방법 수준 성분, 추세 성분 및 계절 성분이 있는 모형입니다. 자세한 내용은 "Winters 방법 (가법 모형)"에서 확인하십시오.

참고 : 단순 이동 평균 모형을 제외한 각 평활 모형은 기존 지수 평활 모형이며 ARIMA 모형과 동일합니다. 일부 평활 모형은 ARIMA 옵션에서 ARIMA 모형 모수를 허용하지 않도록 제한하는 방식이 내재되어 있으므로 ARIMA 옵션을 사용하여 동등한 ARIMA 모형을 지정하지 못할 수도 있습니다.

상태 공간 평활 모형 다양한 상태 공간 평활 모형을 적합시킬 수 있는 "상태 공간 평활 모형 지정" 창을 표시합니다. 이 내용은 Hyndman et al. (2008) 연구 자료에 정의되어 있습니다. 상태 공간 평활 모형은 오차, 추세 성분 및 계절 성분을 기반으로 정의됩니다.

- 오차는 가법 (A) 또는 승법 (M) 일 수 있습니다.
- 추세 성분은 없음 (N), 가법 (A), 진폭감소 가법 (A_d), 승법 (M) 또는 진폭감소 승법 (M_d) 일 수 있습니다.
- 계절 성분은 없음 (N), 가법 (A) 또는 승법 (M) 일 수 있습니다.

특정 모형은 ETS (오차, 추세, 계절) 로 나타낼 수 있습니다. "상태 공간 평활 모형 지정" 창의 체크박스를 사용하여 원하는 모형에 대한 오차, 추세 및 계절성을 선택합니다. 플랫폼에서 권장하는 모형에 해당하는 체크박스를 선택하려면 **권장 선택**을 클릭합니다. 권장 모형이 선택된 상태로 창이 열립니다. **모두 선택**을 클릭하여 모든 체크박스를 선택하거나 **모두 선택 취소**를 클릭하여 모든 체크박스를 선택 취소할 수 있습니다.

이 창에는 다음 옵션도 포함되어 있습니다.

주기 모형 적합 프로세스에서 고려할 계절성 값을 지정합니다.

모수 제한 관측값이 현재에서 멀어질수록 모형의 현재 상태에 미치는 영향이 적어지도록 모수를 제한합니다. 상태 공간 평활 모형에서 모든 이전 관측값이 주어진 경우 시간 t 에서의 예측은 시간 t 까지의 모든 관측값의 가중합과 동일합니다. 가중치는 모수의 함수입니다. 따라서 모수를 제한하면 과거 관측값에 대한 가중치가 0 이 되고 관측값이 현재에서 멀어질수록 가중치가 더 빨리 0 이 됩니다.

확인을 클릭하면 지정된 모형 집합이 적합됩니다. 적합한 각 상태 공간 평활 모형이 모형 비교 테이블에 추가됩니다. 모형 비교 테이블에서 상태 공간 평활 모형에 대한 "보고서" 체크박스를 선택하면 "상태 공간 평활" 보고서가 나타납니다. 자세한 내용은 "[상태 공간 평활 보고서](#)" 에서 확인하십시오.

전이 함수 ("전이 함수 분석" 의 빨간색 삼각형 메뉴에서만 사용 가능) "전이 함수 모형 규격" 창을 표시합니다. 전이 함수 모형을 생성하는 것은 ARIMA 모형을 생성하는 것과 비슷하며 모형 탐색, 적합 및 비교를 반복하는 프로세스입니다. 모형을 생성하기 전 데이터 탐색 프로세스 중에 데이터를 사전백색화하는 것이 유용할 때가 있습니다. 자세한 내용은 "[사전백색화](#)" 에서 확인하십시오. 전이 함수에 대한 자세한 내용은 "[전이 함수에 대한 통계 상세 정보](#)" 에서 확인하십시오.

참고: 현재 "전이 함수" 옵션은 결측값을 제한적으로 지원합니다. 시계열 플랫폼에서는 제외된 행도 결측값으로 처리하므로 제외된 행이 포함된 데이터 테이블에 대해서는 전이 함수 모형을 적합시킬 수 없습니다.

그림 18.7 전이 함수 모형 규격 창

전이 함수 모형 지정

잡음 계열 차수

p, 자기회귀 차수	Output CO2	2
d, 차분 차수		0
q, 이동 평균 차수		0
p, 자기회귀 차수		0
D, 차분 차수		0
Q, 이동 평균 차수		0
S, 기간별 관측값 수		12

입력 선택

☒ Input Gas Rate

입력 계열 차수

s1, 분자 차수 연산자	Input Gas Rate	2
d1, 차분 차수 연산자		0
r1, 분모 차수 연산자		2
s2, 계절 분자 차수 연산자		0
d2, 계절 차분 차수 연산자		0
r2, 계절 분모 차수 연산자		0
S, 기간별 관측값 수		12
L, 입력 시차		3

☒ 절편

☐ 대체 파라미터화

☒ 제약 적합

예측 주기

예측 구간

추정

취소

도움말

" 전이 함수 모형 규격 " 창에는 다음 섹션이 포함되어 있습니다.

잡음 계열 차수 잡음 계열에 대한 규격을 포함합니다 . 소문자는 비계절 다항식의 계수이고 대문자는 계절 다항식의 계수입니다 .

입력 선택 모형에 대한 입력 계열을 선택할 수 있습니다 .

입력 계열 차수 입력 계열에 대한 규격을 포함합니다 . 처음 세 차수는 비계절 다항식과 관련 됩니다 . 다음 네 차수는 계절 다항식과 관련됩니다 . 마지막 옵션은 입력 시차에 대한 것 입니다 .

다음과 같이 모형 적합을 제어하는 세 가지 추가 옵션이 있습니다 .

절편 모형에 절편이 있는지 여부를 지정합니다 . 이 옵션을 선택하지 않으면 μ 가 0 으로 가정됩니다 .

대체 파라미터화 분자 다항식에서 일반 회귀 계수를 인수 분해할지 여부를 지정합니다 .

제약 적합 AR 및 MA 계수의 제약을 전환합니다 .

예측 주기 예측에 사용되는 예측 주기 수를 지정합니다 . 데이터 테이블의 끝에 Y 변수에 대한 결측값과 입력 변수에 대한 비결측 값이 포함된 행이 있으면 이러한 행이 초기 예측 설정에 사용됩니다 . 입력 변수의 값은 입력 변수의 미래 값으로 간주됩니다 .

예측 구간 예측 구간에 대한 신뢰 수준을 지정합니다 .

ARIMA 모형 그룹 차수 범위를 지정하여 다양한 ARIMA 모형 또는 계절 ARIMA 모형을 적합시킬 수 있는 "ARIMA 모형 그룹 " 창을 표시합니다 . 창에 범위를 입력하면 그에 따라 " 총 모형 수 " 가 업데이트됩니다 .

그림 18.8 ARIMA 모형 그룹 규격 창

ARIMA 모형 지정

ARIMA		계절 ARIMA	
p, 자기회귀 차수	0 0	P, 자기회귀 차수	0 0
d, 차분 차수	0 0	D, 차분 차수	0 0
q, 이동 평균 차수	0 0	Q, 이동 평균 차수	0 0
		기간별 관측값 수	12 12

예측 구간 0.95 ▾

☒ 절편
☒ 제약 적합

총 모형 수 0

추정 취소 도움말

모형을 지정하고 **추정**을 클릭하면 지정된 각 모형에 대한 모형 보고서가 보고서 창에 추가됩니다. 자세한 내용은 "[추가 시계열 보고서](#)"에서 확인하십시오.

단순 평활 평균 지정 창

"단순 이동 평균"을 평활 모형으로 선택하면 "단순 평활 평균 지정" 창이 나타납니다. w 를 SMA(단순 이동 평균) 모형의 평활 창 너비라고 하겠습니다. $f_t = (y_t + y_{t-1} + y_{t-2} + \dots + y_{t-(w-2)} + y_{t-(w-1)}) / w$ 를 특정 시점 t 에 대한 연속된 w 개 관측값의 평균이라고 하겠습니다.

그림 18.9 단순 평활 평균 지정 창

평활 창 너비를 입력하십시오.

☒ 중심화 안 함
☐ 중심화됨
☐ 짝수 항에 대해 중심화 및 이중 평활

확인 취소

평활 창 너비를 입력하십시오. 평균을 구할 연속 점 수를 정의하는 평활 창 너비 w 입니다. 창 너비가 클수록 계열이 더 많이 평활됩니다.

중심화 안 함 평활 창은 계열이 추정되는 점인 시점 t 까지 이어지는 연속 점으로 구성됩니다. 즉, f_t 는 시간 t 에 대해 그림에 표시되는 값입니다.

중심화됨 평활 창은 계열이 추정되는 시점을 중심으로 합니다.

- 홀수 w 의 경우 f_t 는 시간 $t-(w-1)/2$ 에 대해 그림에 표시되는 값입니다.
- 짝수 w 의 경우 f_t 는 시간 $t-(w-1)/2$ 에 대해 그림에 표시되는 값입니다. 데이터 테이블에 저장할 때 f_t 는 $t-(w-2)/2$ 입니다.

짝수 항에 대해 중심화 및 이중 평활 짝수 w 의 경우 평활 창은 계열이 추정되는 시점을 중심으로 할 수 없습니다. 이 옵션은 거의 중심에 있는 두 개의 평활 창을 생성하여 평균을 구합니다. 평활 추정값은 다음과 같이 계산됩니다.

$$f_{t-\frac{w}{2}} = \frac{y_t + 2 \sum_{i=1}^{w-1} y_{t-i} + y_{t-w}}{2w}$$

평활 모형 창

"단순 이동 평균" 이외의 평활 모형 옵션 중 하나를 선택하면 평활 모형 규격 창이 나타납니다. 창 제목과 사용 가능한 옵션은 선택한 평활 모형 옵션에 따라 다릅니다.

그림 18.10 평활 모형 규격 창

예측 구간 예측 구간에 대해 예측 수준을 설정할 수 있습니다.

기간별 관측값 수 (계절 평활 모형에만 사용 가능) 계절 평활 모형에서 기간별 관측값 수를 설정할 수 있습니다.

제약 조건 적합 중 평활 가중치에 적용할 제약 조건 유형을 지정할 수 있습니다. 다음과 같은 제약 조건 옵션을 사용할 수 있습니다.

0부터 1까지 평활 가중치 값을 0부터 1까지의 범위로 제한합니다.

무제약 모수의 범위를 자유롭게 지정할 수 있습니다.

안정적 가역적 동등한 ARIMA 모형이 안정적이고 가역적이 되도록 모수를 제한합니다.

사용자 정의 개별 평활 가중치에 대한 제약 조건을 설정할 수 있도록 대화상자를 확장합니다. 각 평활 가중치는 가중치 이름 옆의 팝업 메뉴 설정에 따라 **유계 (bounded)**, **고정** 또는 **무제약**일 수 있습니다. 고정 또는 유계 가중치에 대한 값을 입력할 때 해당 값은 양의 실수 또는 음의 실수일 수 있습니다.

그림 18.11 사용자 평활 가중치

예측 구간 0.95

제약 조건 사용자 정의

사용자 평활 가중치

가중치 제약	고정 가중치	하한	상한
수준 고정	0.3		
추세 유계(bounded)		0.1	0.8

그림 18.11 에 표시된 예제에는 0.3 값으로 고정된 " 수준 " 가중치 (α) 와 0.1 ~ 0.8 범위로 제한된 " 추세 " 가중치 (γ) 가 있습니다. 이 경우 추세 가중치 값은 0.1 ~ 0.8 범위 내에서 달라질 수 있지만 수준 가중치는 0.3 으로 유지됩니다. 이러한 사용자 제약 조건을 사용하여 모든 평활 가중치를 미리 지정할 수 있습니다. 이 경우 예측과 잔차는 계산되지만 가중치는 데이터에서 추정되지 않습니다.

추가 옵션

"시계열"의 빨간색 삼각형 메뉴에는 다음 옵션도 포함되어 있습니다.

모형의 예측 결합 및 저장 보고서에 있는 모든 모형 적합의 결과를 결합하여 새 데이터 테이블을 생성합니다.

스펙트럼 밀도 저장 ($i+1$)번째 행이 빈도 $f_i = i/N$ (즉, $1/N$ 의 i 번째 조화)에 해당하는 주기와 스펙트럼 밀도가 포함된 새 데이터 테이블을 생성합니다. 새 데이터 테이블에는 다음과 같은 열이 있습니다.

주기 i 번째 조화의 주기입니다 ($1/f_i$).

빈도 조화의 빈도입니다 (f_i).

각 빈도수 조화의 각 빈도수입니다 ($2\pi f_i$).

사인 Fourier 사인 계수입니다 (a_i).

코사인 Fourier 코사인 계수입니다 (b_i).

주기도 주기도입니다 ($I(f_i)$).

스펙트럼 밀도 주기도의 평활 버전인 스펙트럼 밀도입니다.

예측 주기 수 적합 모형에 대해 예측되는 미래 기간의 수를 설정할 수 있는 창을 표시합니다. 초기값은 시계열 시작 창에서 설정됩니다. 이 값이 변경되면 모든 기존 및 미래 예측 결과에 새 값이 사용됩니다.

최대 반복 수 ARIMA 모형 적합에 사용되는 미래 최적화에 대한 최대 반복 수를 재설정할 수 있는 창을 표시합니다.

다음 옵션에 대한 자세한 내용은 JMP 사용의 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.

추가 시계열 보고서

시계열 플랫폼에서 많은 옵션을 사용하여 추가 보고서를 생성합니다.

- “차분 보고서”
- “분해 보고서”
- “모형 비교 보고서”
- “모형 보고서”
- “상태 공간 평활 보고서”
- “전이 함수 모형 보고서”
- “스펙트럼 밀도 보고서”

차분 보고서

시계열 플랫폼의 "차분" 보고서에는 차분 계열의 자기상관 및 편자기상관 그래프가 포함됩니다. 이 그래프를 사용하여 차분 계열이 정상인지 여부를 확인할 수 있습니다.

시계열 플랫폼에서 사용할 수 있는 ARIMA 모형과 계절 ARIMA 모형은 차분 연산을 지원합니다. 이러한 모형은 먼저 차분 연산에 따라 시계열을 차분한 후 차분 계열을 적합시키는 2단계 프로세스를 수행합니다. "차분" 옵션은 ARIMA 모형에 지정할 차분 차수를 결정하는 데 유용한 처리 도구입니다.

"차분"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

그래프 차분 계열 그림의 모양을 제어하는 옵션이 포함된 하위 메뉴를 표시합니다. 자세한 내용은 "시계열 플랫폼 옵션"에서 확인하십시오.

자기상관 차분 계열의 자기상관을 표시하거나 숨깁니다.

편자기상관 차분 계열의 편자기상관을 표시하거나 숨깁니다.

변동도 차분 계열의 변동도를 표시하거나 숨깁니다.

저장 차분 계열의 값이 포함된 새 열을 원래 데이터 테이블에 저장합니다. 차분 과정에서 일부 선행 요소가 손실됩니다. 이러한 요소는 저장된 "차분" 열에 결측값으로 나타납니다.

적합 제거 보고서 창에서 "차분" 보고서를 제거합니다.

분해 보고서

시계열 플랫폼에는 세 가지 분해 옵션이 있습니다. 각 옵션은 서로 다른 보고서를 생성합니다.

- "선형 추세 보고서"
- "순환 보고서"
- "X11 보고서"

선형 추세 보고서

데이터에 적합되는 선형 회귀 모형의 β_0 및 β_1 값을 포함합니다.

$$\text{Trend}_t = \beta_0 + \beta_1 \text{time}$$

추세제거 계열은 $D_t = O_t - \text{Trend}_t$ 와 같습니다. 여기서 O_t 는 원래 시계열입니다.

순환 보고서

데이터에 적합되는 순환 성분의 값을 포함합니다.

$$\text{Cycle}_t = C + A \cdot \cos\left(2 \cdot \pi \cdot \left(\left(\frac{1}{U}\right) \cdot t + P\right)\right)$$

모수 값은 다음과 같이 정의됩니다.

- C = 상수 (선택적)
- A = 코사인파의 진폭
- U = 순환별 단위 수
- P = 코사인파의 시기입니다.
- t = 주어진 관측값의 행 번호에서 1을 뺀 값

순환제거 계열은 $D_t = O_t - \text{Cycle}_t$ 와 같습니다. 여기서 O_t 는 원래 시계열입니다.

X11 보고서

분해 유형 선택에 따라 X11 옵션은 "X11 - 승법" 보고서 또는 "X11 - 가법" 보고서를 추가합니다. 보고서에는 다음과 같은 네 개의 동일한 그림이 포함됩니다.

원래 및 조정 원래 시계열 O_t 위에 X11 조정 시계열을 중첩합니다. X11 조정 값은 승법 조정의 경우 O_t/S_t 이고 가법 조정의 경우 $O_t - S_t$ 입니다.

D10 - 최종 계절 요인 시간에 따른 계절 요인 성분 S_t 를 표시합니다.

D12 - 최종 추세 순환 시간에 따른 추세 순환 성분 C_t 를 표시합니다.

D13 - 최종 불규칙 계열 시간에 따른 불규칙 성분 I_t 를 표시합니다.

X11 보고서 옵션

X11 보고서에는 다음과 같은 빨간색 삼각형 옵션이 있습니다.

테이블 표시 X11 요약 테이블을 표시하거나 숨깁니다. 자세한 내용은 Shiskin et. al. (1967) 연구 자료에서 확인하십시오. 테이블은 B~F 라벨이 지정된 5 개의 범주로 그룹화됩니다 (표 18.1의 설명 참조).

열 저장 계절 조정 시계열, 계절 성분 (S_t), 추세 순환 성분 (C_t) 및 불규칙 계열 성분 (I_t) 에 대한 네 개의 열을 데이터 테이블에 저장합니다.

모든 열 저장 보고서에서 **테이블 표시** 옵션으로 생성된 모든 테이블에 대한 열을 데이터 테이블에 저장합니다.

적합 제거 보고서 창에서 X11 보고서를 제거합니다.

표 18.1 X11 출력 테이블의 범주 설명

맨 앞의 문자	범주 설명
B	계절, 추세 순환 및 불규칙 성분의 예비 추정값
C	계절, 추세 순환 및 불규칙 성분의 중간 추정값
D	계절, 추세 순환 및 불규칙 성분의 최종 추정값
E	분석 테이블
F	요약 지표

X11 출력 테이블의 내용에 대한 자세한 정보는 Shiskin et. al. (1967) 연구 자료 또는 SAS/ETS 15.2 User's Guide("The output from PROC X11" 검색)에서 확인하십시오.

모형 비교 보고서

시계열 플랫폼의 "모형 비교" 보고서에는 모형에 대한 적합 통계량 테이블과 그림이 포함되어 있습니다. 새 모형이 적합될 때마다 고유한 색상 코딩을 사용하여 새 행이 모형 비교 테이블에 추가됩니다. 모형 비교 테이블은 각 모형에 대한 적합 통계량의 요약 정보를 제공하며, 동일한 시계열에 적합된 여러 모형을 비교하는 데 사용됩니다. 기본적으로 모형은 AIC 통계량에 따라 오름차순으로 정렬됩니다. 다른 통계량을 기준으로 모형을 정렬하려면 모형 비교 테이블에서 아무 열이나 마우스 오른쪽 버튼으로 클릭하고 "열별 정렬"을 선택합니다. 아니면 열 머리글을 클릭하여 정렬할 수도 있습니다. 적합 통계량에 대한 설명은 "모형 요약 테이블"에서 확인하십시오. 모형 비교 테이블에만 있는 적합 통계량은 "가중치"입니다. 이 적합 통계량은 정규화된 AIC 가중치입니다. 모형의 AIC 가중치는 다음과 같이 계산됩니다.

$$AICWeight = \exp[-0.5(AIC - BestAIC)] / \sum_{k=1}^K (\exp[-0.5(AIC_k - BestAIC)])$$

K 는 총 모형 수, AIC_k 는 모형 k 에 대한 AIC 값, $BestAIC$ 는 AIC 값이 가장 작은 모형에 대한 AIC 값입니다.

홀드백 집합을 지정하면 훈련 데이터에 대해 계산된 적합 통계량이 모형 비교 테이블에 표시되지 않습니다. 대신 홀드백 집합에 대한 개별 모형의 예측 성능을 판별하기 위해 RMSE(제곱근 평균 제곱 오차), MSE(평균 제곱 오차), MAPE(평균 절대 백분율 오차) 및 MAE(평균 절대 오차)와 같은 측정 기준이 계산됩니다. 기본적으로 모형은 RMSE 통계량에 따라 오름차순으로 정렬됩니다. 다른 통계량을 기준으로 모형을 정렬하려면 모형 비교 테이블에서 아무 열이나 마우스 오른쪽 버튼으로 클릭하고 "열별 정렬"을 선택합니다. 아니면 열 머리글을 클릭하여 정렬할 수도 있습니다. 적합 통계량에 대한 설명은 "모형 요약 테이블"에서 확인하십시오.

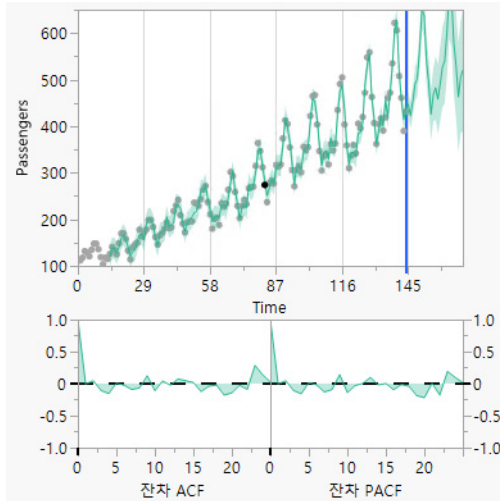
주의: 홀드백 집합을 지정하지 않고 상태 공간 평활 모형과 ARIMA 모형의 조합을 적합시키면 모형 비교 테이블 맨 위에 주의 메시지가 표시됩니다. 서로 다른 두 모형 계층 간에는 가능성도 값 및 가능성도 기반 정보 기준(예: AIC 통계량)을 비교할 수 없으므로 이 경우 상태 공간 평활 모형을 ARIMA 모형과 비교할 수 없습니다. 그러나 홀드백 집합을 지정하면 RMSE 통계량이 사용되므로 모형을 비교할 수 있습니다.

그림 18.12 모형 비교 테이블

모형 비교															
보고서	그래프	모형	DF	분산	AIC ^	SBC	R²	-2*로그 가능도	가중치	.2	.4	.6	.8	MAPE	MAE
☒	☐	계절 ARIMA(0, 1, 1)(0, 1, 1)12	128	138.49122	1020.9047	1029.5302	0.990	1014.9047	1.000000					3.197010	8.990978
☒	☐	ARIMA(1, 1, 1)	140	978.9437	1394.1215	1403.0101	0.932	1388.1215	0.000000					8.687058	24.364166
☒	☐	ARMA(1, 1)	141	989.13835	1407.7483	1416.6577	0.919	1401.7483	0.000000					9.704486	25.348507
☒	☐	AR(1)	142	1134.3871	1426.1794	1432.1190	0.909	1422.1794	0.000000					9.873726	26.556219
☒	☐	MA(1)	142	4264.6121	1616.8626	1622.8022	0.693	1612.8626	0.000000					23.865796	53.716974

보고서 체크박스를 사용하여 보고서 창에 표시할 전체 모형 보고서를 선택할 수 있습니다. 모형 비교 테이블 오른쪽에 두 개의 모형 그림이 나타납니다. 위의 그림은 데이터, 예측 및 예측 한계에 대한 시계열 그림입니다. 아래에는 자기상관 및 편자기상관 함수에 대한 그림이 있습니다. 그래프 체크박스를 사용하여 모형 그림에 표시할 모형을 선택할 수 있습니다.

그림 18.13 모형 그림



모형 비교 보고서 옵션

" 모형 비교 " 보고서의 각 모형에는 다음과 같은 빨간색 삼각형 메뉴 옵션이 있습니다.

새로 적합 모형에 대한 규격 창을 엽니다. 설정을 변경하여 다른 모형을 적합시킬 수 있습니다.

한 번 시뮬레이션 (홀드백 집합을 지정한 경우에는 사용 불가능) 시간 축의 위쪽 끝에 모형 시뮬레이션을 한 번 제공합니다. 모형 비교 시계열 그림에 시뮬레이션이 표시됩니다.

추가 시뮬레이션 (홀드백 집합을 지정한 경우에는 사용 불가능) 시간 축의 위쪽 끝에 지정된 수의 모형 시뮬레이션을 제공합니다. 모형 비교 시계열 그림에 시뮬레이션이 표시됩니다.

모형 시뮬레이션 제거 (홀드백 집합을 지정한 경우에는 사용 불가능) 모형에 대한 시뮬레이션을 제거합니다.

모든 시뮬레이션 제거 (홀드백 집합을 지정한 경우에는 사용 불가능) 모든 모형에 대한 시뮬레이션을 제거합니다.

시뮬레이션 생성 (홀드백 집합을 지정한 경우에는 사용 불가능) 모형에 대한 시뮬레이션을 생성하고 결과를 데이터 테이블에 저장합니다. 난수 시드값, 시뮬레이션 수 및 예측 주기 수를 지정할 수 있습니다.

시드값 설정 (홀드백 집합을 지정한 경우에는 사용 불가능) 시뮬레이션된 궤적을 생성하기 위한 시드값을 지정합니다.

적합 제거 보고서 창에서 해당 모형을 제거합니다.

모든 모형 제거 보고서 창에서 모형 비교 테이블에 나열된 모든 모형을 제거합니다.

선택 취소한 항목 제거 보고서 창에서 " 보고서 " 체크박스가 선택되지 않은 모든 모형을 제거합니다.

선택한 항목 제거 보고서 창에서 " 보고서 " 체크박스가 선택된 모든 모형을 제거합니다.

모든 보고서 숨기기 보고서 창에서 모형 비교 테이블에 나열된 모든 모형을 숨깁니다. 이것은 "보고서" 체크박스를 모두 선택 취소하는 것과 같습니다.

모든 그래프 숨기기 모형 그림에서 모형 비교 테이블에 나열된 모든 모형을 숨깁니다. 이것은 "그래프" 체크박스를 모두 선택 취소하는 것과 같습니다.

참고: 시계열 플랫폼에서는 제외된 행을 결측값으로 처리합니다. 데이터 테이블에서 마지막 몇몇 관측값이 제외되거나 결측인 경우 "모형 비교" 보고서의 시뮬레이션 옵션이 제대로 작동하지 않습니다. 시뮬레이션 절차에는 시계열의 마지막 몇몇 관측값이 사용됩니다. 마지막 몇몇 관측값이 결측인 경우 시뮬레이션을 수행하면 모든 미래 결측값이 생성됩니다.

모형 보고서

시계열 플랫폼에서 "ARIMA", "계절 ARIMA" 및 "평활 모형" 옵션은 비슷한 모형 보고서를 생성합니다. 이러한 모델링 옵션은 이론적 모형을 계열에 적합시키고 적합 모형을 사용하여 계열의 미래 값을 예측하는 데 사용됩니다. 또한 이러한 옵션을 사용하면 선택한 모형의 적합성을 확인할 수 있는 통계량 및 잔차가 생성됩니다. 모델링 옵션을 여러 번 선택할 수 있습니다. 모형을 선택할 때마다 해당 모형이 모형 비교 테이블에 추가됩니다. 모형 비교 테이블에서 모형 옆의 "보고서" 체크박스를 선택하면 해당 모형에 대한 보고서가 생성됩니다. 보고서 제목에 모형이 지정됩니다.

기본적으로 다음 보고서가 표시됩니다.

- 모형 요약 테이블
- 모수 추정값 테이블
- 예측 그림
- 잔차
- 반복 기록

모형 요약 테이블 모형에 대한 적합 통계량을 포함합니다. 아래 계산식에서 n 은 계열의 길이이고 k 는 모형에 있는 적합 모수의 수입니다.

DF 적합의 자유도입니다 ($n - k$).

개변 제곱합 개변 제곱의 합입니다.

잔차 제곱합 잔차 제곱의 합입니다.

분산 추정값 개변 제곱합을 자유도 ($n - k$) 로 나눈 값입니다. 이 값은 "ARIMA 모형" 섹션에 설명된 확률 충격 a_t 의 분산에 대한 표본 추정값입니다.

표준편차 분산 추정값의 제곱근입니다. 이 값은 확률 충격 a_t 의 표준편차에 대한 표본 추정값입니다.

Akaike 정보 기준 $-2 \times \text{로그 가능도} + 2k$ 로 계산되는 AIC 값입니다. AIC 값이 작을수록 더 나은 적합을 나타냅니다.

Schwarz Bayes 기준 $-2 \times \text{로그 가능도} + k \times \ln(n)$ 으로 계산되는 SBC 값입니다. SBC 값이 작을수록 더 나은 적합을 나타냅니다. Schwarz Bayes 기준은 BIC(Bayes 정보 기준)와 같습니다.

R² 다음과 같이 계산되는 R² 값입니다.

$$R^2 = 1 - \frac{SSE}{SST}$$

다음은 각 요소에 대한 설명입니다.

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$\hat{y}_i$ 1 시차 후 예측

$\bar{y}$ 평균

모형이 계열에 잘 적합되지 않으면 모형 SSE(오차 제곱합)가 SST(총 제곱합)보다 클 수 있습니다. 따라서 R²이 음수일 수 있습니다.

Adj-R² 다음과 같이 계산되는 조정 R² 값입니다.

$$1 - \left[\frac{(n-1)}{(n-k)} (1 - R^2) \right]$$

MAPE 다음과 같이 계산되는 평균 절대 백분율 오차 값입니다.

$$\frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

MAE 다음과 같이 계산되는 평균 절대 오차 값입니다.

$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

-2* 로그 가능도 최량 적합 모수 추정값에서 계산된 음의 로그 가능도 함수의 두 배입니다. 값이 작을수록 더 나은 적합을 나타냅니다. 자세한 내용은 선형 모형 적합에서 확인하십시오.

안정적 자기회귀 연산자가 안정적인지 여부를 나타냅니다. 즉, $\phi(z) = 0$ 의 모든 제곱근이 단위 원을 벗어나는지 여부입니다.

가역적 이동 평균 연산자가 가역적인지 여부를 나타냅니다. 즉, $\theta(z) = 0$ 의 모든 제곱근이 단위 원을 벗어나는지 여부입니다.

참고 : ϕ 및 θ 연산자는 "ARIMA 모형" 섹션에 정의되어 있습니다.

모수 추정값 테이블 시계열 모형 모수의 추정값을 표시합니다. 각 모형 유형마다 고유한 모수 집합이 있으며, 이러한 모수는 특정 시계열 모형에 대한 섹션에 설명되어 있습니다. 각 모수 추정값 테이블에는 다음과 같은 열이 포함됩니다.

항 각 모형 유형에 대한 섹션에 설명된 모수의 이름입니다. 일부 모형에는 절편 또는 평균 항이 포함됩니다. 이러한 모형에서는 관련 상수 추정값도 표시됩니다. 상수 추정값에 대한 정의는 ARIMA 모형 설명 아래에 제공됩니다.

요인 (승법 계절 ARIMA 모형에만 표시됨) 모수를 포함하는 모형의 요인입니다. 승법 계절 모형에서 요인 1은 비계절 요인이고 요인 2는 계절 요인입니다.

시차 (ARIMA 및 계절 ARIMA 모형에만 표시됨) 모수를 곱한 항에 적용되는 시차 또는 후방 연산자의 차수입니다.

추정값 시계열 모형의 모수 추정값입니다.

표준 오차 모수 추정값의 표준 오차 추정값입니다. 이 추정값은 검정 및 예측 구간을 계산하는 데 사용됩니다.

t 비 각 모수가 0이라는 가설에 대한 검정 통계량입니다. 모수에 대한 검정 통계량은 표준 오차에 대한 모수 추정값의 비율입니다. 가설이 참이면 이 통계량은 근사 스튜던트 t 분포를 따릅니다. 절대값이 2보다 큰 t 비는 0.05 유의 수준에 가까우므로 이러한 값을 찾는 것이 유의성을 판단하는 일반적인 규칙입니다.

Prob>|t| 각 모수에 대해 계산되는 관측된 p 값입니다. p 값은 가설이 참일 때 t 비 (절대값 기준)가 계산된 값보다 클 확률입니다.

상수 추정값 (절편 또는 평균 항이 포함된 모형에만 표시됨) 상수 추정값에 대한 정의는 ARIMA 모형 아래에 제공됩니다.

평균 (ARIMA 및 계절 ARIMA 모형에만 표시됨) ARIMA 또는 계절 ARIMA 모형의 절편 값에 대한 추정값입니다.

예측 시계열에 대한 관측값과 예측값을 모두 보여 주는 그림입니다. 이 그림은 세로선에 의해 두 영역으로 구분됩니다.

홀드백 집합을 지정한 경우 세로선은 훈련 데이터와 홀드백 데이터를 구분합니다. 세로선 왼쪽에는 1 시차 후 예측이 훈련 데이터 점과 중첩됩니다. 세로선 오른쪽에는 모형에서 예측한 값이 홀드백 데이터 점과 중첩됩니다. 예측 구간도 표시됩니다. 플랫폼 시작 창의 "예측 주기" 옵션을 사용하여 예측 값 수를 제어할 수 있습니다.

홀드백 집합을 지정하지 않은 경우 세로선은 관측된 데이터와 미래 예측을 구분합니다. 세로선 왼쪽에는 1 시차 후 예측이 관측된 데이터 점과 중첩됩니다. 세로선 오른쪽에는 모형에서 예측한 미래 값과 예측 구간이 표시됩니다. 플랫폼 시작 창의 "예측 주기" 옵션을 사용하거나 "시계열"의 빨간색 삼각형 메뉴에서 "예측 주기 수"를 선택하여 미래 예측 값 수를 제어할 수 있습니다.

잔차 적합 모형을 기반으로 한 잔차 값을 보여 주는 그래프입니다. 이러한 값은 시계열의 관측 값에서 1 시차 후 예측값을 뺀 값입니다. 잔차에 대한 자기상관 및 편자기상관 보고서도 표시 됩니다. 이 보고서를 사용하면 적합 모형이 데이터를 설명하기에 적절한지 여부를 결정할 수 있습니다. 잔차 그림의 점이 0 주위에서 정규 분포를 따르고, 0 보다 큰 시차에 대해 잔차의 자기상관 및 편자기상관에 유의한 성분이 없으면 적합 모형이 적절하다고 할 수 있습니다.

홀드백 계열에 대한 예측 오차 (홀드백 집합을 지정한 경우에만 사용 가능) 홀드백 집합의 관측값에 대한 예측 오차 그림입니다. 오차에 대한 자기상관 및 편자기상관 보고서도 표시됩니다.

반복 기록 각 반복에서 목적 함수의 값을 포함합니다. 이 옵션은 적합 절차의 문제를 진단하는데 유용할 수 있습니다. 데이터에 잘 맞지 않는 모형을 적합시키려고 하면 가능도에 대한 최적값에서 수렴하지 못할 수 있습니다. "반복 기록" 테이블에는 다음과 같은 통계량이 포함 됩니다.

반복 반복 번호입니다.

반복 기록 각 단계의 목적 함수 값입니다.

단계 반복 단계의 유형입니다.

목적 함수 - 기준 목적 함수의 그래디언트 norm 입니다.

ARIMA 및 계절 ARIMA 모형 보고서 옵션

각 ARIMA 및 계절 ARIMA 모형 보고서에는 다음 옵션이 포함된 빨간색 삼각형 메뉴가 있습니다.

점 표시 예측 그래프에 데이터 점을 표시하거나 숨깁니다.

예측 구간 표시 예측 그래프에 예측 구간을 표시하거나 숨깁니다.

열 저장 모형 결과를 나타내는 열이 포함된 새 데이터 테이블을 생성합니다.

예측 계산식 저장 (홀드백 집합을 지정한 경우에는 사용 불가능) 데이터 및 예측 계산식을 새 데이터 테이블에 저장합니다.

SAS 작업 생성 SAS 에서 모형 분석을 복제하는 SAS 코드를 생성합니다.

SAS 로 전송 모형 분석을 복제하는 SAS 코드를 SAS 로 전송합니다. SAS 서버에 연결되어 있지 않은 경우 이 옵션을 선택하면 연결 과정을 안내합니다.

잔차 통계량 모형에 대해 나타낼 잔차 통계량의 표시 유형을 제어합니다. 이러한 표시 유형은 "[시계열 플랫폼 옵션](#)" 섹션에서 설명합니다. 그러나 이 설정은 일련의 잔차에 적용됩니다.

적합 제거 보고서에서 개별 모형 적합을 제거합니다.

상태 공간 평활 보고서

시계열 플랫폼의 각 "상태 공간 평활" 보고서에는 모형을 식별하는 고유한 개요 이름이 있습니다. 각 모형이 모형 비교 테이블에 추가됩니다. "모형 비교" 테이블에서 모형에 대한 "보고서" 체크박스를 선택하면 "상태 공간 평활" 보고서가 나타납니다. 다음 보고서가 포함됩니다.

모형 유형 계열의 선택된 모형에 대한 오차, 추세 및 계절성을 설명합니다.

모형 요약 모형에 대한 적합 통계량을 포함합니다. 아래 계산식에서 n 은 계열의 길이이고 k 는 모형에 있는 적합 모수의 수입니다.

-2* 로그 가능도 최량 적합 모수 추정값에서 계산된 음의 로그 가능도 함수의 두 배입니다. 값이 작을수록 더 나은 적합을 나타냅니다. 자세한 내용은 **선형 모형 적합**에서 확인하십시오.

AIC $-2*로그 가능도 + 2k$ 로 계산되는 AIC(Akaike 정보 기준) 값입니다. AIC 값이 작을수록 더 나은 적합을 나타냅니다.

BIC $-2*로그 가능도 + k*\ln(n)$ 으로 계산되는 SBC(Schwarz Bayes 기준) 값입니다. SBC 값이 작을수록 더 나은 적합을 나타냅니다. Schwarz Bayes 기준은 BIC(Bayes 정보 기준)와 같습니다.

모수 수 정보 기준 계산의 모수 수입니다. 이 수는 가능도 함수의 모수 수이며, 예측 계산식의 모수 수에 1 을 더한 값과 같습니다.

참고 : "모수 수" 는 모수 추정값 테이블에 있는 모수의 수가 아닙니다.

시그마 1 시차 후 예측 표준편차의 추정값입니다.

R² 다음과 같이 계산되는 R² 값입니다.

$$R^2 = 1 - \frac{SSE}{SST}$$

다음은 각 요소에 대한 설명입니다.

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$\hat{y}_i$ 1 시차 후 예측

$\bar{y}$ 평균

모형이 계열에 잘 적합되지 않으면 모형 SSE(오차 제곱합)가 SST(총 제곱합)보다 클 수 있습니다. 따라서 R² 이 음수일 수 있습니다.

Adj-R² 다음과 같이 계산되는 조정 R² 값입니다.

$$1 - \left[\frac{(n-1)}{(n-k)} (1 - R^2) \right]$$

MAPE 다음과 같이 계산되는 평균 절대 백분율 오차 값입니다.

$$\frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

여기서,

$\hat{y}_i$ 1 시차 후 예측

MAE 다음과 같이 계산되는 평균 절대 오차 값입니다.

$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

여기서,

$\hat{y}_i$ 1 시차 후 예측

모수 추정값 모형의 각 모수에 대한 추정값, 표준 오차 및 95% 신뢰 구간을 보여 주는 테이블입니다.

예측 시계열에 대한 관측값과 예측값을 모두 보여 주는 그림입니다. 자세한 내용은 "[예측](#)"에서 확인하십시오.

1 시차 후 예측 오차 시간에 따른 1 시차 후 예측 오차에 대한 그림입니다. 오차는 각 시점 t 에 대해 t 에서의 시계열 관측값에서 t 에서의 1 시차 후 예측값을 뺀 값으로 계산됩니다. 홀드백 집합을 지정하면 훈련 데이터에 대해서만 오차가 계산됩니다. 오차에 대한 자기상관 및 편자기상관 보고서도 표시됩니다.

홀드백 계열에 대한 예측 오차 (홀드백 집합을 지정한 경우에만 사용 가능) 홀드백 집합의 관측값에 대한 예측 오차 그림입니다. 오차에 대한 자기상관 및 편자기상관 보고서도 표시됩니다.

성분 상태 모형의 각 상태에 대한 그림과 모형의 해당 계산식을 포함하는 보고서입니다. 각 그림에는 계열의 상태 추정값과 상태 예측이 포함됩니다. 홀드백 집합을 지정한 경우 상태 추정값은 훈련 부분을 대상으로 하고 상태 예측은 홀드백 부분을 대상으로 합니다. 홀드백 집합을 지정하지 않은 경우 상태 추정값은 시계열의 마지막 관측값까지 포함하는 부분을 대상으로 하고 상태 예측은 시계열의 마지막 관측값 이후를 대상으로 합니다. 모형은 수준, 추세 및 계절이라는 최대 세 개의 상태를 가질 수 있습니다.

"성분 상태"의 빨간색 삼각형 메뉴에는 "상태 저장" 옵션이 있습니다. 각 시점의 성분 상태를 별도의 데이터 테이블에 저장하려면 이 옵션을 선택합니다.

승법 오차 모형의 1 시차 후 상대 예측 오차 (승법 오차 모형에만 사용 가능) 시간에 따른 1 시차 후 상대 예측 오차에 대한 그림입니다. 오차는 각 시점 t 에 대해 t 에서의 1 시차 후 예측 오차를 t 에서의 1 시차 후 예측값으로 나누어 계산됩니다.

상태 공간 평활 보고서 옵션

각 상태 공간 평활 보고서에는 다음 옵션이 포함된 빨간색 삼각형 메뉴가 있습니다.

점 표시 예측 그래프에 데이터 점을 표시하거나 숨깁니다.

예측 구간 표시 예측 그래프에 예측 구간을 표시하거나 숨깁니다.

열 저장 모형 결과를 나타내는 열이 포함된 새 데이터 테이블을 생성합니다.

적합 제거 보고서에서 개별 모형 적합을 제거합니다.

전이 함수 모형 보고서

시계열 플랫폼의 각 "전이 함수 모형" 보고서에는 다음 보고서가 포함됩니다.

- 모형 요약
- 모수 추정값
- 잔차
- 홀드백 계열에 대한 예측 오차
- 대화식 예측
- 반복 기록

각 전이 함수 모형이 모형 비교 테이블에 추가됩니다. 모형 비교 테이블에서 전이 함수 모형에 대한 "보고서" 체크박스를 선택하면 "전이 함수 모형" 보고서가 나타납니다. 홀드백 집합이 지정되면 "모형 비교" 보고서에 모형 그림도 제공됩니다.

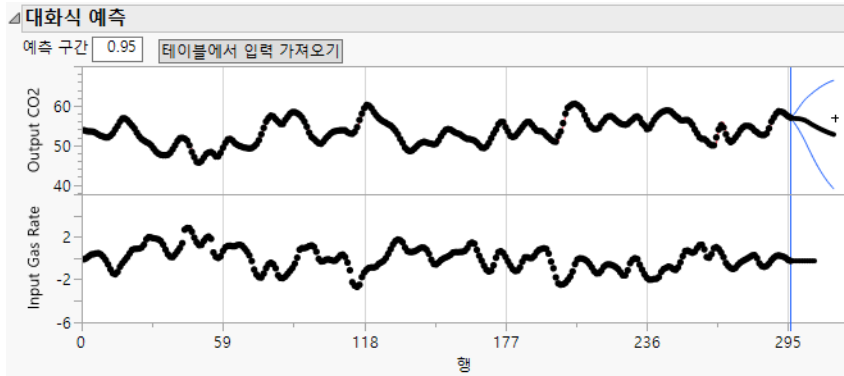
"모형 요약", "모수 추정값", "잔차", "홀드백 계열에 대한 예측 오차" 및 "반복 기록" 보고서의 정보는 ARIMA 및 계절 ARIMA 모형 보고서와 동일합니다. 이러한 보고서에 대한 자세한 내용은 "[모형 보고서](#)"에서 확인하십시오. "모수 추정값" 테이블 다음에는 모형의 계산식이 나옵니다. 여기서 **B**는 후방 연산자입니다.

대화식 예측

"대화식 예측" 보고서에는 "전이 함수 모형 규격" 창에 지정된 예측 구간을 기반으로 예측 그래프가 제공됩니다. 예측 구간은 파란색으로 표시됩니다. 그래프 위의 "예측 구간" 상자에 숫자를 입력하여 이 예측 구간에 대한 신뢰 수준을 변경합니다.

그래프에서 + 기호를 드래그하여 그래프의 예측 주기 수를 변경할 수 있습니다. 예측 주기에서는 "테이블에서 입력 가져오기" 버튼을 사용하거나, 입력 그래프의 점을 다른 값으로 드래그하여 입력 값을 변경할 수 있습니다. 변경 결과는 출력 그래프의 예측 주기에 반영됩니다.

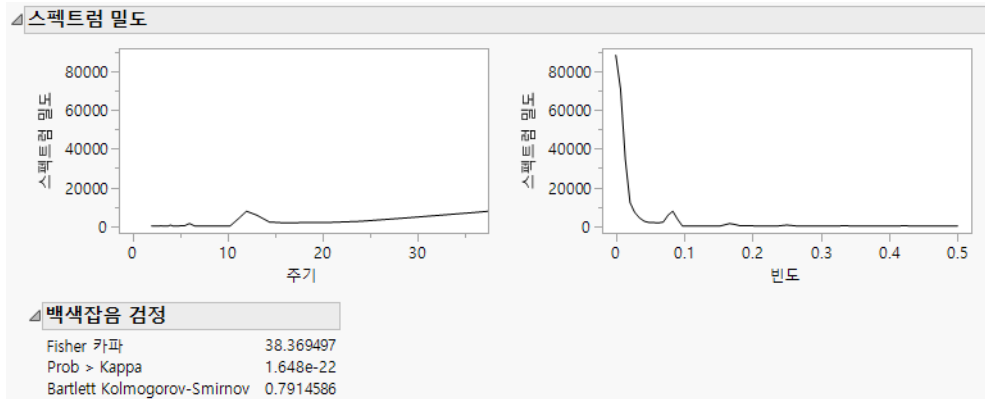
그림 18.14 대화식 예측 그래프



스펙트럼 밀도 보고서

시계열 플랫폼의 "스펙트럼 밀도" 보고서에는 두 개의 스펙트럼 밀도 그림과 "백색잡음 검정" 테이블이 포함되어 있습니다.

그림 18.15 스펙트럼 밀도 그림 및 백색잡음 검정 보고서



" 백색잡음 검정 " 테이블에는 다음 통계량이 포함됩니다 .

Fisher 카파 계열의 값을 분산이 1인 정규 분포에서 가져온다는 귀무가설을 계열에 주기 성분이 있다는 대립가설과 대조하여 검정합니다 . 카파는 주기도의 최대값 $I(f)$ 와 평균 값의 비율입니다 .

Prob > Kappa 귀무가설이 참일 때 카파보다 큰 값을 관측할 확률이며 다음 방정식으로 지정됩니다 .

$$Pr(k > \kappa) = 1 - \sum_{j=0}^q (-1)^j \binom{q}{j} \left[\max\left(1 - \frac{jk}{q}, 0\right) \right]^{q-1}$$

다음은 각 요소에 대한 설명입니다.

$q = N/2$ (N 이 짝수인 경우), $q = (N - 1)/2$ (N 이 홀수인 경우)

κ = 관측된 카파 값

이 확률이 유의 수준 α 보다 작으면 귀무가설이 기각됩니다.

Bartlett Kolmogorov-Smirnov 정규화된 누적 주기도를 (0, 1) 구간에서 균등 분포의 누적 분포 함수와 비교합니다. 검정 통계량은 누적 주기도와 균등 CDF 의 최대 절대 차이와 같습니다. 일반적으로 값이 $a/(\sqrt{q})$ 를 초과하면 계열이 정규 분포를 따른다는 가설을 기각합니다. $a = 1.36$ 및 $a = 1.63$ 값은 각각 5% 와 1% 의 유의 수준에 해당합니다.

시계열 플랫폼의 추가 예

이 섹션에는 시계열 플랫폼을 사용하는 예가 포함되어 있습니다.

- “시간 ID 열 생성의 예”
- “Box-Cox 변환의 예”
- “홀드백 집합을 사용한 예”

시간 ID 열 생성의 예

이 예에서는 시간 ID에 적절한 새 열을 생성한 후 시계열 플랫폼을 사용하여 시계열 분석을 수행합니다.

적절한 시간 ID 열 생성

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Time Series/SeriesP.jmp 를 엽니다.

SeriesP.jmp 데이터 테이블에는 반응이 관측된 기간을 식별하는 Year 열과 Quarter 열이 포함되어 있습니다. 그러나 시계열 플랫폼에서 X 축에 라벨을 지정하려면 일정 간격의 고유한 시점이 포함된 하나의 열이 필요합니다. 시간 ID 를 지정하지 않으면 행 번호를 사용하여 기간을 식별합니다. 이를 방지하고 보고서를 더 쉽게 해석하기 위해 Year 및 Quarter 에서 시간 ID 열을 생성합니다.

2. **열 > 새 열**을 선택합니다. " 열 이름 " 상자에 Year.Quarter 를 입력합니다.
3. " 형식 " 옆에서 **최적 > 날짜 > yyyyQq** 를 선택합니다.
4. **열 특성 > 계산식**을 선택합니다.
5. " 날짜 / 시간 " 옆의 회색 삼각형을 클릭하고 **Informat** 을 클릭합니다.
6. " 문자 " 옆의 회색 삼각형을 클릭합니다.

7. Year 를 선택하고 **Char, Concat** 를 차례로 클릭합니다 .
8. 상자에 "Q"(인용 부호 포함) 를 입력하고 Enter 키를 누릅니다 .
9. **Concat** 를 클릭합니다 .
10. Quarter 를 선택하고 **Char** 을 클릭합니다 .
11. <formatName> 상자를 클릭하고 yyyyQq 를 입력한 후 Enter 키를 누릅니다 .
12. **확인**을 클릭합니다 .

그림 18.16 에 나오는 것과 같은 완료된 " 새 열 " 창이 나타납니다 .

그림 18.16 새 열

'SeriesP' 테이블의 'Year.Quarter'

열 이름: Year.Quarter

☒ 잠금

데이터 유형: 숫자

모델링 유형: 연속형

형식: yyyyQq 너비: 10

입력 형식: yyyyQq

열 특성

계산식

계산식 편집 ☐ 실행 제한 ☐ 오류 무시 ☐

Informat

$$\left(\begin{array}{c} \text{Char}(\text{Year}_{\wedge}) \\ || "Q" \\ || \text{Char}(\text{Quarter}_{\wedge}) \\ \text{yyyyQq} \end{array} \right)_{\wedge}$$

제거

13. **확인**을 클릭합니다 .

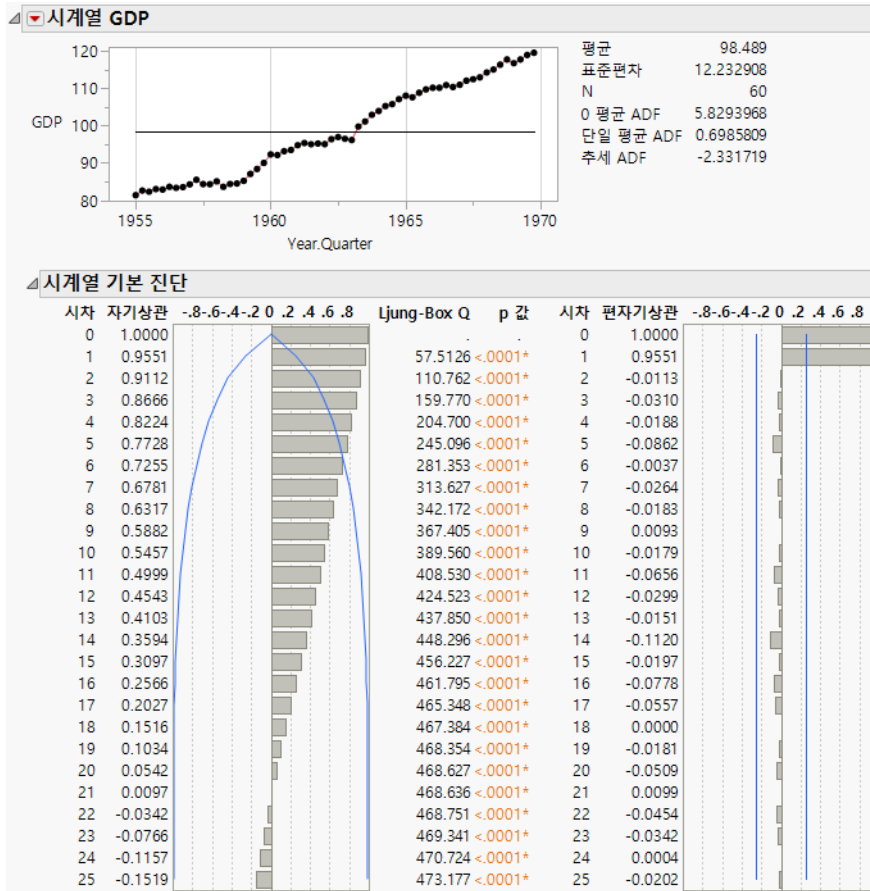
참고 : 이 시간 열은 X11 분석에도 사용됩니다 .

시계열 분석

데이터 테이블에 적절한 시간 ID 열이 포함되었으므로 분석을 진행합니다 .

1. **분석 > 전문 모델링 > 시계열**을 선택합니다 .
2. GDP 를 선택하고 **Y, 시계열**을 클릭합니다 .
3. Year.Quarter 를 선택하고 **X, 시간 ID** 를 클릭합니다 .
4. **확인**을 클릭합니다 .

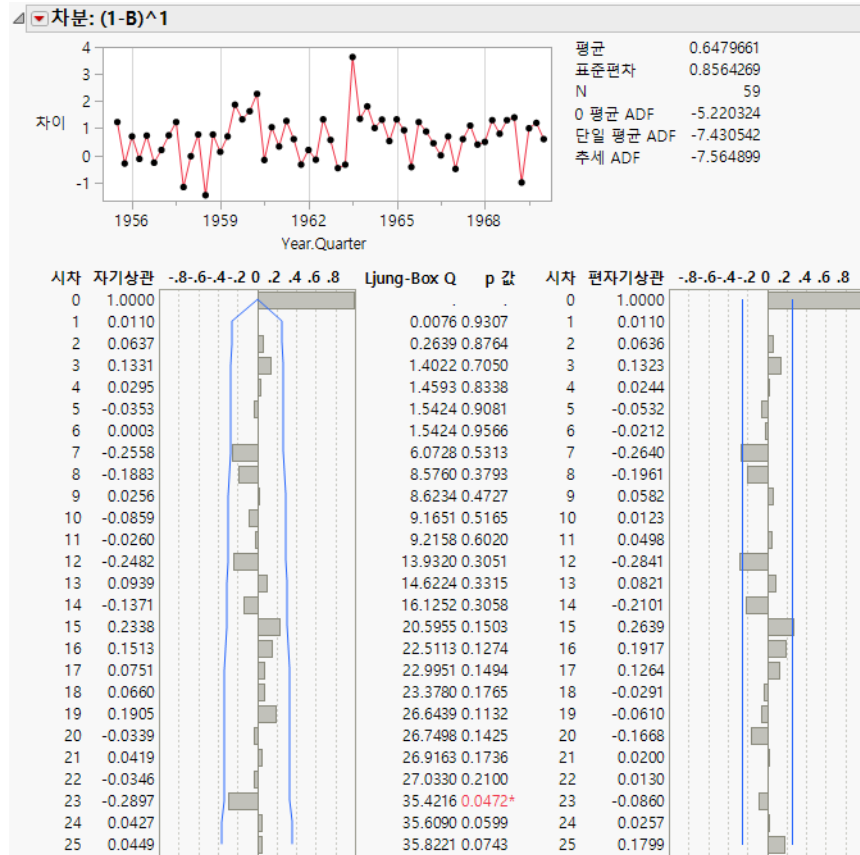
그림 18.17 SeriesPjmp 에 대한 시계열 보고서



계열은 시간이 지남에 따라 선형에 가까운 증가 추세를 보입니다. 또한 자기상관 차트에서는 서로 가까운 점 사이에 강한 상관관계가 있음을 보여 줍니다. 시차가 1, 2, 3 인 점의 자기상관 값은 각각 0.9551, 0.9112, 0.8666 입니다.

- " 시계열 GDP" 의 빨간색 삼각형을 클릭하고 **차분**을 선택합니다.
- " 비계절 차분 차수 " 에 대해 "1" 을 선택하고 **추정**을 클릭합니다.

그림 18.18 SeriesPjmp 에 대한 차분 보고서



"차분" 보고서는 원래 시계열에 적합시킬 적절한 모형을 결정하는 데 도움이 됩니다. 차분 그림에서는 차분 계열에 원래 데이터에서 관측된 추세가 더 이상 없음을 보여 줍니다. 이는 시차 -1 차분이 적절한 선택임을 나타냅니다. 또한 추세를 제거한 후에도 계열에 계절성 징후가 보이지 않습니다. 이러한 이유로 원래 계열을 적합시키는 모형은 선형 추세를 처리할 수 있어야 하지만 계절성을 반드시 처리할 필요는 없습니다. 선형 지수 평활 및 ARIMA 모형이 적합합니다.

7. "시계열 GDP"의 빨간색 삼각형을 클릭하고 **평활 모형 > 선형 지수 평활**을 선택합니다.
8. **추정**을 클릭합니다.
9. "시계열 GDP"의 빨간색 삼각형을 클릭하고 **ARIMA 모형 그룹**을 선택합니다. 이렇게 하면 $(p,d,q)(P,D,Q)$ 의 값 범위에 대해 여러 ARIMA 모형을 적합시킬 수 있습니다.
10. "ARIMA" 상자에서 다음 범위를 설정합니다.
 - 차분 보고서에서 시차 -1 차분이 적절하다는 것을 확인했으므로 1~1 범위를 설정하여 "d, 차분 차수"를 1로 수정합니다.
 - 원래 계열에 자기상관 증거가 보였으므로 "p, 자기회귀 차수"를 0~1 범위로 설정합니다.

- " q , 이동 평균 차수"를 0~1 범위로 설정합니다.

참고: 대부분의 경우 p 와 q 를 작게 유지하면 충분합니다.

- 계열에 계절성 증거가 없었으므로 P, D, Q 는 설정된 값 (0) 그대로 둡니다.
이러한 설정을 통해 총 네 개의 모형이 적합됩니다.

그림 18.19 ARIMA 모형 그룹 규칙

ARIMA 모형 지정

ARIMA

p, 자기회귀 차수 0 1

d, 차분 차수 1 1

q, 이동 평균 차수 0 1

계절 ARIMA

P, 자기회귀 차수 0 0

D, 차분 차수 0 0

Q, 이동 평균 차수 0 0

기간별 관측값 수 12 12

예측 구간 0.95

☒ 절편

☒ 제약 적합

총 모형 수 4

추정 취소 도움말

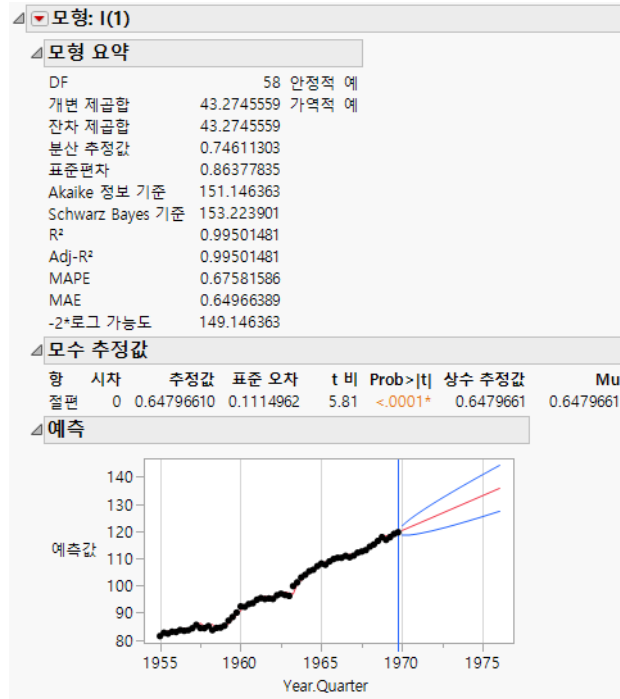
- 추정을 클릭합니다.

그림 18.20 모형 비교 테이블

모형 비교											
보고서	그래프	모형	DF	분산	AIC ^	SBC	R ²	-2*로그 가능도	가중치	.2 .4 .6 .8	MAPE MAE
☒	☐	I(1)	58	0.746113	151.14636	153.22390	0.995	149.14636	0.477865		0.675816 0.649664
☒	☐	ARI(1, 1)	57	0.7591096	153.13924	157.29432	0.995	149.13924	0.176424		0.676216 0.649960
☒	☐	IMA(1, 1)	57	0.7591199	153.14002	157.29510	0.995	149.14002	0.176355		0.676179 0.649933
☒	☐	ARIMA(1, 1, 1)	56	0.748633	153.91882	160.15143	0.995	147.91882	0.119474		0.668189 0.641981
☒	☐	선형 (Holt) 지수 평활	56	0.7878954	155.66571	159.78660	0.994	151.66571	0.049882		0.716452 0.686737

모형 비교 테이블은 AIC 기준에 따라 최량 적합 모형이 목록의 맨 위에 오도록 정렬됩니다. 이 경우 보고서에 "I(1)"로 표시된 ARIMA(0,1,0) 모형이 원래 시계열에 가장 적합합니다. 또한 I(1) 모형이 "최량" 모형이지만 모든 모형의 적합 통계량 값이 매우 비슷하다는 점도 유의해야 합니다. 모두 적절하다고 간주할 수 있습니다.

그림 18.21 ARIMA(0,1,0) 에 대한 모형 보고서



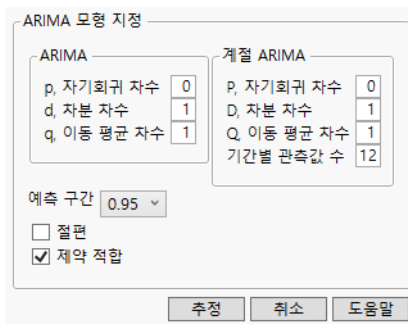
I(1) 모형 보고서에는 예측 그래프가 표시됩니다. 파란색 선은 예측 구간을 나타냅니다. GDP는 선형 비율로 계속 증가한다고 예측됩니다.

Box-Cox 변환의 예

이 예에서는 두 개의 시계열 분석을 비교합니다. 한 분석에서는 Box-Cox 변환을 사용하고 다른 분석에서는 수동으로 변환된 시계열을 분석합니다. 람다 값을 0으로 설정하면 Box-Cox 변환은 근사 로그 변환입니다. 그러나 분석에 Box-Cox 변환을 사용하면 모든 예측이 다시 변환되어 원래 척도로 보고되므로 결과를 더 쉽게 해석할 수 있습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Time Series/Seriesg.jmp 를 엽니다 .
2. **분석 > 전문 모델링 > 시계열**을 선택합니다 .
3. Passengers 를 선택하고 **Y, 시계열**을 클릭합니다 .
4. Time 을 선택하고 **X, 시간 ID** 를 클릭합니다 .
5. **Box-Cox 변환 사용**을 선택합니다 .
람다에 기본값 0 을 사용합니다 .
6. **확인**을 클릭합니다 .

7. " 시계열 Box-Cox 변환 Passengers" 옆의 빨간색 삼각형을 클릭하고 **계절 ARIMA** 를 선택합니다.
8. " 계절 ARIMA 규격 " 창의 "ARIMA" 섹션에서 다음 모수를 설정합니다.
 - " p , 자기회귀 차수 "를 "0" 으로 설정합니다.
 - " d , 차분 차수 "를 "1" 로 설정합니다.
 - " q , 이동 평균 차수 "를 "1" 로 설정합니다.
9. " 계절 ARIMA 규격 " 창의 " 계절 ARIMA" 섹션에서 다음 모수를 설정합니다.
 - " P , 자기회귀 차수 "를 "0" 으로 설정합니다.
 - " D , 차분 차수 "를 "1" 로 설정합니다.
 - " Q , 이동 평균 차수 "를 "1" 로 설정합니다.
10. " 절편 " 을 선택 취소하여 절편이 없는 모형을 실행합니다.
현재 창이 [그림 18.22](#) 에 표시된 창과 일치해야 합니다.

그림 18.22 계절 ARIMA 규격


ARIMA 모델 지정

ARIMA		계절 ARIMA	
p , 자기회귀 차수	0	P , 자기회귀 차수	0
d , 차분 차수	1	D , 차분 차수	1
q , 이동 평균 차수	1	Q , 이동 평균 차수	1
		기간별 관측값 수	12

예측 구간 0.95 ▾

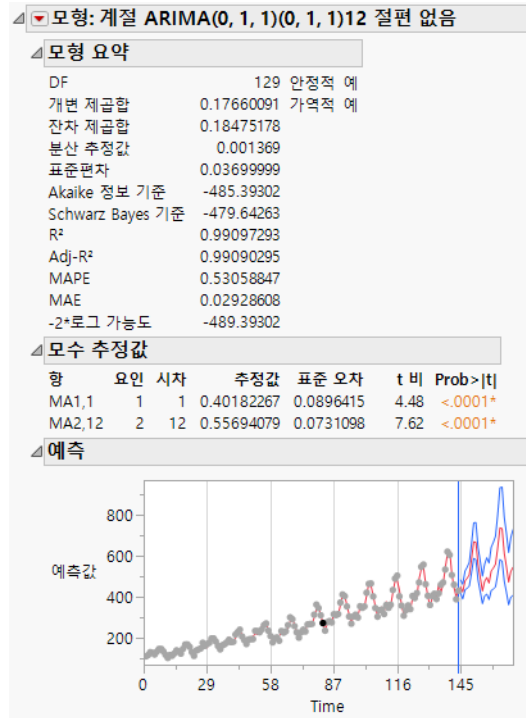
☐ 절편

☒ 제약 적합

추정 취소 도움말

11. **추정**을 클릭합니다.

그림 18.23 Box-Cox 변환 Passengers 데이터에 대한 계절 ARIMA 모형



모수 추정값은 Log Passengers 열에 대해 동일한 모형을 실행하여 얻은 값과 동일합니다. 그러나 [그림 18.23](#)의 예측 및 예측 구간은 원래 척도로 표시됩니다.

홀드백 집합을 사용한 예

이 시계열 예에서는 홀드백 집합을 사용하여 월별 납 생산 데이터에 대한 예측을 평가합니다. 홀드백 집합이 사용되므로 상태 공간 평활 모형과 ARIMA 모형을 모두 적합시키고 비교하여 데이터에 대한 최량 적합 모형을 찾을 수 있습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Time Series/Lead Production.jmp 를 엽니다.
2. **분석 > 전문 모델링 > 시계열**을 선택합니다.
3. LEADPROD 를 선택하고 **Y, 시계열**을 클릭합니다.
4. DATE 를 선택하고 **X, 시간 ID**를 클릭합니다.
5. **홀드백에 대한 예측**을 선택합니다.

이 옵션을 선택하면 미래 관측값 대신 홀드백 집합에 대한 예측이 실행됩니다.

6. "예측 주기" 옆에 "12"를 입력합니다.

이렇게 하면 데이터의 마지막 12개 관측값이 홀드백 집합에 할당됩니다. 월별 데이터이므로 이 값은 작년에 해당합니다.

7. **확인**을 클릭합니다.
8. "시계열 LEADPROD"의 빨간색 삼각형을 클릭하고 **상태 공간 평활**을 선택합니다.
9. "주기" 옆에 "12"를 입력합니다.
10. "모수 제한"을 선택합니다.
11. **확인**을 클릭합니다.
12. "시계열 LEADPROD"의 빨간색 삼각형을 클릭하고 **ARIMA 모형 그룹**을 선택합니다.
13. 모든 모수의 범위를 0~1로 설정합니다. "ARIMA 모형 지정" 창은 [그림 18.24](#)에 나오는 것과 같습니다.

그림 18.24 ARIMA 모형 그룹 규격

ARIMA 모형 지정

ARIMA

p, 자기회귀 차수

0	1

d, 차분 차수

0	1

q, 이동 평균 차수

0	1

계절 ARIMA

P, 자기회귀 차수

0	0

D, 차분 차수

0	0

Q, 이동 평균 차수

0	0

기간별 관측값 수

12	12

예측 구간 0.95

☒ 절편 ☒ 제약 적합

총 모형 수 8

추정
취소
도움말

14. **추정**을 클릭합니다.

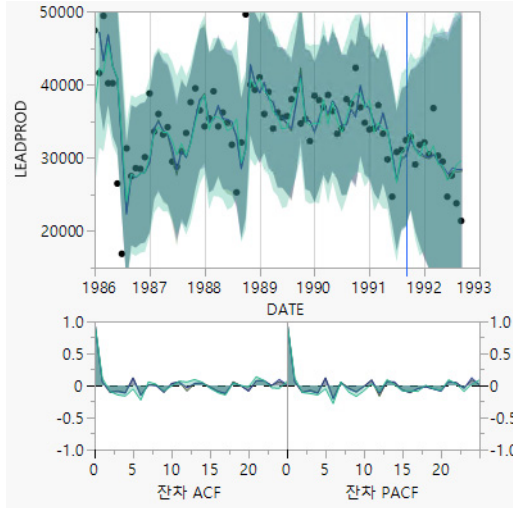
그림 18.25 모형 비교

모형 비교				홀드백 평가			
보고서	그래프	모형 계층	모형	RMSE ^	MSE	MAPE	MAE
☒	☐	ARIMA	계절 ARIMA(0, 1, 1)(1, 0, 0)12	3219.4131	10364621	8.900164	2416.3319
☐	☐	ARIMA	계절 ARIMA(0, 1, 1)(0, 0, 1)12	3316.8746	11001657	9.148390	2456.0873
☒	☐	상태 공간 평활	AAA12 제약 있음	3350.9303	11228734	8.623519	2245.5657
☒	☐	ARIMA	계절 ARIMA(1, 1, 0)(1, 0, 0)12	3397.0855	11540190	9.082427	2392.5823
☐	☐	ARIMA	계절 ARIMA(0, 1, 1)(1, 1, 0)12	3448.6372	11893099	9.829459	2649.8142
☐	☐	ARIMA	계절 ARIMA(1, 1, 0)(1, 1, 0)12	3484.3764	12140879	9.708056	2604.6290
☒	☐	ARIMA	계절 ARIMA(1, 1, 0)(0, 0, 1)12	3491.2648	12188930	9.302681	2435.7441
☐	☐	상태 공간 평활	MAM12 제약 있음	3512.1513	12335206	10.552196	2813.4655
☒	☐	ARIMA	계절 ARIMA(0, 1, 0)(1, 0, 0)12	3525.0858	12426230	9.374006	2440.6892
☒	☐	ARIMA	계절 ARIMA(0, 1, 0)(1, 1, 0)12	3551.3973	12612423	9.367947	2480.3210
☐	☐	ARIMA	계절 ARIMA(0, 1, 1)(1, 0, 1)12	3579.0242	12809414	10.520274	2739.9652
☒	☐	ARIMA	계절 ARIMA(0, 1, 0)(0, 1, 0)12	3597.3098	12940638	11.622868	3219.0476
☒	☐	ARIMA	계절 ARIMA(0, 1, 0)(0, 0, 1)12	3631.8176	13190099	9.792052	2544.4269
☒	☐	ARIMA	계절 ARIMA(1, 1, 0)(0, 1, 0)12	3661.8395	13409068	12.030672	3366.0935
☒	☐	ARIMA	계절 ARIMA(0, 1, 1)(0, 1, 0)12	3673.6369	13495608	12.063523	3379.4644
☒	☐	ARIMA	계절 ARIMA(1, 1, 0)(1, 0, 1)12	3700.6621	13694900	11.077791	2890.6426
☐	☐	ARIMA	계절 ARIMA(0, 1, 0)(1, 0, 1)12	3723.7664	13866437	11.222827	2935.8075

상태 공간 평활 모형과 ARIMA 모형이 홀드백 집합에 대해 평가되므로 두 모형의 적합 통계량을 비교할 수 있습니다.

15. "모형 비교" 보고서에서 처음 세 모형 옆의 "그래프" 체크박스를 선택합니다.

그림 18.26 모형 비교 그래프



평가 통계량과 예측 그림을 보면 상위 모형 간의 예측 성능에 큰 차이가 없음을 알 수 있습니다. 보고서에 따르면 ARIMA 모형 또는 상태 공간 평활 모형이 이 데이터에 적절합니다.

시계열 플랫폼에 대한 통계 상세 정보

이 섹션에는 시계열 플랫폼에 대한 통계 상세 정보가 포함되어 있습니다.

- “스펙트럼 밀도에 대한 통계 상세 정보”
- “X-11 분해에 대한 통계 상세 정보”
- “지수 평활 모형에 대한 통계 상세 정보”
- “ARIMA 모형에 대한 통계 상세 정보”
- “전이 함수에 대한 통계 상세 정보”

스펙트럼 밀도에 대한 통계 상세 정보

시계열 플랫폼에서 Fourier 계열의 계수에 대한 최소 제곱 추정값은 다음과 같이 계산됩니다.

$$a_t = \frac{2}{N} \sum_{i=1}^N y_t \cos(2\pi f_i t)$$

$$b_t = \frac{2}{N} \sum_{i=1}^N y_t \sin(2\pi f_i t)$$

그런 다음 $f_i = i/N$ 가 결합되어 빈도 f_i 에서의 강도를 나타내는 주기도 $I(f_i) = (N/2)(a_i^2 + b_i^2)$ 를 구성합니다.

그런 다음 주기도가 평활되고 $1/(4\pi)$ 로 척도가 조정되어 스펙트럼 밀도를 구성합니다.

X-11 분해에 대한 통계 상세 정보

시계열 플랫폼에서 X-11 방법은 승법 또는 가법 분해를 사용하여 원래 시계열을 조정합니다. 세 가지 X-11 성분(추세 순환, 계절, 불규칙)을 추정하기 위해 반복 프로세스를 사용하여 모형이 적합됩니다. 추세 순환 성분은 장기 추세 및 장기 순환 효과를 모두 포함합니다. 불규칙 성분은 추세 성분과 계절 성분으로 설명되지 않는 변동 효과를 포함합니다. 과거 X-11 방법 개발에 대한 개요를 보려면 **SAS/ETS 15.2 User's Guide**("Historical Development of X-11" 검색)를 참조하십시오.

승법 조정은 다음 모형을 적합시킵니다.

$$O_t = C_t \cdot S_t \cdot I_t$$

다음은 각 요소에 대한 설명입니다.

O_t = 원래 시계열

C_t = 추세 순환 성분

S_t = 계절 성분

I_t = 불규칙 성분

조정 승법 추세는 O_t/S_t 입니다.

가법 조정은 다음 모형을 적합시킵니다.

$$O_t = C_t + S_t + I_t$$

조정 가법 추세는 $O_t - S_t$ 입니다.

지수 평활 모형에 대한 통계 상세 정보

평활 모형은 다음과 같이 정의됩니다.

$$y_t = \mu_t + \beta_t t + s(t) + a_t$$

다음은 각 요소에 대한 설명입니다.

μ_t = 시간에 따라 변하는 평균 항

β_t = 시간에 따라 변하는 기울기 항

$s(t)$ = 시간에 따라 변하는 계절 항 중 하나

a_t = 확률 충격

추세가 없는 모형은 $\beta_t = 0$ 이고 비계절 모형은 $s(t) = 0$ 입니다 . 시간에 따라 변하는 이러한 항의 추정량은 다음과 같이 정의됩니다 .

$L_t = \mu_t$ 를 추정하는 평활 수준

$T_t = \beta_t$ 를 추정하는 평활 추세

$S_{t-j}(j=0, 1, \dots, s-1) = s(t)$ 의 추정값

각 평활 모형은 이러한 추정량의 진화를 설명하는 재귀 평활 방정식 집합을 정의합니다 . 평활 방정식은 **평활 가중치**라고 하는 다음 모형 모수로 작성됩니다 .

α = 수준 평활 가중치

γ = 추세 평활 가중치

φ = 진폭감소 추세 가중치

δ = 계절 평활 가중치

이러한 모수는 각 모형에 서로 다른 방식으로 입력(또는 전혀 입력되지 않음)되지만 가중치가 클수록 최근 데이터에 영향을 많이 주고 가중치가 작을수록 최근 데이터에 영향을 적게 준다는 공통된 특성이 있습니다 . 평활 가중치에 대한 자세한 내용은 **SAS/ETS 15.2 User's Guide**("Forecasting Process Details" 장)에서 확인하십시오 .

단순 지수 평활

단순 지수 평활 모형은 $y_t = \mu_t + \alpha_t$ 입니다 .

평활 방정식 $L_t = \alpha y_t + (1 - \alpha)L_{t-1}$ 은 단일 평활 가중치 α 로 정의됩니다 . 이 모형은 다음 조건이 참일 때 **ARIMA(0, 1, 1)** 모형과 동일합니다 .

$$(1 - B)y_t = (1 - \theta B)\alpha_t \text{ 여기서 } \theta = 1 - \alpha$$

모형의 이동 평균 형식은 다음과 같이 정의됩니다 .

$$y_t = a_t + \sum_{j=1}^{\infty} \alpha a_{t-j}$$

이중 (Brown) 지수 평활

이중 지수 평활 모형은 $y_t = \mu_t + \beta_1 t + a_t$ 입니다 .

평활 방정식은 단일 평활 가중치 α 를 사용하여 다음과 같이 정의됩니다.

$$L_t = \alpha y_t + (1 - \alpha)L_{t-1}$$

$$T_t = \alpha(L_t - L_{t-1}) + (1 - \alpha)T_{t-1}$$

이 모형은 다음 조건이 참일 때 ARIMA(0, 1, 1)(0, 1, 1)₁ 모형과 동일합니다.

$$(1 - B)^2 y_t = (1 - \theta B)^2 a_t \text{ 여기서 } \theta_{1,1} = \theta_{2,1}, \theta = 1 - \alpha$$

모형의 이동 평균 형식은 다음과 같이 정의됩니다.

$$y_t = a_t + \sum_{j=1}^{\infty} (2\alpha + (j-1)\alpha^2) a_{t-j}$$

선형 (Holt) 지수 평활

선형 지수 평활 모형은 $y_t = \mu_t + \beta t + a_t$ 입니다.

평활 방정식은 α 및 γ 평활 가중치를 사용하여 다음과 같이 정의됩니다.

$$L_t = \alpha y_t + (1 - \alpha)(L_{t-1} + T_{t-1})$$

$$T_t = \gamma(L_t - L_{t-1}) + (1 - \gamma)T_{t-1}$$

이 모형은 다음 조건이 참일 때 ARIMA(0, 2, 2) 모형과 동일합니다.

$$(1 - B)^2 y_t = (1 - \theta B - \theta_2 B^2) a_t \text{ 여기서 } \theta = 2 - \alpha - \alpha\gamma, \theta_2 = \alpha - 1$$

모형의 이동 평균 형식은 다음과 같이 정의됩니다.

$$y_t = a_t + \sum_{j=1}^{\infty} (\alpha + j\alpha\gamma) a_{t-j}$$

진폭감소 추세 선형 지수 평활

진폭감소 추세 선형 지수 평활 모형은 $y_t = \mu_t + \beta t + a_t$ 입니다.

평활 방정식은 α, γ 및 ϕ 평활 가중치를 사용하여 다음과 같이 정의됩니다.

$$L_t = \alpha y_t + (1 - \alpha)(L_{t-1} + \phi T_{t-1})$$

$$T_t = \gamma(L_t - L_{t-1}) + (1 - \gamma)\phi T_{t-1}$$

이 모형은 다음 조건이 참일 때 ARIMA(1, 1, 2) 모형과 동일합니다.

$$(1 - \phi B)(1 - B)y_t = (1 - \theta_1 B - \theta_2 B^2) a_t$$

다음은 각 요소에 대한 설명입니다 .

$$\theta_1 = 1 + \phi - \alpha - \alpha\gamma\phi$$

$$\theta_2 = (\alpha - 1)\phi$$

모형의 이동 평균 형식은 다음과 같이 정의됩니다 .

$$y_t = \alpha_t + \sum_{j=1}^{\infty} \left(\frac{\alpha + \alpha\gamma\phi(\phi^j - 1)}{\phi - 1} \right) \alpha_{t-j}$$

계절 지수 평활

계절 지수 평활 모형은 $y_t = \mu_t + s(t) + a_t$ 입니다.

평활 방정식은 α 및 δ 평활 가중치를 사용하여 다음과 같이 정의됩니다 .

$$L_t = \alpha(y_t - S_{t-s}) + (1 - \alpha)L_{t-1}$$

$$S_t = \delta(y_t - L_{t-s}) + (1 - \delta)S_{t-s}$$

이 모형은 계절 ARIMA(0, 1, s+1)(0, 1, 0)_s 모형과 동일합니다 .

$$(1-B)(1-B^s)y_t = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^{s+1})a_t$$

다음은 각 요소에 대한 설명입니다 .

$$\theta_1 = 1 - \alpha$$

$$\theta_2 = (1 - \delta)(1 - \alpha)$$

$$\theta_3 = (1 - \alpha)(\delta - 1)$$

모형의 이동 평균 형식은 다음과 같이 정의됩니다 .

$$y_t = a_t + \sum_{j=1}^{\infty} \psi_j a_{t-j} \text{ 여기서, } \psi = \begin{cases} \alpha \text{ 단, } j \bmod s \neq 0 \\ \alpha + \delta(1 - \alpha) \text{ 단, } j \bmod s = 0 \end{cases}$$

Winters 방법 (가법 모형)

Winters 방법의 가법 버전 모형은 $y_t = \mu_t + \beta_t t + s(t) + a_t$ 입니다.

평활 방정식은 α , γ 및 δ 가중치를 사용하여 다음과 같이 정의됩니다 .

$$L_t = \alpha(y_t - S_{t-s}) + (1 - \alpha)(L_{t-1} + T_{t-1})$$

$$T_t = \gamma(L_t - L_{t-1}) + (1 - \gamma)T_{t-1}$$

$$S_t = \delta(y_t - L_t) + (1 - \delta)S_{t-s}$$

이 모형은 다음과 같이 정의된 계절 ARIMA(0, 1, s+1)(0, 1, 0)s 모형과 동일합니다.

$$(1-B)(1-B^2)y_t = \left(1 - \sum_{i=1}^{s+1} \theta_i B^i\right) a_t$$

모형의 이동 평균 형식은 다음과 같이 정의됩니다.

$$y_t = a_t + \sum_{j=1}^{\infty} \Psi_j a_{t-j}$$

다음은 각 요소에 대한 설명입니다.

$$\Psi = \begin{cases} \alpha + j\alpha\gamma, & j \bmod s \neq 0 \\ \alpha + j\alpha\gamma + \delta(1 - \alpha), & j \bmod s = 0 \end{cases}$$

ARIMA 모형에 대한 통계 상세 정보

시계열 플랫폼에서는 ARIMA 모형과 계절 ARIMA 모형을 모두 적합시킵니다.

ARIMA 모형

반응 계열 $\{y_t\}$ 에 대한 ARIMA 모형의 일반 형식은 다음과 같이 정의됩니다.

$$\phi(B)(w_t - \mu) = \theta(B)a_t$$

다음은 각 요소에 대한 설명입니다.

t = 시간 인덱스

$B = By_t = y_{t-1}$ 로 정의된 후방 연산자

$w_t = (1 - B)^d y_t$ = 차분 후의 반응 계열

μ = 절편 또는 평균 항

$\phi(B)$ 및 $\theta(B)$ = 각각 자기회귀 연산자와 이동 평균 연산자이며 다음과 같이 정의됩니다.

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$$

다음은 각 요소에 대한 설명입니다.

a_t = 확률 충격 시퀀스

a_t 는 평균이 0 이고 분산이 일정한 독립 정규 분포를 따른다고 가정합니다 .

모형을 다음과 같이 다시 작성할 수 있습니다 .

$$\phi(B)w_t = \delta + \theta(B)a_t$$

상수 추정값 δ 는 다음과 같은 관계식으로 제공됩니다 .

$$\delta = \phi(B)\mu = \mu - \phi_1\mu - \phi_2\mu - \dots - \phi_p\mu$$

계절 ARIMA 모형

계절 ARIMA 모델링의 경우 차분 , 자기회귀 및 이동 평균 연산자는 계절 다항식과 비계절 다항식의 곱입니다 .

$$w_t = (1-B)^d(1-B^s)^D y_t$$

$$\phi(B) = (1 - \phi_{1,1}B - \phi_{1,2}B^2 - \dots - \phi_{1,p}B^p)(1 - \phi_{2,s}B^s - \phi_{2,2s}B^{2s} - \dots - \phi_{2,Ps}B^{Ps})$$

$$\theta(B) = (1 - \theta_{1,1}B - \theta_{1,2}B^2 - \dots - \theta_{1,q}B^q)(1 - \theta_{2,s}B^s - \theta_{2,2s}B^{2s} - \dots - \theta_{2,Qs}B^{Qs})$$

여기서 s 는 기간별 관측값 수입니다. 계수의 첫 번째 인덱스는 요인 번호(1 - 비계절, 2 - 계절)이고 두 번째 인덱스는 항의 시차입니다.

전이 함수에 대한 통계 상세 정보

m 개의 입력이 있는 일반적인 전이 함수 모형은 다음과 같이 나타낼 수 있습니다 .

$$Y_t - \mu = \frac{\omega_1(B)}{\delta_1(B)}X_{1,t-d1} + \dots + \frac{\omega_m(B)}{\delta_m(B)}X_{m,m-dm} + \frac{\theta(B)}{\phi(B)}e_t$$

다음은 각 요소에 대한 설명입니다 .

Y_t = 출력 계열

$X_1 \sim X_m = m$ 개의 입력 계열

e_t = 잡음 계열

$X_{1,t-d1} = X_1$ 계열이 $d1$ 단계 시차의 t 에 의해 인덱싱됨

μ = 모형의 평균 수준

$\phi(B)$ 및 $\theta(B)$ = ARIMA 모형의 자기회귀 다항식 및 이동 평균 다항식

$\omega_k(B)$ 및 $\delta_k(B)$ = 개별 전이 함수의 분자 및 분모 요인 (다항식). 여기서, k 는 1 ~ m 개의 개별 입력에 대한 인덱스를 나타냅니다 .

위 모형의 각 다항식은 계절 ARIMA 에서와 같이 두 부분(비계절과 계절 또는 두 성분의 곱)을 포함할 수 있습니다. 모형을 지정할 때 모형에 포함하지 않을 부분에는 기본값 0을 그대로 둡니다.

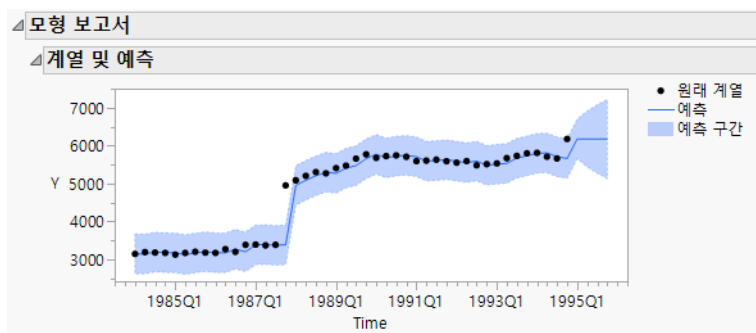
19 장

시계열 예측 다중 시계열 예측

시계열 예측 플랫폼을 사용하면 계열 수가 매우 많은 다중 시계열을 예측할 수 있습니다. 이 플랫폼에서는 다중 시계열에 대한 간단한 요약 표시하고, 지수 평활족에서 최대 30 개의 서로 다른 예측 모델을 효율적으로 적합시킬 수 있는 옵션을 제공합니다. 최량 적합 모델이 자동으로 선택 됩니다.

시계열 예측 플랫폼은 지수 평활 모델 집합에서 최량 적합 모델을 자동으로 선택하기 위해 만들어졌습니다. 이는 최량 적합 모델을 수동으로 선택해야 하고 ARIMA 모델을 사용하는 시계열 플랫폼과 대조적입니다.

그림 19.1 원래 계열 및 예측



목차

시계열 예측의 예	373
시계열 예측 플랫폼 시작	375
데이터 형식	375
시계열 예측 보고서	376
시간 패턴 분석	376
시간에 대한 그룹 요약	377
모델링 규격	378
시계열 예측 플랫폼 옵션	381
시계열 예측 모형 보고서	382

시계열 예측의 예

시계열 예측 플랫폼을 사용하여 여러 시계열이 있는 데이터에 다양한 상태 공간 평활 모델을 적합시킵니다. 홀드백 집합을 사용하는 "예측 성능" 옵션을 통해 최종 모형이 선택됩니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Time Series/M3C Quarterly.jmp 를 엽니다 .
2. **분석 > 전문 모델링 > 시계열 예측**을 선택합니다 .
3. Y 를 선택하고 **Y** 를 클릭합니다 .
4. **Series** 를 선택하고 **그룹화**를 클릭합니다 .
5. **Time** 을 선택하고 **시간**을 클릭합니다 .
6. **확인**을 클릭합니다 .

초기 " 시계열 예측 " 보고서가 나타납니다. 여기에는 시간 및 Y 변수에 대한 요약이 포함되어 있습니다. 예측 모형 적합 옵션을 제공하는 " 모델링 규격 " 보고서도 있습니다.

그림 19.2 시계열 예측의 초기 보고서

시계열 예측

시간 패턴 분석

날짜 분석

분기별 데이터입니다.

암시된 계절성 4

시간에 대한 그룹 요약

Series	행 수	결측값	고유 타임스탬프	시작	끝	결측 레코드
N 646	44	아니요	예	1984Q1	1994Q4	아니요
N 647	44	아니요	예	1984Q1	1994Q4	아니요
N 648	46	아니요	예	1983Q3	1994Q4	아니요
N 649	44	아니요	예	1984Q1	1994Q4	아니요
N 650	44	아니요	예	1984Q1	1994Q4	아니요
N 651	44	아니요	예	1984Q1	1994Q4	아니요
N 652	43	아니요	예	1984Q1	1994Q3	아니요
N 653	43	아니요	예	1984Q1	1994Q3	아니요
N 654	45	아니요	예	1983Q4	1994Q4	아니요
N 655	44	아니요	예	1984Q1	1994Q4	아니요

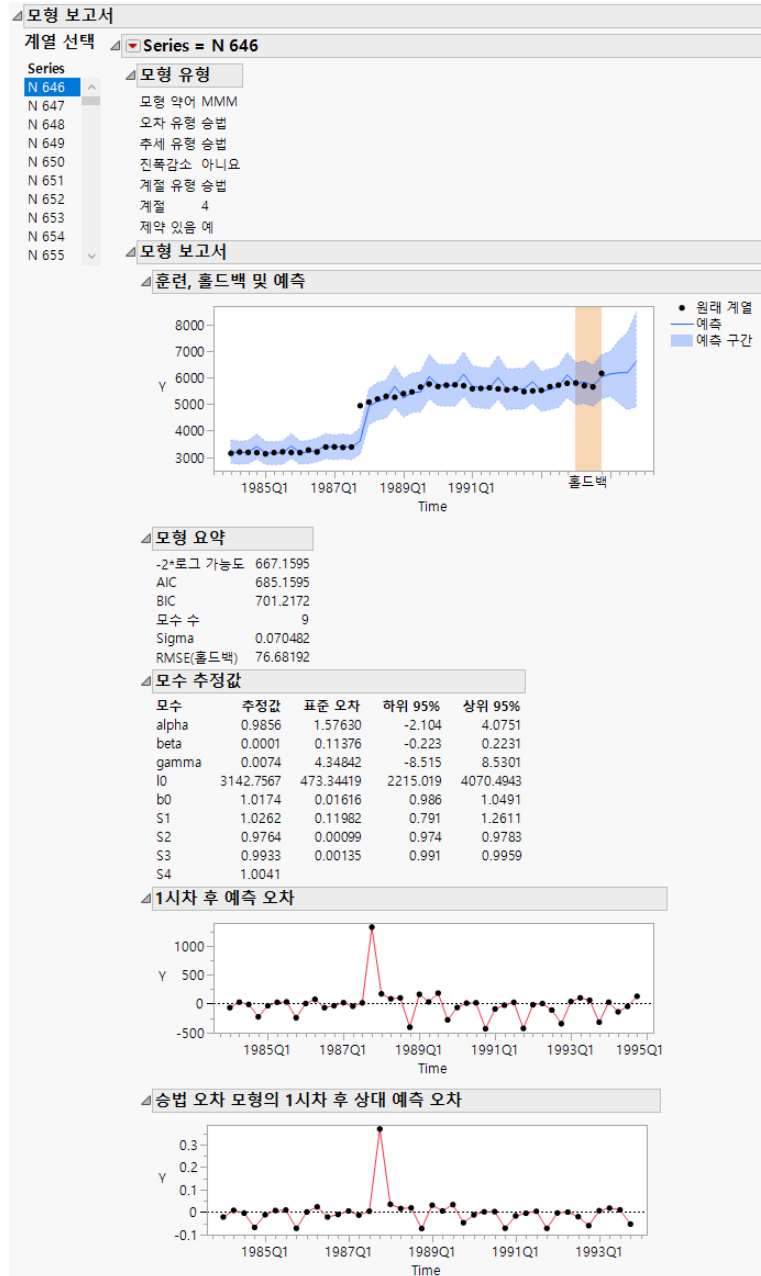
Y에 대한 그룹 요약

Series	최소값	최대값	길이	관측값 수	결측값 수	모두 양수	주석
N 646	3124.38	6176.6	44	44	0	예	
N 647	1522	4320.8	44	44	0	예	
N 648	1549.72	5970	46	46	0	예	
N 649	2136.7	5023.2	44	44	0	예	
N 650	875.51	6908	44	44	0	예	
N 651	7512.4	8526.2	44	44	0	예	
N 652	4861.63	9913.57	43	43	0	예	
N 653	2797	7202	43	43	0	예	
N 654	2372.74	6735.1	45	45	0	예	
N 655	2449.06	8904.28	44	44	0	예	

7. " 모델링 규격 " 보고서에서 " 완전한 규격 " 탭을 클릭합니다 .
8. " 모형 선택 전략 " 아래의 메뉴를 클릭하고 " 예측 성능 " 을 선택합니다 .
9. **실행**을 클릭합니다 .

이렇게 하면 권장 모형 집합을 데이터에 적합시킵니다. 그룹화 변수의 각 수준에 대한 최적 모형이 선택되고 "모형 보고서"에 보고됩니다. 자세한 내용은 "모형 보고서"에서 확인하십시오.

그림 19.3 모형 보고서

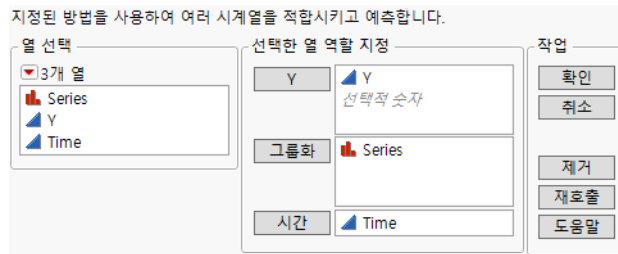


N646 계열에 대한 모형 보고서가 표시됩니다. 이것은 "계열 선택" 목록에서 선택한 계열 이름에 해당합니다. 아래쪽 화살표 키를 사용하여 각 계열에 대한 모형 보고서를 스캔할 수 있습니다. 또는 목록에서 다른 이름을 클릭하여 선택한 계열에 대한 모형 보고서를 표시할 수도 있습니다.

시계열 예측 플랫폼 시작

분석 > 전문 모델링 > 시계열 예측을 선택하여 시계열 예측 플랫폼을 시작합니다.

그림 19.4 시계열 예측 시작 창



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

시계열 예측 플랫폼 시작 창에는 다음 옵션이 포함되어 있습니다.

- Y** 하나 이상의 열을 시계열 변수로 할당합니다. 다중 시계열을 테이블의 개별 열로 구성하거나, 하나 이상의 그룹화 변수를 사용하여 단일 열에 쌓을 수 있습니다.
- 그룹화** 하나 이상의 열을 그룹화 변수로 할당합니다. 그룹화 변수는 쌓인 형식의 개별 시계열을 식별하는 데 사용됩니다.
- 시간** 시간 변수에 대한 열을 할당합니다. 개별 시계열이 개별 열로 구성된 경우 시간 열의 타임 스탬프에 중복 항목이 포함되지 않아야 합니다.

데이터 형식

시계열 예측 플랫폼을 사용하는 경우 다중 시계열을 테이블의 개별 열로 구성하거나, 개별 계열을 식별하기 위한 하나 이상의 그룹화 변수를 사용하여 단일 열에 쌓을 수 있습니다. 여러 Y 열과 여러 그룹화 열을 조합할 수도 있습니다.

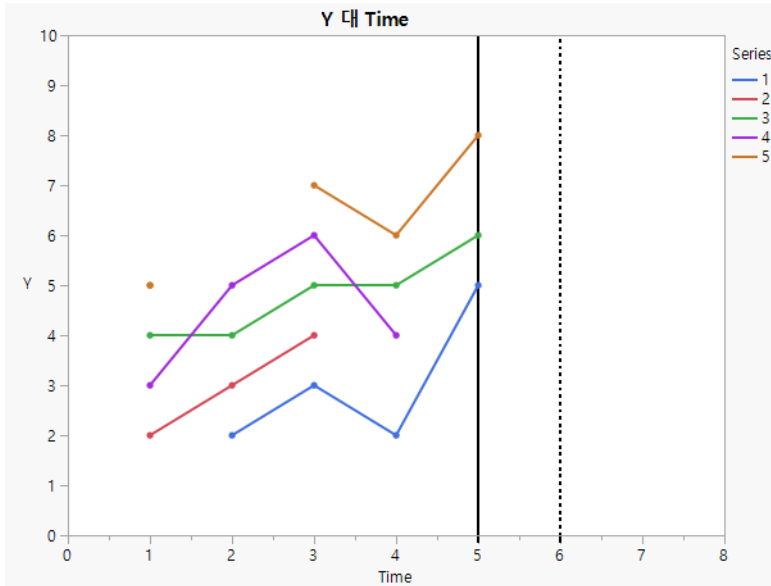
이 플랫폼에서는 데이터에 대해 다음과 같이 가정합니다.

- 모든 시계열은 동일한 시간 해상도로 기록됩니다.
- 모든 시계열은 동일한 최종 관측 시점까지 수집됩니다. 이 최종 관측 시점은 개별 시계열에서 발생하는 맨 마지막 시점으로 정의됩니다. 최종 관측 시점까지 데이터가 없는 계열은 결측값이 있는 것으로 가정합니다. 예를 들어 **그림 19.5**에서 최종 관측 시점은 검은색 실선으로

로 표시된 5 입니다. 계열 2 는 시점 4 와 5 에 결측 관측값이 있고 계열 4 는 시점 5 에 결측 관측값이 있다고 간주됩니다.

- 모든 예측은 동일한 시점에서 시작합니다. 이 시작 시점은 최종 관측 시점 이후의 다음 시점입니다. 예를 들어 [그림 19.5](#) 에서 최종 관측 시점은 검은색 실선으로 표시된 5 입니다. 따라서 모든 예측은 검은색 파선으로 표시된 시점 6 에서 시작합니다.

그림 19.5 가정 데모



시계열 예측 보고서

"시계열 예측" 보고서에는 시작 창에 지정된 변수를 기반으로 한 요약 테이블이 포함됩니다. 그룹별 시계열 요약 테이블이 항상 제공됩니다. 시간 열이 지정된 경우에는 시간 패턴 요약 및 그룹별 시간 요약 테이블도 포함됩니다. 이 보고서에는 각 시계열에 모델을 적합시킬 수 있는 "모델링 규칙" 섹션도 포함되어 있습니다.

- "시간 패턴 분석"
- "시간에 대한 그룹 요약"
- "모델링 규칙"

시간 패턴 분석

시간 열이 지정되면 시계열 예측 플랫폼에서 열 항목을 분석하여 패턴을 결정합니다. 이 플랫폼은 일별, 영업일별, 월별, 격주별, 분기별 등 다양한 시간 패턴을 인식하도록 되어 있습니다. 패턴은 "날짜 분석"으로 보고됩니다. 시간 패턴에 따라 "암시된 계절성"도 제공됩니다.

시간에 대한 그룹 요약

시간 열이 지정되면 "시계열 예측" 보고서에 개별 시계열의 요약 섹션이 포함됩니다. 요약에는 각 시계열에 대한 다음 정보가 포함됩니다.

< 그룹화 열 이름 > (그룹화 변수를 지정한 경우에만 사용 가능) 데이터가 쌓인 형식으로 제공될 때 개별 계열을 정의하는 그룹화 변수의 수준입니다.

행 수 계열의 관측값 수입니다.

결측값 시간 값에 결측값이 있는지 여부를 나타냅니다.

고유 타임스탬프 타임스탬프가 고유한지 여부를 나타냅니다.

시작 시작 타임스탬프입니다.

끝 종료 타임스탬프입니다.

결측 레코드 시작 및 종료 타임스탬프와 데이터 해상도/패턴을 기반으로 데이터 테이블에 결측 관측값이 있는지 여부를 나타냅니다.

팁 : "시간에 대한 그룹 요약" 테이블은 "계열 선택" 목록에 연결되어 있습니다. "시간에 대한 그룹 요약" 테이블에서 계열을 선택하면 "계열 선택" 목록에서도 해당 항목이 선택됩니다. Y에 대한 그룹 요약

"시계열 예측" 보고서에는 각 Y 변수에 대한 요약 테이블이 포함되어 있습니다. 요약에는 각 시계열에 대한 다음 정보가 포함됩니다.

변수 (여러 개의 Y 열과 하나 이상의 그룹화 열이 있는 경우에만 사용 가능) 시계열 변수의 열 이름입니다.

시계열 (계열이 여러 개 있는 경우에만 사용 가능) 개별 시계열의 이름입니다. 데이터가 쌓인 형식으로 제공된 경우 이 값은 그룹화 변수의 수준입니다. 데이터가 여러 개의 열과 함께 가로형 형식으로 제공된 경우에는 개별 계열의 열 이름입니다.

참고 : 데이터가 쌓인 형식으로 제공되는 경우 테이블의 열 머리글은 그룹화 열의 이름입니다.

최소값 개별 계열에 대한 Y의 최소값입니다.

최대값 개별 계열에 대한 Y의 최대값입니다.

길이 계열 기간입니다. 이 값은 시작 및 종료 타임스탬프와 시간 패턴 분석 결과를 기반으로 결정됩니다.

관측값 수 관측값 수입니다. 결측 레코드가 없으면 "길이" 값과 같습니다.

결측값 수 결측 Y 값 수입니다.

모두 양수 모든 Y 값이 양수인지 여부를 나타냅니다.

주석 계열에 문제 (예 : 고유하지 않은 타임스탬프)가 있는지 여부를 나타냅니다.

팁 : "Y 에 대한 그룹 요약 " 테이블은 " 계열 선택 " 목록에 연결되어 있습니다. "Y 에 대한 그룹 요약 " 테이블에서 계열을 선택하면 " 계열 선택 " 목록에서도 해당 항목이 선택됩니다.

모델링 규격

" 시계열 예측 " 보고서의 " 모델링 규격 " 섹션을 사용하면 모형을 개별 시계열에 적합시킬 수 있습니다. 다양한 모형이 고려됩니다. Hyndman et al. (2008) 연구 자료에서는 오차, 추세 성분 및 계절 성분을 기반으로 상태 공간 평활 모형을 정의합니다.

- 오차는 가법 (A) 또는 승법 (M) 일 수 있습니다.
- 추세 성분은 없음 (N), 가법 (A), 진폭감소 가법 (A_d), 승법 (M) 또는 진폭감소 승법 (M_d) 일 수 있습니다.
- 계절 성분은 없음 (N), 가법 (A) 또는 승법 (M) 일 수 있습니다.

특정 모형은 ETS (오차, 추세, 계절)로 나타낼 수 있습니다. 이러한 모형이 시계열 예측에 사용됩니다.

"모델링 규격" 보고서에는 두 개의 탭이 있습니다. 플랫폼에서 선택한 상태 공간 평활 모형 집합을 적합시키려면 "권장 규격" 탭을 사용합니다. 적합시킬 특정 상태 공간 평활 모형을 선택하려면 "완전한 규격" 탭을 사용합니다. 그런 다음 각 개별 시계열에 대해 주어진 집합의 최량 적합 모형이 예측에 사용됩니다.

모형 선택

"**완전한 규격 탭**"에 표시된 "완전한 규격" 탭에서만 사용할 수 있습니다. "모형 선택" 보고서를 사용하면 개별 시계열에 적합시킬 상태 공간 평활 모형을 지정할 수 있습니다. 체크박스를 사용하여 모형에 대한 오차, 추세 및 계절성을 선택합니다. 플랫폼에서 권장하는 모형에 해당하는 체크박스를 선택하려면 **권장 선택**을 클릭합니다. **모두 선택**을 클릭하여 모든 체크박스를 선택하거나 **모두 선택 취소**를 클릭하여 모든 체크박스를 선택 취소할 수 있습니다. "모수 제한" 옵션은 기본적으로 선택되며, 관측값이 현재에서 멀어질수록 모형의 현재 상태에 미치는 영향이 적어지도록 모수를 제한합니다. 상태 공간 평활 모형에서 모든 이전 관측값이 주어진 경우 시간 t 에서의 예측은 시간 t 까지의 모든 관측값의 가중합과 동일합니다. 가중치는 모수의 함수입니다. 따라서 모수를 제한하면 과거 관측값에 대한 가중치가 0이 되고 관측값이 현재에서 멀어질수록 가중치가 더 빨리 0이 됩니다.

그림 19.6 완전한 규격 탭

모형 선택

상태 공간 평활 모형

가법 오차 모형

	계절	없음	가법	승법
추세	☒	☒	☐	☐
없음	☒	☒	☒	☒
가법	☒	☒	☐	☐
진폭감소 가법	☒	☒	☐	☐
승법	☐	☐	☐	☐
진폭감소 승법	☐	☐	☐	☐

권장 선택 | 모두 선택 | 모두 선택 취소

승법 오차 모형

	계절	없음	가법	승법
추세	☒	☒	☒	☒
없음	☒	☒	☒	☒
가법	☒	☒	☒	☒
진폭감소 가법	☒	☒	☒	☒
승법	☒	☐	☐	☒
진폭감소 승법	☒	☐	☐	☒

권장 선택 | 모두 선택 | 모두 선택 취소

☒ 모수 제한

예측 설정

이름	값	설명
NAhead	4	NAhead >= 0, 선택 사항. n시차 후 미래 관측값을 예측합니다.
주기	4	주기 > 0, 선택 사항. 정수이거나 실수로 구분된 정수입니다.

모형 선택 전략

예측 성능 ▾ 측정 기준 RMSE ▾ NHoldback 4

기타 옵션

모형 선택 기준 유지 ☐

예측 구간 수준 0.95

적용 가능한 모형에 대한 대치 없음 ▾

실행

모델링 옵션

"모수 제한" 옵션을 선택하거나 선택 취소할 수 있습니다.

모수 제한 이 옵션을 선택하면 관측값이 현재에서 멀어질수록 모형의 현재 상태에 미치는 영향이 적어지도록 모수를 제한합니다. "모수 제한"은 기본적으로 선택됩니다.

예측 설정

다음과 같은 선택적 설정을 지정할 수 있습니다.

NAhead 예측 전 단계 수를 지정합니다. 단계 수는 음수가 아니어야 합니다.

주기 모형 적합 프로세스에서 고려할 계절성 값을 지정합니다. 기본적으로 주기는 "시간 패턴 분석" 보고서의 암시된 계절성으로 설정됩니다. 계절성이 다른 모형을 고려하려면 여기에 추가 주기를 지정하고 실수로 구분합니다. 주기는 0보다 커야 합니다.

모형 선택 전략

최적 모형을 선택하는 방법을 지정합니다.

정보 기준 지정된 정보 기준에 따라 최적 모형을 결정합니다. 사용 가능한 정보 기준은 AIC와 BIC입니다.

팁 : 시계열이 긴 경우 BIC 를 사용하십시오 .

예측 성능 홀드백 집합을 포함하는 성능 선택 알고리즘에 따라 최적 모델을 결정합니다 . 먼저 시계열이 훈련 데이터 집합과 홀드백 집합으로 분할됩니다 . "NHoldback" 값은 홀드백 집합의 관측값 수를 지정합니다 . 그런 다음 알고리즘이 훈련 데이터 집합의 데이터에 대해 모든 권장 모델 또는 사용자 지정 모델을 적합시킵니다 . 예측은 개별 적합 모델을 사용하여 홀드백 집합에 대해 수행됩니다 . 이러한 예측은 실제 홀드백 관측값과 비교되며 지정된 측정 기준을 사용하여 모델이 평가됩니다 . 측정 기준이 가장 적합한 모델이 최종 모델으로 선택됩니다 . 마지막으로 , 선택된 최종 모델이 전체 시계열 (훈련 데이터 집합과 홀드백 집합) 에 적합되고 재적합 모델은 시계열의 마지막 관측값 이후를 예측하는 데 사용됩니다 .

측정 기준 개별 모델 적합에 의한 예측을 평가할 측정 기준을 지정합니다 . 사용 가능한 측정 기준은 RMSE(제공된 평균 제공 오차) , MSE(평균 제공 오차) 및 MAE(평균 절대 오차) 입니다 .

NHoldback 홀드백 집합에 사용되는 관측값 수를 지정합니다 .

기타 옵션

다음 옵션을 지정할 수 있습니다 .

모델 선택 기준 유지 데이터에 적합한 모델 집합에 대한 모델 선택 기준을 저장합니다 . 모델 적합 후 이 정보는 " 모든 계열에 대한 모든 모델의 정보 기준 " 보고서에 표시됩니다 .

예측 구간 수준 예측에 대한 예측 구간 수준입니다 . 이 값은 예측 그림에서 음영 영역의 너비를 변경합니다 .

적용 가능한 모델에 대한 대치 모델 적합 프로세스 중에 사용되는 결측값 대체법을 지정합니다 .

참고 : 이러한 옵션은 모델 적합에 사용되는 데이터만 변경합니다 . 원시 데이터는 변경되지 않습니다 .

없음 결측값을 대체하지 않습니다 .

마지막 값 결측 관측값 시퀀스 전에 나오는 마지막 값을 사용하여 결측값을 대체합니다 .

중간 값 결측 관측값 시퀀스 전에 나오는 마지막 값과 결측 관측값 시퀀스 후에 나오는 첫 번째 값의 평균을 구하여 결측값을 대체합니다 .

선형 보간 결측 관측값 시퀀스 전에 나오는 마지막 값과 결측 관측값 시퀀스 후에 나오는 첫 번째 값 간의 선형 보간을 생성하여 결측값을 대체합니다 .

실행을 클릭하면 지정된 모델 집합이 적합됩니다 . 활성 스톱드 수, 완료된 작업 수, 총 작업 수 및 완료된 작업의 백분율을 보고하는 진행률 표시줄이 나타납니다 . 작업은 하나의 계열과 반응 변수 조합에 대해 지정된 모든 모델을 적합시키는 것으로 정의됩니다 . 적합 프로세스가 완료되면 "모델 보고서"가 표시됩니다 . 자세한 내용은 "[시계열 예측 모델 보고서](#)"에서 확인하십시오 . "모델링 규격" 보고서의 옵션을 변경하고 **실행**을 다시 클릭하면 "모델 보고서"가 새 보고서로 대체됩니다 .

시계열 예측 플랫폼 옵션

"시계열 예측"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

보고서 옵션 (모형을 적합시킨 후에만 적용) 다음과 같은 보고서 옵션을 포함합니다.

예측 구간 수준 설정 예측 구간 수준을 변경할 수 있습니다. 개별 모형 보고서가 자동으로 업데이트됩니다.

NAhead 설정 예측 전 단계 수를 변경할 수 있습니다. 개별 모형 보고서가 자동으로 업데이트됩니다.

NHoldout 설정 홀드백 집합의 기간을 변경할 수 있습니다. 개별 모형 보고서가 자동으로 업데이트됩니다.

결과 저장 (모형을 적합시킨 후에만 적용) "예측"이라는 별도의 데이터 테이블에 예측을 저장합니다. "예측" 데이터 테이블은 원래 데이터 테이블과 레이어아웃 및 형식이 동일합니다. NHoldout으로 홀드아웃 집합을 지정한 경우 홀드아웃 데이터의 실제값과 예측을 저장하는 추가 데이터 테이블이 생성됩니다. 이 테이블을 사용하여 여러 유형의 오차 추정값을 계산할 수 있습니다.

다음과 같은 추가 저장 옵션이 "출력 요청" 창에 표시됩니다.

예측 구간 저장 하한 및 상한 예측 구간을 "예측" 데이터 테이블의 열로 저장합니다.

1 시차 후 예측 저장 1 시차 후 예측과 예측을 쌓고 이를 별도의 데이터 테이블에 저장합니다.

원래 계열 저장 원래 계열과 예측을 쌓고 이를 별도의 데이터 테이블에 저장합니다.

예측 결과를 원래 테이블에 저장 결과를 원래 데이터 테이블의 새 열에 저장합니다.

다음 옵션에 대한 자세한 내용은 JMP 사용의 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.

시계열 예측 모형 보고서

"모형 보고서" 섹션에는 "계열 선택" 목록과 하나 이상의 개별 계열 모델링 보고서가 포함되어 있습니다. 표시되는 개별 보고서는 "계열 선택" 목록에 의해 결정됩니다. 목록에서 계열 이름을 클릭하면 선택한 계열에 대한 보고서가 표시됩니다. Ctrl 키 또는 Shift 키를 사용하여 계열 이름을 두 개 이상 선택하면 선택한 각 계열에 대한 개별 모형 적합 보고서가 표시됩니다. "계열 선택" 목록에서 하나 이상의 계열을 선택 취소하면 해당 계열에 대한 보고서가 제거됩니다.

"계열 선택" 목록은 데이터 테이블에 연결되어 있습니다. "계열 선택" 목록에서 계열을 선택하면 해당 계열의 관측값과 반응 변수 조합도 데이터 테이블에서 선택됩니다. "계열 선택" 목록은 "Y에 대한 그룹 요약" 테이블에도 연결됩니다. 시간 변수를 지정한 경우에는 "계열 선택" 목록이 "시간에 대한 그룹 요약" 테이블에도 연결됩니다. "계열 선택" 목록에서 계열을 선택하면 해당 항목이 테이블에서도 선택됩니다.

각 개별 모형 보고서에는 다음 정보가 포함되어 있습니다.

모형 유형 계열의 최량 적합 모형에 대한 오차, 추세 및 계절성을 설명합니다.

모형 보고서 계열에 적합된 모형을 기반으로 한 보고서 그룹입니다.

계열 및 예측 ("정보 기준"을 모형 선택 전략으로 지정한 경우에만 사용 가능) 원래 계열과 예측에 대한 그림입니다. 원래 계열의 기간 동안 1 시차 후 예측이 표시됩니다. 원래 계열의 기간이 지나면 예측이 표시됩니다. 음영 영역은 예측 구간을 나타냅니다.

참고 : 모든 예측은 모든 개별 계열에서 마지막 시점의 바로 다음 시점에서 시작됩니다. 자세한 내용은 "**데이터 형식**"에서 확인하십시오.

훈련, 홀드백 및 예측 ("예측 성능"을 모형 선택 전략으로 지정한 경우에만 사용 가능) 훈련 계열, 홀드백 계열 및 예측 계열에 대한 그림입니다. 세로 음영 영역은 홀드백 계열을 나타냅니다. 전체 시계열에 대한 모형 재적합이므로 전체 원래 시계열의 기간 동안 1 시차 후 예측이 표시됩니다. 마지막 관측값 이후의 기간 동안 예측이 표시됩니다. 음영 영역은 예측 구간을 나타냅니다.

모형 요약 $-2 \times$ 로그 가능도, AIC, 모수 수 및 시그마 값을 제공합니다. 시그마 값은 1 시차 후 예측 표준편차의 추정값입니다. "예측 성능"이 모형 선택 전략으로 지정된 경우 홀드백 집합의 RMSE(제곱근 평균 제곱 오차)도 표시됩니다. 자세한 내용은 **선형 모형 적합**에서 확인하십시오.

모수 추정값 모형에 대한 모수 추정값을 제공합니다. 모수는 "모형 유형" 보고서에 설명된 최량 적합 모형의 모수에 해당합니다.

1 시차 후 예측 오차 시간에 따른 1 시차 후 예측 오차에 대한 그림입니다. 각 시점 t 에 대해 오차는 $e(t) = y(t) - \mu(t)$ 로 계산됩니다.

승법 오차 모형의 1 시차 후 상대 예측 오차 (승법 오차 모형에만 사용 가능) 시간에 따른 1 시차 후 상대 예측 오차에 대한 그림입니다. 각 시점 t 에 대해 오차는 $e(t) = [y(t) - \mu(t)]/\mu(t)$ 로 계산됩니다.

20 장

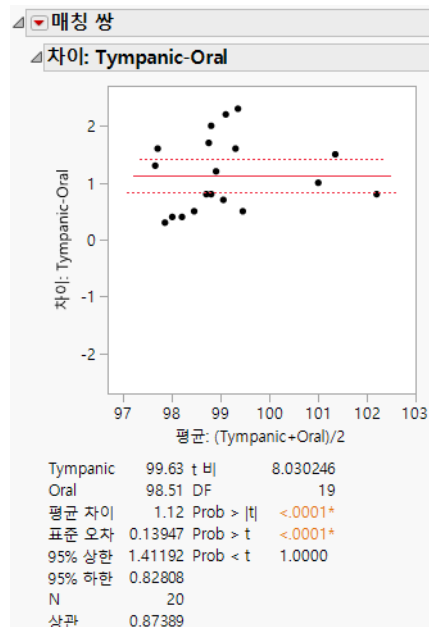
매칭 쌍 분석

동일한 개체에 대한 측정값 비교

매칭 쌍 플랫폼에서는 상관관계가 있는 두 개 이상의 변수 간 평균을 비교하고 차이를 평가합니다. 예를 들어 동일한 피실험자에 대해 치료 전과 치료 후의 혈압 측정값을 비교할 수 있습니다. 쌍체 t -검정이라는 통계적 방법은 상관된 반응을 고려합니다.

이 플랫폼에서는 세 가지 모든 대립가설에 대한 쌍체 t -검정 결과와 쌍체 평균별 쌍체 차이 그래프를 제공합니다. 세 개 이상의 매칭 반응과 그룹화 열에 대해 단순 반복 측정 분석으로 표본 간 검정을 수행할 수 있는 추가 기능도 제공합니다.

그림 20.1 매칭 쌍 분석의 예



목차

매칭 쌍 플랫폼 개요.....	385
매칭 쌍 비교의 예.....	385
매칭 쌍 플랫폼 시작.....	386
여러 Y 열.....	387
매칭 쌍 보고서.....	388
차이 그림 및 보고서.....	388
그룹 간.....	389
매칭 쌍 플랫폼 옵션.....	389
그룹 간 매칭 쌍 비교의 예.....	390
매칭 쌍 플랫폼에 대한 통계 상세 정보.....	391
매칭 쌍의 그래픽에 대한 통계 상세 정보.....	391
반응 상관에 대한 통계 상세 정보.....	392

매칭 쌍 플랫폼 개요

매칭 쌍 플랫폼에서는 쌍체 t -검정을 사용하여 두 반응 열 간의 행별 차이를 비교합니다. 두 열은 동일한 피실험자에 대한 처리 전과 처리 후의 측정값을 나타내는 경우가 많습니다. 또는 측정값이 두 개의 다른 측정기로 동일한 개체에서 수집한 데이터를 나타낼 수 있습니다.

데이터 테이블의 두 열에 쌍체 데이터가 배열된 경우 매칭 쌍 플랫폼을 사용할 수 있습니다. 그러나 모든 측정값이 단일 열에 있는 경우 다음 작업 중 하나를 수행하십시오.

- **테이블** 메뉴의 **분할** 옵션을 사용하여 측정값 열을 두 개의 열로 분할합니다. 그런 다음 매칭 쌍 플랫폼을 사용할 수 있습니다.
- 반응 열이 두 개인 경우 두 반응 간의 차이를 계산하는 세 번째 열을 생성합니다. 그런 다음 분포 플랫폼을 사용하여 차이 열의 평균이 0 인지 검정합니다.
- 두 반응이 단일 열에 저장된 경우 이원 분산 분석을 수행할 수 있습니다. 한 요인 (ID 변수) 은 두 반응을 식별하고 다른 요인은 개체를 식별합니다. 블록 변수 (개체 열) 와 함께 X 로 Y 적합 일원 분석 플랫폼을 사용하거나, 모형 적합 플랫폼을 사용하여 이원 ANOVA 를 수행합니다. ID 요인에 대한 검정은 쌍체 t -검정과 동일합니다.

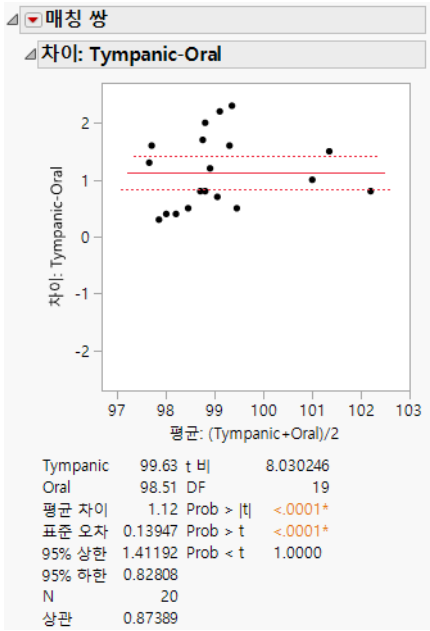
참고: 쌍체 데이터의 경우 일반적인 독립 t -검정을 수행하지 마십시오. 데이터를 한 열에 쌓지 말고, 블록 변수 없이 ID 에 대해 X 로 Y 적합 일원 ANOVA 를 사용합니다. 이렇게 하면 반응 간의 상관을 무시하는 효과가 있습니다. 이 경우 검정을 수행할 때 반응에 음의 상관관계가 있으면 효과를 과대추정하고 반응에 양의 상관관계가 있으면 효과를 과소추정합니다.

매칭 쌍 비교의 예

이 예에서는 두 가지 유형의 체온계 간 차이를 검정하기 위해 매칭 쌍 분석을 사용합니다. 데이터에는 20 명의 체온 측정값이 포함되어 있습니다. 체온은 구강 체온계와 고막 (귀) 체온계 두 종류를 사용하여 측정되었습니다. 두 종류의 체온계로 측정한 체온이 동일한지 여부를 확인하려고 합니다. 다른 사람과의 체온 차이는 중요하지 않습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Therm.jmp 를 엽니다.
2. **분석 > 전문 모델링 > 매칭 쌍**을 선택합니다.
3. Oral 과 Tympanic 을 선택하고 **Y, 쌍체 반응**을 클릭합니다.
4. **확인**을 클릭합니다.

그림 20.2 매칭 쌍 보고서 창



여기에서는 대체로 고막 체온계가 구강 체온계보다 1.12도 높게 측정된다는 결과를 보여 줍니다. p 값(Prob > |t|)이 작으므로 이 차이는 우연히 발생한 것이 아니라 통계적으로 유의하다는 것을 나타냅니다.

이 매칭 쌍 분석에서는 어떤 체온계가 정확한지(둘 중 하나인 경우)는 나타내지 않고 체온계 간에 차이가 있다는 것만 나타냅니다.

매칭 쌍 플랫폼 시작

분석 > 전문 모델링 > 매칭 쌍을 선택하여 매칭 쌍 플랫폼을 시작합니다.

그림 20.3 매칭 쌍 시작 창

매칭된 변수에 대한 쌍제 t-검정을 제공합니다.

열 선택

4개 열

- Name
- Oral
- Tympanic
- difference

선택한 열 역할 지정

Y, 쌍제 반응	필수 숫자 필수 숫자 선택적 숫자
X, 그룹화	선택적
가중치	선택적 숫자
빈도	선택적 숫자
기준	선택적

작업

확인
취소
제거
재호출
도움말

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

Y, 쌍체 반응 두 개의 반응 열을 제공합니다. 세 개 이상의 반응을 분석하는 방법에 대한 자세한 내용은 "**여러 Y 열**"에서 확인하십시오.

X, 그룹화 그룹 간 차이를 비교하기 위한 그룹화 변수를 제공합니다. 자세한 내용은 "**그룹 간**"에서 확인하십시오.

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 매칭 쌍 분석을 수행합니다.

확인을 클릭하면 "매칭 쌍" 보고서 창이 나타납니다. 자세한 내용은 "**매칭 쌍 보고서**"에서 확인하십시오.

여러 Y 열

매칭 쌍 플랫폼에서는 세 개 이상의 반응을 지정할 수 있습니다. 반응 수가 홀수이면 가능한 모든 쌍이 분석됩니다. 다음 표에서는 세 가지 반응의 예를 보여 줍니다.

Y1 대 Y2	Y1 대 Y3
	Y2 대 Y3

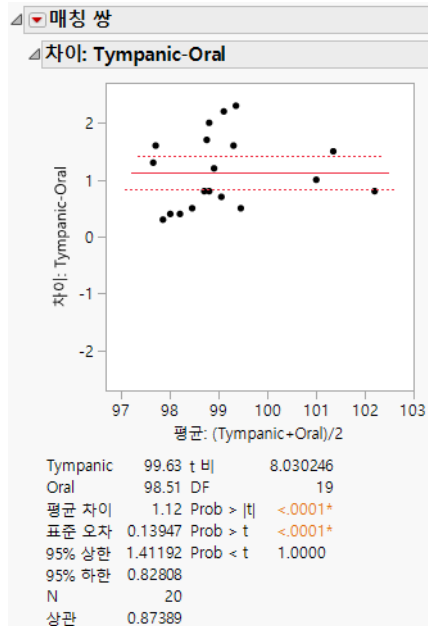
반응 수가 짝수이면 매칭 쌍 플랫폼에서 가능한 모든 쌍을 수행할지 묻습니다. 가능한 모든 쌍을 수행하지 않으면 인접한 반응이 쌍으로 분석됩니다. 다음 표에서는 네 개 반응에 대한 분석 배열을 보여 줍니다.

Y1 대 Y2	Y3 대 Y4

매칭 쌍 보고서

"매칭 쌍" 보고서에는 Tukey 평균 - 차이 그림, 요약 통계량 및 쌍체 t -검정 결과가 표시됩니다. 자세한 내용은 "[차이 그림 및 보고서](#)"에서 확인하십시오. X_i 그룹화 변수를 지정한 경우 보고서에 "그룹 간" 보고서도 포함됩니다. 자세한 내용은 "[그룹 간](#)"에서 확인하십시오.

그림 20.4 매칭 쌍 보고서의 예



참고: 빨간색 삼각형 메뉴에는 초기 보고서 창에 보고서를 추가할 수 있는 다른 옵션이 더 있습니다. 자세한 내용은 "[매칭 쌍 플랫폼 옵션](#)"에서 확인하십시오.

차이 그림 및 보고서

"매칭 쌍" 보고서의 "차이" 그림에는 평균별 차이가 표시됩니다. "차이" 그림에서 평균 차이는 가로선으로 표시됩니다. 95% 신뢰 구간은 평균 차이의 위아래에 점선으로 표시됩니다. 신뢰 영역에 0이 포함되면 평균이 0.05 수준에서 유의한 차이가 없음을 나타냅니다. 참조 프레임을 추가하면 쌍의 평균이 세로선으로 표시됩니다. 참조 프레임에 대한 자세한 내용은 "[매칭 쌍 플랫폼 옵션](#)"에서 확인하십시오.

"차이" 보고서에는 각 반응의 평균, 평균 차이 및 차이에 대한 신뢰 구간이 표시됩니다. 쌍체 t -검정에 대한 결과도 "차이" 보고서에 표시됩니다.

그룹 간

매칭 쌍 플랫폼의 그룹 간 분석은 단순 반복 측정 분석에 해당합니다. 모형 적합 플랫폼의 MANOVA 분석법을 사용하여 동일한 검정 결과를 얻을 수 있습니다.

참고 : " 그룹 간 " 보고서는 X, 그룹화 변수를 지정한 경우에만 나타납니다.

평균 차이 두 쌍체 열 사이에 각 그룹의 전체 행 차이의 평균을 표시합니다. 다시 말하면, 개체 내와 개체 간 교호작용 또는 분할구와 주구 교호작용입니다.

평균 평균 두 쌍체 열에 대해 각 그룹의 전체 행 평균의 평균을 표시합니다. 다시 말하면, 개체 간 또는 주구 효과입니다.

그룹 간 검정 다음 두 F-검정은 그룹 간 값이 다른지 여부를 판별합니다.

평균 차이 반응 쌍 간의 변화가 그룹마다 다른지 검정합니다.

평균 평균 개체에 대한 평균 반응이 그룹마다 다른지 검정합니다.

관련 정보

- " 그룹 간 매칭 쌍 비교의 예 "

매칭 쌍 플랫폼 옵션

" 매칭 쌍 " 의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

평균별 차이 그림 쌍체 평균별 쌍체 차이 그림을 표시하거나 숨깁니다. 이 그림에 대한 자세한 설명은 " 차이 그림 및 보고서 " 에서 확인하십시오.

행별 차이 그림 행 번호별 쌍체 차이 그림을 표시하거나 숨깁니다.

참조 프레임 평균별 차이 그림에 참조 프레임을 표시하거나 숨깁니다. 이 그림 안에서 Y2 대 Y1 그림의 위치를 보여 주는 직사각형으로 기울어져 있고, 눌린 모양일 수 있습니다. 평균의 평균을 나타내는 빨간색 세로선이 표시됩니다. 차이 범위가 데이터 범위의 절반보다 클 때 참조 프레임이 처음부터 표시됩니다.

Wilcoxon 부호 순위 Wilcoxon 부호 순위 검정을 표시하거나 숨깁니다. Wilcoxon 부호 순위 검정은 쌍체 차이에 적용됩니다. 이 검정은 양수 차이의 크기를 음수 차이의 크기와 비교하는 비모수 검정입니다. 이 검정에서는 Pratt 방법을 사용하여 0 차이 값을 처리합니다. 또한 차이의 분포가 대칭적이라고 가정합니다. 자세한 내용은 기본 분석에서 확인하십시오. Lehman(2006), Conover(1999, p. 350) 및 Cureton(1967) 연구 자료도 확인하십시오.

부호 검정 부호 검정을 표시하거나 숨깁니다. 이 검정은 차이 값의 부호 (양수 또는 음수) 만 검정에 사용하는 쌍체 t-검정의 비모수 버전입니다.

Bland-Altman 분석 Bland-Altman 보고서를 표시하거나 숨깁니다. 이 보고서에는 X 및 Y 변수의 평균 차이 또는 편향이 포함됩니다. 합치도 상한 및 하한은 편향 $\pm z_{1-\alpha/2}^*$ (편향의 표준편차) 로 설정됩니다.

α 수준 설정 분석에 사용되는 유의 수준을 변경합니다. 보고서와 그림의 신뢰 구간에 영향을 줍니다.

다음 옵션에 대한 자세한 내용은 **JMP 사용의** 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.

그룹 간 매칭 쌍 비교의 예

이 예에서는 매칭 쌍 그룹 간 보고서 및 모형 적합을 사용한 해당 MANOVA 보고서를 생성합니다.

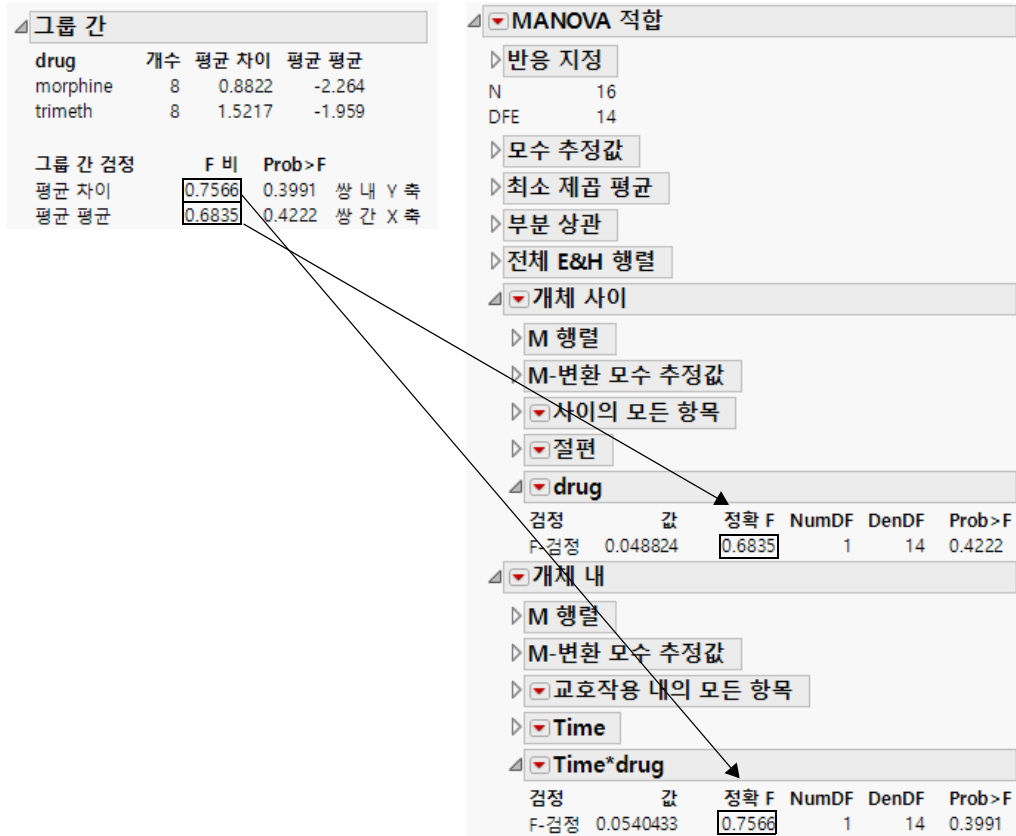
매칭 쌍 그룹 간 보고서 생성

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Dogs.jmp 를 엽니다.
2. **분석 > 전문 모델링 > 매칭 쌍**을 선택합니다.
3. LogHist0 및 LogHist1 을 선택하고 **Y, 쌍체 반응**을 클릭합니다.
4. drug 를 선택하고 **X, 그룹화**를 클릭합니다.
5. **확인**을 클릭합니다.

모형 적합 보고서 생성

1. **분석 > 모형 적합**을 선택합니다.
2. LogHist0 및 LogHist1 을 선택하고 **Y** 를 클릭합니다.
3. drug 를 선택하고 **추가**를 클릭합니다.
4. **Manova** 분석법을 선택합니다.
5. **실행**을 클릭합니다.
6. "반응 지정" 보고서의 **반응 선택** 메뉴에서 **반복 측정**을 선택합니다.
7. **확인**을 클릭합니다.

그림 20.5 그룹 간 매칭 쌍 및 반복 측정을 사용한 모형 적합 MANOVA 의 예



"그룹 간" 보고서에서 "평균 차이"의 "F 비"는 "개체 내" 보고서에서 "시간*drug"의 F 비에 해당합니다. "그룹 간" 보고서에서 "평균 평균"의 "F 비"는 "MANOVA 적합" 보고서의 "개체 사이"에서 "drug"의 F 비에 해당합니다.

매칭 쌍 플랫폼에 대한 통계 상세 정보

이 섹션에는 매칭 쌍 플랫폼에 대한 통계 상세 정보가 포함되어 있습니다.

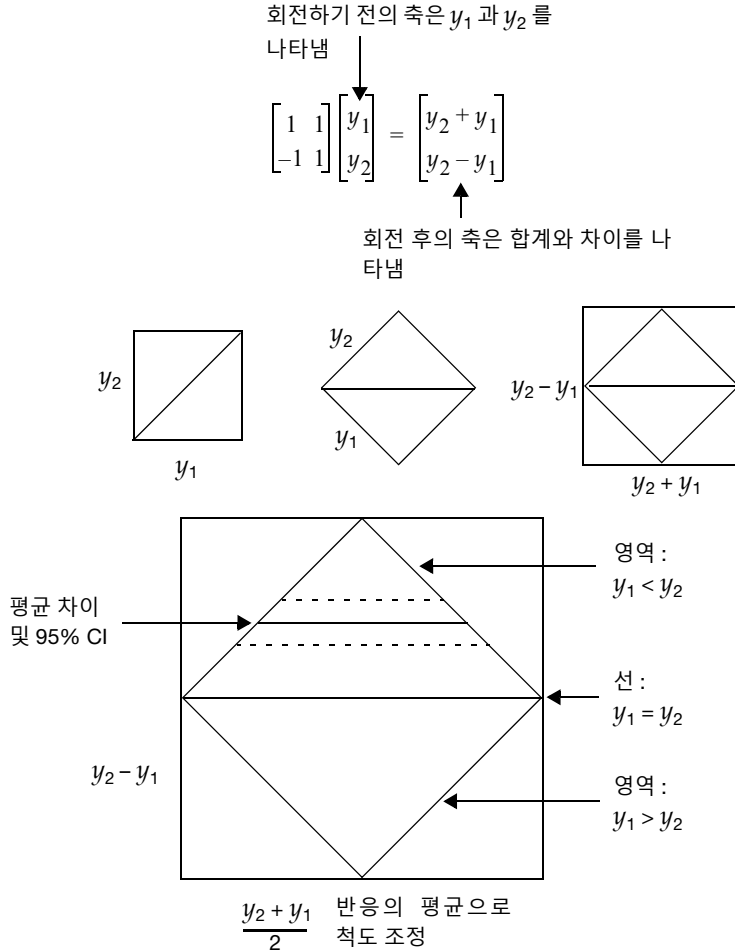
- "매칭 쌍의 그래픽에 대한 통계 상세 정보"
- "반응 상관에 대한 통계 상세 정보"

매칭 쌍의 그래픽에 대한 통계 상세 정보

매칭 쌍 플랫폼의 기본 그래프는 Tukey 평균-차이입니다(Cleveland 1994, p. 130 참조). 이 그래프는 두 반응의 차이가 세로 축에 표시되고 두 반응의 평균이 가로 축에 표시됩니다. 이 그래프는

두 원래 변수의 산점도를 45도 회전한 것과 같습니다. 45도 회전하고 척도를 조정하면 원래 좌표가 차이와 평균이 됩니다.

그림 20.6 평균별 차이로 변환 (45도 회전)의 예

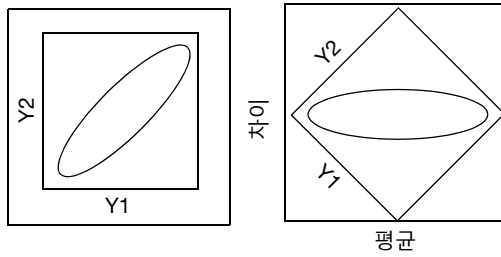


반응 상관에 대한 통계 상세 정보

매칭 쌍 플랫폼에서 분석을 수행할 때 측정값 쌍 간의 상관을 알아야 합니다. 동일한 개체에서 서로 다른 시간에 한 쌍의 측정값을 가져오는 대부분의 경우 측정값은 양의 상관관계가 있습니다. 그러나 측정값이 경쟁 반응을 나타내는 경우 음의 상관관계가 있을 수 있습니다.

그림 20.7에서는 두 반응의 양의 상관관계가 차이(세로 축)에 대한 작은 분산이 되는 방식을 보여 줍니다. 음의 상관관계가 있으면 타원이 다른 방향을 향하고 회전한 그래프의 분산이 세로 축에서 큼니다.

그림 20.7 회전 전후의 양의 상관관계 예제



21 장

이상치 탐색

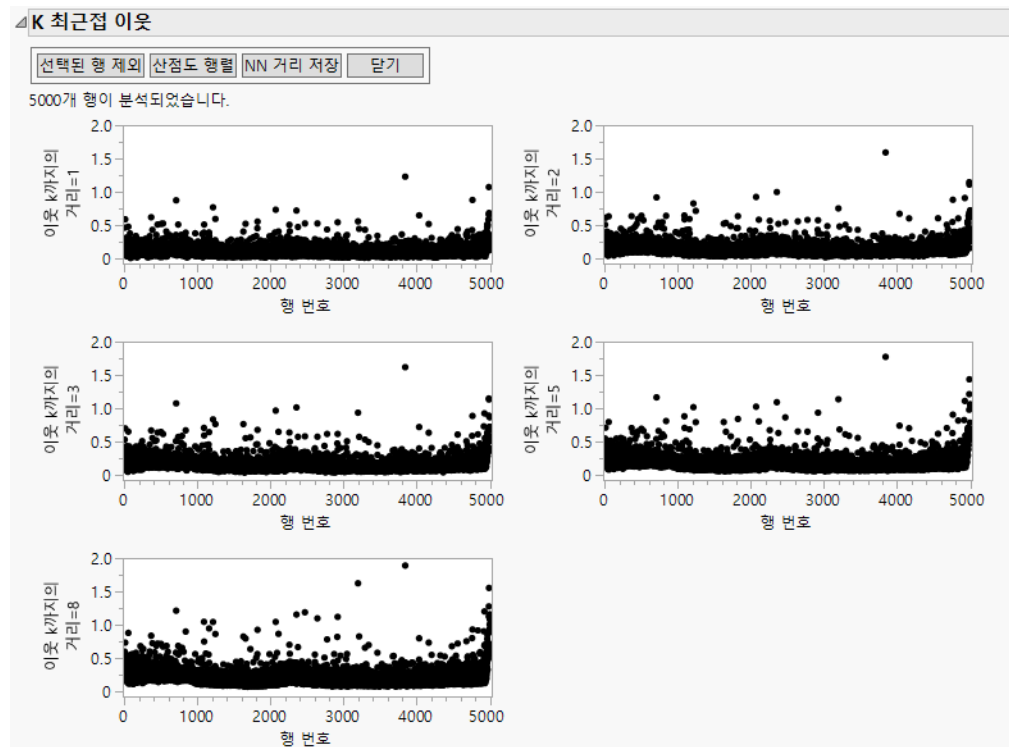
단변량 및 다변량 데이터에서 이상치 찾기

이상치 탐색 플랫폼을 사용하면 이상치를 식별, 탐색 및 관리할 수 있습니다. 데이터에서 이상치를 탐색하고 이해하는 것은 분석의 중요한 부분입니다. 데이터의 이상치는 데이터 수집 또는 보고 단계에서의 실수, 측정 시스템 오류, 데이터 집합에 오류 또는 결측값 포함, 단순 비정상 값 등이 원인이 될 수 있습니다. 이상치가 있으면 추정값이 왜곡되고 결과가 해당 이상치로 치우칠 수 있습니다.

이상치는 표본 분산을 부풀립니다. 경우에 따라 데이터의 이상치를 유지해야 하는데 이를 제거하면 표본 분산이 과소추정되고 데이터가 반대 방향으로 편향될 수 있습니다.

이상치는 제거 여부에 상관없이 찾는 것이 좋습니다. 데이터를 시각적으로 검사할 수 있는 여러 가지가 있습니다. 예를 들어 상자 그림, 히스토그램 및 산점도에서는 이러한 극단값을 쉽게 표시할 수 있습니다. 자세한 내용은 JMP 살펴보기에서 확인하십시오.

그림 21.1 다변량 k-최근접 이웃 이상치 예



목차

이상치 탐색 개요	397
분위수 범위 이상치	397
로버스트 적합 이상치	397
로버스트 PCA 이상치	397
K 최근접 이웃 이상치	398
이상치 탐색의 예	399
이상치 탐색 플랫폼 시작	402
이상치 탐색 보고서	405
분위수 범위 이상치 보고서	406
로버스트 적합 이상치 보고서	410
로버스트 PCA 이상치 보고서	413
K 최근접 이웃 보고서	414
이상치 탐색 플랫폼 옵션	415
이상치 탐색 플랫폼의 추가 예	416

이상치 탐색 개요

이상치 탐색 플랫폼에는 데이터에서 잠재적 이상치를 식별하는 네 가지 방법이 있습니다. 단변량 방법과 다변량 방법이 각각 두 개씩 있습니다.

- "분위수 범위 이상치"
- "로버스트 적합 이상치"
- "로버스트 PCA 이상치"
- "K 최근접 이웃 이상치"

분위수 범위 이상치

이상치 탐색 플랫폼의 분위수 범위 이상치 방법은 열 값의 분위수 분포를 사용하여 극단값을 찾습니다. 분위수는 관련된 분포 가정이 없으므로 이상치를 감지하는 데 유용합니다. 데이터는 가장 작은 것부터 오름차순으로 정렬됩니다. 예를 들어 20번째 분위수는 값의 20%가 더 작은 위치의 기준 값입니다. 극단값은 분위수 간 범위, 즉 지정된 두 분위수 간 거리의 승수를 사용하여 구합니다. 분위수 계산 방법에 대한 자세한 내용은 기본 분석에서 확인하십시오.

분위수 범위 이상치 방법은 데이터 내에 저장된 결측값 코드를 식별하는 데도 유용합니다. 일부 업계에서는 결측값이 9모음값(예: 999 또는 9999)으로 입력됩니다. 이 기법은 3사분위수보다 큰 9모음값을 의심되는 결측값 코드로 찾습니다. 그런 다음 해당 결측값 코드를 데이터 테이블에 열 특성으로 추가하거나, 이러한 값을 데이터 테이블의 결측값으로 변경할 수 있습니다.

로버스트 적합 이상치

이상치 탐색 플랫폼의 로버스트 적합 이상치 방법은 데이터의 중심 및 산포에 대한 여러 유형의 로버스트 추정값을 제공하여 이상치 식별을 위한 임계값을 결정합니다. 모수의 로버스트 추정값은 비로버스트 추정값보다 이상치에 덜 민감합니다.

로버스트 PCA 이상치

이상치 탐색 플랫폼의 로버스트 PCA 이상치 방법을 사용하여 상관된 다변량 데이터에서 이상치 셀을 식별할 수 있습니다. 다른 많은 다변량 방법은 이상치 행만 식별하므로 이 방법이 유용합니다. 이 방법을 데이터에 적용하기 전에 먼저 열을 중심화 및 척도화할 수 있습니다. 척도 인수는 다음과 같이 정의됩니다.

$$\max [Q(0.75) - Q(0.50), Q(0.50) - Q(0.25)] / [\text{정규 분위수 } (0.75)]$$

다음은 각 요소에 대한 설명입니다.

$Q(p)$ 는 p 번째 분위수입니다.

참고 : $Q(0.75)$ 또는 $Q(0.25)$ 가 중앙값과 같으면 0 이 아닌 범위가 될 때까지 더 극단적인 분위수가 사용됩니다.

데이터 중심화 및 척도화가 완료되면 로버스트 PCA 이상치 방법에서 특이값 분해 및 임계 지정 단계를 순차적으로 수행하여 데이터 행렬을 분해합니다. 데이터는 낮은 계수 행렬과 잔차 최소 행렬로 분해됩니다. 임계값은 잔차가 이상치에 대해 매우 크거나 이상치가 아닌 경우 0에 매우 가깝도록 지정됩니다. 이 알고리즘은 이상치나 작은 잡음 없이 체계적인 변동을 포착하는 데 적절한 행렬 계수를 결정합니다. 낮은 계수 공간에 없는 이상치는 잔차를 기반으로 감지됩니다. 자세한 내용은 Candes et al(2009) 및 Lin et al(2013) 연구 자료에서 확인하십시오. 결측값이 있는 경우 처음에는 중심화 및 척도화 단계 후에 0으로 대체됩니다. 그런 다음 각 SVD(특이값 분해) 반복 후 SVD의 예측값으로 결측값이 업데이트됩니다.

K 최근접 이웃 이상치

이상치 탐색 플랫폼의 K 최근접 이웃 이상치 방법을 사용하여 최근접 이웃과의 거리를 기반으로 이상치를 식별할 수 있습니다. K 최근접 이웃 이상치 방법은 각 k 값에 대해 각 점에서 k 번째 최근접 이웃까지의 유클리드 거리 그림을 표시합니다. K 로 표시되는 가장 큰 k 값을 지정합니다. 그림이 너무 많이 표시되지 않도록 피보나치 수열을 사용하여 $k = 1, 2, 3, 5, \dots, K$ 에 대한 그림이 제공됩니다.

최근접 이웃을 계산하기 전에 열을 중심화하고 척도화합니다. 척도 인수는 다음과 같습니다.

$$\max [Q(0.75) - Q(0.50), Q(0.50) - Q(0.25)] / [\text{정규 분위수}(0.75)]$$

다음은 각 요소에 대한 설명입니다.

$Q(p)$ 는 p 번째 분위수입니다.

참고 : $Q(0.75)$ 또는 $Q(0.25)$ 가 중앙값과 같으면 0 이 아닌 범위가 될 때까지 더 극단적인 분위수가 사용됩니다.

이 방법은 지정된 k 값에 민감합니다. k 값이 작으면 점을 이상치로 식별하지 못할 수 있고 k 값이 크면 점을 이상치로 잘못 분류할 수 있습니다.

- 소수의 이웃만 연구하기 위해 K 값을 작게 지정하는 경우를 가정해 보겠습니다. K 개를 초과하는 점을 포함하고 나머지 점에서 멀리 떨어져 있는 군집의 경우 군집 내의 점은 최근접 이웃과의 거리가 짧습니다. 이상치 군집을 감지하지 못할 수 있습니다.
- 많은 수의 이웃을 연구하기 위해 K 값을 크게 지정하는 경우를 가정해 보겠습니다. K 개 미만의 데이터 점을 포함하는 군집의 경우 이러한 군집 내의 점은 이상치일 수 있습니다. 군집이 점으로 구성된다는 점을 간과하고 대신 개별 군집 멤버를 이상치로 해석할 수 있습니다.

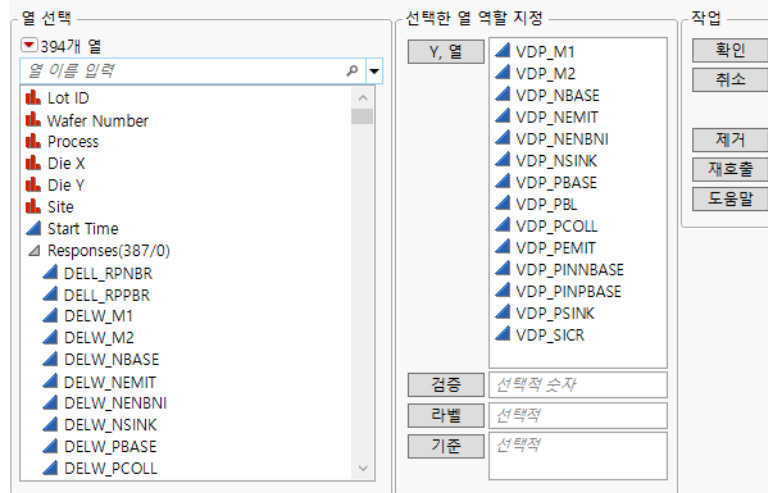
이상치 탐색의 예

이상치 탐색을 사용하여 데이터 집합의 선택된 열 그룹 내에서 이상치를 식별할 수 있습니다. 식별된 이상치는 분포 플랫폼을 사용하여 검토할 수 있습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Probe.jmp 샘플 데이터 테이블을 엽니다.
2. **분석 > 선별 > 이상치 탐색**을 선택합니다.
3. Responses(387/0) 옆의 삼각형을 클릭하여 그룹의 모든 열을 표시합니다.
4. VDP_M1 ~ VDP_SICR 열을 선택하고 **Y, 열**을 클릭합니다. 14 개 열이 선택되어야 합니다.

그림 21.2 이상치 탐색 시작 창

단변량 또는 다변량 데이터에서 이상치를 식별, 탐색 및 관리합니다.



5. **확인**을 클릭합니다.
6. **분위수 범위 이상치**를 클릭합니다.
"분위수 범위 이상치" 보고서는 탭을 사용하여 결과를 구성합니다. "분위수 범위 이상치" 보고서의 "열별 이상치" 탭에 각 열이 표시되고 발견된 이상치의 수와 ID가 나열됩니다.
7. "분위수 범위 이상치" 보고서의 "열별 이상치" 탭에서 **이상치를 포함하는 열만 표시**를 선택합니다. 이렇게 하면 이상치를 포함하는 열만 표시되도록 열 목록이 제한됩니다.
몇몇 열에 이상치 값 9999가 포함되어 있습니다. 많은 업계에서 9 모음값을 결측값 코드로 사용합니다.
8. "9 모음값" 탭을 클릭합니다.
9. "9 모음값" 보고서에서 각 열을 선택합니다.
10. **결측값 코드에 가장 큰 9 모음값 추가**를 클릭합니다.

원래 데이터를 유지하려면 **다른 이름으로 저장** 명령을 사용해야 함을 나타내는 "JMP 경고"가 표시됩니다.

11. **확인**을 클릭합니다.
12. "분위수 범위 이상치" 보고서에서 **다시 스캔**을 클릭합니다.
13. **검색을 정수로 제한**을 선택합니다.

연속형 데이터에서 정수 값은 종종 오류 코드 또는 코드화된 다른 데이터 값을 나타냅니다. 이 열 집합에는 추가 오류 코드가 포함되어 있지 않습니다.

14. **검색을 정수로 제한**을 선택 취소합니다.

데이터 검토

보고서의 "열별 이상치" 탭으로 돌아갑니다.

1. 나머지 열을 모두 선택합니다.
2. **행 선택하기**를 클릭합니다.
3. **분석 > 분포**를 선택합니다.
4. 선택된 열을 **Y, 열** 역할에 할당합니다. "분위수 범위 이상치" 보고서에서 이러한 열 이름을 선택했으므로 "분포" 시작 창에 이미 선택되어 있습니다. **Y, 열**을 클릭합니다.
5. **확인**을 클릭합니다.

그림 21.3 이상치가 선택된 열의 분포



VDP_M1 및 VDP_PEMIT 열에서는 선택된 이상치 중 일부가 대부분의 데이터에 다소 가깝습니다. 나머지 열의 경우 선택된 이상치가 대부분의 데이터에서 멀리 떨어져 있습니다. 데이터 점을 조사했으므로 분석에서 제외할 열을 결정할 수 있습니다.

제외된 이상치 구체화

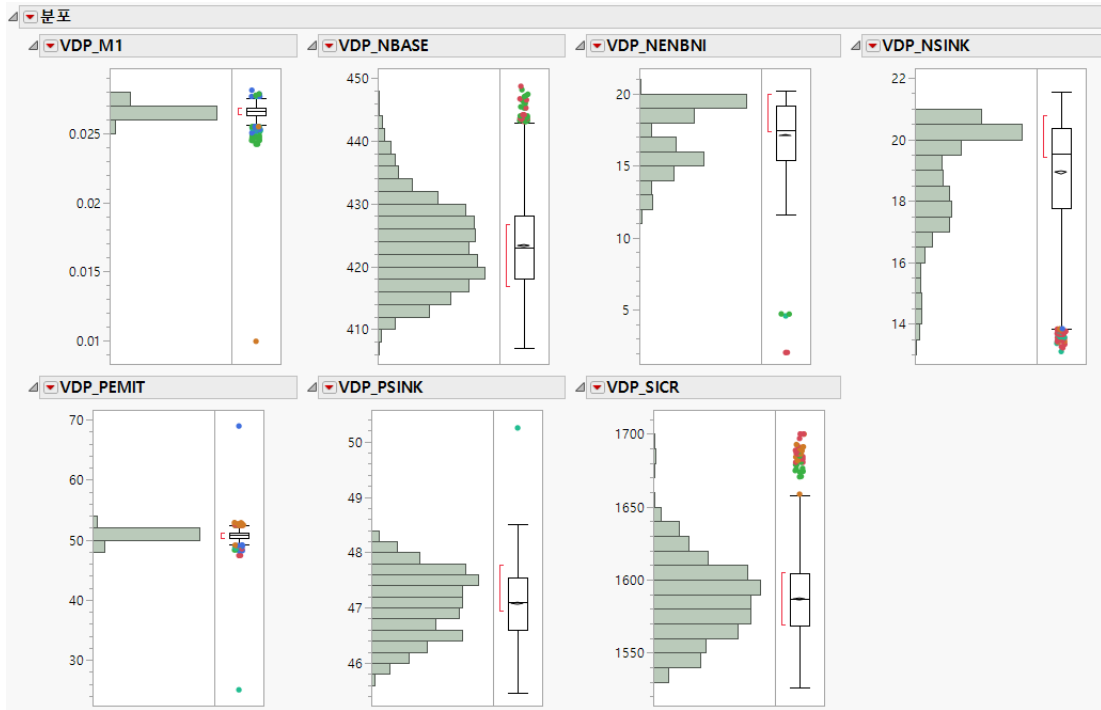
보고서의 "열별 이상치" 탭으로 돌아갑니다.

1. Ctrl 키를 누른 채 VDP_M1 및 VDP_PEMIT 열을 선택 취소합니다.
2. 나머지 열은 보고서에서 선택된 상태로 **행 제외**를 클릭합니다.
3. "Q"를 "20"으로 변경합니다.
4. **다시 스캔**을 클릭합니다.
5. 보고서에서 VDP_M1 및 VDP_PEMIT 열을 선택합니다.
6. **행 선택하기**를 클릭합니다.

데이터 재검토

1. "분포" 보고서를 다시 검토합니다. 선택된 이상치가 대부분의 데이터에서 충분히 떨어져 있어 분석에서 선택 및 제외할 수 있습니다.
2. "분위수 범위 이상치" 보고서에서 **행 제외**를 클릭합니다.
3. "분포" 보고서에서 "분포"의 빨간색 삼각형을 클릭하고 **다시 실행 > 분석 다시 실행**을 선택합니다.

그림 21.4 이상치가 제외된 열의 분포



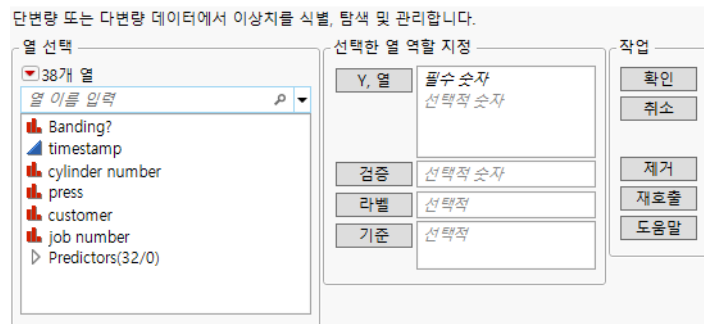
이제 이상치 없이 더 유용한 정보가 데이터 분포에 표시됩니다.

이상치 탐색 플랫폼 시작

이상치 탐색을 시작하려면 **분석 > 선별 > 이상치 탐색**을 선택합니다.

참고: "이상치 탐색" 명령은 연속형 모델링 유형의 열만 분석합니다. 다른 열을 시작 창에 입력할 수 있지만 무시됩니다.

그림 21.5 이상치 탐색 플랫폼 시작 창



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

Y, 열 분석할 열을 지정합니다.

검증 로버스트 PCA 이상치에 사용되는 검증 열을 지정합니다.

라벨 다변량 분석 보고서의 행 번호를 라벨로 바꾸는 열을 지정합니다.

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.

팁: 기준 변수의 모든 수준에 대해 이상치 분석을 실행하려면 **Ctrl** 키를 누른 채 원하는 이상치 분석 명령 버튼을 클릭합니다.

확인을 클릭하면 "이상치 탐색" 보고서가 나타납니다. 이 보고서에는 단변량 및 다변량 데이터에서 이상치를 찾는 여러 가지 방법이 포함되어 있습니다. 선택하기 전에 각 방법에 대해 옵션을 지정할 수 있습니다.

단변량

단변량 데이터에서 이상치를 탐색하기 위한 두 가지 옵션이 있습니다.

분위수 범위 이상치 각 열의 분위수 분포를 사용하여 이상치를 극단값으로 식별합니다. 이 도구는 데이터 내에서 결측값 또는 오류 코드를 검색하는 데 유용합니다. 이 방법은 데이터의 이상치 탐색을 시작하는 데 권장됩니다. 자세한 내용은 "**분위수 범위 이상치**"에서 확인하십시오. 다음 옵션을 지정할 수 있습니다.

꼬리 분위수 분위수 간 범위를 계산하는 데 사용되는 1 사분위수에 대한 확률입니다. 3 사분위수 확률은 1 - 꼬리 분위수로 간주됩니다. 예를 들어 꼬리 분위수 값이 0.1이면 분위수 간 범위는 데이터의 0.1 분위수에서 0.9 분위수 사이임을 의미합니다. 기본값은 0.1입니다.

Q 이상치 임계값을 결정하는 승수입니다. 꼬리 분위수 또는 1 - 꼬리 분위수 값을 넘어 분위수 간 범위의 Q 배를 초과하는 값이 이상치로 식별됩니다. Q 값이 크면 작은 값보다 더 보수적인 이상치 집합이 제공됩니다. 기본값은 3 입니다.

로버스트 적합 이상치 각 열의 중심 및 산포에 대한 로버스트 추정값을 찾고 이상치를 해당 값에서 멀리 떨어진 데이터 점으로 식별합니다. 자세한 내용은 "**로버스트 적합 이상치**"에서 확인하십시오. 다음 옵션을 지정할 수 있습니다.

K 시그마 중심에서 산포의 K 배 이상 떨어진 값을 이상치로 결정하는 승수입니다. K 값이 크면 작은 값보다 더 보수적인 이상치 집합이 제공됩니다. 기본값은 4 입니다.

Huber Huber M- 추정을 사용하여 중심과 산포를 추정합니다. 이 옵션이 기본값입니다. 자세한 내용은 Huber 와 Ronchetti 연구 자료 (2009)에서 확인하십시오.

Cauchy Cauchy 분포를 가정하여 중심 및 산포 추정값을 계산합니다. Cauchy 추정값은 높은 봉피점을 가지며 일반적으로 Huber 추정값보다 더 로버스트합니다. 그러나 데이터가 군집으로 분리된 경우 Cauchy 분포는 더 가깝게 군집화된 데이터의 절반만 고려하고 나머지는 무시하는 경향이 있습니다.

사분위수 중앙값을 중심 측도로 사용하고, IQR(사분위수 범위) 을 1.34898 로 나눈 값을 산포 측도로 사용합니다. 데이터가 정규 분포를 따르는 경우 IQR 을 1.34898 인수로 나누면 산포가 1 표준편차에 해당합니다.

다변량

다변량 데이터에서 이상치를 탐색하기 위한 두 가지 옵션이 있습니다.

로버스트 PCA 이상치 데이터를 낮은 계수 행렬과 잔차로 분해하고 잔차를 사용하여 이상치를 감지합니다. 자세한 내용은 "**로버스트 PCA 이상치**"에서 확인하십시오. 람다 값을 지정하고 데이터를 중심화할지 여부를 선택할 수 있습니다. 고급 옵션의 경우 Shift 키를 누르고 "로버스트 PCA 이상치" 버튼을 클릭하여 "로버스트 PCA 이상치" 옵션 창에 액세스합니다.

람다 잔차 행렬의 희소성을 결정하는 값을 지정합니다. 람다 값이 클수록 잔차 행렬의 희소성도 커집니다. n 개의 훈련 행과 p 개의 열이 있는 데이터 테이블의 경우 기본 람다 값은 다음과 같이 정의됩니다.

$$\lambda = \frac{2}{\sqrt{\max(n, p)}}$$

최대 반복 최대 SVD 반복 수를 지정합니다. 기본 반복 수는 100 입니다. 시작 창에 20,000 개가 넘는 열을 지정한 경우 기본 반복 수는 50 입니다.

참고 : 최대 반복 수에 도달한 후에도 알고리즘이 수렴하지 않으면 JMP 경고가 표시됩니다. 계속해서 반복하거나 취소할 수 있습니다. "취소"를 클릭한 경우 덜 엄격한 수렴 기준이 충족되면 결과가 표시됩니다. "취소"를 클릭한 경우 덜 엄격한 수렴 기준이 충족되지 않으면 결과를 수락할지 여부를 묻는 다른 JMP 경고가 표시됩니다.

수렴 기준 알고리즘의 중지 시기를 결정합니다. 기본 수렴 기준 값은 시작 창에 지정된 열 수를 기준으로 설정됩니다.

- 열 수가 2,000 개 미만이면 기본값은 1e-7 입니다.
- 열 수가 2,000 개 이상이면 기본값은 1e-6 입니다.
- 열 수가 20,000 개 이상이면 기본값은 1e-5 입니다.
- 덜 엄격한 수렴 기준은 원래 수렴 기준의 1000 배로 설정됩니다.

이상치 임계값 " 큰 잔차 셀 " 테이블에 표시되는 이상치를 결정하는 이상치 임계값을 지정합니다. 척도화된 잔차가 다음 값보다 크면 관측값이 표시됩니다.

$\min[0.99 \times \max\{\text{abs}(\text{척도화된 잔차})\}, \text{이상치 임계값}]$

이상치 임계값의 기본값은 2 입니다. 2 를 이상치 임계값으로 사용하여 이상치가 백만 개가 넘는 경우 이상치 임계값이 3 으로 변경됩니다.

중심 로버스트 PCA 이상치 알고리즘을 수행하기 전에 데이터를 중심화할지 여부를 결정합니다.

참고 : 행 수가 10 보다 작거나 같으면 데이터가 중심화되지 않습니다.

척도화 로버스트 PCA 이상치 알고리즘을 수행하기 전에 데이터를 척도화할지 여부를 결정합니다.

참고 : 행 수가 10 보다 작거나 같으면 데이터가 척도화되지 않습니다.

변수의 수가 매우 많은 와이드 문제의 경우 랜덤화 SVD (시작 창에 지정된 열 수가 1000 개 이상인 경우에만 사용 가능) Lanczos 방법 대신 랜덤화 SVD 방법을 사용하여 데이터를 분해합니다. 이 옵션을 사용하면 가로형 데이터에 대한 로버스트 PCA 이상치 계산 속도가 빨라집니다. 자세한 내용은 **다변량 방법**에서 확인하십시오.

랜덤화 차원 (시작 창에 지정된 열 수가 1000 개 이상인 경우에만 사용 가능) 랜덤화 SVD 에 사용되는 차원 수를 지정합니다.

K 최근접 이웃 이상치 이상치를 k 개 최근접 이웃에서 멀리 떨어진 값으로 식별합니다. 자세한 내용은 "**K 최근접 이웃 이상치**" 에서 확인하십시오. 다음 옵션을 지정할 수 있습니다.

K 고려할 가장 먼 이웃의 상한을 지정합니다. 기본값은 8 입니다.

결측값 대치 결측값 대치 여부를 지정합니다.

이상치 탐색 보고서

"이상치 탐색" 보고서에는 선택한 각 이상치 감지 방법에 대한 보고서가 포함됩니다.

- "**분위수 범위 이상치 보고서**"

- " 로버스트 적합 이상치 보고서 "
- " 로버스트 PCA 이상치 보고서 "
- "K 최근접 이웃 보고서 "

분위수 범위 이상치 보고서

이상치 탐색 플랫폼의 "분위수 범위 이상치" 보고서에는 일련의 컨트롤 및 여러 탭에 구성된 결과가 포함됩니다.

분위수 범위 이상치 컨트롤

분위수 범위 이상치 컨트롤을 사용하면 이상치를 식별하는 방법을 지정할 수 있습니다. 이상치는 1사분위수와 3사분위수의 분위수 간 범위의 Q 배를 넘는 값으로 간주됩니다. Q 값과 분위수 간 범위 크기를 조정할 수 있습니다.

그림 21.6 분위수 범위 이상치 초기 옵션

꼬리 분위수 분위수 간 범위를 계산하는 데 사용되는 1사분위수에 대한 확률입니다. 3사분위수 확률은 1 - 꼬리 분위수로 간주됩니다. 예를 들어 꼬리 분위수 값이 0.1 이면 분위수 간 범위는 데이터의 0.1 분위수에서 0.9 분위수 사이임을 의미합니다. 기본값은 0.1 입니다.

Q 이상치 임계값을 결정하는 승수입니다. 꼬리 분위수 또는 1 - 꼬리 분위수 값을 넘어 분위수 간 범위의 Q 배를 초과하는 값이 이상치로 식별됩니다. Q 값이 크면 작은 값보다 더 보수적인 이상치 집합이 제공됩니다. 기본값은 3 입니다.

검색을 정수로 제한 이상치 값을 정수 값으로만 제한합니다. 이 설정은 업종별 결측값 코드 및 오류 코드를 찾기 위해 이상치 검색을 제한합니다.

다시 스캔 이상치 처리 동작을 수행한 후 데이터를 다시 스캔합니다.

팁 : 열려 있는 모든 이상치 방법에서 다시 스캔하려면 Ctrl 키를 누른 채 **다시 스캔**을 클릭합니다.

닫기 " 분위수 범위 이상치 " 보고서를 닫습니다.

팁 : 모든 이상치 보고서를 닫으려면 Ctrl 키를 누른 채 **닫기**를 클릭합니다.

열별 이상치

"분위수 범위 이상치" 보고서의 "열별 이상치" 탭에는 컨트롤에 지정된 설정으로 찾은 이상치가 있는 열 테이블이 포함되어 있습니다. 이 테이블에는 1사분위수/3사분위수 및 낮은 임계/높은 임계에 대한 값이 표시됩니다. 이러한 임계값 한계를 벗어나는 값을 이상치로 간주합니다. 각 열의 이상치 수가 표시됩니다. 각 이상치 값이 테이블의 마지막 열에 나열됩니다. 한 열에 두 번 이상 나오는 이상치는 괄호 안에 개수가 표시됩니다.

"열별 이상치" 탭에는 이상치 테이블에서 하나 이상의 행을 선택한 경우 적용할 수 있는 다음 옵션이 포함되어 있습니다.

이상치를 포함하는 열만 표시 "열별 이상치" 탭의 테이블에서 이상치가 없는 열을 제거합니다.

테이블에서 이상치 식별 이상치 요약 테이블에서 선택한 행의 원래 데이터 테이블에 작업을 적용합니다.

행 선택하기 이상치를 포함하는 행을 선택합니다.

행 제외 제외 행 상태를 적용합니다. "분위수 범위 이상치" 보고서를 업데이트하려면 **다시 스캔**을 클릭합니다.

셀에 색상 적용 이상치를 포함하는 셀에 색상을 적용합니다. 낮은 값의 이상치는 파란색으로 표시되고 높은 값의 이상치는 빨간색으로 표시됩니다.

행에 색상 적용 이상치를 포함하는 행에 색상을 적용합니다.

테이블에서 이상치 지우기 이상치 요약 테이블에서 선택한 행의 원래 데이터 테이블에 작업을 적용합니다.

결측값 코드에 추가 이상치를 결측값 코드 열 특성에 추가합니다. 이 옵션을 사용하면 데이터 내에서 알려진 결측값 또는 오류 코드를 식별할 수 있습니다. 결측값 및 오류 코드는 대개 정수이고 9 모음값일 때도 있습니다. "분위수 범위 이상치" 보고서를 업데이트하려면 **다시 스캔**을 클릭합니다.

참고: 시작 창에서 기준 변수를 지정한 경우 "분위수 범위 이상치" 에서 "결측값 코드에 추가" 를 사용할 수 없습니다.

결측으로 변경 이상치 값을 결측값으로 변경합니다. 값을 결측값으로 변경할 때 주의해야 합니다. 데이터가 올바르지 않거나 정확하지 않을 때만 결측값으로 변경합니다. "분위수 범위 이상치" 보고서를 업데이트하려면 **다시 스캔**을 클릭합니다.

참고: 선택한 이상치가 결측값 코드에 추가된 경우 이상치가 결측값으로 변경되지 않습니다.

계산식 열 이상치를 결측으로 설정하기 위해 각 열에 대한 새 계산식 열을 생성합니다. 새 열은 원래 열과 구분하기 위해 사용자가 지정한 이름을 접두사 또는 접미사로 사용합니다. 기본적으로 접미사는 "선별" 로 설정됩니다.

계산식 스크립트 데이터 테이블에 추가되는 스크립트를 생성합니다. 스크립트가 실행되면 이상치를 결측으로 설정하기 위해 각 열에 대해 새 계산식 열이 생성됩니다. 새 열은 원래 열과 구분하기 위해 사용자가 지정한 이름을 접두사 또는 접미사로 사용합니다. 기본적으로 접미사는 " 선별 "로 설정됩니다.

셀별 이상치

"분위수 범위 이상치" 보고서의 "셀별 이상치" 탭에는 컨트롤에 지정된 설정으로 찾은 개별 이상치 테이블이 포함되어 있습니다. 테이블에는 열 이름, 행 번호, 이상치 거리 및 개별 이상치의 실제 값이 표시됩니다. 이상치 거리는 이상치가 얼마나 극단적인지를 나타내는 척도이며 다음 방정식을 사용하여 계산됩니다.

$$\text{이상치 거리} = \frac{|x - m|}{IQR}$$

다음은 각 요소에 대한 설명입니다.

x = 이상치의 실제 값

m = 이상치가 있는 열의 중앙값

IQR = 지정된 꼬리 분위수를 사용한 분위수 간 범위

이상치 거리가 클수록 더 극단적인 이상치를 나타냅니다.

"셀별 이상치" 탭에는 이상치 테이블에서 하나 이상의 행을 선택한 경우 적용할 수 있는 다음 옵션이 포함되어 있습니다.

테이블에서 이상치 식별 이상치 요약 테이블에서 선택한 행의 원래 데이터 테이블에 작업을 적용합니다.

행 / 열 선택 선택한 이상치에 해당하는 행과 열을 선택합니다.

셀에 색상 적용 데이터 테이블에서 이상치 셀에 색상을 적용합니다. 낮은 값의 이상치는 파란색으로 표시되고 높은 값의 이상치는 빨간색으로 표시됩니다.

테이블에서 이상치 지우기 이상치 요약 테이블에서 선택한 행의 원래 데이터 테이블에 작업을 적용합니다.

결측값 코드에 추가 선택한 이상치를 결측값 코드 열 특성에 추가합니다. 이 옵션을 사용하면 데이터 내에서 알려진 결측값 또는 오류 코드를 식별할 수 있습니다. 결측값 및 오류 코드는 대개 정수이고 9 모음값일 때도 있습니다. "분위수 범위 이상치" 보고서를 업데이트하려면 **다시 스캔**을 클릭합니다.

참고 : 시작 창에서 기준 변수를 지정한 경우 "분위수 범위 이상치"에서 "결측값 코드에 추가"를 사용할 수 없습니다.

결측으로 변경 이상치 값을 결측값으로 변경합니다. 값을 결측값으로 변경할 때 주의해야 합니다. 데이터가 올바르게 않거나 정확하지 않을 때만 결측값으로 변경합니다. "분위수 범위 이상치" 보고서를 업데이트하려면 **다시 스캔**을 클릭합니다.

참고 : 선택한 이상치가 결측값 코드에 추가된 경우 이상치가 결측값으로 변경되지 않습니다.

행별 이상치

"행별 이상치" 탭에는 컨트롤에 지정된 설정으로 찾은 이상치가 있는 행 테이블이 포함되어 있습니다. 이 테이블에는 행 번호와 해당 행에서 발견된 이상치 수가 표시됩니다.

"행별 이상치" 탭에는 이상치 테이블에서 하나 이상의 행을 선택한 경우 적용할 수 있는 다음 옵션이 포함되어 있습니다.

테이블에서 이상치 식별 이상치 요약 테이블에서 선택한 행의 원래 데이터 테이블에 작업을 적용합니다.

행 선택하기 이상치가 있는 행을 선택합니다.

행 제외 제외 행 상태를 적용합니다. "분위수 범위 이상치" 보고서를 업데이트하려면 **다시 스캔**을 클릭합니다.

행에 색상 적용 이상치를 포함하는 행에 색상을 적용합니다.

9 모음값

"분위수 범위 이상치" 보고서의 "9모음값" 탭에는 가능성 있는 결측값 코드가 포함된 열 테이블이 표시됩니다. 이러한 결측값 코드는 연속된 9(보통 9999)로 표시되며, 모두 9이고 3사분위수보다 높은 가장 큰 숫자입니다. 그 수가 많으면 이러한 이상치가 실제로 결측값 코드일 가능성이 높습니다. 그 수가 매우 적으면 추가 분석을 통해 해당 값이 이상치인지 아니면 결측값 코드인지 판별해야 합니다. 테이블에는 3사분위수 값도 포함됩니다.

"9 모음값" 탭은 가능성 있는 결측값 코드가 식별된 경우에만 표시됩니다.

"9모음값" 탭에는 테이블에서 하나 이상의 행을 선택한 경우 선택할 수 있는 다음 옵션이 포함되어 있습니다.

결측값 코드에 가장 큰 9 모음값 추가 선택한 이상치 값을 결측값 코드 열 특성에 추가합니다. "분위수 범위 이상치" 보고서를 업데이트하려면 **다시 스캔**을 클릭해야 합니다.

참고 : 시작 창에서 기준 변수를 지정한 경우 "분위수 범위 이상치" 에서 "결측값 코드에 가장 큰 9 모음값 추가" 를 사용할 수 없습니다.

가장 높은 9 모음값을 결측값으로 변경 선택한 이상치 값을 데이터 테이블의 결측값으로 바꿉니다.

참고 : 데이터 변경을 위해 처음으로 작업 (예 : 결측으로 변경 또는 행 제외) 을 선택하면 원래 데이터의 복사본을 유지하기 위해 다른 이름으로 저장 명령을 사용하여 데이터 테이블을 새 파일로 저장하라는 경고 창이 표시됩니다. 이 창이 나타나면 확인을 클릭합니다. 새 데이터 파일을 저장하기로 결정하면 파일을 새 이름으로 저장하라는 메시지가 자동으로 표시됩니다.

로버스트 적합 이상치 보고서

이상치 탐색 플랫폼의 "로버스트 적합 이상치" 보고서에는 일련의 컨트롤 및 여러 탭에 구성된 결과가 포함됩니다.

로버스트 적합 이상치 초기 옵션

"로버스트 적합 이상치" 컨트롤은 로버스트 추정값 및 승수 K 를 계산하는 데 사용되는 방법을 지정합니다. 중심 및 산포에 대한 로버스트 추정값이 주어지면 이상치는 로버스트 중심에서 로버스트 산포의 K 배 이상 떨어진 값으로 정의됩니다.

그림 21.7 로버스트 적합 이상치 컨트롤

Huber Huber M- 추정을 사용하여 중심과 산포를 추정합니다. 이 옵션이 기본값입니다. 자세한 내용은 Huber와 Ronchetti 연구 자료 (2009)에서 확인하십시오.

Cauchy Cauchy 분포를 가정하여 중심 및 산포 추정값을 계산합니다. Cauchy 추정값은 높은 봉피점을 가지며 일반적으로 Huber 추정값보다 더 로버스트합니다. 그러나 데이터가 균집으로 분리된 경우 Cauchy 분포는 더 가깝게 균집화된 데이터의 절반만 고려하고 나머지는 무시하는 경향이 있습니다.

사분위수 중앙값을 중심 측도로 사용하고, IQR(사분위수 범위)을 1.34898로 나눈 값을 산포 측도로 사용합니다. 데이터가 정규 분포를 따르는 경우 IQR을 1.34898 인수로 나누면 산포가 1 표준편차에 해당합니다.

K 시그마 중심에서 산포의 K 배 이상 떨어진 값을 이상치로 결정하는 승수입니다. K 값이 크면 작은 값보다 더 보수적인 이상치 집합이 제공됩니다. 기본값은 4입니다.

다시 스캔 이상치 처리 동작을 수행한 후 데이터를 다시 스캔합니다.

팁: 열려 있는 모든 이상치 방법에서 다시 스캔하려면 Ctrl 키를 누른 채 **다시 스캔**을 클릭합니다.

닫기 "로버스트 적합 이상치" 패널을 닫습니다.

팁: 모든 이상치 보고서를 닫으려면 Ctrl 키를 누른 채 **닫기**를 클릭합니다.

열별 이상치

"로버스트 적합 이상치" 보고서의 "열별 이상치" 탭에는 시작 창에서 선택한 각 열에 대한 행이 있는 테이블이 포함되어 있습니다. 테이블의 열은 데이터의 중심과 산포를 추정하는 데 사용되는 기법 (Huber, Cauchy 또는 사분위수)에 따라 달라집니다. 각 기법에 대해 추정 중심, 추정 산포, 중심과 산포에 기반한 이상치 수에 해당하는 열이 있습니다.

"열별 이상치" 탭에는 이상치 테이블에서 하나 이상의 행을 선택한 경우 적용할 수 있는 다음 옵션이 포함되어 있습니다.

이상치를 포함하는 열만 표시 "열별 이상치" 탭의 테이블에서 이상치가 없는 열을 제거합니다.

테이블에서 이상치 식별 이상치 요약 테이블에서 선택한 행의 원래 데이터 테이블에 작업을 적용합니다.

행 선택하기 이상치를 포함하는 행을 선택합니다.

행 제외 제외 행 상태를 적용합니다. "로버스트 적합 이상치" 보고서를 업데이트하려면 **다시 스캔**을 클릭합니다.

셀에 색상 적용 이상치를 포함하는 셀에 색상을 적용합니다. 낮은 값의 이상치는 파란색으로 표시되고 높은 값의 이상치는 빨간색으로 표시됩니다.

행에 색상 적용 이상치를 포함하는 행에 색상을 적용합니다.

테이블에서 이상치 지우기 이상치 요약 테이블에서 선택한 행의 원래 데이터 테이블에 작업을 적용합니다.

결측값 코드에 추가 선택한 이상치를 결측값 코드 열 특성에 추가합니다. 이 옵션을 사용하면 데이터 내에서 알려진 결측값 또는 오류 코드를 식별할 수 있습니다. "로버스트 적합 이상치" 보고서를 업데이트하려면 **다시 스캔**을 클릭합니다.

참고: 시작 창에서 기준 변수를 지정한 경우 "로버스트 적합 이상치" 에서 "결측값 코드에 추가" 를 사용할 수 없습니다.

결측으로 변경 이상치 값을 결측값으로 변경합니다. "로버스트 적합 이상치" 보고서를 업데이트하려면 **다시 스캔**을 클릭합니다.

계산식 열 이상치를 결측으로 설정하기 위해 각 열에 대한 새 계산식 열을 생성합니다. 새 열은 원래 열과 구분하기 위해 사용자가 지정한 이름을 접두사 또는 접미사로 사용합니다. 기본적으로 접미사는 "선별"로 설정됩니다.

계산식 스크립트 데이터 테이블에 추가되는 스크립트를 생성합니다. 스크립트가 실행되면 이상치를 결측으로 설정하기 위해 각 열에 대해 새 계산식 열이 생성됩니다. 새 열은 원래 열과 구분하기 위해 사용자가 지정한 이름을 접두사 또는 접미사로 사용합니다. 기본적으로 접미사는 "선별"로 설정됩니다.

셀별 이상치

"로버스트 적합 이상치" 보고서의 "셀별 이상치" 탭에는 컨트롤에 지정된 설정으로 찾은 개별 이상치 테이블이 포함되어 있습니다. 테이블에는 열 이름, 행 번호, 이상치 거리 및 개별 이상치의 실제값이 표시됩니다. 이상치 거리는 이상치가 얼마나 극단적인지를 나타내는 척도이며 다음 방정식을 사용하여 계산됩니다.

$$\text{이상치 거리} = \frac{|x - c|}{s}$$

다음은 각 요소에 대한 설명입니다.

x = 이상치의 실제값

c = 이상치를 포함하는 열의 중심. 지정된 이상치 방법 (Huber, Cauchy 또는 사분위수) 으로 추정됩니다.

s = 이상치를 포함하는 열의 산포. 지정된 이상치 방법 (Huber, Cauchy 또는 사분위수) 으로 추정됩니다.

이상치 거리가 클수록 더 극단적인 이상치를 나타냅니다.

"셀별 이상치" 탭에는 이상치 테이블에서 하나 이상의 행을 선택한 경우 적용할 수 있는 다음 옵션이 포함되어 있습니다.

테이블에서 이상치 식별 이상치 요약 테이블에서 선택한 행의 원래 데이터 테이블에 작업을 적용합니다.

행 / 열 선택 선택한 이상치에 해당하는 행과 열을 선택합니다.

셀에 색상 적용 이상치를 포함하는 셀에 색상을 적용합니다. 낮은 값의 이상치는 파란색으로 표시되고 높은 값의 이상치는 빨간색으로 표시됩니다.

테이블에서 이상치 지우기 이상치 요약 테이블에서 선택한 행의 원래 데이터 테이블에 작업을 적용합니다.

결측값 코드에 추가 선택한 이상치를 결측값 코드 열 특성에 추가합니다. 이 옵션을 사용하면 데이터 내에서 알려진 결측값 또는 오류 코드를 식별할 수 있습니다. 결측값 및 오류 코드는 대개 정수이고 9 모음값일 때도 있습니다. "로버스트 적합 이상치" 보고서를 업데이트하려면 **다시 스캔**을 클릭합니다.

참고: 시작 창에서 기준 변수를 지정한 경우 "로버스트 적합 이상치" 에서 "결측값 코드에 추가" 를 사용할 수 없습니다.

결측으로 변경 데이터 테이블에서 이상치 값을 결측값으로 변경합니다. 값을 결측값으로 변경할 때 주의해야 합니다. 데이터가 올바르게 않거나 정확하지 않을 때만 결측값으로 변경합니다. "로버스트 적합 이상치" 보고서를 업데이트하려면 **다시 스캔**을 클릭합니다.

참고 : 선택한 이상치가 결측값 코드에 추가된 경우 이상치가 결측값으로 변경되지 않습니다.

로버스트 PCA 이상치 보고서

이상치 탐색 플랫폼의 "로버스트 PCA 이상치" 보고서에는 알고리즘에 대한 정보 테이블과 여러 결과 테이블이 포함되어 있습니다. 초기 테이블에는 낮은 계수 행렬의 계수, SVD 반복 수, 수렴 기준, 람다 값 및 대치된 결측값 수가 포함됩니다. "랜덤화 SVD" 옵션이 활성화된 경우에는 "랜덤화 SVD" 옵션에 사용되는 차원 수도 요약 테이블에 포함됩니다. 다음은 보고서에 포함되는 다른 테이블입니다.

큰 잔차 셀 열 이름과 행 번호로 식별되는 가장 큰 이상치 관측값을 보여 주는 테이블입니다. 표시되는 관측값 수는 이상치 임계값에 의해 결정됩니다. 이 테이블에는 관측값의 열 이름과 행 번호, 잔차 값 및 척도화된 잔차 값이 포함됩니다.

팁 : 데이터 테이블의 특정 이상치 셀에 색상을 적용하려면 "큰 잔차 셀" 테이블에서 행을 선택하고 **색상 지정**을 클릭합니다.

행 제공근 평균 제공 데이터 테이블의 각 행에 대한 제공근 평균 제공 값을 보여 주는 테이블입니다. 제공근 평균 제공은 척도화된 잔차를 사용하여 계산됩니다.

팁 : "행 제공근 평균 제공" 테이블에서 행을 선택하면 데이터 테이블에서 해당 행이 선택됩니다.

열 제공근 평균 제공 시작 창에 지정된 각 열에 대한 제공근 평균 제공 값을 보여 주는 테이블입니다. 제공근 평균 제공은 척도화된 잔차를 사용하여 계산됩니다.

팁 : "열 제공근 평균 제공" 테이블에서 행을 선택하고 **열 선택**을 클릭하면 데이터 테이블에서 해당 열이 선택됩니다.

스냅샷 데이터 테이블에 있는 이상치 셀의 그래픽 표현입니다. 이상치 셀은 빨간색으로 표시됩니다.

잔차 행렬 분해의 잔차 행렬입니다. 척도화된 잔차의 절대값이 다음 값보다 크면 셀에 색상이 적용됩니다.

$\min[0.99 \times \max\{\text{abs}(\text{residuals})\}, \text{이상치 임계값}]$

낮은 계수 근사 행렬 분해의 척도화된 잔차 행렬입니다.

특이값 SVD의 특이값 벡터입니다.

로버스트 PCA 이상치 옵션

"로버스트 PCA 이상치" 보고서 아래쪽에 보고서의 각 항목을 저장하는 옵션을 제공하는 버튼이 있습니다.

닫기 "로버스트 PCA 이상치" 보고서를 닫습니다.

큰 이상치 저장 "큰 잔차 셀" 테이블의 정보를 새 데이터 테이블에 저장합니다.

정리된 항목 저장 임계값을 기반으로 이상치를 정리하고 새 열을 데이터 테이블에 저장하는 몇 가지 기법을 제공하는 창을 엽니다.

절사 해당하는 척도화된 절대 잔차가 지정된 임계값보다 크면 이상치 셀을 제거합니다. 기본적으로 임계값은 10 입니다. 이상치 셀을 빨간색으로 표시하려면 "색상 ■" 을 선택합니다. 절사 셀은 비척도화 임계값으로 설정됩니다.

결측값 대치 해당하는 척도화된 절대 잔차가 지정된 임계값보다 크면 이상치 셀을 낮은 계수 근사의 값으로 설정합니다. 기본적으로 임계값은 100 입니다. 해당 셀을 녹색으로 표시하려면 "색상 ■" 을 선택합니다.

결측값 생성 해당하는 척도화된 절대 잔차가 지정된 임계값보다 크면 이상치 셀을 결측으로 설정합니다. 기본적으로 임계값은 1000 입니다. 해당 셀을 파란색으로 표시하려면 "색상 ■" 을 선택합니다.

결측값의 대치된 값 색상 적용 ■ 이 옵션을 선택하면 원래 있던 결측값이 대치된 셀에 색상을 적용합니다.

잔차 저장 잔차를 원래 데이터 테이블의 새 열에 저장합니다.

척도화된 잔차 저장 척도화된 잔차를 원래 데이터 테이블의 새 열에 저장합니다.

낮은 계수 근사 저장 낮은 계수 근사를 원래 데이터 테이블의 새 열에 저장합니다.

K 최근접 이웃 보고서

"이상치 탐색" 플랫폼의 "K 최근접 이웃 이상치" 보고서에는 최대 K 값까지 선택한 k 값에 대한 그림이 포함되어 있습니다. 각 그림의 k 값은 세로 축 라벨에 표시됩니다. 라벨은 "이웃 k 까지의 거리 $= \langle a \rangle$ " 형식으로 표시되며, 여기서 " a "는 가장 가까운 a 번째 이웃을 나타내는 정수입니다. 각 그림에는 i 번째 행의 점에서 a 번째 최근접 이웃까지의 거리가 표시됩니다. 여러 k 값에 대해 이웃과의 거리가 먼 점은 이상치일 가능성이 높습니다.

그림 위의 버튼으로 다음을 수행합니다.

선택된 행 제외 선택한 점에 해당하는 행을 추가 분석에서 제외합니다. 데이터 테이블에서 해당 행에 "제외" 행 상태가 할당됩니다. "K 최근접 이웃" 보고서를 다시 실행할지 아니면 닫을지 묻는 메시지가 표시됩니다. 분석을 다시 실행하면 새로운 최근접 이웃이 식별됩니다. 그림이 업데이트되고 제외된 점이 표시되지 않습니다.

산점도 행렬 분석의 모든 열에 대한 산점도 행렬이 포함된 별도의 창을 엽니다. 잠재적 이상치를 "K 최근접 이웃" 그림에서 선택하고 산점도 행렬에서 확인하여 탐색할 수 있습니다.

NN 거리 저장 각 행에서 n 번째 최근접 이웃까지의 거리를 데이터 테이블의 새 열로 저장합니다.

닫기 "K 최근접 이웃" 보고서를 닫습니다.

가장 큰 이상치

"K 최근접 이웃" 보고서에는 "가장 큰 이상치" 테이블도 포함됩니다. 이 테이블에는 K 번째 최근접 이웃과의 거리가 가장 큰 20 개의 관측값이 포함됩니다. 테이블에 다음 열이 있습니다.

행 관측값의 행 번호입니다.

거리 지정된 행의 관측값과 K 번째 최근접 이웃의 거리입니다. 테이블은 이 열을 기준으로 내림차순으로 정렬됩니다.

근접 이웃 k 개의 최근접 이웃에 대한 행 번호를 나열합니다. 첫 번째 행 번호가 가장 가까운 이웃입니다. 마지막 행 번호는 K 번째 최근접 이웃이며 이 관측값과 지정된 행 사이의 거리가 "거리" 열에 표시됩니다.

열 <n> 해당 RMS 값의 열 이름을 지정합니다.

RMS<n> 각 열에 대한 k 개의 최근접 이웃에서 제공된 평균 제공 차이입니다. 가장 큰 RMS 값이 순서대로 표시되며, 여기서 RMS1 이 최대 RMS 값입니다. p 번째 RMS 값은 다음과 같이 계산됩니다.

$$RMS_p = \sqrt{\frac{\sum_{k=1}^K (D_{p,i} - D_{p,i_k})^2}{K}}$$

다음은 각 요소에 대한 설명입니다.

D_p = p 번째 열

$D_{p,i}$ = i 행에 대한 p 번째 열의 값

$D_{p,ik}$ = i 행의 k 번째 최근접 이웃에 대한 p 번째 열의 값

참고: 가장 큰 이상치 테이블에 표시된 "열" 및 "RMS" 열의 수는 시작 창에 지정된 열 수와 5 중 더 작은 값입니다.

이상치 탐색 플랫폼 옵션

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

다음 옵션에 대한 자세한 내용은 JMP 사용의 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.

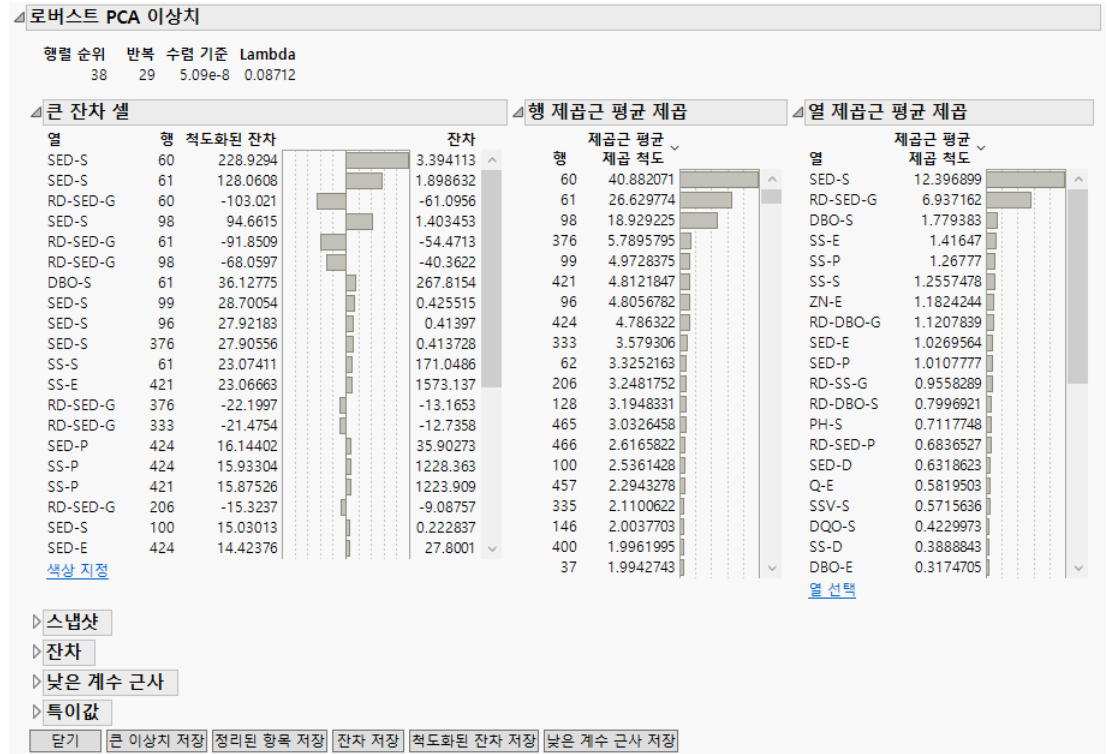
그룹별 스크립트 저장 기준 변수의 모든 수준에 대한 플랫폼 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다. 시작 창에서 기준 변수를 지정한 경우에만 사용할 수 있습니다.

이상치 탐색 플랫폼의 추가 예

로버스트 PCA 이상치 방법을 사용하여 도시 폐수 처리장의 데이터에서 잠재적 이상치를 탐색합니다. 잠재적 이상치에는 센서 고장, 악천후 및 기타 상황이 포함될 수 있습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Water Treatment.jmp 를 엽니다.
2. **분석 > 선별 > 이상치 탐색**을 선택합니다.
3. **Sensor Measurements** 열 그룹을 선택하고 **Y, 열**을 클릭합니다.
4. **확인**을 클릭합니다.
5. Shift 키를 누른 채 **로버스트 PCA 이상치**를 클릭합니다.
6. "이상치 임계값" 옆에 "10" 을 입력합니다.
다른 옵션에는 기본값을 사용합니다.
7. **확인**을 클릭합니다.

그림 21.8 로버스트 PCA 이상치 보고서



"큰 잔차 셀" 테이블에는 척도화된 잔차가 10 보다 크거나 -10 보다 작은 셀이 표시됩니다. SED-S 열에 척도화된 잔차가 큰 셀이 여러 개 있습니다. "열 제공근 평균 제공" 테이블을 살펴보면 실제로 SED-S 열의 제공근 평균 제공 척도 값이 가장 크다는 것을 알 수 있습니다. 제공근 평균 제공 척도 값이 가장 큰 행은 60 행입니다.

8. 보고서 맨 아래에서 **정리된 항목 저장**을 클릭합니다.

이 옵션을 사용하면 값을 절사하고, 새 값을 대치하거나, 셀을 결측으로 설정하여 이상치가 큰 셀을 조정할 수 있습니다. "정리된 열 저장" 보고서에 이러한 옵션이 제공됩니다.

9. "결측값 대치"를 선택합니다. "절사"는 기본적으로 선택되어 있습니다.

10. **확인**을 클릭합니다.

Water Treatment.jmp 데이터 테이블에 38 개의 "정리된" 열이 새로 추가되었습니다. 60 행의 제공근 평균 제공 척도 값이 가장 크다는 것을 알고 있으므로 이 행으로 스크롤하여 절사 및 결측값 대치된 일부 셀을 확인합니다. 이제 모든 유형의 분석에서 정리된 열을 사용할 수 있습니다.

그림 21.9 로버스트 PCA 이상치의 정리된 열

Water Treatment		RD-DQO-S	RD-DBO-G	RD-DQO-G	RD-SS-G 정리됨	RD-SED-G 정리됨
장기 파일 C:\Program Files\MS Reference Bache K. & Lichma Note These data come from t K Means Cluster		정리됨	정리됨	정리됨		
	57	53.7	92.1	66.9	94.6	100.0
	58	58.2	92.2	73.8	90.2	99.4
	59	66.1	89.0	69.0	86.5	99.6
	60	25.7	31.3	36.0	43.0	97.5
	61	43.3	42.3	50.1	24.5	91.4
	62	20.4	34.8	31.7	20.9	95.4
	63	52.9	84.3	54.1	43.7	100.0
	64	61.3	87.0	71.4	78.2	99.2
	65	36.6	85.9	60.4	90.7	100.0
	66	96.8	94.4	98.1	92.1	100.0
	67	70.4	91.3	73.4	89.4	99.4
	68	49.3	80.8	60.8	76.1	99.6
	69	68.1	87.4	72.2	88.3	100.0
행						
모든 행	527					
선택됨	0					
제외	0					
숨김	0					
라벨 항목	0					

결측값 탐색

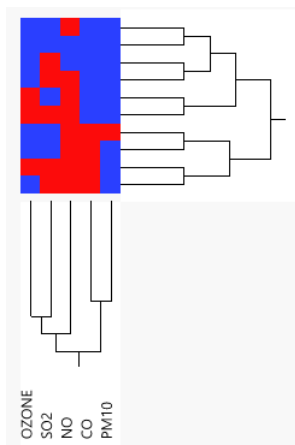
결측값 탐색 및 대처

결측값 탐색 플랫폼은 데이터의 결측값을 식별하고 파악할 수 있는 여러 가지 방법을 제공합니다. 결측값에 대해 다변량 대치를 수행하는 방법도 있습니다. 이러한 결측값 대처법은 데이터의 결측값이 랜덤이라고 가정합니다. 즉, 결측 데이터와 비결측 데이터 간의 차이를 연구에 포함된 다른 변수의 값으로 설명할 수 없다는 의미입니다. 결측값이 랜덤하지 **않다**고 의심되면 여러 플랫폼에서 사용할 수 있는 "결측값 정보화" 절차를 사용해 보십시오. 자세한 내용은 선형 모형 적합에서 확인하십시오.

데이터 집합에 결측값이 있으면 해당 데이터를 사용하여 결론을 내릴 때 영향을 줄 수 있습니다. 예를 들어 건강한 몇몇 참가자가 종단 연구에서 빠지고 그들의 데이터가 계속 누락된 경우 연구 결과는 남아 있는 건강하지 않은 개인에게 편향될 수 있습니다. 결측 데이터 값을 식별해야 하며 추가 분석을 수행하기 전에 해당 값을 파악해야 합니다.

주의: 결측값 대처 후 데이터를 분석할 때는 결과가 편향될 가능성이 있으므로 주의해야 합니다. 원래 데이터 테이블의 복사본을 만들거나, 대처된 값이 있는 데이터를 새 파일에 저장하는 것이 가장 좋습니다.

그림 22.1 결측값 탐색의 예



목차

결측값 탐색 플랫폼 개요.....	421
다변량 정규 대치법	421
다변량 SVD 대치법	421
다변량 RPCA 대치법.....	422
자동 데이터 결측값 대치.....	422
결측값 탐색 플랫폼의 예.....	423
결측값 탐색 플랫폼 시작.....	425
결측값 탐색 보고서	425
명령	426
결측값 대치 보고서	428
결측값 탐색 플랫폼 옵션.....	428

결측값 탐색 플랫폼 개요

결측값 탐색 플랫폼은 데이터의 결측값을 식별하고 파악할 수 있는 여러 가지 방법을 제공합니다.

- "다변량 정규 대치법"
- "다변량 SVD 대치법"
- "다변량 RPCA 대치법"
- "자동 데이터 결측값 대치"

다변량 정규 대치법

결측값 탐색 플랫폼의 다변량 정규 대치법은 다변량 정규 분포를 기반으로 결측값을 대치합니다. 절차를 수행하려면 모든 변수의 모델링 유형이 연속형이어야 합니다. 이 알고리즘은 최소 제곱 대치법을 사용합니다. 공분산 행렬은 쌍별 공분산을 사용하여 구성됩니다. 대각 항목(분산)은 각 변수에 대한 모든 비결측 값을 사용하여 계산됩니다. 두 변수의 비대각 항목은 두 변수에 대해 비결측인 모든 관측값을 사용하여 계산됩니다. 공분산 행렬이 특이 행렬인 경우 이 알고리즘은 Moore-Penrose 유사역행렬을 기반으로 하는 최소-노름 최소 제곱 대치법을 사용합니다.

다변량 정규 대치법 옵션을 사용하면 공분산에 대한 축소 추정량을 사용할 수 있습니다. 축소 추정량 사용은 공분산 행렬의 추정을 개선할 수 있는 방법입니다. 축소 추정량에 대한 자세한 내용은 Schäfer and Strimmer(2005) 연구 자료에서 확인하십시오.

참고: 검증 열이 지정되면 훈련 데이터 집합의 관측값을 사용하여 공분산 행렬이 계산됩니다.

다변량 SVD 대치법

결측값 탐색 플랫폼의 다변량 SVD 대치법은 SVD(특이값 분해)를 사용하여 결측값을 대치합니다. 이 방법은 변수의 수가 수백 또는 수천 개인 데이터에 유용합니다. SVD 계산에는 공분산 행렬 계산이 필요하지 않으므로 변수가 많이 포함된 광범위한 문제의 경우 SVD 방법이 권장됩니다. 절차를 수행하려면 모든 변수의 모델링 유형이 연속형이어야 합니다.

특이값 분해는 관측값 행렬 $\mathbf{X}$ 를 $\mathbf{X} = \mathbf{UDV}'$ 로 나타냅니다. 여기서 $\mathbf{U}$ 와 $\mathbf{V}$ 는 직교 행렬이고 $\mathbf{D}$ 는 대각 행렬입니다.

다변량 SVD 대치법에서 사용되는 SVD 알고리즘은 IRLBA(Implicitly Restarted Lanczos Bidiagonalization) 방법이라고도 하는 희소 Lanczos 방법입니다. 자세한 내용은 Baglama와 Reichel(2005) 연구 자료에서 확인하십시오. 다변량 SVD 대치법 알고리즘은 다음을 수행합니다.

1. 각 결측값이 해당 열의 평균으로 대체됩니다.
2. 관측값 행렬 $\mathbf{X}$ 에 대해 SVD 분해가 수행됩니다.
3. 결측값이 있었던 각 셀이 SVD 분해를 통해 얻은 $\mathbf{UDV}'$ 행렬의 해당 요소로 대체됩니다.

4. SVD가 행렬 X 로 수렴되거나 최대 반복 수에 도달할 때까지 2 단계와 3 단계를 반복합니다.

다변량 RPCA 대체법

결측값 탐색 플랫폼의 다변량 로버스트 PCA 방법은 로버스트 주성분을 사용하여 결측값을 대체합니다. 즉, 이상치에 로버스트한 SVD(낮은 계수 행렬 분해)를 사용하여 결측값을 대체합니다. 자세한 내용은 "[로버스트 PCA 이상치](#)"에서 확인하십시오. 이 방법은 변수의 수가 많은 와이드 문제에 유용하지만 매우 큰 규모에서 계산 비용이 많이 들 수 있습니다.

JMP® 자동 데이터 결측값 대체

결측값 탐색 플랫폼의 ADI(자동 데이터 결측값 대체) 방법은 행렬 완성이라고도 하는 낮은 계수 행렬 근사 방법을 사용하여 결측값을 대체합니다. 훈련된 ADI 모형은 스코어링 계산식을 통해 스트리밍 데이터에 대한 결측 데이터 대체를 수행할 수 있습니다. 스트리밍 데이터는 시간 경과에 따라 이용 가능해지고, 대체된 모형의 조정 또는 검증에 사용되지 않은 관측값 행이 추가된 것입니다. 이 방법은 낮은 계수 근사에 대한 최선의 차원을 선택할 수 있도록 유연하고, 로버스트하며, 자동화되었습니다. 이러한 기능을 통해 ADI가 다양한 유형의 데이터 집합에 대해 잘 작동할 수 있습니다.

행렬의 낮은 계수 근사는 $X = UDV'$ 형식이며 SVD(특이값 분해)의 확장으로 간주할 수 있습니다. ADI는 Soft-Impute 방법을 결측값 대체 모형으로 사용하며, 데이터가 낮은 계수 근사의 순위를 결정하도록 되어 있습니다.

ADI 알고리즘은 다음 단계를 수행합니다.

1. 데이터가 훈련 데이터 집합과 검증 데이터 집합으로 분할됩니다.
2. 훈련 데이터 집합의 관측값을 사용하여 각 집합이 중심화 및 척도화됩니다.
3. 분할된 각 데이터 집합에 대해 IM(유도된 결측값)이라고 하는 추가 결측값이 각 열 내에 더해집니다.
4. 조정 모수의 해 경로를 따라 결측값 대체 모형이 훈련 데이터 집합에 대해 적합됩니다. IM 값은 조정 모수의 최량 값을 결정하는 데 사용됩니다.
5. 4 단계에서 선택한 결측값 대체 모형 결과의 편향을 제거하여 훈련 데이터 집합을 사용한 순위 축소가 추가로 수행됩니다.
6. 스트리밍 데이터에 대한 모형을 보정하고 과대적합을 방지하기 위해 최종 순위 축소가 수행됩니다. 5 단계에서 결정된 순위를 상한으로 사용하고 검증 데이터 집합에 대해 결측값 대체 모형을 적합시켜 이 작업을 완료합니다.

이 알고리즘으로 낮은 계수 근사를 찾을 수 없으면 연속형 변수의 경우 평균이 대치되고 범주형 변수의 경우 최빈값이 대치됩니다. 이러한 값은 훈련 데이터 집합에서 계산됩니다.

결측값 탐색 플랫폼의 예

데이터에서 결측값을 탐색하고, 해당하는 경우 이러한 값을 대치하려고 합니다. 먼저 데이터에서 결측값을 탐색한 후 연속형 모델링 유형의 열에 대해 결측값을 대치합니다.

결측값 검토

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Arrhythmia.jmp** 를 엽니다.
2. **분석 > 선별 > 결측값 탐색**을 선택합니다.
3. 모든 열 (총 280 개) 을 선택하고 **Y, 열**을 클릭합니다.
4. **확인**을 클릭합니다. **결측값을 포함하는 열만 표시** 체크박스를 선택합니다.

그림 22.2 결측값 보고서

결측값 탐색

명령

결측값 보고서

결측값 군집화

결측값 스냅샷

다변량 정규 대치법

다변량 SVD 대치법

다변량 RPCA 대치법

자동 데이터 결측값 대치

각 열의 결측값 수를 나열합니다.

열 그룹에 유사한 결측 패턴이 있는지 여부를 판별하기 위해 결측 데이터의 계층적 군집화 분석을 제공합니다.

그래픽 맵을 사용하여 결측값 패턴을 찾습니다.

비결측 열에서 최소 제곱 예측을 사용하여 결측값을 대치합니다. 열의 수가 수백 개인 경우에는 이 방법을 사용하지 마십시오. 축소 옵션을 사용하여 공분산 행렬 추정을 향상시킬 수 있습니다.

반복된 낮은 계수 SVD 행렬 완성 방법을 사용하여 큰 문제에 대해 빠르게 결측값을 대치합니다.

로버스트 주성분을 사용하여 결측값을 대치합니다. 즉, 이상치에 로버스트한 낮은 계수 행렬 분해(SVD)를 사용하여 결측값을 대치합니다. 이 방법은 변수의 수가 많은 와이드 문제에 유용합니다.

낮은 계수 행렬 근사 방법을 사용하여 결측값을 대치합니다. 이 방법은 데이터를 기반으로 낮은 계수 근사에 대한 최선의 자원을 자동으로 선택합니다.

자동 데이터 결측값 대치 컨트롤

결측 열

결측값을 포함하는 열만 표시

닫기

열을 선택하고 작업을 선택하십시오.

행 선택하기

행에 색상 적용

행 제외

행에 색상 적용

열	결측 수
T	8
P	22
QRST	1
J	376
Heart rate	1

그림 22.2 에 표시된 " 결측 열 " 보고서에서는 5 개 열에만 결측 데이터가 있음을 나타냅니다. 총 452 개 행 중에서 J 열에 376 개의 결측값이 있습니다. 대부분의 값이 결측이므로 대치된 값을 사용한 분석 결과가 의미 없을 수 있습니다. 이러한 데이터의 경우 " 결측값 정보화 " 옵션

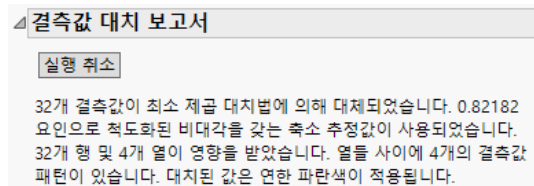
션을 지원하는 플랫폼에서 J 열에 대해 "결측값 정보화" 옵션을 사용하여 모형을 탐색할 수 있습니다.

결측값 대치

결측값이 있는 5 개 열이 연속형입니다. 데이터 테이블의 연속형 열에 대해 다변량 대치법을 사용하여 J 열을 제외한 4 개 열의 결측값을 대치합니다. 이렇게 하면 값이 결측일 확률은 제외된 명목형 변수의 값이 아니라 연속형 변수의 값에만 의존한다고 암묵적으로 가정합니다. 이 새 분석을 수행하려면 결측값 탐색 플랫폼을 다시 시작해야 합니다.

1. **분석 > 선별 > 결측값 탐색**을 선택합니다.
2. 시작 창에서 **280 개 열** 옆의 빨간색 삼각형을 클릭합니다.
열 필터 메뉴를 사용하여 "열 선택" 목록에서 연속형 모델링 유형의 열만 표시합니다.
3. **모델링 유형 > 모두 선택 해제**를 선택합니다.
이렇게 하면 "열 선택" 목록에서 모든 열이 제거됩니다.
4. **모델링 유형 > 연속형**을 선택합니다.
이제 "열 선택" 목록에 207 개의 연속형 열만 포함됩니다.
5. 207 개 열을 모두 선택합니다. 그런 다음 Ctrl 키를 누른 채 J 열을 클릭하여 선택 취소하고 Y, 열을 클릭합니다.
6. **확인**을 클릭합니다.
7. **다변량 정규 대치법**을 클릭합니다.
공분산에 대해 축소 추정량을 사용할지 묻는 창이 나타납니다.
8. **예 (축소)**를 클릭합니다.
원래 데이터를 유지하려면 **다른 이름으로 저장** 명령을 사용해야 함을 알리는 "JMP 경고"가 나타납니다.
9. **확인**을 클릭합니다.

그림 22.3 결측값 대치 보고서



"결측값 대치 보고서"에는 대치된 결측값 수와 결측값 대치 관련 상세 정보가 표시됩니다. 결측값이 있었던 4 개 열에 결측 데이터가 남아 있지 않습니다.

결측값 탐색 플랫폼 시작

분석 > 선별 > 결측값 탐색을 선택하여 결측값 탐색 플랫폼을 시작합니다.

그림 22.4 결측값 탐색 시작 창

결측값의 패턴을 찾고 대치를 수행합니다.

The image shows the 'Missing Value Explorer' start window. It is divided into three main sections: '열 선택' (Select Columns), '선택한 열 역할 지정' (Assign Roles to Selected Columns), and '작업' (Actions).
1. '열 선택' (Select Columns): A list of 16 variables is shown, with 'city' selected. The variables are: city, Latitude, Longitude, State, Region, pop- m, POP, Max deg. F Jan, OZONE, CO, SO2, NO, PM10, Lead, X, and Y.
2. '선택한 열 역할 지정' (Assign Roles to Selected Columns): This section contains three rows of controls. The first row has a 'Y, 열' (Y, Column) label and a '필수 선택적' (Required Optional) dropdown. The second row has a '검증' (Validation) label and a '선택적 숫자' (Optional Number) dropdown. The third row has a '기준' (Criteria) label and a '선택적' (Optional) dropdown.
3. '작업' (Actions): A vertical column of buttons including '확인' (OK), '취소' (Cancel), '제거' (Remove), '재호출' (Re-run), and '도움말' (Help).

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

Y, 열 분석할 열을 지정합니다. 각 열의 모델링 유형은 수치형이어야 합니다.

검증 일부 결측값 대체법에 사용되는 검증 열을 지정합니다.

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.

결측값 탐색 보고서

"결측값 탐색" 보고서에는 명령 섹션, 결측 열 보고서 및 결측값 대체 보고서(결측값 대체법이 선택된 경우)가 포함됩니다. "명령" 섹션에는 추가 보고서 및 결측값 대체법에 대한 몇 가지 옵션이 포함되어 있습니다.

- "명령"
- "결측값 대체 보고서"

명령

결측값 보고서 각 열의 이름과 해당 열의 결측값 수를 나열하는 "결측 열" 보고서를 표시합니다. "결측 열" 보고서에는 다음 옵션도 포함되어 있습니다.

결측값을 포함하는 열만 표시 결측값이 없는 열을 목록에서 제거합니다.

닫기 "결측 열" 보고서를 닫습니다.

행 선택하기 데이터 테이블에서 "결측 열" 보고서에서 선택한 열의 결측값이 포함된 행을 선택합니다.

행 제외 데이터 테이블에서 "결측 열" 보고서에서 선택한 열의 결측값이 포함된 행에 "제외" 행 상태를 적용합니다.

셀에 색상 적용 데이터 테이블에서 "결측 열" 보고서에서 선택한 열의 결측값이 포함된 셀에 색상을 적용합니다.

행에 색상 적용 데이터 테이블에서 "결측 열" 보고서에서 선택한 열의 결측값이 포함된 행에 색상을 적용합니다.

"결측 열" 보고서를 제거하려면 "닫기" 버튼을 클릭합니다.

결측값 군집화 결측 데이터의 계층적 군집화 분석을 제공합니다. 보고서에는 그림과 두 개의 덴드로그램이 포함됩니다. 그림의 행은 결측 데이터 패턴으로 정의되며 각 패턴에 대한 행이 있습니다. 열은 변수에 해당합니다. 빨간색 각 셀은 그림 아래에 나열된 열의 결측값 그룹을 나타냅니다. 셀을 커서로 가리키면 표시된 값 목록을 볼 수 있습니다. 그림을 클릭하면 결측 데이터 패턴 행이 선택됩니다. 선택된 패턴을 나타내는 세로 막대가 표시됩니다.

- 그림 오른쪽의 덴드로그램에는 결측 데이터 패턴 행의 군집이 표시됩니다. 이러한 행은 테이블 > 결측 데이터 패턴을 사용하여 얻을 수 있습니다.
- 그림 아래의 덴드로그램에는 변수 군집이 표시됩니다.

이 보고서를 사용하여 특정 열 그룹의 결측값 패턴이 비슷한지 여부를 확인할 수 있습니다. "결측값 군집화" 보고서를 제거하려면 "닫기" 버튼을 클릭합니다.

결측값 스냅샷 결측값에 대한 셀 그림을 표시합니다. 열은 변수를 나타냅니다. 검은색 셀은 결측값을 나타냅니다. 이 그림은 데이터 수집 기간이 끝나기 전에 피실험자가 연구에서 빠질 수 있는 중단 데이터의 결측을 이해하는 데 특히 유용합니다. "결측값 스냅샷" 보고서를 제거하려면 "닫기" 버튼을 클릭합니다.

다변량 정규 대치법 비결측 열에서 최소 제곱 예측을 사용하여 결측값을 대치합니다. 축소 옵션을 사용하여 공분산 행렬 추정을 향상시킬 수 있습니다.

주의 : 열의 수가 수백 개인 경우에는 이 방법을 사용하지 마십시오.

다변량 SVD 대치법 반복된 낮은 계수 SVD 행렬 완성 방법을 사용하여 큰 문제에 대해 빠르게 결측값을 대치합니다. "다변량 SVD 대치법" 을 클릭하면 "결측값 대치법" 창에 조정 가능한 권장 설정이 표시됩니다.

특이 벡터 수 계산된 후 결측값 대치에 사용되는 특이 벡터의 수입니다.

참고: 특이 벡터를 너무 많이 지정하지 않아야 합니다. 그렇지 않으면 반복되는 동안 SVD와 결측값 대치가 변하지 않습니다.

최대 반복 수 결측값 대치에 사용되는 반복 수입니다.

반복 로그 표시 반복 횟수를 표시하고 기준에 대한 정보를 제공하는 " 상세 정보 " 보고서를 엽니다.

규모가 큰 문제의 경우 SVD가 완료된 차원 수를 보여 주는 진행률 표시줄이 나타납니다. 언제든지 결측값 대치를 중지하고 해당 차원 수를 사용할 수 있습니다.

다변량 RPCA 대치법 로버스트 주성분을 사용하여 결측값을 대치합니다. 즉, 이상치에 로버스트한 낮은 계수 행렬 분해를 사용하여 결측값을 대체합니다.

팁: 이 방법은 변수의 수가 많은 와이드 문제에 유용합니다.

자동 데이터 결측값 대치 낮은 계수 행렬 근사 방법을 사용하여 결측값을 대치합니다. 이 방법은 데이터를 기반으로 낮은 계수 근사에 대한 최선의 차원을 자동으로 선택합니다. 이 방법을 선택하기 전에 대치된 값 저장 옵션 및 기타 고급 컨트롤을 지정할 수 있습니다.

새 데이터 테이블 생성 원래 데이터 테이블과 동일한 차원의 새 데이터 테이블을 생성합니다. 새 데이터 테이블에는 시작 창에서 선택한 열에 대치된 값이 포함됩니다.

스코어링 계산식을 현재 데이터 테이블에 저장 시작 창에 지정한 대치된 열을 포함하는 결측값 대치된 열 그룹을 현재 데이터 테이블에 저장합니다. ADI 결측값 대치 열이라는 숨겨진 열도 현재 데이터 테이블에 추가됩니다. 이 열에는 데이터 결측값 대치에 사용된 스코어링 계산식과 대치된 벡터가 포함됩니다. 데이터 테이블에 행이 추가되면 열 계산식이 자동으로 업데이트되므로 스트리밍 데이터에 대한 결측 데이터 대치가 가능합니다. 이 옵션은 기본 옵션입니다.

현재 위치에서 결측값 대치 현재 데이터 테이블에서 결측값을 대치합니다. 대치된 값은 연한 파란색으로 표시됩니다.

결측값 정보화 열 포함 (" 스코어링 계산식을 현재 데이터 테이블에 저장 " 을 선택한 경우에만 사용 가능) 시작 창에 지정된 각 대치 열에 대해 Imputed_ 열 그룹에 추가 표시자 열을 추가합니다. 이 열은 대치된 각 Y 열에 대해 각 행이 결측인지 여부를 지정합니다.

차원 상한 낮은 계수 근사에 허용되는 최대 순위를 결정합니다. 기본값은 선택한 열로 구성되는 행렬의 차원에 의해 결정됩니다.

최대 반복 수 결측값 대치 모형의 조정 모수를 결정하기 위해 반복되는 값의 수를 결정합니다. 기본값은 10 입니다.

결측값으로 유도할 관측값 비율 훈련 및 검증 데이터 집합에 추가되는 IM 값의 비율을 결정합니다. 각 집합에 대한 기본 비율은 0.2 입니다.

검증에 사용할 행 비율 훈련 및 검증 데이터 집합에 사용할 행 비율을 결정합니다. 검증 데이터 집합에 대한 기본 비율은 0.3 입니다.

난수 시드값 설정 ADI 에 사용할 난수 시드값을 결정합니다 . 재현 가능한 결과를 얻으려면 이 옵션을 사용합니다 .

팁 : 기준 변수의 모든 수준에서 결측값 명령을 실행하려면 **Ctrl** 키를 누른 채 원하는 명령 버튼을 클릭합니다 .

결측값 대치 보고서

결측값 대치법 중 하나를 선택하면 "결측값 탐색" 보고서 창에 "결측값 대치 보고서"도 포함됩니다. 결측값 대치 보고서에는 선택한 대치법 프로세스의 결과가 설명됩니다. 선택한 방법에 따라 다음 결과가 포함됩니다.

- 대체된 결측값 수
- 선택한 결측값 대치법 및 선택한 방법에 대한 상세 정보
- 영향을 받은 행과 열의 수
- (다변량 정규 대치법에만 해당) 발견된 다양한 결측값 패턴의 수
- 데이터 테이블에서 대치된 값의 색상
- (다변량 RPCA 대치법에만 해당) "상세 정보" 보고서에 반복 수, 행렬의 계수, 수렴 기준 값 및 척도화된 절대 잔차의 최대값이 표시됩니다.

결측값 대치가 완료되면 데이터 테이블에서 대치된 값에 해당하는 셀에 색상이 지정됩니다. "결측 열" 보고서가 열려 있는 경우 결측값을 표시하지 않도록 업데이트됩니다.

결측값 대치를 취소하고 대치된 데이터를 결측값으로 바꾸려면 **실행 취소**를 클릭합니다.

결측값 탐색 플랫폼 옵션

"결측값 탐색"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

ADI 적재 행렬 (ADI 방법에만 사용 가능) 각 성분에 대한 요인 적재에 해당하는 열이 포함된 보고서를 표시하거나 숨깁니다.

다음 옵션에 대한 자세한 내용은 **JMP 사용**의 에서 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 .

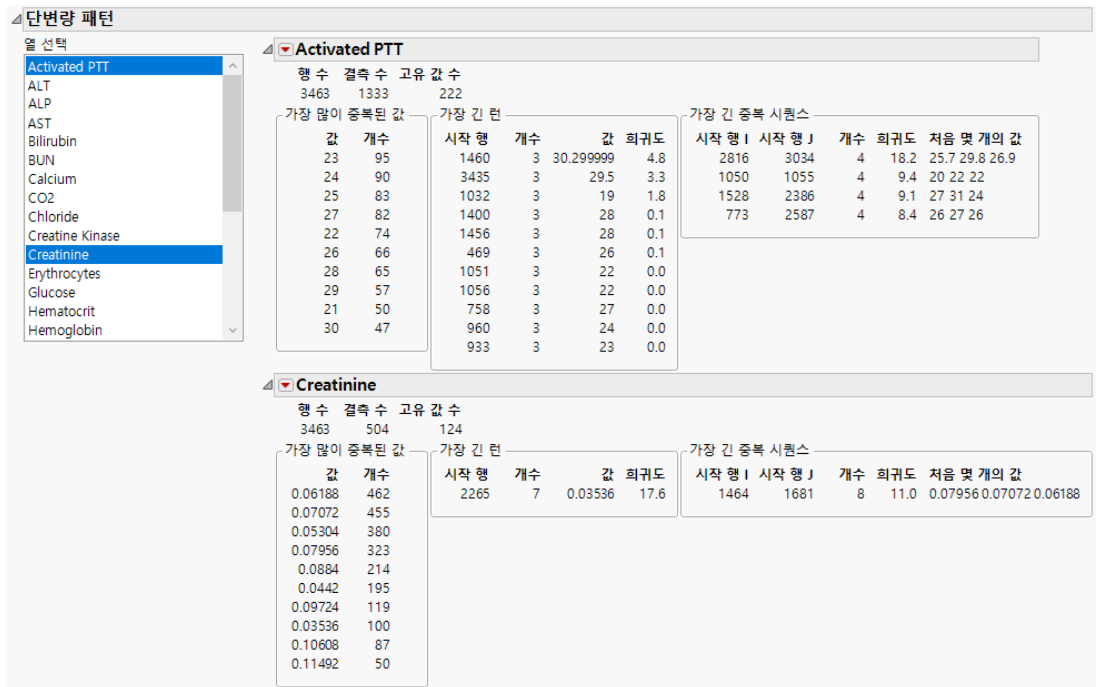
그룹별 스크립트 저장 기준 변수의 모든 수준에 대한 플랫폼 보고서를 재생성하는 스크립트를 여러 대상에 저장할 수 있는 옵션이 포함되어 있습니다. 시작 창에서 기준 변수를 지정한 경우에만 사용할 수 있습니다.

패턴 탐색

데이터에서 일반적이지 않은 특징이나 패턴 탐색

패턴 탐색 플랫폼을 사용하면 데이터에서 비정상적이거나 예기치 않은 패턴을 감지할 수 있습니다. 우연히 또는 정상 메커니즘에 의해 발생된 패턴이지만 위조되거나 무단 변경된 데이터를 나타낼 수도 있습니다. 패턴 탐색 플랫폼은 값이 정확하여 매칭과 패턴이 우연히 발생할 가능성이 매우 낮을 때 가장 유용합니다. 또한 패턴 탐색 플랫폼을 사용하면 큰 데이터 테이블을 효율적으로 처리할 수 있습니다.

그림 23.1 패턴 탐색 보고서의 예



목차

패턴 탐색 플랫폼 개요.....	433
패턴 탐색 플랫폼의 예.....	433
패턴 탐색 플랫폼 시작.....	435
패턴 탐색 보고서	436
제어판	436
단변량 요약.....	436
단변량 패턴.....	436
패턴 탐색 옵션	437
단변량 패턴 보고서 옵션.....	440
패턴 탐색 플랫폼의 추가 예	441
패턴 탐색 플랫폼에 대한 통계 상세 정보	442

패턴 탐색 플랫폼 개요

패턴 탐색 플랫폼을 사용하면 데이터에서 비정상적이거나 예기치 않은 패턴을 감지할 수 있습니다. 우연히 또는 정상 메커니즘에 의해 발생된 패턴이지만 위조되거나 무단 변경된 데이터를 나타낼 수도 있습니다. 이 플랫폼은 값이 정확하여 매칭과 패턴이 우연히 발생할 가능성이 매우 낮을 때 가장 유용합니다. 또한 패턴 탐색 플랫폼을 사용하면 큰 데이터 테이블을 효율적으로 처리할 수 있습니다.

패턴 탐색 플랫폼은 중복된 값과 중복된 값 시퀀스를 찾습니다. 이러한 유형의 패턴은 동일한 값을 여러 번 붙여넣기, 동일한 값을 전체 셀 범위에 붙여넣기 또는 데이터 범위를 복사하여 다른 곳에 붙여넣기와 같은 데이터 번조 전략을 나타낼 수 있습니다. 또한 이 플랫폼은 형식이 지정된 값의 특성을 분석하여 데이터가 실제 측정값인지 아니면 난수 생성기로 생성되었을 수 있는지를 판단합니다. 분포 특성을 검토하고 행 그룹에서 열 간의 선형 관계를 감지하는 옵션도 있습니다.

참고: 패턴 탐색 플랫폼은 제외된 행을 무시하지만 행을 삭제한 경우와 결과가 다를 수 있습니다. "가장 긴 런" 및 "가장 긴 중복 시퀀스" 옵션과 같은 행 런의 경우 제외된 행이 있으면 런이 중지됩니다.

패턴 탐색 플랫폼의 예

패턴 탐색 플랫폼을 사용하여 약물 연구 실험 결과에 예기치 않은 패턴이 있는지 여부를 확인합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Nicardipine Lab Patterns.jmp 샘플 데이터 테이블을 엽니다.
2. **분석 > 선별 > 패턴 탐색**을 선택합니다.
3. Laboratory Results 열 그룹을 선택하고 **Y, 열**을 클릭합니다.
4. **확인**을 클릭합니다.
5. "단변량 요약" 옆의 회색 표시 아이콘을 클릭하여 "단변량 요약" 보고서를 엽니다.

그림 23.2 패턴 탐색 단변량 요약 보고서

단변량 요약				
가장 긴 런				
열	시작 행	개수	값	회귀도
Creatinine	2265	7	0.03536	17.6
Hematocrit	3048	4	28.999999	12.3
BUN	2267	5	2.142	7.8
Prothrombin Time	2931	6	12	6.7
LDH	3456	3	602	5.8
Activated PTT	1460	3	30.299999	4.8
Platelet	521	3	150	4.8
Platelet	1092	3	336	4.1
Activated PTT	3435	3	29.5	3.3
CO2	1390	4	82.612	3.1
Creatine Kinase	3456	3	63	3.1
Erythrocytes	815	3	4.3000002	3.0
가장 긴 중복 시퀀스				
열	시작 행 I	시작 행 J	개수	회귀도
Activated PTT	2816	3034	4	18.2
Prothrombin Time	2942	2966	6	14.9
Hematocrit	3051	3057	3	14.6
Leukocytes	547	1275	3	11.1
Creatinine	1464	1681	8	11.0
Leukocytes	975	1485	3	10.8
ALT	840	2425	3	10.5
Sodium	2339	2976	6	9.8
Leukocytes	456	2805	3	9.8
Activated PTT	1050	1055	4	9.4
Potassium	1046	1067	6	9.4
Hemoglobin	1247	3273	3	9.2

"단변량 요약" 보고서에는 모든 열에서 발견된 가장 긴 런과 가장 긴 중복 시퀀스가 요약됩니다. 각 테이블은 런 또는 시퀀스가 우연히 발생할 가능성을 계산한 회귀도에 따라 정렬됩니다. 가장 긴 런에서 회귀도 스코어가 가장 높은 열은 **Creatinine** 인 반면, 가장 긴 중복 시퀀스에서 회귀도 스코어가 가장 높은 열은 **Activated PTT** 입니다. "단변량 패턴" 보고서에서 이러한 열에 대해 자세히 살펴보겠습니다.

- 처음에는 첫 번째 열인 **Activated PTT** 에 대한 보고서가 "단변량 패턴" 보고서에 표시됩니다. Ctrl 키를 누른 채 "열 선택" 목록에서 **Creatinine** 을 선택합니다.

그림 23.3 패턴 탐색 단변량 패턴 보고서

단변량 패턴

열 선택

Activated PTT

ALT

ALP

AST

Bilirubin

BUN

Calcium

CO2

Chloride

Creatine Kinase

Creatinine

Erythrocytes

Glucose

Hematocrit

Hemoglobin

Activated PTT

행 수

결측 수

고유 값 수

3463

1333

222

가장 많이 중복된 값

값

개수

23

95

24

90

25

83

27

82

22

74

26

66

28

65

29

57

21

50

30

47

가장 긴 런

시작 행

개수

값

회귀도

1460

3

30.299999

4.8

3435

3

29.5

3.3

1032

3

19

1.8

1400

3

28

0.1

1456

3

28

0.1

469

3

26

0.1

1051

3

22

0.0

1056

3

22

0.0

758

3

27

0.0

960

3

24

0.0

933

3

23

0.0

가장 긴 중복 시퀀스

시작 행 I

시작 행 J

개수

회귀도

처음 몇 개의 값

2816

3034

4

18.2

25.7 29.8 26.9

1050

1055

4

9.4

20 22 22

1528

2386

4

9.1

27 31 24

773

2587

4

8.4

26 27 26

Creatinine

행 수

결측 수

고유 값 수

3463

504

124

가장 많이 중복된 값

값

개수

0.06188

462

0.07072

455

0.05304

380

0.07956

323

0.0884

214

0.0442

195

0.09724

119

0.03536

100

0.10608

87

0.11492

50

가장 긴 런

시작 행

개수

값

회귀도

2265

7

0.03536

17.6

가장 긴 중복 시퀀스

시작 행 I

시작 행 J

개수

회귀도

처음 몇 개의 값

1464

1681

8

11.0

0.07956 0.07072 0.06188

"열 선택" 목록에서 열 이름을 선택하면 해당 열 보고서가 표시됩니다. **Activated PTT** 보고서에서는 이 열에 개수가 3 인 런과 4 인 시퀀스가 여러 개 포함되어 있음을 보여 줍니다. **Creatinine** 보고서에는 가장 긴 런이 7 카운트이고 가장 긴 중복 시퀀스가 8 카운트임을 보여 줍니다.

패턴 탐색 플랫폼 시작

분석 > 선별 > 패턴 탐색을 선택하여 패턴 탐색 플랫폼을 시작합니다.

그림 23.4 패턴 탐색 플랫폼 시작 창

데이터에서 일반적이지 않은 패턴을 감지합니다.

The screenshot shows the 'Pattern Search Platform Start' window. It is divided into several sections:

- 열 선택 (Column Selection):** A list of 37 columns is shown, including STUDYID, SITEID, USUBJID, AGE, SEX, RACE, Description of Planned Arm, Vasoconstriction Flag, Patient Died, Measurement, and Laboratory Results(27/0). A search bar '열 이름 입력' is at the top.
- 선택한 열 역할 지정 (Selected Column Role Assignment):** A table with two columns: 'Y, 열' (Y, Column) and '선택한 열 역할 지정' (Selected Column Role Assignment). The first row has 'Y, 열' set to '필수 숫자' (Required Number) and '선택한 열 역할 지정' set to '선택적 숫자' (Optional Number). The second row has 'Y, 열' set to '기준' (Criteria) and '선택한 열 역할 지정' set to '선택적' (Optional).
- 작업 (Action):** Buttons for '확인' (OK), '취소' (Cancel), '제거' (Remove), '재호출' (Re-call), and '도움말' (Help).
- 최소 런 크기 (Minimum Run Size):** A text box with the value '2'.
- 가장 긴 중복 값의 최소 크기 (Minimum Size of Longest Duplicate Value):** A text box with the value '2'.
- 열 간 중복 값의 최소 런 크기 (Minimum Run Size of Duplicate Values Between Columns):** A text box with the value '2'.
- 선형 관계의 최소 행 수 (Minimum Number of Rows for Linear Relationship):** A text box with the value '10'.
- 결측값 포함 (Include Missing Values):** An unchecked checkbox.

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

Y, 열 분석할 열을 할당합니다. 변수는 수치형이고 연속형이어야 합니다.

기준 개별 분석을 정의하는 수준이 포함된 하나 이상의 열을 할당합니다. 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.

최소 런 크기 행에서 시퀀스를 런으로 보고할 최소 값 수를 지정합니다.

가장 긴 중복 값의 최소 크기 열 내에서 중복 시퀀스로 보고할 시퀀스의 최소 길이를 지정합니다.

열 간 중복 값의 최소 런 크기 열 간에 중복 시퀀스로 보고할 시퀀스의 최소 길이를 지정합니다.

선형 관계의 최소 행 수 행 그룹에서 열 간의 선형 관계를 감지하는 데 사용되는 최소 행 수를 지정합니다.

참고 : 최소 행 수는 3 보다 크거나 같아야 합니다 .

결측값 포함 결측값을 패턴 감지에 포함하도록 지정합니다 .

패턴 탐색 보고서

초기 " 패턴 탐색 " 보고서에는 " 제어판 ", " 단변량 요약 " 보고서 및 " 단변량 패턴 " 보고서가 포함됩니다 . 이러한 보고서의 여러 테이블에는 선택 가능한 행이 포함되어 있습니다 . 행을 선택하고 첫 번째 열을 마우스 오른쪽 버튼으로 클릭하면 데이터 테이블에서 관련 행과 열을 선택하거나 색상을 지정할 수 있는 메뉴가 나타납니다 .

- “제어판”
- “단변량 요약”
- “단변량 패턴”

제어판

제어판에는 보고서 창에 표시할 보고서를 선택할 수 있는 옵션이 있습니다 . 이러한 옵션은 "패턴 탐색"의 빨간색 삼각형 메뉴에 있는 옵션과 동일합니다 . 자세한 내용은 "패턴 탐색 옵션"에서 확인하십시오 . 보고서 설정을 지정하는 제어판 섹션도 있습니다 . 이러한 설정은 시작 창의 설정과 동일합니다 . 자세한 내용은 "패턴 탐색 플랫폼 시작"에서 확인하십시오 . 제어판에서 설정을 변경하면 보고서가 자동으로 업데이트됩니다 .

단변량 요약

"단변량 요약" 보고서는 전체 열에 대한 여러 패턴 보고서의 결과를 결합합니다 . 기본적으로 "가장 긴 런"과 "가장 긴 중복 시퀀스"에 대한 요약 테이블이 있습니다 . 회귀도 ≥ 2 인 가장 긴 런 및 가장 긴 중복 시퀀스가 테이블에 표시됩니다 .

"분수 길이" 옵션을 선택하면 "분수 길이" 요약 테이블도 표시됩니다 . 이 테이블에는 각 열에서 분수 길이가 15보다 크거나 같은 값의 비율이 표시됩니다 . 분수 길이가 15보다 크거나 같은 값을 포함하지 않는 열은 "분수 길이" 요약 테이블에 표시되지 않습니다 .

하나 이상의 Y 열에 "규격 한계" 열 특성이 있고 "규격 한계에 대한 분포" 옵션을 선택한 경우 "규격 한계 분포" 요약 테이블도 표시됩니다 . 이 테이블에는 통계적으로 유의한 수의 관측값이 규격 한계를 벗어나는 열이 표시됩니다 .

"단변량 요약" 보고서 내용에 대한 자세한 설명은 "패턴 탐색 옵션"에서 확인하십시오 .

단변량 패턴

"단변량 패턴" 보고서에는 "열 선택" 목록과 하나 이상의 열별 패턴 보고서가 포함되어 있습니다 . 표시되는 열별 보고서는 "열 선택" 목록에서 결정됩니다 . 각 열의 보고서를 검토하려면 첫

번째 열을 선택한 후 아래쪽 화살표 키를 연속 사용하여 다음 열로 이동합니다. 아니면 이 목록에서 열 이름을 클릭하여 선택한 열의 패턴 보고서를 표시할 수도 있습니다. Ctrl 키 또는 Shift 키를 사용하여 둘 이상의 열 이름을 선택하면 선택한 각 열에 대한 패턴 보고서가 표시됩니다. "열 선택" 목록에서 하나 이상의 열을 선택 취소하면 해당 열에 대한 보고서가 제거됩니다.

각 개별 패턴 보고서에는 지정된 열에 대한 행 수, 결측 관측값 수 및 고유 값 수를 포함하는 요약 테이블이 표시됩니다. "가장 많이 중복된 값", "가장 긴 런" 및 "가장 긴 중복 시퀀스" 테이블도 기본적으로 제공됩니다. 제어판이나 "패턴 탐색"의 빨간색 삼각형 메뉴에서 "형식 지정 값 너비", "분수 길이" 또는 "선행 및 후행 숫자" 옵션을 선택하면 해당 테이블이 각 개별 패턴 보고서에 추가됩니다.

"단변량 패턴" 보고서의 테이블에서 행을 마우스 오른쪽 버튼으로 클릭하면 마우스 오른쪽 버튼 클릭 옵션 메뉴가 나타납니다.

행 및 열 선택 "단변량 패턴" 보고서의 특정 테이블에서 선택한 행에 해당하는 열과 행을 원래 데이터 테이블에서 선택합니다. 예를 들어 특정 열에 대한 "선행 숫자" 테이블의 두 번째 행을 마우스 오른쪽 버튼으로 클릭하면 해당 열에서 선행 숫자가 2 인 행이 원래 데이터 테이블에서 선택됩니다.

셀 색상 지정 "단변량 패턴" 보고서의 특정 테이블에서 선택한 행에 해당하는 원래 데이터 테이블의 셀에 색상을 지정합니다. 또한 색상이 지정된 셀의 행이 선택됩니다.

열별 패턴 보고서에는 빨간색 삼각형 옵션이 있습니다. 자세한 내용은 ["단변량 패턴 보고서 옵션"](#)에서 확인하십시오.

패턴 탐색 옵션

제어판과 "패턴 탐색"의 빨간색 삼각형 메뉴에는 요약 및 패턴 테이블에 대한 옵션이 포함되어 있습니다.

가장 많이 중복된 값 데이터에서 가장 많이 나오는 값을 보여 주는 테이블을 표시하거나 숨깁니다. 각 열에 대해 최대 10 개의 값을 표시합니다.

가장 긴 런 동일한 값의 가장 긴 런을 보여 주는 테이블을 표시하거나 숨깁니다. 이 테이블에는 다음 정보가 포함됩니다.

시작 행 런 시작에 해당하는 행 번호입니다.

개수 런 길이입니다.

값 런의 항목 값입니다.

희귀도 런이 우연히 발생할 가능성을 측정한 값입니다. 희귀도가 높으면 해당 런이 우연히 발생했을 가능성이 낮다는 의미입니다. 자세한 내용은 ["가장 긴 런의 희귀도"](#)에서 확인하십시오.

가장 긴 중복 시퀀스 같은 열에 두 번 이상 나오는 값의 가장 긴 시퀀스를 보여 주는 테이블을 표시하거나 숨깁니다. 반복 값 시퀀스도 고려됩니다. 시퀀스가 두 번 이상 중복되면 각 인스턴스가 테이블에 나타납니다. 이 테이블에는 다음 정보가 포함됩니다.

시작 행 I 첫 번째 시퀀스의 시작에 해당하는 행 번호입니다.

시작 행 J 중복 시퀀스의 시작에 해당하는 행 번호입니다.

개수 시퀀스 길이입니다.

희귀도 중복 시퀀스가 우연히 발생할 가능성을 측정한 값입니다. 희귀도가 높으면 해당 중복 시퀀스가 우연히 발생했을 가능성이 낮다는 의미입니다. 자세한 내용은 "[가장 긴 시퀀스의 희귀도](#)"에서 확인하십시오.

처음 몇 개의 값 중복 시퀀스의 처음 몇 개 값입니다. 각 시퀀스에 대해 최대 3 개의 값을 표시합니다.

형식 지정 값 너비 형식이 지정된 값의 전체 너비 및 소수점 이하 너비를 보여 주는 테이블을 표시하거나 숨깁니다. 이 테이블에는 다음 정보가 포함됩니다.

너비 너비 크기입니다.

전체 기준 개수 해당 열에서 전체 너비가 지정된 너비와 같은 관측값의 수입니다.

소수점 이하 기준 개수 해당 열에서 소수점 이하 너비가 지정된 너비와 같은 관측값의 수입니다.

분수 길이 값의 연속된 소수 부분 길이를 보여 주는 테이블을 표시하거나 숨깁니다. 연속된 소수 부분은 수를 연속해서 나눈 항의 시퀀스로 표현한 것입니다. 분수 길이는 어떤 수가 유리수인지 아닌지를 나타낼 수 있습니다. 길이가 15 이상이면 값이 무리수 (예: 루트 함수의 경우), 무리수를 포함하는 함수 또는 난수 생성기 (Benford, 1938)에 의한 값을 나타냅니다.

참고 : 패턴 탐색에서 최대 분수 길이는 15 입니다. 분수 길이가 15 이상인 숫자는 모두 15 로 보고됩니다.

길이 연속된 소수 부분의 길이입니다.

개수 해당 열에서 연속된 소수 부분 길이가 지정된 길이와 같은 관측값의 수입니다.

선행 및 후행 숫자 1~9 의 각 선행 및 후행 숫자에 대한 개수 테이블을 표시하거나 숨깁니다. 선행 숫자의 경우 해당 선행 숫자를 포함하는 최소값 및 최대값과 기대 비율을 나타내는 열도 있습니다. 기대 비율은 값의 첫 번째 숫자가 a 와 동일한 빈도가 $F_a = \log_{10}(a+1/a)$ 라는 Benford 법칙을 기반으로 합니다. 이 법칙은 자릿수가 모두 네 개 이상인 큰 범위의 숫자를 포함하는 데이터에 적용됩니다.

규격 한계 매칭 (하나 이상의 열에 "규격 한계" 열 특성이 있는 경우 사용 가능) 정확히 규격 하한, 규격 상한 또는 목표값에 해당하는 값 테이블을 표시하거나 숨깁니다. 정확히 매칭되는 항목이 없으면 보고서가 표시되지 않습니다.

규격 한계에 대한 분포 (하나 이상의 열에 "규격 한계" 열 특성이 있는 경우에만 사용 가능) 규격 한계를 벗어나는 관측값 대 기대값 테이블을 표시하거나 숨깁니다. 카이제곱 및 p 값 열은 실제 관측값 수가 기대 관측값 수보다 훨씬 많은 경우에만 표시됩니다.

열 간 중복 값 "열 간 중복 값" 보고서를 표시하거나 숨깁니다. 이 보고서에는 둘 이상의 열에서 동일한 행에 나타나는 값 시퀀스 테이블이 포함되어 있습니다. 보고서 맨 위에는 중복으로 간주할 최소 시퀀스 길이를 나타내는 노트가 있습니다. 이 값은 시작 창 또는 제어판에서 "열 간 중복 값의 최소 런 크기"로 지정됩니다. "색상 지정" 옵션도 보고서 맨 위에 있습니다. 자세한 내용은 "**열 간 중복 값 색상 지정**"에서 확인하십시오.

열 간 중복 값 색상 지정 ("열 간 중복 값"을 선택한 후에만 사용 가능) 데이터 테이블에서 "열 간 중복 값" 보고서의 중복 매칭 항목에 해당하는 셀에 색상을 지정합니다. 또한 중복 항목을 포함하는 행이 선택됩니다.

선형 관계 "변수 간 선형 관계" 보고서를 표시하거나 숨깁니다. 이 보고서에는 지정된 수의 순차적 행에 대해 열 간의 정확 선형 관계를 보여 주는 테이블이 포함됩니다. 정확 선형 관계는 R^2 값이 0.999999 보다 크거나 같은 선형 관계로 정의됩니다. 보고서 맨 위에는 고려된 최소 행 수를 나타내는 노트가 있습니다. 이 값은 시작 창 또는 제어판에서 "선형 관계의 최소 행 수"로 지정됩니다. "색상 지정" 옵션도 보고서 맨 위에 있습니다. 자세한 내용은 "**선형 관계 색상 지정**"에서 확인하십시오.

선형 관계 색상 지정 ("선형 관계"를 선택한 후에만 사용 가능) 데이터 테이블에서 "변수 간 선형 관계" 테이블의 선형 관계에 해당하는 셀에 색상을 지정합니다. 또한 선형 관계를 포함하는 행이 선택됩니다.

열 저장 다음 저장 옵션을 포함합니다.

런 테이블 저장 모든 열의 가장 긴 런을 새 데이터 테이블에 저장합니다. 이 데이터 테이블은 "단변량 패턴" 보고서의 "가장 긴 런" 테이블과 동일한 열을 포함하고 "시작 행"을 기준으로 정렬됩니다.

중복 시퀀스 테이블 저장 모든 열의 가장 긴 중복 시퀀스를 새 데이터 테이블에 저장합니다. 이 데이터 테이블은 "단변량 패턴" 보고서의 "가장 긴 중복 시퀀스" 테이블과 동일한 열을 포함하고 "시작 행 I"를 기준으로 정렬됩니다.

열 간 중복 값 저장 "열 간 중복 값" 테이블을 새 데이터 테이블에 저장합니다.

선형 관계 저장 "변수 간 선형 관계" 테이블을 새 데이터 테이블에 저장합니다.

열 순서화 정렬 옵션이 포함되어 있습니다. 정렬을 변경하면 "단변량 패턴" 보고서의 열 목록이 재정렬됩니다. 맨 위 열이 자동으로 선택되고 해당 열에 대한 보고서만 표시됩니다.

런 회귀도별로 정렬 가장 긴 런의 회귀도 값을 기준으로 열 목록을 정렬합니다. 더 높은 회귀도 값이 맨 위에 나옵니다.

시퀀스 회귀도별로 정렬 가장 긴 중복 시퀀스의 회귀도 값을 기준으로 열 목록을 정렬합니다. 더 높은 회귀도 값이 맨 위에 나옵니다.

열 이름별로 정렬 열 이름을 기준으로 열 목록을 정렬합니다 (사전순).

원래 순서 시작 창에서 열이 지정된 원래 순서대로 열 목록을 정렬합니다.

셀 색상 지우기 데이터 테이블에서 선택한 열 또는 모든 열 (데이터 테이블에서 열이 선택되지 않은 경우)에 대해 셀 색상을 지웁니다.

다음 옵션에 대한 자세한 내용은 **JMP 사용**의 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.

단변량 패턴 보고서 옵션

단변량 패턴의 각 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

가장 많이 중복된 값 선택 데이터 테이블에서 " 가장 많이 중복된 값 " 테이블의 값에 해당하는 관측값을 선택합니다.

가장 긴 런 선택 데이터 테이블에서 " 가장 긴 런 " 테이블의 값에 해당하는 관측값을 선택합니다.

가장 긴 중복 시퀀스 선택 데이터 테이블에서 " 가장 긴 중복 시퀀스 " 테이블의 값에 해당하는 관측값을 선택합니다.

형식 지정 값 너비별로 선택 데이터 테이블에서 " 형식 지정 값 너비 " 테이블의 값 중 사용자가 지정한 전체 너비의 값에 해당하는 관측값을 선택합니다.

소수점 이하 너비별로 선택 데이터 테이블에서 " 형식 지정 값 너비 " 테이블의 값 중 사용자가 지정한 소수점 이하 너비의 값에 해당하는 관측값을 선택합니다.

분수 길이별로 선택 데이터 테이블에서 " 분수 길이 " 테이블의 값 중 사용자가 지정한 분수 길이의 값에 해당하는 관측값을 선택합니다.

선행 숫자별로 선택 데이터 테이블에서 " 선행 및 후행 숫자 " 테이블의 값 중 사용자가 지정한 선행 숫자의 값에 해당하는 관측값을 선택합니다.

패턴 탐색 플랫폼의 추가 예

패턴 탐색 플랫폼을 사용하여 도시 폐수 처리장의 센서에서 센서 측정값에 예기치 않은 패턴이 있는지 여부를 확인합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Water Treatment.jmp** 샘플 데이터 테이블을 엽니다.
2. **분석 > 선별 > 패턴 탐색**을 선택합니다.
3. **Sensor Measurements** 열 그룹을 선택하고 **Y, 열**을 클릭합니다.
4. **확인**을 클릭합니다.
5. "제어판"의 "열 간"에서 "열 간 중복 값" 옆에 있는 상자를 클릭합니다.

그림 23.5 열 간 중복 값 보고서

열 간 중복 값				
2회 이상의 런 사용 색상 지정				
시작 행	끝 행	행 수	동일 값 수	중복 집합
219	230	12	12	PH-P, PH-D
207	217	11	11	PH-E, PH-P
214	217	4	8	PH-E, PH-P, PH-D
330	333	4	8	PH-E, PH-P, PH-D
198	204	7	7	PH-P, PH-D
91	96	6	6	PH-P, PH-D
94	96	3	6	PH-E, PH-P, PH-D
208	210	3	6	PH-E, PH-P, PH-D
216	217	2	6	PH-E, PH-P, PH-D, PH-S
228	229	2	6	PH-E, PH-P, PH-D, PH-S
247	252	6	6	PH-E, PH-P
290	295	6	6	PH-E, PH-P
293	295	3	6	PH-E, PH-P, PH-D
427	429	3	6	PH-E, PH-P, PH-D
472	477	6	6	PH-E, PH-P
26	30	5	5	PH-E, PH-P
63	67	5	5	PH-P, PH-D
141	145	5	5	PH-E, PH-D
299	303	5	5	PH-E, PH-P
352	356	5	5	PH-P, PH-D

"열 간 중복 값" 보고서에 중복 열이 포함된 긴 목록이 표시됩니다. 그러나 최소 런 크기가 2 이고 작은 런이 목록에 많이 있습니다. 10 이상의 더 큰 런에 초점을 맞추려고 합니다.

6. "제어판"의 "설정"에서 "열 간 중복 값의 최소 런 크기" 옆에 있는 상자에 "10"을 입력합니다.

팁: 상자 바깥쪽의 아무 곳이나 클릭하면 보고서가 자동으로 업데이트됩니다.

그림 23.6 런 크기가 10 이상인 열 간 중복 값 보고서

열 간 중복 값				
10회 이상의 런 사용 색상 지정				
시작 행	끝 행	행 수	동일 값 수	중복 집합
219	230	12	12	PH-P, PH-D
207	217	11	11	PH-E, PH-P

업데이트된 "열 간 중복 값" 보고서에서는 런 횟수가 10 이상인 중복 집합이 두 개 있음을 보여 줍니다. 두 집합 모두 PH-P 열을 포함합니다.

7. "열 간 중복 값" 보고서에서 **색상 지정**을 클릭합니다.

그림 23.7 색상이 지정된 Water Treatments.jmp 데이터 테이블

		PH-E	DBO-E	DQO-E	SS-E	SSV-E	SED-E	COND-E	PH-P	DBO-P	SS-P	SSV-P	SED-P	COND-P	PH-D
●	205	7.7	192.0	253.0	186.0	58.7	5.0	1818.0	7.6	250.0	194.0	59.4	4.5	1745.0	7.5
●	206	7.7	124.0	234.0	120.0	61.7	1.8	1550.0	7.9	123.0	220.0	59.0	4.0	1651.0	7.8
●	207	7.6	172.0	394.0	226.0	51.3	5.5	1324.0	7.6	197.0	302.0	47.0	4.0	1306.0	7.7
●	208	7.6	268.0	324.0	226.0	53.1	4.0	1390.0	7.6	265.0	236.0	53.4	4.5	1389.0	7.6
●	209	7.6	176.0	200.0	176.0	50.0	3.0	859.0	7.6	149.0	172.0	51.2	2.0	860.0	7.6
●	210	7.6	146.0	243.0	164.0	57.3	3.2	1095.0	7.6	193.0	208.0	55.8	3.0	1089.0	7.6
●	211	7.8	139.0	380.0	256.0	53.1	5.5	2040.0	7.8	201.0	340.0	47.6	8.5	2000.0	7.9
●	212	7.9	205.0	529.0	264.0	61.4	6.5	2750.0	7.9	220.0	324.0	60.5	6.0	2850.0	7.8
●	213	7.8	117.0	366.0	500.0	27.2	4.2	3230.0	7.8	135.0	620.0	32.3	7.0	3170.0	7.7
●	214	7.6	148.0	400.0	564.0	31.6	7.0	1554.0	7.6	235.0	798.0	28.6	11.0	1515.0	7.6
●	215	7.8	158.0	256.0	194.0	47.4	3.0	1831.0	7.8	218.0	240.0	46.7	4.0	1860.0	7.8
●	216	7.7	123.0	246.0	186.0	39.8	2.5	1288.0	7.7	241.0	300.0	25.3	4.5	1328.0	7.7
●	217	7.7	133.0	410.0	204.0	62.7	4.5	1490.0	7.7	161.0	256.0	57.8	4.5	1421.0	7.7
●	218	7.5	237.0	388.0	178.0	57.3	4.0	1620.0	7.6	300.0	240.0	53.3	4.5	1571.0	7.7
●	219	7.7	203.0	357.0	222.0	58.6	3.0	1828.0	7.6	215.0	208.0	73.1	5.0	1817.0	7.6
●	220	7.8	199.0	380.0	244.0	55.7	5.0	1453.0	7.8	110.0	196.0	60.2	3.5	1430.0	7.8
●	221	7.8	214.0	941.0	304.0	53.3	6.5	1454.0	7.8	396.0	320.0	57.5	7.0	1528.0	7.8
●	222	7.7	208.0	298.0	260.0	37.7	4.0	1236.0	7.8	245.0	284.0	46.5	5.0	1285.0	7.8
●	223	7.9	206.0	305.0	188.0	56.4	4.0	1200.0	7.9	220.9	272.0	50.0	6.0	1228.0	7.9
●	224	7.7	247.0	361.0	286.0	51.0	4.5	1180.0	7.7	146.0	212.0	56.4	2.5	1251.0	7.7
●	225	7.5	438.0	681.0	370.0	56.8	6.0	1396.0	7.6	212.0	348.0	64.4	6.0	1446.0	7.6
●	226	7.8	255.0	282.0	166.0	67.5	2.0	950.0	7.8	284.0	172.0	65.1	2.0	965.0	7.8
●	227	7.7	283.0	310.0	170.0	61.2	2.5	1065.0	7.8	235.0	270.0	51.1	5.5	1100.0	7.8
●	228	7.9	313.0	341.0	512.0	33.6	6.0	915.0	7.9	205.0	520.0	33.8	7.5	976.0	7.9
●	229	7.7	300.0	610.0	452.0	45.1	8.0	1313.0	7.7	375.0	556.0	43.9	10.0	1312.0	7.7
●	230	7.8	177.0	412.0	196.0	69.4	4.5	2190.0	7.8	330.0	476.0	52.1	7.5	2330.0	7.8

"색상 지정" 기능은 원래 데이터 테이블에서 보고서의 중복 집합에 해당하는 셀에 색상을 지정합니다.

패턴 탐색 플랫폼에 대한 통계 상세 정보

이 섹션에는 패턴 탐색 플랫폼에서 가장 긴 런과 가장 긴 시퀀스의 회귀도를 계산하기 위한 통계 상세 정보가 포함되어 있습니다.

가장 긴 런의 회귀도

가장 긴 런의 회귀도를 계산하려면 먼저 다음 변수를 정의합니다.

n = 열의 행 수

k = 열에서 특정 값이 나타나는 횟수

$p = k/n$ = 열에서 특정 값을 관측할 확률

m = 런 길이

N = 고유 런의 수

가장 긴 런의 희귀도는 다음과 같이 계산됩니다 .

$$\text{희귀도} = -\log_2(1 - (1 - p^{m-1})^N)$$

가장 긴 시퀀스의 희귀도

가장 긴 시퀀스의 희귀도를 계산하려면 먼저 다음 변수를 정의합니다 .

p = 열에서 특정 시퀀스를 한 번 관측할 확률

k = 열에서 시퀀스의 시작 값이 나타나는 횟수

가장 긴 시퀀스의 희귀도는 다음과 같이 계산됩니다 .

$$\text{희귀도} = -\log_2(1 - (1 - p)^k)$$

반응 변수 선별

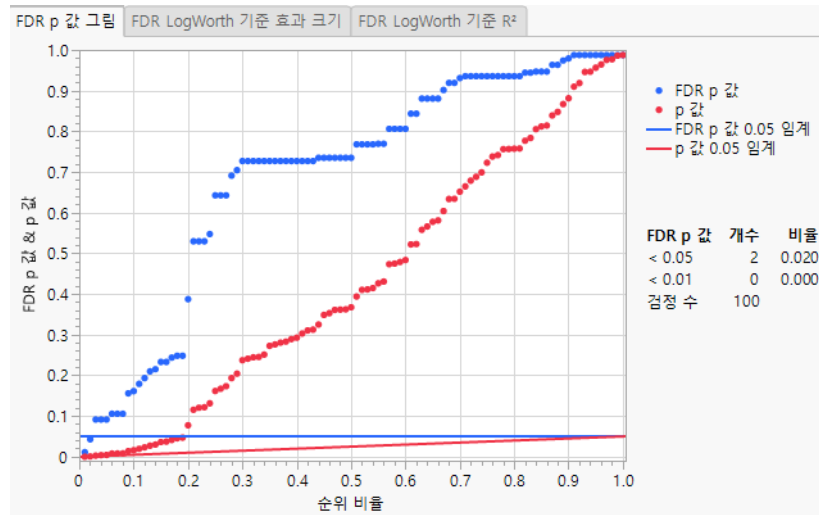
대규모 데이터의 여러 반응 변수 검정

반응 변수 선별 플랫폼에서는 다수의 반응 변수에 대한 검정 수행 과정을 자동화합니다. 테스트 결과와 요약 통계량은 보고서에 직접 표시되거나, 추가 데이터 탐색을 위해 데이터 테이블에 표시될 수 있습니다. 오해의 소지가 없더라도 요인 효과에 대해 많은 반응 변수를 검정하는 것은 적절한 방법론 없이는 어려울 수 있습니다. False Discovery Rate 방법은 유의성이 잘못 선언되지 않도록 보호합니다. p 값 그림은 LogWorth를 사용하여 척도화되므로 쉽게 해석할 수 있습니다.

대규모 데이터 집합은 정리되지 않은 경우가 많으므로 반응 변수 선별에서는 불규칙하게 분포된 데이터 및 결측 데이터를 처리하는 방법을 제공합니다. 로버스트 추정값 방법을 사용하면 이상치가 데이터에 남아 있을 수 있지만 이러한 이상치에 대한 검정 민감도가 줄어듭니다. 결측 데이터 옵션을 사용하면 결측값을 분석에 포함할 수 있습니다. 이러한 기능을 통해 먼저 광범위한 데이터 품질 분석을 수행하지 않고도 데이터를 분석할 수 있습니다.

관측값이 많으면 실제적 관심 영역이 아닌 차이도 통계적으로 유의할 수 있습니다. 반응 변수 선별에서는 감지하려는 차이를 지정할 수 있는 실제적 차이 검정을 제공합니다. 다른 한편으로는 차이가 주어진 크기를 초과하지 않는지, 즉 평균이 동일한지 여부를 알고 싶을 수도 있습니다. 이를 위해 반응 변수 선별에서 동등성 검정을 제공합니다.

그림 24.1 반응 변수 선별 그림의 예



목차

반응 변수 선별 플랫폼 개요	447
반응 변수 선별의 예	448
반응 변수 선별 플랫폼 시작	452
반응 변수 선별 보고서	455
반응 변수 선별 그림	455
반응 변수 선별 보고서 테이블	457
반응 변수 선별 플랫폼 옵션	460
평균 데이터 테이블	463
평균 차이 데이터 테이블	464
p 값 데이터 테이블	466
모형 적합의 반응 변수 선별 분석법	466
모형 적합의 반응 변수 선별 시작	467
반응 변수 선별 적합 보고서	468
반응 변수 선별 적합 옵션	470
반응 변수 선별의 추가 예	472
실제적 유의성 및 동등성 검정의 예	472
최대 LogWorth 옵션의 예	474
로버스트 적합의 예	476
반응 변수 선별 분석법의 예	479
화산(Volcano) 그림의 예	480
반응 변수 선별 플랫폼에 대한 통계 상세 정보	482

반응 변수 선별 플랫폼 개요

반응 변수 선별 플랫폼에서는 다수의 반응 변수에 대한 검정 수행 과정을 자동화합니다. 지정된 각 반응 변수를 지정된 각 요인에 대해 검정합니다. 반응 변수 선별은 대규모 데이터와 관련된 두 가지 주요 문제를 해결합니다. 검정을 많이 수행해야 한다는 것과 이상치 및 결측값을 효과적으로 처리하기 위한 요구 사항이 이러한 문제입니다.

반응 변수 선별은 플랫폼 및 모형 적합 분석법으로 사용할 수 있습니다. 표 24.1에 표시된 대로 두 경우 모두 X로 Y 적합 플랫폼의 검정과 유사한 검정을 수행합니다. 분석법으로 사용될 경우에는 개별 모형 효과에 대한 반응을 검정합니다.

반응 변수 선별에서는 필요한 여러 추론을 지원하기 위해 다음과 같은 기능을 제공합니다.

데이터 테이블 결과가 보고서뿐만 아니라 데이터 테이블에도 표시되므로 결과를 탐색, 정렬, 검색 및 그림에 표시할 수 있습니다. p 값의 $\text{LogWorth}(-\log_{10}(p \text{ 값}))$ 와 같이 그림을 쉽게 해석할 수 있는 통계량이 제공됩니다.

False Discovery Rate 많은 수의 검정을 수행하므로 검정이 유의하다고 선언하는 전체 비율을 제어해야 합니다. 반응 변수 선별은 **False Discovery Rate**를 제어합니다. **FDR**(False Discovery Rate)은 유의한 것으로 잘못 선언할 유의성 검정의 기대 비율입니다 (Benjamini and Hochberg 1995 및 Westfall et al. 2011 참조).

실제적 유의성 검정 데이터 테이블이 많은 수의 행으로 구성된 경우, 즉 n 이 큰 경우 검정에 사용되는 표준 오차는 매우 작을 수 있습니다. 이로 인해 사실은 관측된 차이가 너무 작아서 실제로 중요하지 않지만 검정 결과는 통계적으로 유의할 수 있습니다. 이 문제를 해결하기 위해 실제로 유의하다고 간주할 효과 크기를 정의할 수 있습니다. 그런 다음 실제적 유의성 검정을 수행하여 현실적으로 관심을 끌 수 있는 큰 효과만 감지합니다.

동등성 검정 많은 요인을 연구할 때 반응에 근본적으로 동일한 영향을 미치는 요인에 관심을 가질 수 있습니다. 이 경우 실제적 동등성을 정의하는 효과 크기를 지정된 후 동등성 검정을 수행할 수 있습니다.

반응 변수 선별에서는 정리되지 않은 데이터를 처리할 때 발생하는 문제를 해결하기 위해 이상치 및 결측 데이터를 처리하는 기능을 제공합니다. 이러한 기능을 사용하면 데이터 품질 문제를 해결하기 위해 노력할 필요 없이 데이터를 직접 분석할 수 있습니다.

로버스트 추정 데이터에 이상치가 있으면 표준 오차 추정값이 커지고 이로 인해 검정이 실제 효과에 민감하지 않게 됩니다. 로버스트 옵션을 선택하여 Huber M- 추정을 수행합니다. 이상치가 데이터에 남아 있지만 이러한 이상치에 대한 검정 민감도가 감소합니다.

결측값 옵션 이 플랫폼에는 범주형 예측 변수의 결측값을 정보가 포함된 것으로 간주하는 옵션이 포함되어 있습니다.

표 24.1 반응 변수 선별에서 수행되는 분석

반응	요인	X로 Y 적합 분석	설명
연속형	범주형	일원 분석	분산 분석
연속형	연속형	이변량	단순 선형 회귀
범주형	범주형	분할 분석	카이제곱
범주형	연속형	로지스틱	단순 로지스틱 회귀

반응 변수 선별 플랫폼에서는 그림 섹션과 보고서 섹션을 생성합니다. "FDR p 값" 그림과 "결과 테이블"이 기본적으로 표시됩니다. "반응 변수 선별" 분석법을 사용하면 "효과 검정" 테이블, "효과에 대한 FDR p 값 그림" 및 "FDR LogWorth 대 효과 크기" 그림이 포함된 보고서가 생성됩니다.

JSL 명령 `Summarize Y by X`는 반응 변수 선별 플랫폼과 동일한 기능을 수행하지만 플랫폼 창을 생성하지 않습니다. 자세한 내용은 [JSL Syntax Reference](#)에서 확인하십시오.

반응 변수 선별의 예

반응 선별 플랫폼을 사용하여 공정 변경 시 다른 값을 표시하는 반응 특성을 확인합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 `Probe.jmp`를 엽니다.
2. **분석 > 선별 > 반응 변수 선별**을 선택합니다.
3. **Responses** 열 그룹을 선택하고 **Y, 반응**을 클릭합니다.
4. **Process**를 선택하고 **X**를 클릭합니다.
5. "고급 옵션" 옆의 회색 삼각형을 클릭합니다.
6. **최대 LogWorth** 옆의 상자에 "100"을 입력합니다.

100 이상의 $\log_{10}(-\log_{10}(p \text{ 값}))$ 는 극도로 작은 p 값에 해당합니다. "최대 LogWorth" 값을 설정하면 보고되는 LogWorth 값이 제한되므로 그림의 척도를 제어하는 데 도움이 됩니다.

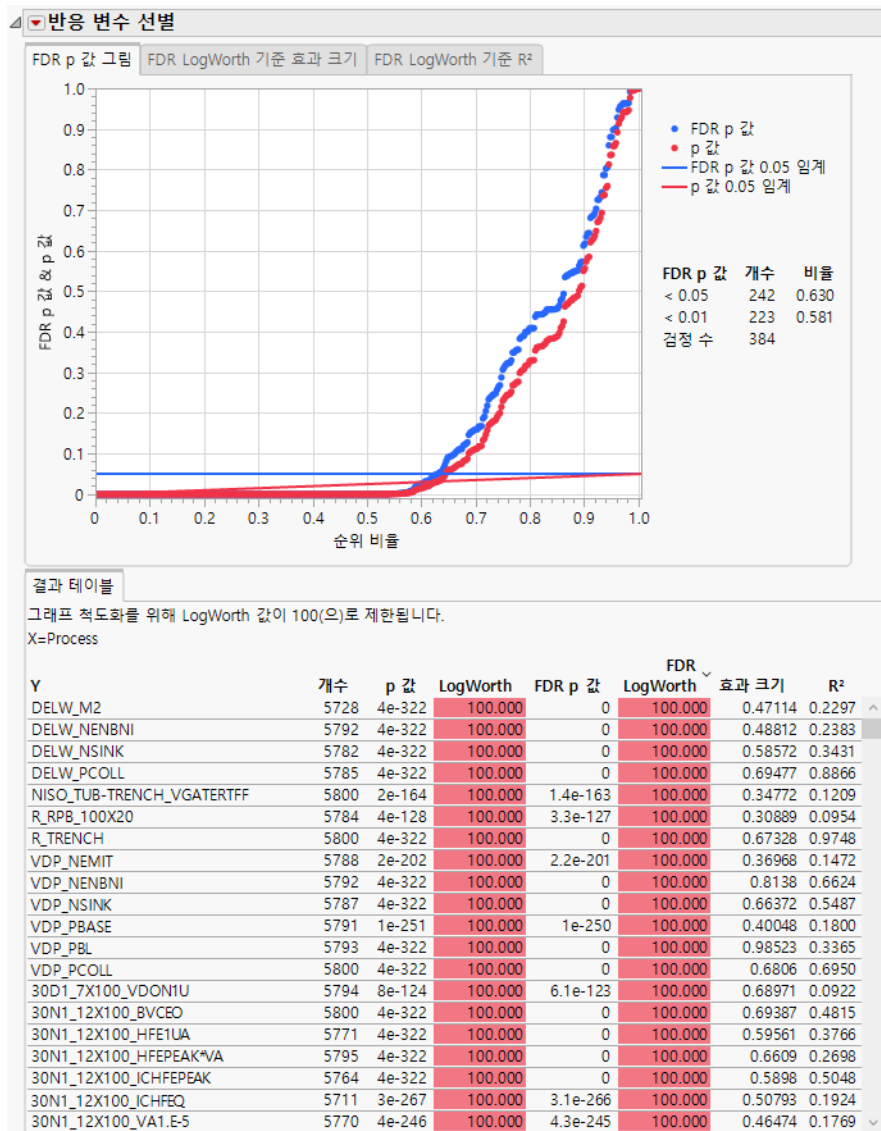
7. **확인**을 클릭합니다.

보고서에는 p 값 결과 테이블과 탭 형식의 세 가지 그림이 표시됩니다. "FDR p 값 그림"이 기본적으로 표시됩니다.

"FDR p 값 그림"에는 384개의 각 검정에 대해 "FDR p 값"과 " p 값"이라는 두 가지 유형의 p 값이 표시됩니다. "순위 비율"에 대한 p 값이 그림에 표시됩니다. " p 값"은 **Process**에 대한 단일 Y 검정의 일반적인 p 값입니다. "FDR p 값"은 주어진 FDR(False Discover Rate)을 보장하기 위해 조정된 p 값입니다. 이 예에서 False Discovery Rate는 0.05입니다. 그림에 FDR p 값은 파란색으로 표시되고 p 값은 빨간색으로 표시됩니다. "순위 비율"은 FDR p 값을 가장 작은 것부터 순서대로, 즉 유의성이 큰 것부터 내림차순으로 정렬합니다.

그림의 파란색 가로선과 경사진 빨간색 선은 모두 유의성에 대한 임계입니다. 파란색 선 아래의 FDR p 값을 사용한 검정은 False Discover Rate에 대해 조정된 0.05 수준에서 유의합니다. 빨간색 선 아래의 보통 p 값을 사용한 검정은 False Discover Rate에 대해 조정되지 않은 0.05 수준에서 유의합니다. 이런 방식으로 그림을 사용하여 두 p 값 집합 중 하나에서 FDR 유의성을 이해할 수 있습니다.

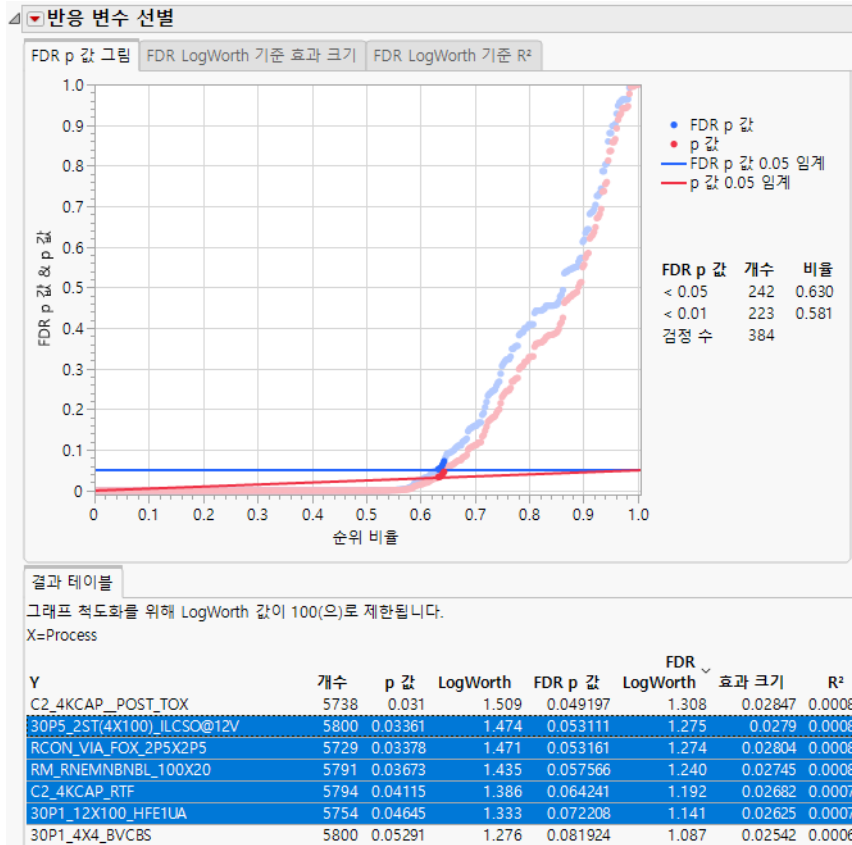
그림 24.2 Process에 대한 384개 검정의 반응 변수 선별 보고서



"FDR p 값 그림"에서는 유의한 검정이 60%가 넘는다는 것을 보여 줍니다. 조정되지 않은 p 값을 사용할 경우 유의하지만 FDR p 값을 사용할 경우에는 유의하지 않은 소수의 검정이 있습니다.

이러한 검정은 빨간색 선 위에 있지만 파란색 선 아래에 있는 빨간색 점에 해당합니다."결과 테이블"에서 0.05보다 큰 첫 번째 FDR p 값까지 아래로 스크롤하여 이러한 검정을 볼 수도 있습니다. 해당하는 보통 p 값은 0.03361입니다.

그림 24.3 반응 변수 선별 보고서 검정



Process 간에 유의하게 다른 특성을 식별하기 위해 그림에서 적절한 점 주위를 직사각형 형태로 드래그할 수 있습니다. 그러면 첫 번째 열에 특성 이름이 제공된 "결과 테이블"에서 이러한 점에 해당하는 행이 선택됩니다.

"결과 테이블"에는 Responses 그룹의 각 반응 측도당 하나씩 총 387개 행이 포함되어 있습니다. 반응은 첫 번째 열, Y에 표시되고 각 반응은 X 열, 즉 Process의 효과에 대해 검정됩니다."결과 테이블"의 마지막 세 행에는 결측값이 포함되어 있습니다. 이 때문에 그림에 384개의 검정이 보고됩니다. 마지막 세 행에 해당하는 반응 열에 상수 측정값이 포함되어 있으므로 분석이 수행되지 않습니다.

나머지 열은 X에 대한 Y검정의 정보를 제공합니다. 여기서 검정은 일원 분산 분석입니다. 이 테이블에는 검정의 p 값, LogWorth, FDR(False Discovery Rate) p 값 및 FDR LogWorth 등이 제공

됩니다. 이 테이블을 사용하여 다양한 통계량을 기준으로 정렬하고, 행을 선택하거나 관심 통계량을 그림에 표시할 수 있습니다.

시작 창에서 "최대 LogWorth"를 "100"으로 설정했으므로 $1e-100$ 이하의 p 값에 해당하는 LogWorth 및 FDR LogWorth 값은 100으로 보고됩니다. 또한 2보다 큰 FDR LogWorth 값에 해당하는 셀은 강도 그래디언트가 적용된 색상이 지정됩니다.

보고서 및 p 값 테이블에 대한 자세한 내용은 "[반응 변수 선별 보고서](#)"에서 확인하십시오.

반응 변수 선별 플랫폼 시작

분석 > 선별 > 반응 변수 선별을 선택하여 반응 변수 선별 플랫폼을 시작합니다.

그림 24.4 반응 변수 선별 시작 창

False Discovery Rate를 제어하면서 다수의 Y 대 X 조합을 적합시킵니다.

열 선택

394개 열

열 이름 입력

- Lot ID
- Wafer Number
- Process
- Die X
- Die Y
- Site
- Start Time
- Responses(387/0)

선택한 열 역할 지정

Y, 반응

필수 선택적

X

필수 선택적

그룹화

선택적

가중치

선택적 숫자

빈도

선택적 숫자

기준

선택적

작업

확인

취소

제거

재호출

도움말

☐ 공통 Y 적도

☐ 공통 X 적도

평균 비교

각 수준과 대조군

비율 조정

조정 없음

☐ 결측값을 범주로 처리

☐ 쌍제 X와 Y

☐ 시작 시 p 값 테이블

☐ 보고서 없음

고급 옵션

☐ 로버스트

☐ Cauchy

☐ Poisson Y

☐ Empirical Bayes Shrinkage

☐ 카파

☐ 상관

최대 LogWorth

☐ X 범주형 강제 적용

☐ X 연속형 강제 적용

☐ Y 범주형 강제 적용

☐ Y 연속형 강제 적용

☐ 언스레드

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

시작 창 역할

Y, 반응 분석할 측정값이 포함된 반응 열을 식별합니다.

X 반응을 검정할 열을 식별합니다.

그룹화 지정된 열의 각 수준에 할당된 행을 개별적으로 분석합니다. 모든 결과는 단일 테이블과 보고서에 제공됩니다.

가중치 각 행에 가중치를 할당하는 데 사용할 값이 있는 열을 식별합니다. 이러한 값은 분석에서 가중치로 사용됩니다. 자세한 내용은 **선형 모형 적합**에서 확인하십시오.

빈도 각 행에 빈도를 할당하는 데 사용할 값이 있는 열을 식별합니다. 이러한 값을 사용하면 사전 요약된 데이터를 고려할 수 있습니다. 자세한 내용은 **선형 모형 적합**에서 확인하십시오.

기준 지정된 열의 각 수준에 대해 해당하는 Y 및 X를 분석하고 결과를 별도의 테이블과 보고서에 제공합니다.

시작 창 옵션

공통 Y 척도 (연속형 반응이 두 개 이상인 경우에만 해당) 모든 연속형 반응의 척도가 동일하다는 것을 플랫폼에 알립니다. 이는 평균 차이 또는 기울기를 비교하는 데 필요합니다. 이 옵션을 선택하고 하나 이상의 범주형 X 변수가 있는 경우 "반응 변수 선별" 보고서 창에 "LogWorth 대 차이" 그림과 "평균 차이" 보고서가 기본적으로 표시됩니다. 이 옵션을 선택하지 않으면 "반응 변수 선별"의 빨간색 삼각형 메뉴를 통해 "LogWorth 대 차이" 그림에 액세스할 수 있습니다.

공통 X 척도 (연속형 X 변수가 두 개 이상인 경우에만 해당) 모든 연속형 X 변수의 척도가 동일하다는 것을 플랫폼에 알립니다. 이는 여러 X 변수의 기울기를 비교하는 데 필요합니다. "공통 X 척도" 옵션과 "공통 Y 척도" 옵션을 둘 다 선택한 경우 "반응 변수 선별" 보고서 창에 "LogWorth 대 기울기" 그림이 기본적으로 표시됩니다. 이 옵션을 선택하지 않으면 "반응 변수 선별"의 빨간색 삼각형 메뉴를 통해 "LogWorth 대 기울기" 그림에 액세스할 수 있습니다.

평균 비교 (연속형 반응과 범주형 X 변수가 각각 하나 이상인 경우에만 해당) X 변수의 여러 수준에 대한 평균을 비교하는 방법을 지정합니다.

각 수준과 대조군 각 범주형 X 변수에 대해 각 수준을 대조군 수준과 비교합니다. 기본 대조군 수준은 첫 번째 수준입니다.

팁: "대조 수준" 열 특성을 요인 열에 추가하여 기본 대조군과 다른 대조군을 지정할 수 있습니다. 자세한 내용은 **JMP 사용**에서 확인하십시오.

모든 조합 각 범주형 X 변수에 대해 가능한 모든 수준 조합을 비교합니다.

비율 조정 위험비, 승산비 및 위험도 차이를 계산할 때 셀 개수에 0.5를 추가하는 옵션을 제공합니다. 이렇게 조정하면 셀 개수가 0인 경우 0으로 나누기에서 발생하는 문제를 방지할 수 있습니다. 자세한 내용은 [Haldane\(1956\)](#) 연구 자료에서 확인하십시오. 하나 이상의 명목형 반응과 하나 이상의 범주형 예측 변수가 있는 경우 비율 조정 옵션을 사용합니다.

조정 없음 셀 개수를 조정하지 않습니다.

0 일 때 0.5 추가 위험비, 승산비 및 위험도 차이를 계산할 때 개수가 0 인 셀에 0.5 를 추가합니다.

항상 0.5 추가 위험비, 승산비 및 위험도 차이를 계산할 때 모든 셀 개수에 0.5를 추가합니다.

결측값을 범주로 처리 범주형 X 변수에 대해 X의 결측값을 범주로 처리합니다.

쌍체 X와 Y Y, 반응 및 X 목록의 순서에 따라 X 열과 쌍을 이룬 Y 열에 대해서만 검정을 수행합니다. 첫 번째 Y는 첫 번째 X와 쌍이 되고 두 번째 Y는 두 번째 X와 쌍이 되는 식입니다.

참고 : " 쌍체 X와 Y" 옵션을 선택한 경우 평균 차이 표시, 실제적 차이 및 동등성, 기울기 표시 및 2 x M 보고서 기능이 지원되지 않습니다.

시작 시 p 값 테이블 p 값 및 개별 모형 적합 통계량에 대한 데이터 테이블을 생성합니다. 이 데이터 테이블은 반응 변수 선별 보고서의 " 결과 테이블 "에 연결됩니다.

보고서 없음 보고서 창을 표시하지 않습니다. 이 옵션을 사용하면 보고서 창을 나타내지 않고 저장 명령을 실행하여 결과를 얻을 수 있습니다.

고급 옵션

로버스트 (로버스트 결과는 연속형 반응에 대해서만 표시됨) Huber M- 추정 방법 (Huber 와 Ronchetti 2009)을 사용하여 회귀 및 ANOVA 모형을 적합시킵니다. 이 방법은 Huber 손실 함수를 최소화하여 모수를 추정합니다. Huber 손실 함수는 이상치에 벌점을 부과하여 이상치가 검정에 미치는 영향을 줄입니다. 이상치가 없으면 추정값이 최소 제곱 추정값에 가깝습니다. " 로버스트 " 옵션을 선택한 경우 여러 로버스트 측도 및 모형 적합 통계량 열이 " 결과 테이블 "에 추가됩니다. Huber M- 추정에 대한 자세한 내용은 기본 분석에서 확인하십시오. 예는 " 로버스트 적합의 예 "에서 확인하십시오.

주의 : " 로버스트 " 옵션을 사용하면 처리 시간이 길어질 수 있습니다.

Cauchy (Cauchy 결과는 연속형 반응에 대해서만 표시됨) 최대 가능도와 Cauchy 연결 함수를 사용하여 모수를 추정합니다. 이 추정 방법은 오차가 Cauchy 분포를 따른다고 가정합니다. Cauchy 분포는 정규 분포보다 꼬리가 더 두꺼우므로 이상치의 영향이 줄어듭니다. 이 옵션은 데이터에 포함된 이상치의 비율이 높은 경우에 유용할 수 있습니다. 하지만 데이터가 정규 분포에 가깝고 이상치가 일부만 있는 경우에 이 옵션을 사용하면 추론이 잘못될 수 있습니다. "Cauchy" 옵션을 선택하면 여러 Cauchy 측도 및 모형 적합 통계량 열이 " 결과 테이블 "에 추가됩니다.

Poisson Y 각 Y 반응을 Poisson 분포를 따르는 개수로 적합시킵니다. 반응이 연속형이고 X 변수가 범주형인 경우에만 검정이 수행됩니다. 이 옵션은 반응이 개수일 때 적절합니다.

경험적 Bayes 축소 (" 공통 Y 척도 " 옵션을 선택한 경우에만 적용 가능) 잔차 분산 추정값을 추정된 사전 최빈값으로 축소, 모든 추정값에서 강도 차용 (borrowing strength). 이는 공통 척도를 사용하는 많은 연속형 반응을 선별할 때 유용합니다. 사전 값을 적용하면 각 개별 적합에 대한 자유도와 오차 제곱합이 증가합니다. 이로 인해 검정에 사용할 검정력이 약간 증가합니다. 자세한 내용은 Smyth(2004) 연구 자료에서 확인하십시오.

카파 (Y 와 X 가 모두 범주형이고 수준 수가 같은 경우에만 해당) Y 와 X 간의 합치도 측도인 카파를 계산합니다. 결과는 결과 테이블의 " 카파 " 열에 보고됩니다.

상관 Pearson 곱적률 상관계수를 계산합니다. 결과는 결과 테이블의 " 상관 " 열에 보고됩니다. 범주형 변수의 경우 값 순서화로 정의된 인덱스에 따라 상관이 계산됩니다.

X 와 Y 가 모두 이항인 경우 Pearson 곱적률을 계산하면 Spearman Rho 및 Kendall Tau-B 가 제공됩니다. 그렇지 않은 경우 " 상관 " 값이 크면 연관성을 나타내고 " 상관 " 값이 작으면 연관성을 배제하지 않습니다.

최대 LogWorth LogWorth 값 (p 값의 $-\log_{10}$) 과 관련된 그림의 척도를 제어합니다. 최대 LogWorth를 초과하는 LogWorth 값은 LogWorth 그림에서 극단 척도를 방지하기 위해 최대 LogWorth 로 표시됩니다. 예는 " [최대 LogWorth 옵션의 예](#) " 에서 확인하십시오.

X 범주형 강제 적용 모델링 유형을 무시하고 모든 X 열을 범주형 열로 처리합니다.

X 연속형 강제 적용 모델링 유형을 무시하고 모든 X 열을 연속형 열로 처리합니다.

Y 범주형 강제 적용 모델링 유형을 무시하고 모든 Y 열을 범주형 열로 처리합니다.

Y 연속형 강제 적용 모델링 유형을 무시하고 모든 Y 열을 연속형 열로 처리합니다.

언스레드 계산 속도를 위해 멀티스레딩 사용을 제한합니다.

반응 변수 선별 보고서

" 반응 변수 선별 " 보고서는 그림 섹션과 보고서 섹션으로 구성됩니다.

- " [반응 변수 선별 그림](#) "
- " [반응 변수 선별 보고서 테이블](#) "

팁: 그림 섹션 또는 보고서 섹션의 레이아웃을 변경하려면 원하는 섹션의 맨 위를 마우스 오른쪽 버튼으로 클릭하고 **스타일 설정**을 선택합니다. 기본 레이아웃은 최대 5 개 그림에 대한 탭 형식 보고서입니다. 그림이 5 개를 초과하면 콤보 레이아웃이 기본값입니다. 콤보 레이아웃에는 드롭다운 그림 선택기가 사용됩니다.

반응 변수 선별 그림

" 반응 변수 선별 " 보고서 창에 포함되는 기본 그림은 시작 창에서 선택한 내용에 따라 달라집니다. 그림은 FDR(False Discovery Rate) 통계량에 초점을 맞춥니다. 자세한 내용은 " [반응 변수 선별 플랫폼에 대한 통계 상세 정보](#) " 에서 확인하십시오. 다음 목록에서는 가능한 모든 그림과 해당 그림을 사용할 수 있는 시기를 설명합니다.

FDR p 값 그림 세로 축에 FDR p 값과 표준 p 값을 표시하고 가로 축에 순위 비율을 표시합니다. 이 그림은 False Discovery Rate 조정 여부와 상관없이 유의한 검정 수를 표시하여 조정이 적용되는 정도를 나타냅니다. 순위 비율은 p 값을 유의성이 큰 것부터 순서대로 정렬합니다. FDR p 값은 파란색으로 표시되고 p 값은 빨간색으로 표시됩니다. 파란색 가로선은 0.05 유

의 수준을 나타냅니다. 증가하는 빨간색 선은 조정되지 않은 p 값에 대한 FDR 임계를 제공합니다. FDR 조정 p 값이 파란색 선 아래에 있으면 p 값이 빨간색 선 아래에 있습니다. 이를 통해 조정되거나 조정되지 않은 p 값 중 하나를 사용하여 FDR 을 기준으로 유의성을 이해할 수 있습니다. 또한 그림 옆에는 검정 수 및 유의 수준 0.05 와 0.01 에서 유의한 검정의 비율을 나타내는 범례가 있습니다.

FDR LogWorth 대 효과 크기 세로 축에 FDR LogWorth 를 표시하고 가로 축에 효과 크기를 표시합니다. 이 그림은 큰 효과와 연관된 매우 작은 p 값을 그래픽으로 시각화하는 데 도움이 됩니다. $\text{LogWorth}(-\log_{10}(p \text{ 값}))$ 척도로 변환하면 매우 유의한 p 값은 LogWorth 값이 크고 유의하지 않은 p 값은 LogWorth 값이 낮습니다. LogWorth 값이 0 이면 유의하지 않은 p 값 1 에 해당합니다. LogWorth 값이 2 를 초과하면 p 값이 0.01 미만입니다. 빠른 시각화를 위해 그림에서 $\text{LogWorth} = 2$ 에 참조선이 있습니다.

일반적으로 효과가 클수록 p 값이 더 유의하고 LogWorth 값이 더 큼니다. 그러나 유의성은 오차 분산에도 영향을 받으므로 이 관계가 항상 성립하는 것은 아닙니다. 사실 오차 분산의 크기 때문에 큰 LogWorth 값은 작은 효과와 연관될 수 있고 작은 LogWorth 값은 큰 효과와 연관될 수 있습니다. "FDR LogWorth 대 효과 크기" 그림을 사용하여 이 관계를 살펴볼 수 있습니다.

FDR LogWorth 대 R^2 (연속형 Y 변수에만 사용 가능) 세로 축에 FDR LogWorth 를 표시하고 가로 축에 R^2 값을 표시합니다. LogWorth 값이 클수록 더 큰 R^2 값과 연관되는 경향이 있지만 이 관계는 관측값 수에도 영향을 받습니다.

FDR LogWorth 대 카파 그림 (시작 창에서 "카파" 옵션을 선택하고 수준 수가 동일한 범주형 X 및 Y 열이 있는 경우에만 사용 가능) 세로 축에 FDR LogWorth 를 표시하고 가로 축에 카파를 표시합니다.

FDR LogWorth 대 상관 그림 (시작 창에서 "상관" 옵션을 선택하고 범주형 X 및 Y 열이 있는 경우에만 사용 가능) 세로 축에 FDR LogWorth 를 표시하고 가로 축에 상관을 표시합니다.

LogWorth 대 차이 (기본적으로 하나 이상의 연속형 반응과 하나 이상의 범주형 X 변수가 있고 시작 창에서 "공통 Y 척도" 옵션을 선택한 경우에만 사용 가능) 세로 축에 LogWorth 를 표시하고 가로 축에 평균 차이를 표시합니다. 각 수준 대 대조 수준에 대한 평균 차이가 계산됩니다. 기본 대조 수준은 첫 번째 수준입니다. 변수가 하나뿐이고 수준이 세 개 이상인 경우 그림에 표시할 수준을 선택할 수 있는 체크박스가 있습니다.

유의하지 않은 점 ($\text{FDR } p \text{ 값} \geq 0.01$) 은 회색으로 표시됩니다. 유의한 점 ($\text{FDR } p \text{ 값} < 0.01$) 은 평균 차이가 양수이면 빨간색으로 표시되고 평균 차이가 음수이면 파란색으로 표시됩니다.

참고 : 이 그림을 화산 (Volcano) 그림이라고도 합니다.

LogWorth 대 기울기 (기본적으로 하나 이상의 연속형 반응과 하나 이상의 연속형 X 변수가 있고 시작 창에서 "공통 Y 척도" 옵션과 "공통 X 척도" 옵션을 모두 선택한 경우에만 사용 가능) 세로 축에 FDR LogWorth 차이를 표시하고 가로 축에 기울기를 표시합니다. 기울기는 각 반응과 각 연속형 변수 간의 단순 선형 회귀에서 계산됩니다.

유의성을 나타내는 참조선이 있습니다. 참조선은 $\text{FDR } p \text{ 값} 0.01$ 에 해당하는 $\text{FDR LogWorth} = 2$ 에 있습니다. LogWorth 값이 2 를 초과하면 p 값이 0.01 미만이라는 의미입니다. 유의하

지 않은 점 (FDR p 값 ≥ 0.01) 은 회색으로 표시됩니다. 유의한 점 (FDR p 값 < 0.01) 은 기울기가 양수이면 빨간색으로 표시되고 기울기가 음수이면 파란색으로 표시됩니다.

참고 : 이 그림을 화산 (Volcano) 그림이라고도 합니다.

실제적 차이 FDR LogWorth 차이와 FDR LogWorth 동등성 중 더 큰 값을 세로 축에 표시합니다. 가로 축은 "실제적 차이 및 동등성" 대화상자에서 지정한 규격에 따라 다릅니다. 비율을 지정한 경우 가로 축은 차이를 감지할 차이로 나눈 값입니다. 특정 차이를 지정한 경우 가로 축은 해당 차이입니다.

로버스트 FDR p 값 그림 (시작 창에서 "로버스트" 옵션을 선택한 경우에만 사용 가능) 세로 축에 로버스트 FDR p 값과 표준 p 값을 표시하고 가로 축에 로버스트 순위 비율을 표시합니다. 이 그림은 FDR p 값 그림과 비슷하지만 표준 값 대신 로버스트 값을 사용합니다.

로버스트 FDR LogWorth 대 FDR LogWorth 그림 (시작 창에서 "로버스트" 옵션을 선택한 경우에만 사용 가능) 세로 축에 로버스트 FDR LogWorth 를 표시하고 가로 축에 FDR LogWorth 를 표시합니다.

로버스트 FDR LogWorth 대 효과 크기 (시작 창에서 "로버스트" 옵션을 선택한 경우에만 사용 가능) 세로 축에 로버스트 FDR LogWorth 를 표시하고 가로 축에 효과 크기를 표시합니다. 이 그림은 FDR LogWorth 대 효과 크기 그림과 비슷하지만 표준 FDR LogWorth 대신 로버스트 FDR LogWorth 를 사용합니다.

Cauchy FDR p 값 그림 (시작 창에서 "Cauchy" 옵션을 선택한 경우에만 사용 가능) 세로 축에 Cauchy FDR p 값과 표준 p 값을 표시하고 가로 축에 Cauchy 순위 비율을 표시합니다. 이 그림은 FDR p 값 그림과 비슷하지만 표준 값 대신 Cauchy 값을 사용합니다.

Cauchy FDR LogWorth 대 FDR LogWorth 그림 (시작 창에서 "Cauchy" 옵션을 선택한 경우에만 사용 가능) 세로 축에 Cauchy FDR LogWorth 를 표시하고 가로 축에 FDR LogWorth 를 표시합니다.

Cauchy FDR LogWorth 대 효과 크기 (시작 창에서 "Cauchy" 옵션을 선택한 경우에만 사용 가능) 세로 축에 Cauchy FDR LogWorth 를 표시하고 가로 축에 효과 크기를 표시합니다. 이 그림은 Cauchy LogWorth 대 효과 크기 그림과 비슷하지만 표준 FDR LogWorth 대신 Cauchy FDR LogWorth 를 사용합니다.

반응 변수 선별 보고서 테이블

"반응 변수 선별" 보고서 창에 포함되는 기본 테이블은 시작 창에서 선택한 내용에 따라 달라집니다. "결과 테이블"은 항상 표시되고, "2 x M 결과" 테이블은 하나 이상의 명목형 반응과 하나 이상의 범주형 예측 변수가 있는 경우에만 표시됩니다.

결과 테이블

"반응 변수 선별" 보고서의 "결과 테이블"에는 Y 및 X 변수의 각 쌍에 대한 행이 포함되어 있습니다. 테이블의 열에는 선택한 적합과 Y 및 X 모델링 유형에 고유한 측도와 모형 적합 통계량이 포함됩니다.

그룹 (그룹화 변수가 있는 경우에만 표시됨) 그룹화 열의 수준입니다 .

Y 지정된 반응 열입니다 .

X 지정된 요인 열입니다 .

개수 검정에 사용된 행 수 또는 빈도 변수나 가중치 변수의 해당 합입니다 .

p 값 Y 와 X 변수 쌍에 해당하는 유의성 검정의 p 값입니다 . X 로 Y 적합 통계량에 대한 자세한 내용은 기본 분석에서 확인하십시오 .

LogWorth $-\log_{10}(p \text{ 값})$ 의 계산 값입니다 . 이 변환은 그래프 작성에 적절한 척도를 제공하도록 p 값을 조정합니다 . 2 를 초과하는 값은 0.01 수준에서 유의합니다 ($-\log_{10}(0.01) = 2$).

FDR p 값 Benjamini-Hochberg 기법을 사용하여 계산된 False Discovery Rate p 값입니다 . 이 기법은 여러 검정에 대한 FDR(False Discovery Rate) 을 제어하기 위해 p 값을 조정합니다 . 그룹 변수가 없으면 테이블에 표시된 모든 검정이 여러 검정 집합에 포함됩니다 . 그룹 변수가 있으면 그룹 변수의 각 수준에 대해 수행된 모든 검정이 여러 검정 집합에 포함됩니다 . FDR 수정에 대한 자세한 내용은 Benjamini and Hochberg 연구 자료 (1995) 에서 확인하십시오 . False Discovery Rate 에 대한 자세한 내용은 " 반응 변수 선별 플랫폼에 대한 통계 상세 정보 " 에서 확인하십시오 .

FDR LogWorth $-\log_{10}(\text{FDR } p \text{ 값})$ 의 계산 값입니다 . 이 값은 유의성을 그림에 표시하고 평가하는 데 사용되는 통계량입니다 . p 값이 작으면 FDR LogWorth 값이 커집니다 . 2 보다 큰 FDR LogWorth 값 (0.01 보다 작은 p 값) 에 해당하는 셀은 강도 그래디언트가 적용된 색상이 지정됩니다 .

효과 크기 X 의 수준 또는 값에 따라 반응 값이 달라지는 정도를 나타냅니다 . 효과 크기는 척도 불변입니다 .

- Y 가 연속형인 경우 효과 크기는 가설 검정에서 구한 평균 제곱합의 제곱근을 반응 표준 편차의 로버스트 추정값으로 나눈 값입니다 . IQR(사분위수 범위) 이 0 이 아니고 $\text{IQR} > \text{범위} / 20$ 이면 표준편차 추정값은 $\text{IQR} / 1.3489795$ 입니다 . 그렇지 않으면 표본 표준편차가 사용됩니다 .
- Y 가 범주형이고 X 가 연속형인 경우 효과 크기는 전체 모형 검정에 대한 평균 카이제곱 값의 제곱근입니다 .
- Y 와 X 가 모두 범주형인 경우 효과 크기는 평균 Pearson 카이제곱 값의 제곱근입니다 .

순위 비율 FDR LogWorth 순위를 검정 수의 비율로 나타낸 것입니다 . 검정 수가 m 이면 가장 큰 FDR LogWorth 값의 순위 비율은 $1/m$ 이고 가장 작은 값의 순위 비율은 1 입니다 . 순위 비율은 p 값과 FDR p 값을 순위별로 유의성이 큰 것부터 그림에 표시하는 데 사용됩니다 .

R² (Y 가 연속형인 경우에만 표시됨) 모형에 의해 설명되는 총 변동의 비율을 측정하는 결정 계수입니다 .

카파 (시작 창에서 지정한 경우에만 표시되고 , 수준 수가 동일한 범주형 Y 및 X 에만 사용 가능) Y 와 X 간의 합치도 척도입니다 .

상관 (시작 창에서 지정한 경우에만 표시됨) Pearson 곱적률 상관계수입니다 . 범주형 변수의 경우 값 순서화로 정의된 인덱스에 따라 상관이 계산됩니다 .

시작 창에서 "로버스트" 옵션을 선택한 경우 다음 열이 "결과 테이블"에 추가됩니다. "로버스트" 옵션은 Y가 연속형일 때만 적용되므로 Y가 범주형이면 로버스트 열 셀이 비어 있습니다.

로버스트 p 값 로버스트 적합을 사용한 Y와 X 변수 쌍에 해당하는 유의성 검정의 p 값입니다.

로버스트 LogWorth $-\log_{10}(\text{로버스트 } p \text{ 값})$ 의 계산 값입니다.

로버스트 FDR p 값 Benjamini-Hochberg 기법을 사용하여 로버스트 p 값에 대해 계산된 False Discovery Rate입니다. 그룹 변수가 없으면 테이블에 표시된 모든 검정에 다중 검정 조정이 적용됩니다. 그룹 변수가 있으면 그룹 변수의 각 수준에 대해 수행된 모든 검정에 다중 검정 조정이 적용됩니다.

로버스트 FDR LogWorth $-\log_{10}(\text{로버스트 FDR } p \text{ 값})$ 의 계산 값입니다.

로버스트 순위 비율 로버스트 FDR LogWorth 순위를 검정 수의 비율로 나타낸 것입니다.

로버스트 카이제곱 로버스트 검정과 관련된 카이제곱 값입니다.

로버스트 시그마 오차 표준편차의 로버스트 추정값입니다.

로버스트 이상치 비율 로버스트 평균에서의 거리가 로버스트 시그마의 3 배를 초과하는 값의 비율입니다.

시작 창에서 "Cauchy" 옵션을 선택하면 다음 열이 "결과 테이블"에 추가됩니다. "Cauchy" 옵션은 Y가 연속형일 때만 적용되므로 Y가 범주형이면 Cauchy 열 셀이 비어 있습니다.

Cauchy p 값 Cauchy 적합을 사용한 Y와 X 변수 쌍에 해당하는 유의성 검정의 p 값입니다.

Cauchy LogWorth $-\log_{10}(\text{Cauchy } p \text{ 값})$ 의 계산 값입니다.

Cauchy FDR p 값 Benjamini-Hochberg 기법을 사용하여 Cauchy p 값에 대해 계산된 False Discovery Rate입니다. 그룹 변수가 없으면 테이블에 표시된 모든 검정에 다중 검정 조정이 적용됩니다. 그룹 변수가 있으면 그룹 변수의 각 수준에 대해 수행된 모든 검정에 다중 검정 조정이 적용됩니다.

Cauchy FDR LogWorth $-\log_{10}(\text{Cauchy FDR } p \text{ 값})$ 의 계산 값입니다.

Cauchy 순위 비율 Cauchy FDR LogWorth 순위를 검정 수의 비율로 나타낸 것입니다.

Cauchy 카이제곱 Cauchy 적합의 검정과 관련된 카이제곱 값입니다.

Cauchy 시그마 Cauchy 적합으로부터의 오차 표준편차의 추정값입니다.

Cauchy 이상치 비율 Cauchy 적합의 평균에서의 거리가 Cauchy 시그마의 3 배를 초과하는 값의 비율입니다.

2 x M 결과

2 x M 테이블에는 2수준 범주형 반응과 범주형 예측 변수 간의 관계를 조사하는 데 사용할 수 있는 결과가 포함되어 있습니다. 각 반응과 예측 변수 쌍에 대해 상대 위험도, 위험도 차이 및 승산 비가 기본적으로 보고됩니다. 테이블을 마우스 오른쪽 버튼으로 클릭하고 **열**을 선택하여 개수, 신뢰 구간, 라벨 또는 FDR LogWorth 값을 테이블에 추가할 수 있습니다. 첫 번째 수준이 기준으로 설정되고 다른 각 수준은 이 기준과 비교됩니다.

참고 : "2xM 결과 " 보고서는 수준이 두 개인 하나 이상의 범주형 반응과 하나 이상의 범주형 예측 변수가 있는 경우에만 사용할 수 있습니다 .

평균 차이

"평균 차이" 테이블의 각 행은 범주형 요인의 두 수준에 대한 반응을 비교합니다. 시작 창의 "평균 비교" 옵션을 사용하여 각 수준을 대조 수준과 비교할지 아니면 가능한 모든 수준 조합과 비교할지 지정할 수 있습니다. 이 테이블은 비교되는 수준과 X 변수를 보여 줍니다.

"평균 차이" 테이블의 "차" 열에는 두 수준에서 반응 평균의 추정된 차이가 포함됩니다. "로버스트" 옵션을 선택한 경우 평균의 로버스트 추정값이 사용됩니다. 두 평균 간의 차이가 0인지 여부를 검정하여 얻은 FDR p 값 및 FDR LogWorth에 대한 열도 있습니다.

참고 : " 평균 차이 " 테이블은 하나 이상의 연속형 반응과 하나 이상의 범주형 예측 변수가 있는 경우에만 사용할 수 있습니다 .

반응 변수 선별 플랫폼 옵션

" 반응 변수 선별 "의 빨간색 삼각형 메뉴에는 화면 표시를 사용자 정의하고 계산된 데이터를 저장하는 옵션이 포함되어 있습니다 .

그림 표시 보고서 창에서 그림을 표시하거나 숨깁니다 .

보고서 테이블 표시 보고서 창에서 결과 테이블을 표시하거나 숨깁니다 .

평균 차이 표시 (연속형 반응과 범주형 X 변수가 각각 하나 이상인 경우에만 사용 가능) " 반응 변수 선별 " 보고서 창에 "LogWorth 대 차이 " 그림과 " 평균 차이 " 보고서를 추가합니다 . 자세한 내용은 "LogWorth 대 차이 " 및 " 평균 차이 " 에서 확인하십시오 .

참고 : 그림과 테이블이 보고서에 이미 포함되어 있으므로 시작 창에서 " 공통 Y 척도 "를 선택하는 경우 이 옵션을 사용할 수 없습니다 .

기울기 표시 " 반응 변수 선별 " 보고서 창에 "LogWorth 대 기울기 " 그림을 추가합니다 . 자세한 내용은 "LogWorth 대 기울기 " 에서 확인하십시오 .

참고 : 그림이 보고서에 이미 포함되어 있으므로 시작 창에서 " 공통 Y 척도 " 및 " 공통 X 척도 "를 선택하는 경우 이 옵션을 사용할 수 없습니다 .

실제적 차이 및 동등성 범주형 변수의 두 수준에서 평균의 실제적 차이 및 동등성 검정을 수행합니다. 시작 창의 " 평균 비교 " 옵션을 사용하여 각 수준을 대조 수준과 비교할지 아니면 가능한 모든 수준 조합과 비교할지 지정할 수 있습니다. 실제적 유의성 검정은 평균 간의 차이가 0 인지 검정하는 대신 평균의 차이가 지정된 실제적 차이를 초과하는지 여부를 검정합니다. 동등성 검정을 사용하면 반응에 대한 두 수준의 효과가 본질적으로 동일한지 여부를 실제적

관점에서 판단할 수 있습니다. 즉, 동등성 검정은 차이가 실제적 차이보다 작은지 여부를 결정합니다.

이 옵션을 선택하면 감지할 차이의 크기를 지정합니다. 다음 두 옵션을 사용하여 실제적 차이를 지정할 수 있습니다.

비율 Y 변수에 "규격 한계" 특성을 할당하는 경우 실제적 차이는 규격 상한과 규격 하한 사이의 차이에 지정된 비율을 곱한 값으로 계산됩니다. Y 변수에 "규격 한계" 특성을 할당하지 않는 경우 실제적 차이는 $6\hat{\sigma}$ 에 지정된 비율을 곱한 값으로 계산됩니다. 여기서 $\hat{\sigma}$ 은 표준편차의 로버스트 추정값입니다. IQR(사분위수 범위)이 0이 아니고 $IQR > \text{범위}/20$ 이면 표준편차 추정값은 $\hat{\sigma} = (IQR)/(1.3489795)$ 입니다. 그렇지 않으면 표본 표준편차가 $\hat{\sigma}$ 으로 사용됩니다. 기본적으로 비율은 0.1입니다.

실제 차이 모든 비교 검정에 대한 실제적 차이를 지정합니다. 이 옵션은 모든 반응의 척도가 동일한 경우에 적절합니다.

"확인"을 클릭하면 "실제적 차이" 그림과 "실제적 차이" 테이블이 보고서 창에 추가됩니다. 그림에 대한 자세한 내용은 "[실제적 차이](#)"에서 확인하십시오. "실제적 차이" 테이블은 비교되는 수준과 X 변수를 보여 줍니다. 수준 간의 실제 차이, 감지하도록 지정된 실제적 차이, 실제적 차이 및 동등성 검정의 p 값에 대한 열이 있습니다. 또한 검정 결과에 따라 수준이 다른지, 동등한지 또는 판단할 수 없는지를 나타내는 열도 있습니다.

선택한 항목 적합 선택한 관계에 대해 적절한 X로 Y 적합 보고서를 "반응 변수 선별" 보고서에 추가합니다. 그림의 점, 결과 테이블의 행 또는 p 값 데이터 테이블(열린 경우)의 행을 선택하여 관계를 선택할 수 있습니다.

열 선택 결과 테이블에서 선택한 행, 반응 변수 선별 보고서 창의 그림에서 선택한 점 또는 p 값 테이블(열린 경우)의 행에 해당하는 열을 원래 데이터 테이블에서 선택합니다. 먼저 행 또는 점을 선택한 후 "열 선택"을 선택합니다. 데이터 테이블에서 해당 열이 선택됩니다. 먼저 p 값 테이블의 추가 행 또는 그림의 점을 선택한 후 다시 "열 선택"을 선택하면 이에 해당하는 열을 선택할 수 있습니다. 다른 행이나 점에 해당하는 열을 선택하려면 먼저 원래 데이터 테이블에서 현재 열 선택을 지웁니다.

선택 조건 "선택 조건" 창을 엽니다. "비교" 메뉴와 "값" 텍스트 상자를 사용하여 결과 테이블에서 특정 조건에 해당하는 특정 반응을 선택할 수 있습니다. 예를 들어 "효과 크기 > 0.80"에 해당하는 모든 반응을 선택할 수 있습니다. "확인"을 클릭하면 "결과 테이블"에서 반응이 선택됩니다.

팁: "결과 테이블"의 아무 곳이나 마우스 오른쪽 버튼으로 클릭하여 "선택 조건" 창에 액세스할 수도 있습니다.

테이블 저장 새 데이터 테이블을 생성하는 옵션이 포함된 하위 메뉴를 표시합니다.

p 값 저장 p 값 데이터 테이블을 생성합니다. 자세한 내용은 "[반응 변수 선별 플랫폼 옵션](#)"에서 확인하십시오.

평균 저장 (연속형 Y 변수 및 범주형 X 변수에 사용 가능) 범주형 변수의 각 수준에 대한 개수, 평균 및 표준편차를 포함하는 데이터 테이블을 생성합니다. "로버스트" 옵션을 선택한 경우 로버스트 평균이 포함됩니다. 자세한 내용은 "[평균 데이터 테이블](#)"에서 확인하십시오.

평균 차이 저장 (연속형 Y 변수 및 범주형 X 변수에 사용 가능) 범주형 변수의 수준 간 비교 검정을 위한 새 데이터 테이블을 생성합니다. 여기에는 "평균 차이" 보고서 테이블의 열과 추가 열이 포함됩니다. 실제적 차이 검정이 지정된 경우 이러한 검정 결과가 새 데이터 테이블에 포함됩니다. 자세한 내용은 "[평균 차이 데이터 테이블](#)"에서 확인하십시오. 예는 "[실제적 유의성 및 동등성 검정의 예](#)"에서 확인하십시오.

2 x M 저장 (범주형 X 변수 및 2 수준 범주형 Y 변수에 사용 가능) 2 x M 결과 보고서의 정보와 기타 검정 통계량을 포함하는 데이터 테이블을 생성합니다. 데이터 테이블에는 다음과 같은 열 그룹이 포함됩니다.

셀 개수 일반적으로 분할표에 나타나는 셀 개수 열을 포함합니다.

승산비 승산비의 추정값, 신뢰 구간 하한 / 상한 및 유의성 상태에 대한 열을 포함합니다.

상대 위험도 상대 위험도의 추정값, 신뢰 구간 하한 / 상한 및 유의성 상태에 대한 열을 포함합니다.

위험도 차이 위험도 차이의 추정값, 신뢰 구간 하한 / 상한 및 유의성 상태에 대한 열을 포함합니다.

검정 통계량 여러 개의 검정 통계량 열을 포함합니다.

Fisher 정확 검정 오른쪽 꼬리, 왼쪽 꼬리 및 양쪽 꼬리 Fisher 정확 검정에 대한 여러 열을 포함합니다.

2 x M 테이블의 결과에 대한 자세한 내용은 [기본 분석](#)에서 확인하십시오.

열 저장 (연속형 Y 변수에 사용 가능) 원래 데이터 테이블에 새 열을 저장하는 옵션이 포함된 하위 메뉴를 표시합니다.

표준화 잔차 저장 "잔차 그룹"이라는 새 열 그룹을 원래 데이터 테이블에 저장하고 이 열 그룹에 열을 배치합니다. 각 연속형 Y와 범주형 X에 대해 잔차를 추정된 표준편차로 나눈 값을 포함하는 열이 생성됩니다. 즉, 표준화 잔차가 열에 포함됩니다. 열은 계산식으로 정의됩니다.

"로버스트" 옵션을 선택한 경우 로버스트 적합 및 로버스트 추정값을 사용하여 표준화 잔차 열이 생성됩니다.

이상치 표시자 저장 "이상치 그룹"이라는 새 열 그룹을 원래 데이터 테이블에 저장하고 이 열 그룹에 열을 배치합니다. "이상치 표시자 저장"은 "로버스트" 옵션을 선택한 경우에 가장 효과적입니다.

각 연속형 Y와 범주형 X에 대해 이상치를 나타내는 열이 생성됩니다. 이상치는 예측값까지의 거리가 시그마 추정값의 3 배를 초과하는 점입니다. 즉, 이상치는 표준화 잔차가 3을 초과하는 점입니다. 열은 계산식으로 정의됩니다.

"로버스트" 옵션을 선택한 경우 로버스트 적합 및 로버스트 추정값이 사용됩니다. 이상치는 예측값까지의 거리가 로버스트 시그마 추정값의 3 배를 초과하는 점입니다.

"이상치 군집화" 스크립트가 원래 데이터 테이블에 추가됩니다. 이 스크립트는 데이터의 계층적 군집화 그림에 이상치를 표시합니다.

다음 옵션에 대한 자세한 내용은 **JMP 사용**의 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.

평균 데이터 테이블

"반응 변수 선별"의 빨간색 삼각형 메뉴에서 "평균 저장" 옵션을 선택하면 "평균" 데이터 테이블이 생성됩니다. 데이터 테이블에는 반응과 X 수준의 각 조합에 대한 행이 포함되어 있습니다. Probe.jmp 샘플 데이터 테이블의 경우 387개의 반응 변수가 있고 각 변수가 두 수준의 Process에 대해 검정되었습니다. 따라서 "평균" 테이블에는 $387 \times 2 = 774$ 개 행이 포함됩니다.

그림 24.5 평균 데이터 테이블

Save Means		Y	X	수준	개수	평균	표준편차
		1 DELL_RPNBR	Process	New	3044	0.2035840106	0.1916031302
		2 DELL_RPNBR	Process	Old	2750	0.2346329436	0.1278299414
		3 DELL_RPPBR	Process	New	3044	-0.068072506	0.1784161107
		4 DELL_RPPBR	Process	Old	2750	-0.043321135	1.9165048167
		5 DELW_M1	Process	New	3039	-0.04387197	1.0774346893
		6 DELW_M1	Process	Old	2750	-0.071400861	0.0389764703
		7 DELW_M2	Process	New	3039	0.8014275806	0.0698264259
		8 DELW_M2	Process	Old	2689	0.7274958056	0.0649383327
		9 DELW_NBASE	Process	New	3028	0.1957655669	77.801881218
		10 DELW_NBASE	Process	Old	2750	1.4765185729	0.1326787074
		11 DELW_NEMIT	Process	New	3038	0.3411755212	0.0940762694
		12 DELW_NEMIT	Process	Old	2750	0.3813085954	0.0976561441
		13 DELW_NENBNI	Process	New	3043	3.7172025841	1.1419433505
		14 DELW_NENBNI	Process	Old	2749	4.6556238745	0.1893626325
		15 DELW_NSINK	Process	New	3032	9.2876152816	1.2840573691
		16 DELW_NSINK	Process	Old	2750	7.8902297629	0.3781489659
		17 DELW_PBASE	Process	New	3036	1.1651478779	0.1351193171
		18 DELW_PBASE	Process	Old	2750	1.2087969753	0.1486923744

평균 데이터 테이블에는 다음 열이 포함됩니다.

Y 연속형 반응 변수입니다.

X 범주형 변수입니다.

수준 범주형 X 변수의 수준입니다.

개수 해당 수준의 값 개수입니다.

평균 지정된 수준에 대한 Y 변수의 평균입니다.

표준편차 지정된 수준에 대한 Y 변수의 표준편차입니다.

로버스트 평균 평균의 로버스트 M- 추정값입니다. 시작 창에서 "로버스트" 옵션을 선택한 경우 표시됩니다.

평균 차이 데이터 테이블

"반응 변수 선별"의 빨간색 삼각형 메뉴에서 "평균 차이 저장" 옵션을 선택하면 "평균 차이" 데이터 테이블이 생성됩니다. 데이터 테이블에는 표준 차이 검정에 대한 결과가 제공됩니다. 실제적 차이 및 동등성 옵션을 지정한 경우에는 실제적 차이 검정과 실제적 동등성 검정도 모두 새 데이터 테이블에 제공됩니다. 각 행은 범주형 요인의 두 수준에 대한 반응을 비교합니다. 쌍별 비교 결과는 쉽게 해석할 수 있도록 색상으로 구분됩니다. 실제적 차이를 지정하는 방법에 대한 설명은 "**실제적 차이**"에서 확인하십시오. 예는 "**실제적 유의성 및 동등성 검정의 예**"에서 확인하십시오.

Y 연속형 반응 변수입니다.

X 범주형 변수입니다.

수준 범주형 X 변수의 수준입니다.

- 수준 초기 수준과 비교되는 범주형 X 변수의 수준입니다.

현재 평균 "수준" 열의 수준에 대한 반응 변수의 평균입니다. "로버스트" 옵션을 선택한 경우 평균의 로버스트 추정값이 사용됩니다.

기준 평균 "- 수준" 열의 수준에 대한 반응 변수의 평균입니다. "로버스트" 옵션을 선택한 경우 평균의 로버스트 추정값이 사용됩니다.

차이 두 수준에서 추정된 평균 차이이며, 현재 평균 - 기준 평균으로 계산됩니다.

차이 표준 오차 평균 차이의 표준 오차입니다. "로버스트" 옵션을 선택한 경우 이 값은 로버스트 추정값입니다.

차이 L95 평균 간 차이에 대한 95% 신뢰 구간의 하한입니다.

차이 U95 평균 간 차이에 대한 95% 신뢰 구간의 상한입니다.

p 값 쌍별 비교를 위한 일반적인 스튜던트 t -검정의 p 값입니다. "로버스트" 옵션을 선택한 경우 이 값은 t -검정의 로버스트 버전입니다. 0.05 수준에서 유의한 검정이 강조 표시됩니다.

LogWorth $-\log_{10}(p \text{ 값})$ 의 계산 값입니다. 이 변환은 그래프 작성에 적절한 척도를 제공하도록 p 값을 조정합니다.

FDR p 값 쌍별 비교를 위한 스튜던트 t -검정의 FDR p 값입니다. 자세한 내용은 "[반응 변수 선별 플랫폼에 대한 통계 상세 정보](#)"에서 확인하십시오.

FDR LogWorth $-\log_{10}(\text{FDR } p \text{ 값})$ 의 계산 값입니다.

감지할 차이 실제적 관심 영역으로 간주할 평균 차이입니다. "실제적 차이 및 동등성"의 빨간색 삼각형 메뉴 옵션을 선택할 때 지정됩니다. 자세한 내용은 "[실제적 차이 및 동등성](#)"에서 확인하십시오.

실제적 차이 p 값 수준과 -수준 간 Y 의 평균 차이 절대값이 감지할 차이보다 작거나 같은지 여부를 검정하기 위한 p 값입니다. p 값이 작으면 절대 차이가 감지할 차이를 초과한다는 것을 나타냅니다. 이는 수준과 -수준이 실제로 중요한 차이를 설명한다는 의미입니다.

실제적 동등 p 값 TOST(Two One-Sided Tests: 두 개의 단측 검정) 방법을 사용하여 평균 간의 실제적 차이를 검정합니다 (Schuirmann 1987). 감지할 차이는 더 작은 차이가 실제로 동등한 것으로 간주되는 임계 차이를 지정합니다. 두 개의 귀무가설, 즉 실제 차이가 감지할 차이를 초과한다는 가설과 실제 차이가 감지할 차이의 음의 값보다 작다는 가설에 대해 단측 t -검정이 생성됩니다. 두 검정이 모두 기각되면 평균의 절대 차이가 감지할 차이 내에 있음을 나타냅니다. 따라서 그룹은 실제로 동등하다고 간주됩니다.

실제적 동등 p 값은 단측 t -검정에서 구한 가장 큰 p 값입니다. 실제적 동등 p 값이 작으면 실제로 수준의 평균 반응이 -수준의 평균과 같다는 것을 나타냅니다.

FDR 실제적 차이 p 값 쌍별 비교를 위한 실제적 차이 검정의 FDR p 값입니다. 자세한 내용은 "[반응 변수 선별 플랫폼에 대한 통계 상세 정보](#)"에서 확인하십시오.

FDR 실제적 동등 p 값 쌍별 비교를 위한 TOST 검정의 FDR p 값입니다. 자세한 내용은 "[반응 변수 선별 플랫폼에 대한 통계 상세 정보](#)"에서 확인하십시오.

실제적 결과 실제적 차이 및 동등성 검정 결과에 대한 설명입니다. 값은 유의한 결과를 식별하는데 도움이 되도록 색상으로 구분됩니다.

- 다름 (분홍색): 절대 차이가 실제적 차이보다 유의하게 크다는 것을 나타냅니다.

- 동등 (녹색): 절대 차이가 실제적 차이 내에 유의하게 포함된다는 것을 나타냅니다 .
- 판단할 수 없음 (회색): 실제적 차이 검정과 실제적 동등성 검정이 모두 유의하지 않다는 것을 나타냅니다 .

p 값 데이터 테이블

"반응 변수 선별" 시작 창에서 "시작 시 p 값 테이블" 옵션을 선택하면 p 값 데이터 테이블이 생성됩니다. 이 데이터 테이블에는 "반응 변수 선별" 보고서의 "결과 테이블"과 동일한 정보가 포함됩니다. p 값 테이블에는 다음과 같은 모형 관련 열이 더 포함됩니다.

Y 평균 Y의 평균입니다.

SSE Y가 연속형일 때 표시됩니다. 오차의 제곱합입니다.

DFE Y가 연속형일 때 표시됩니다. 오차의 자유도입니다.

MSE Y가 연속형일 때 표시됩니다. 평균 제곱 오차입니다.

F 비 Y가 연속형일 때 표시됩니다. 분산 분석 또는 회귀 검정의 F 비입니다.

절편 Y와 X가 모두 연속형일 때 표시됩니다. 해당하는 Y 및 X 변수 쌍과 관련된 회귀 모형의 절편입니다.

기울기 Y와 X가 모두 연속형일 때 표시됩니다. 해당하는 Y 및 X 변수 쌍과 관련된 회귀 모형의 기울기입니다.

DF Y와 X가 모두 범주형일 때 표시됩니다. 카이제곱 검정의 자유도입니다.

LR 카이제곱 Y와 X가 모두 범주형일 때 표시됩니다. 가능도비 카이제곱 통계량의 값입니다.

로버스트 경과 시간 (시작 창에서 " 로버스트 " 를 지정한 경우에만 표시됨) 로버스트 보고서를 생성하는 데 필요한 계산 시간 (초) 입니다.

p 값 데이터 테이블 스크립트

관련 스크립트가 반응 변수 선별 p 값 데이터 테이블에 저장됩니다. 이 중 하나를 제외한 모든 스크립트는 보고서에 제공된 그림을 재현합니다. p 값 테이블에서 행을 선택하는 경우 "선택한 항목 적합" 스크립트가 적절한 X로 Y 적합 분석을 생성합니다.

모형 적합의 반응 변수 선별 분석법

선형 모형 효과에 대한 단변량 검정에 관심이 있는 경우 모형 적합 플랫폼의 반응 변수 선별 분석법을 적합시킬 수 있습니다. 생성된 보고서 및 테이블은 모든 모형 효과에 대해 모든 반응을 검정합니다. 반응 변수 선별 분석법을 사용한 예는 "[반응 변수 선별 분석법의 예](#)"에서 확인하십시오.

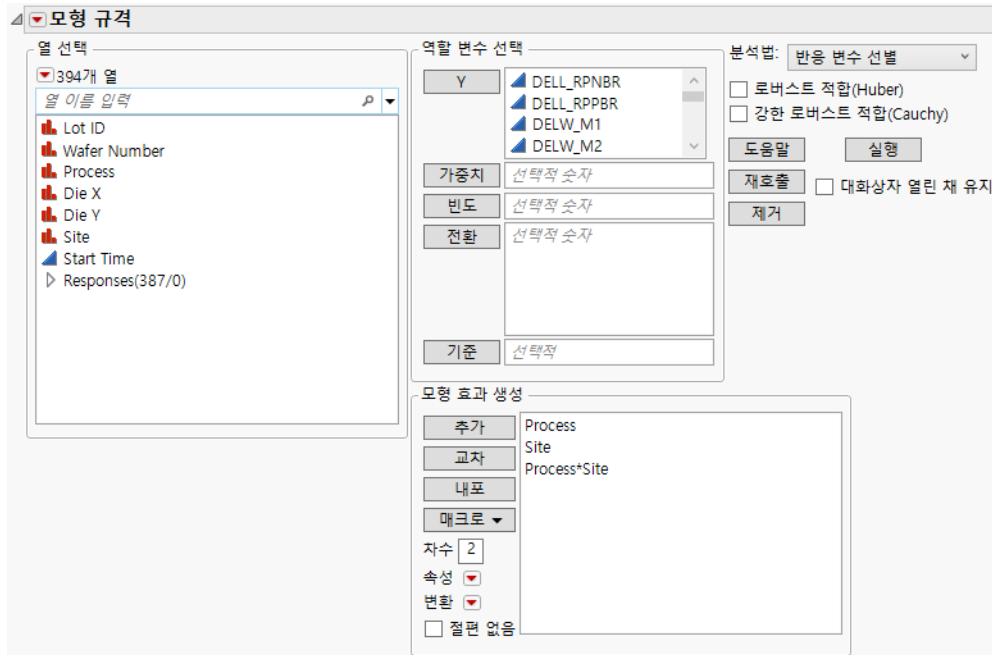
- "[모형 적합의 반응 변수 선별 시작](#)"
- "[반응 변수 선별 적합 보고서](#)"

- "반응 변수 선별 적합 옵션"

모형 적합의 반응 변수 선별 시작

분석 > 모형 적합을 선택하고 "분석법" 목록에서 **반응 변수 선별**을 선택하여 반응 변수 선별 적합을 시작합니다.

그림 24.6 모형 적합 창의 반응 변수 선별



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

다음 열 역할은 반응 변수 선별 분석법에만 있습니다.

전환 한 번에 하나씩 모형으로 전환할 열을 지정합니다.

로버스트 추정을 위한 다음 두 가지 옵션이 있습니다.

로버스트 적합 (Huber) 연속형 반응의 이상치에 대한 비중 강하 (down weight) 를 위해 로버스트 (Huber) 추정을 지정합니다. 이상치가 없으면 이러한 추정값이 최소 제곱 추정값에 가깝습니다. 이 옵션을 사용하면 처리 시간이 늘어납니다.

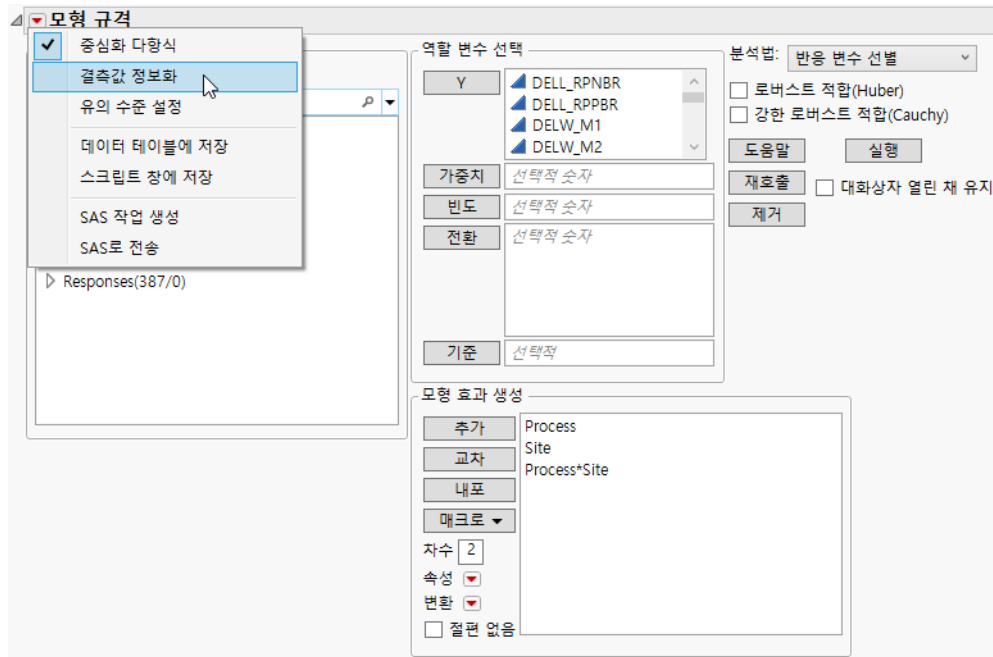
강한 로버스트 적합 (Cauchy) 연속형 반응의 이상치에 대한 비중 강하 (down weight) 를 위해 강한 로버스트 (Cauchy) 추정을 지정합니다. 오차가 Cauchy 분포를 따른다고 가정합니다. Cauchy 분포는 정규 분포보다 꼬리가 더 두꺼우므로 이상치의 영향이 줄어듭니다. 이 옵션은 데이터에 포함된 이상치의 비율이 높은 경우에 유용할 수 있습니다. 하지만 데이터가 정

규 분포에 가깝고 이상치가 일부만 있는 경우에 이 옵션을 사용하면 추론이 잘못될 수 있습니다. Cauchy 옵션은 최대 가능도와 Cauchy 연결 함수를 사용하여 모수를 추정합니다.

팁 : 두 로버스트 옵션을 모두 선택하면 플랫폼에서 Cauchy 추정만 사용합니다.

결측값 정보화 옵션은 결측값에 대한 코딩 시스템을 제공합니다. 결측값 정보화 코딩을 사용하면 결측값이 있어도 예측 모형을 추정할 수 있습니다. 이 옵션은 결측 데이터에 정보가 포함된 경우 유용합니다. "모형 규격"의 빨간색 삼각형 메뉴에서 이 옵션을 선택합니다.

그림 24.7 결측값 정보화 옵션



모형 적합 창에 대한 자세한 내용은 선형 모형 적합에서 확인하십시오.

반응 변수 선별 적합 보고서

모형 적합 플랫폼을 통해 반응 변수 선별을 시작하면 "반응 변수 선별 적합" 보고서에 "효과 검정" 테이블과 두 개의 그림이 포함됩니다. 표시된 그림은 "효과에 대한 FDR p 값 그림"과 "FDR LogWorth 대 효과 크기" 그림입니다. 두 그림 모두 플랫폼 자체와 동일한 방식으로 해석됩니다. 자세한 내용은 "반응 변수 선별 그림"에서 확인하십시오.

효과 검정 테이블

"효과 검정" 테이블에는 Y 변수와 모형 효과로 구성된 각 쌍에 대한 행이 포함되어 있습니다. 시작 창에서 "로버스트 적합" 옵션을 선택하면 Huber M-추정을 사용하여 모형이 적합됩니다. 기

준 변수를 지정한 경우 기준 변수의 각 수준에 대해 "효과 검정" 테이블이 생성됩니다. 테이블에는 다음 열이 포함됩니다.

Y 지정된 반응 열입니다.

전환 (시작할 때 "전환" 열을 지정한 경우 사용 가능) 모형에 포함되는 열을 지정합니다. 이는 행이 설명하는 모형을 식별합니다. "효과" 열에 "전환"이 포함된 경우 "전환" 열은 효과를 지정합니다.

팁 : "전환" 열 머리글을 클릭하면 "효과 검정" 테이블이 "전환" 열을 기준으로 정렬됩니다. 이렇게 하면 검정되는 각 모형 내의 효과를 더 명확하게 확인할 수 있습니다. 여러 Y의 경우 "전환"을 정렬한 후 "Y"를 정렬하여 각 반응에 대한 모형 내 효과를 정렬합니다.

효과 지정된 모형 효과입니다.

F 비 "효과" 검정에 대한 검정 통계량입니다. 이 값은 최소 제곱 적합의 "효과 검정" 보고서에 나오는 값과 같습니다.

카이제곱 (시작할 때 로버스트 적합을 지정한 경우에만 사용 가능) "효과" 검정에 대한 검정 통계량입니다.

p 값 "F 비"에 해당하는 유의성 검정의 p 값입니다. 효과 검정에 대한 자세한 내용은 선형 모형 적합에서 확인하십시오.

LogWorth $-\log_{10}(p \text{ 값})$ 의 계산 값입니다. 이 변환은 그래프 작성에 적절한 척도를 제공하도록 p 값을 조정합니다. $-\log_{10}(0.01) = 2$ 이므로 2를 초과하는 값은 0.01 수준에서 유의합니다.

FDR p 값 Benjamini-Hochberg 기법을 사용하여 계산된 False Discovery Rate p 값입니다. 이 기법은 여러 검정에 대한 FDR(False Discovery Rate)을 제어하기 위해 p 값을 조정합니다. FDR 수정에 대한 자세한 내용은 Benjamini and Hochberg 연구 자료(1995)에서 확인하십시오. False Discovery Rate에 대한 자세한 내용은 "반응 변수 선별 플랫폼에 대한 통계 상세 정보" 또는 Westfall et al. 연구 자료(2011)에서 확인하십시오.

FDR LogWorth $-\log_{10}(\text{FDR } p \text{ 값})$ 의 계산 값입니다. 이 통계량은 유의성을 그림에 표시하고 평가하는 데 가장 적합합니다. p 값이 작으면 FDR LogWorth 값이 커집니다.

순위 비율 (기본적으로 표시되지 않음) FDR LogWorth 순위를 검정 수의 비율로 나타낸 것입니다. 검정 수가 m 이면 가장 큰 FDR LogWorth 값의 순위 비율은 $1/m$ 이고 가장 작은 값의 순위 비율은 1입니다. 마찬가지로 "순위 비율"은 p 값을 검정 수의 비율로 작은 것부터 순서대로 정렬합니다. 순위 비율은 p 값과 FDR p 값을 순위별로 유의성이 큰 것부터 그림에 표시하는 데 사용됩니다.

효과 크기 X의 수준 또는 값에 따라 반응 값이 달라지는 정도를 나타냅니다. 효과 크기는 척도 불변입니다.

- Y가 연속형인 경우 효과 크기는 가설 검정에서 구한 평균 제곱합의 제곱근을 반응 표준편차의 로버스트 추정값으로 나눈 값입니다. IQR(사분위수 범위)이 0이 아니고 $\text{IQR} > \text{범위}/20$ 이면 표준편차 추정값은 $\text{IQR}/1.3489795$ 입니다. 그렇지 않으면 표본 표준편차가 사용됩니다.

- Y가 범주형이고 X가 연속형인 경우 효과 크기는 전체 모형 검정에 대한 평균 카이제곱 값의 제곱근입니다.
- Y와 X가 모두 범주형인 경우 효과 크기는 평균 Pearson 카이제곱의 제곱근입니다.
- 시작할 때 로버스트 적합을 지정한 경우 효과 크기는 (카이제곱 검정 통계량 / m)의 제곱근으로 계산됩니다. 여기서 n 은 관측값 수입니다.

검정 DF 효과 검정에 대한 자유도입니다.

반응 변수 선별 적합 옵션

"반응 변수 선별 적합"의 빨간색 삼각형 메뉴에는 계산된 데이터를 저장하는 옵션이 포함되어 있습니다.

효과 검정 효과 검정 테이블을 표시하거나 숨깁니다. 자세한 내용은 "**효과 검정 테이블**"에서 확인하십시오.

효과 그림 "효과에 대한 FDR p 값 그림"과 "FDR LogWorth 대 효과 크기" 그림을 표시하거나 숨깁니다.

전체 보고서 각 Y 변수에 대한 행이 포함된 "전체 적합" 테이블을 표시하거나 숨깁니다. 테이블의 열에는 각 Y에 대해 모형 적합에 대한 정보가 요약되어 있습니다. 시작 창에서 "로버스트 적합" 옵션을 선택하면 Huber M-추정을 사용하여 모형이 적합됩니다. 이 테이블에는 다음 열이 포함됩니다.

Y 지정된 반응 열입니다.

R² 다중 상관계수입니다.

RMSE 제곱근 평균 제곱 오차입니다.

개수 관측값 수 또는 가중치 변수의 합입니다.

전체 F 비 최소 제곱 적합의 "분산 분석" 보고서에 나오는 모형 적합에 대한 검정 통계량입니다.

전체 p 값 전체 모형 유의성 검정에 대한 p 값입니다.

전체 LogWorth 전체 모형 유의성 검정에 대한 p 값의 LogWorth입니다.

전체 FDR p 값 False Discover Rate에 대해 조정된 전체 p 값입니다. 자세한 내용은 "**반응 변수 선별 보고서**"에서 확인하십시오.

전체 FDR LogWorth 전체 FDR p 값의 LogWorth입니다.

전체 순위 비율 (기본적으로 표시되지 않음) 전체 FDR LogWorth 순위를 검정 수의 비율로 나타낸 것입니다. 검정 수가 m 이면 가장 큰 전체 FDR LogWorth 값의 순위 비율은 $1/m$ 이고 가장 작은 값의 순위 비율은 1입니다.

전체 그림 "전체 FDR p 값 그림"과 "FDR LogWorth 대 R² 그림"을 표시하거나 숨깁니다.

효과 선택 조건 " 선택 조건 " 창을 엽니다. " 비교 " 메뉴와 " 값 " 텍스트 상자를 사용하여 효과 검정 테이블에서 특정 조건에 해당하는 특정 반응을 선택할 수 있습니다. 예를 들어 " 효과 크기 > 0.80 " 에 해당하는 모든 효과를 선택할 수 있습니다. " 확인 " 을 클릭하면 " 결과 테이블 " 에서 반응이 선택됩니다.

팁 : " 효과 검정 " 테이블의 아무 곳이나 마우스 오른쪽 버튼으로 클릭하여 " 선택 조건 " 창에 액세스할 수도 있습니다.

선택한 효과에 대한 반응 선택 " 효과 검정 " 테이블에서 선택한 효과에 해당하는 반응 열을 원래 데이터 테이블에서 선택합니다.

효과 검정 저장 " 효과 검정 " 테이블을 새 데이터 테이블에 저장합니다. " 효과 검정 " 데이터 테이블에는 분석에 사용된 데이터 테이블의 이름을 제공하는 " 원래 데이터 " 라는 테이블 변수가 포함되어 있습니다. 기준 변수를 지정한 경우 " 원래 데이터 " 변수에 해당 기준 변수와 수준도 표시됩니다.

전체 적합 저장 " 전체 적합 " 테이블을 새 데이터 테이블에 저장합니다. " 전체 적합 " 데이터 테이블에는 분석에 사용된 데이터 테이블의 이름을 제공하는 " 원래 데이터 " 라는 테이블 변수가 포함되어 있습니다. 기준 변수를 지정한 경우 " 원래 데이터 " 변수에 해당 기준 변수와 수준도 표시됩니다.

추정값 저장 각 행이 반응에 해당하고 열이 모형 항에 해당하는 데이터 테이블을 엽니다. 항목은 지정된 모형을 적합시켜 구한 모수 추정값입니다. 이 데이터 테이블에는 분석에 사용된 데이터 테이블 이름을 제공하는 " 원래 데이터 " 라는 테이블 변수도 포함되어 있습니다. 기준 변수를 지정한 경우 JMP 에서 기준 변수의 각 수준에 대한 추정값 테이블이 생성되고 " 원래 데이터 " 변수는 기준 변수와 해당 수준을 제공합니다.

예측 계산식 저장 모든 반응에 대한 예측 방정식이 포함된 열을 원래 데이터 테이블에 추가합니다.

최소 제곱 평균 저장 각 행이 반응 및 효과 설정 조합에 해당하는 데이터 테이블을 엽니다. 이 행은 해당 설정 조합의 표준 오차 및 최소 제곱 평균을 포함합니다.

모형 대화상자 제공된 보고서를 얻기 위해 실행한 모형 대화상자가 포함된 창을 엽니다.

다음 옵션에 대한 자세한 내용은 **JMP 사용** 의 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.

반응 변수 선별의 추가 예

이 섹션에는 반응 변수 선별 플랫폼을 사용하는 예가 포함되어 있습니다.

- "실제적 유의성 및 동등성 검정의 예"
- "최대 LogWorth 옵션의 예"
- "로버스트 적합의 예"
- "반응 변수 선별 분석법의 예"
- "화산 (Volcano) 그림의 예"

실제적 유의성 및 동등성 검정의 예

반응 변수 선별 플랫폼을 사용하여 실제적 차이를 검정합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Probe.jmp 를 엽니다.
2. **분석 > 선별 > 반응 변수 선별**을 선택합니다.
3. Responses 열 그룹을 선택하고 **Y, 반응**을 클릭합니다.
4. Process 를 선택하고 **X**를 클릭합니다.
5. **확인**을 클릭합니다.
6. "반응 변수 선별"의 빨간색 삼각형을 클릭하고 **실제적 차이 및 동등성**을 선택합니다.
7. "6* 시그마 또는 규격 범위의 비율 입력" 옆의 상자에 "0.15"를 입력합니다.
8. **확인**을 클릭합니다.
9. "반응 변수 선별"의 빨간색 삼각형을 클릭하고 **테이블 저장 > 평균 차이 저장**을 선택합니다.

그림 24.8에는 데이터 테이블의 일부가 표시됩니다. Y의 각 반응에 대해 해당 행은 Process의 New 수준과 Old 수준에 대한 검정 정보를 제공합니다.

그림 24.8 평균 차이 테이블 (일부분)

	Y	FDR p 값	FDR LogWorth	감지할 차이	실제적 차이 p 값	실제적 동등 p 값	FDR 실제적 차이 p 값	FDR 실제적 동등 p 값	실제적 결과
1	DELL_RPNBR	2.39467e-12	11.620754237	0.1486469486	1	1.57079e-153	1	2.37474e-153	동등
2	DELL_RPPBR	0.5449559037	0.2636386382	1.1939335307	1	2.40983e-225	1	1.77957e-224	동등
3	DELW_M1	0.2459236742	0.6091996613	0.7030518329	1	1.40119e-217	1	5.12434e-217	동등
4	DELW_M2	0	325.14036157	0.0704822866	0.0269479498	0.9730520502	0.27967602	1	판단할 수 없음
5	DELW_NBASE	0.4570716144	0.3400157489	50.689291592	1	6.62236e-223	1	3.48355e-222	동등
6	DELW_NEMIT	4.065967e-55	54.390836173	0.0848381437	1	7.74847e-69	1	9.445754e-69	동등
7	DELW_NENBNI	0	342.83731846	0.864024229	0.0003726256	0.9996273744	0.0039746735	1	다름
8	DELW_NSINK	0	528.07666335	1.0723100972	3.228438e-37	1	3.999097e-36	1	다름
9	DELW_PBASE	1.227409e-30	29.911010679	0.1296380428	1	8.21675e-113	1	1.10711e-112	동등
10	DELW_PCOLL	0	2733.3647756	1.17539613	0	1	0	1	다름
11	DELW_PEMIT	0.7257260746	0.1392272729	1.8265075627	1	1.2499e-228	1	1.59987e-227	동등
12	DELW_PSINK	7.709736e-40	39.112960512	2.3558597461	1	3.147655e-97	1	4.153607e-97	동등
13	DELW_RPNBR	3.391334e-23	22.469629432	0.3464883523	1	1.6441e-125	1	2.28744e-125	동등
14	DELW_RPPBR	4.583412e-26	25.338811147	0.9698077792	1	7.48088e-120	1	1.02595e-119	동등
15	DELW_SICR	0.1667193458	0.7780140026	0.6024110643	1	7.77189e-215	1	2.57276e-214	동등
16	M1_COMB_VGAT...	0.4548355067	0.3421456395	20.662273741	1	1.83138e-223	1	1.09883e-222	동등
17	M1_TRENCH_VG...	0.0005193703	3.2845228853	5.8087989555	1	2.39165e-191	1	5.18867e-191	동등
18	M2/M1_CAP_VG...	0.6866587979	0.1632590111	3.3496974301	1	2.54456e-228	1	3.05347e-227	동등
19	M2_COMB_BB_V...	0.999486039	0.0002232678	4.3470387459	1	3.07472e-234	1	1.50782e-232	동등
20	M2_COMB_VGAT...	0.3827958663	0.4170327607	9.7914324315	1	1.42412e-221	1	6.35887e-221	동등
21	NISO_TUB-TREN...	1.38575e-163	162.85831566	9.0621071628	1	9.077234e-17	1	1.037398e-16	동등
22	NISO_TUB-TUB_...	0.5509658664	0.2588753059	3.7013990674	1	8.97914e-226	1	7.6622e-225	동등
23	PS_RPNBR	0.0000493267	4.3069182395	530.27422027	1	3.12381e-189	1	6.44916e-189	동등

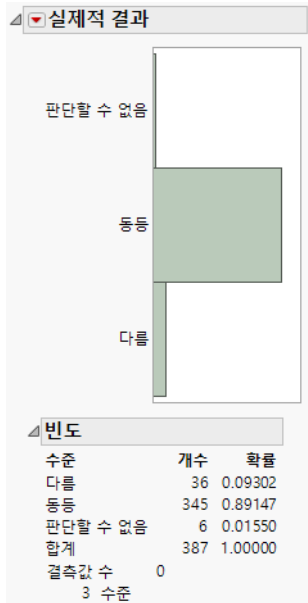
Probe.jmp 에 규격 한계가 열 특성으로 저장되지 않았으므로 JMP 에서 각 반응의 실제적 차이에 대한 값이 계산됩니다. 사용자가 지정한 실제적 차이 비율 0.15 에 반응의 6σ 범위 추정 값을 곱합니다. 이 값을 사용하여 실제적 차이 및 동등성을 검정합니다. 감지할 차이 열에 값이 표시됩니다.

FDR p 값 열에는 p 값이 유의성을 나타내는 반응이 표시됩니다. FDR 실제적 차이 p 값 및 FDR 실제적 동등 p 값 열은 실제적 차이와 실제적 동등성 검정에 대한 p 값을 제공합니다. 많은 열이 통계적으로 유의한 차이를 보이지만 실제적으로 유의한 차이는 보이지 않습니다.

10. "평균 차이" 데이터 테이블을 표시하고 분석 > 분포를 선택합니다.
11. 실제적 결과를 선택하고 Y, 열을 클릭합니다.
12. 확인을 클릭합니다.

그림 24.9 에서는 실제적 유의성 결과의 분포를 보여 줍니다. 지정된 실제적 차이에 대한 검정에 따라 36 개의 검정만 다릅니다. 6 개의 반응은 검정으로 판단할 수 없었습니다. 즉, 반응이 Process 간에 실제적 차이를 나타내는지 여부를 구별할 수 없습니다.

그림 24.9 실제적 유의성 결과의 분포



그림에서 해당 막대를 클릭하여 추가 연구에 사용할 36개 반응을 선택할 수 있습니다.

최대 LogWorth 옵션의 예

이 예에서는 반응 변수 선별 플랫폼의 "최대 LogWorth" 옵션을 사용하여 LogWorth 척도가 매우 큰 값에 의해 왜곡되지 않도록 제어합니다. 데이터 집합에 관측값이 많은 경우 p 값이 매우 작을 수 있습니다. LogWorth 값은 이러한 경우 p 값을 그래픽으로 연구하는 유용한 방법을 제공합니다. 일부 데이터 집합에서는 p 값이 너무 작아 LogWorth 척도가 매우 큰 값에 의해 왜곡됩니다.

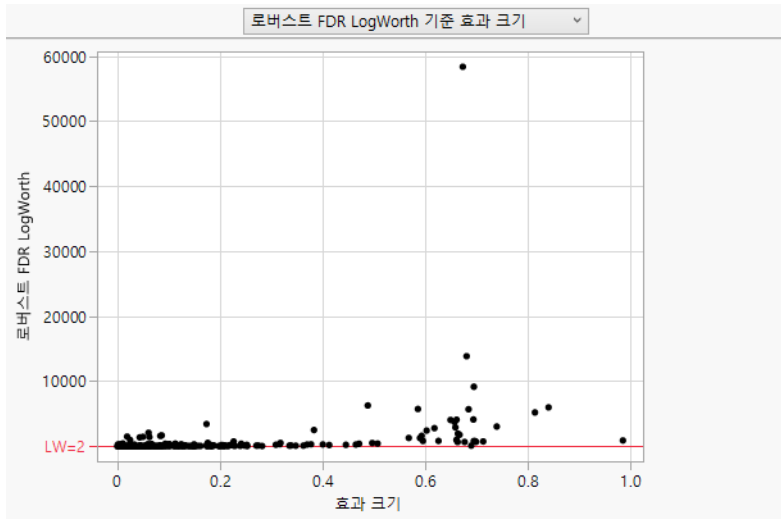
1. **도움말 > 샘플 데이터 폴더**를 선택하고 Probe.jmp 를 엽니다.
2. **분석 > 선별 > 반응 변수 선별**을 선택합니다.
3. "반응 변수 선별" 시작 창에서 **Responses** 열 그룹을 선택하고 **Y, 반응**을 클릭합니다.
4. **Process** 를 선택하고 **X** 를 클릭합니다.
5. "고급 옵션" 옆의 회색 삼각형을 클릭합니다.
6. **로버스트** 체크박스를 선택합니다.
7. **확인**을 클릭합니다.

수치가 많아서 분석을 완료하는 데 시간이 걸릴 수 있습니다.

8. "반응 변수 선별"의 그림 제목 목록을 클릭하고 **로버스트 FDR LogWorth 대 효과 크기**를 선택합니다.

"로버스트 FDR LogWorth" 값 중 하나가 너무 커서 (약 58,000) 그림의 상세 정보를 보기가 어렵습니다. 그래프에 상세 정보가 충분히 표시되게 하려면 LogWorth의 최대값을 설정합니다.

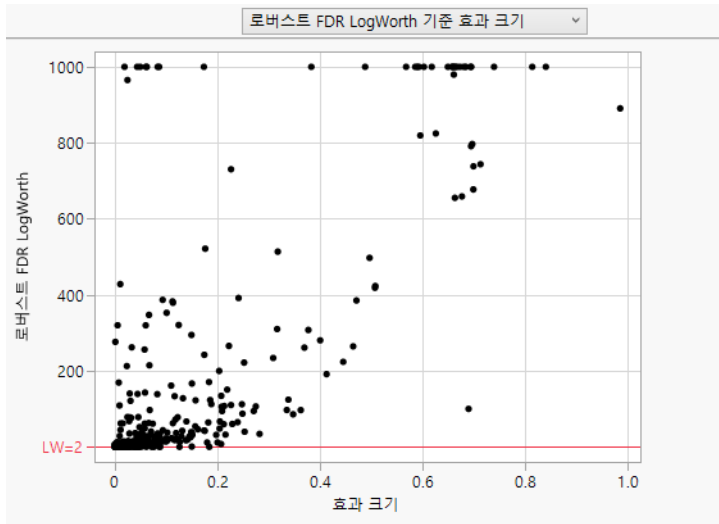
그림 24.10 로버스트 FDR LogWorth 대 효과 크기 (최대 LogWorth 설정 안 함)



9. "반응 변수 선별"의 빨간색 삼각형을 클릭하고 **다시 실행 > 분석 다시 시작**을 선택합니다.
10. "고급 옵션" 옆의 회색 삼각형을 클릭합니다.
11. "최대 LogWorth" 옆의 상자에 "1000"을 입력합니다.
12. **확인**을 클릭합니다.
분석을 완료하는 데 시간이 걸릴 수 있습니다.
13. "반응 변수 선별"의 그림 제목 목록을 클릭하고 **로버스트 FDR LogWorth 대 효과 크기**를 선택합니다.

이제 그림의 상세 정보가 분명하게 보입니다. 1000보다 큰 로버스트 FDR LogWorth 값은 1000으로 표시됩니다.

그림 24.11 로버스트 FDR LogWorth 대 효과 크기 (최대 LogWorth = 1000)

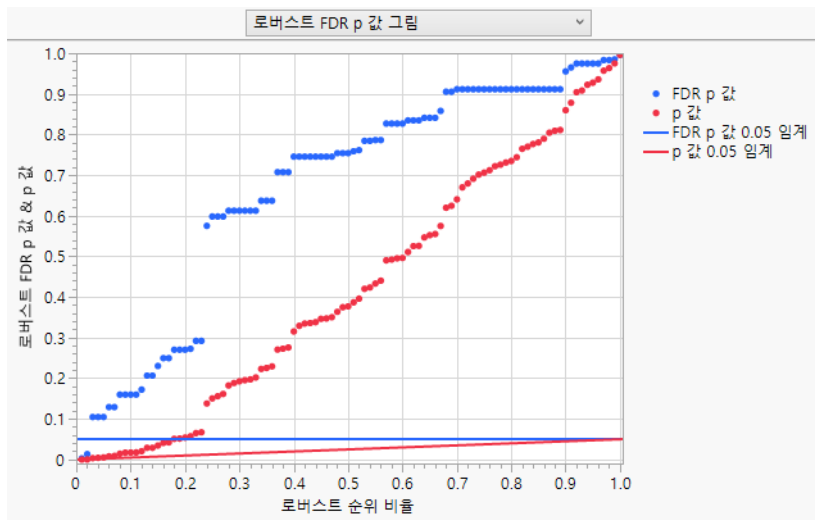


로버스트 적합의 예

반응 변수 선별 플랫폼의 "로버스트" 옵션을 사용하여 이상치가 많은 반응을 분석합니다. 이상치는 기존 검정의 오차 분산을 부풀려 유의한 효과를 알아보기 어려울 수 있으므로 로버스트 적합을 사용해야 합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Drosophila Aging.jmp 를 엽니다 .
2. **분석 > 선별 > 반응 변수 선별**을 선택합니다 .
3. 연속형 열을 모두 선택하고 **Y, 반응**을 클릭합니다 .
4. **line** 을 선택하고 **X** 를 클릭합니다 .
5. " 고급 옵션 " 옆의 회색 삼각형을 클릭합니다 .
6. **로버스트**를 선택합니다 .
7. **확인**을 클릭합니다 .

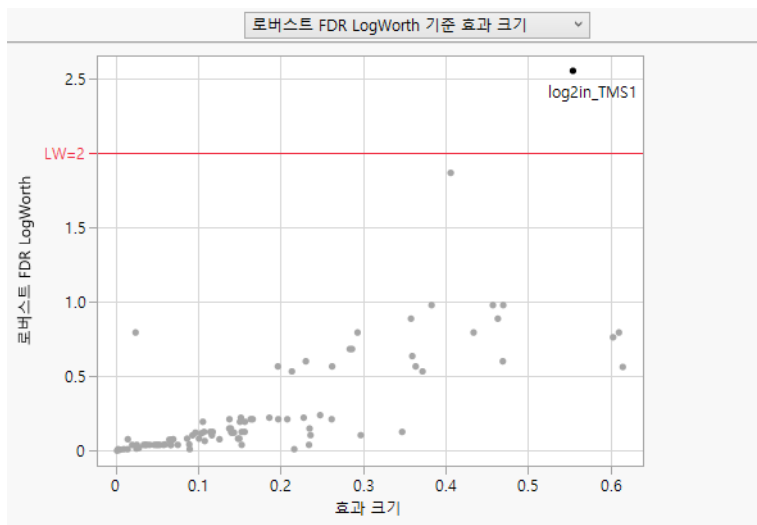
그림 24.12 Drosophila 데이터에 대한 로버스트 FDR p 값 그림



0.05 미만의 빨간색 점으로 표시된 것처럼 조정되지 않은 로버스트 p 값을 사용하여 많은 검정이 유의합니다. 그러나 로버스트 FDR p 값에 따르면 두 가지 검정만 유의합니다. 이 두 점은 FDR LogWorth 값을 보여 주는 그림에서 더 쉽게 식별됩니다.

8. "반응 변수 선별"의 그림 제목 목록을 클릭하고 **로버스트 FDR LogWorth 대 효과 크기**를 선택합니다.
9. 로버스트 FDR LogWorth 값이 2를 초과하는 점을 클릭하고 마우스 오른쪽 버튼을 클릭한 후 **행 라벨**을 선택합니다.

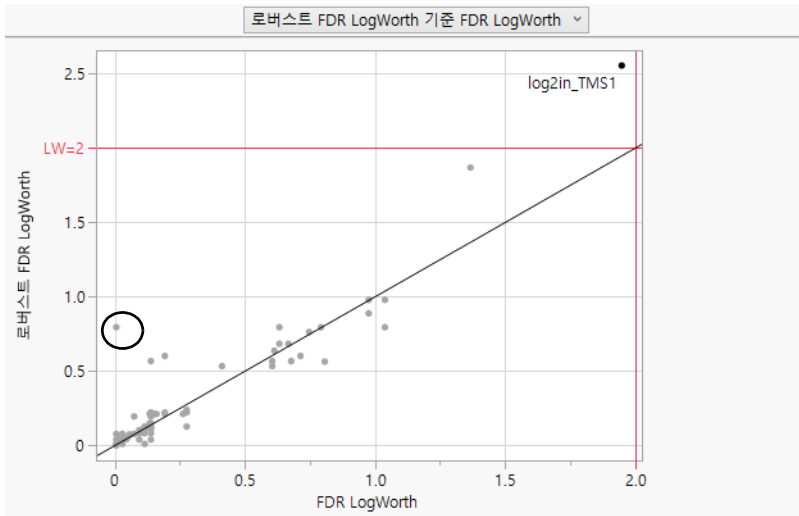
그림 24.13 Drosophila 데이터에 대한 로버스트 LogWorth 대 효과 크기



빨간색 선 (2) 위의 점은 유의 수준이 $0.01(-\log_{10}(0.01) = 2)$ 미만입니다. 가로선 (약 1.3) 은 0.05 유의 수준에 해당합니다.

10. "반응 변수 선별" 의 그림 제목 목록을 클릭하고 **로버스트 FDR LogWorth 대 FDR LogWorth** 를 선택합니다.

그림 24.14 Drosophila 데이터에 대한 로버스트 FDR LogWorth 대 FDR LogWorth



반응에 대한 로버스트 검정이 일반 검정과 동일할 경우 해당 점은 **그림 24.14** 에서와 같이 대각선 위에 분포합니다. 그림에서 원 안의 점은 로버스트 LogWorth 값이 LogWorth 값을 초과하므로 선에서 떨어져 있습니다.

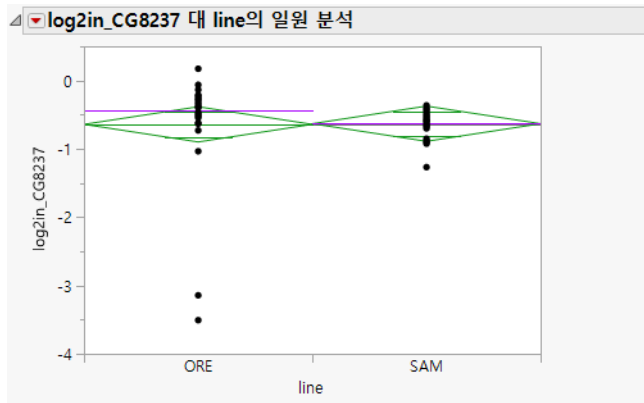
11. 그림에서 이 점을 클릭하여 선택합니다.
12. "결과 테이블" 에서 이 점에 해당하는 행을 찾습니다.

반응 `log2in_CG8237` 의 p 값은 0.9568 이고 로버스트 p 값은 0.0176 입니다.

13. "반응 변수 선별" 의 빨간색 삼각형을 클릭하고 **선택한 항목 적합** 을 선택합니다.

반응 `log2in_CG8237` 에 대한 일원 분석이 포함된 "선택한 항목 적합" 보고서가 표시됩니다. 그림에 ORE line 에 대한 두 개의 이상치가 표시됩니다 (**그림 24.15**). 이러한 이상치는 로버스트 검정과 일반 검정의 결과가 서로 다른 이유를 나타냅니다. 이상치는 비로버스트 검정의 오차 분산을 부풀리므로 유의한 효과를 알아보기가 더 어렵습니다. 반면, 로버스트 적합은 이러한 이상치의 비중을 낮추어 오차 분산에 대한 기여도를 줄입니다.

그림 24.15 log2in_CG8237 에 대한 일원 분석

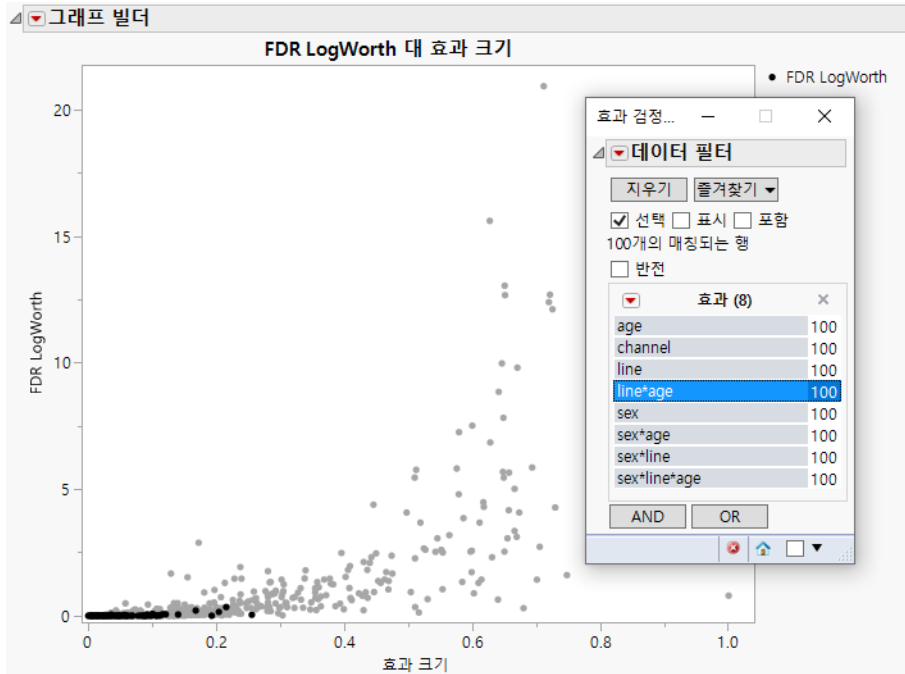


반응 변수 선별 분석법의 예

이 예에서는 모형 적합의 반응 변수 선별 분석법을 사용하여 선형 모형 효과에 대한 다중 반응 검정을 연구합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Drosophila Aging.jmp** 를 엽니다.
2. **분석 > 모형 적합**을 선택합니다.
3. 연속형 열을 모두 선택하고 **Y**를 클릭합니다.
4. **channel** 을 선택하고 **추가**를 클릭합니다.
5. **line, sex** 및 **age** 를 선택하고 **매크로 > 완전 요인**을 선택합니다.
6. "분석법" 목록에서 **반응 변수 선별**을 선택합니다.
7. **실행**을 클릭합니다.
8. "반응 변수 선별 적합"의 빨간색 삼각형을 클릭하고 "효과 검정 저장"을 선택합니다.
9. "효과 검정" 데이터 테이블에서 **FDR LogWorth** 대 **효과 크기** 스크립트를 실행합니다.
10. **행 > 데이터 필터**를 선택합니다.
11. "데이터 필터" 창에서 **효과**를 선택하고 **+**을 클릭합니다.
12. "데이터 필터"에서 각 모형 효과를 순차적으로 클릭하면서 "FDR LogWorth 대 효과 크기"에서 선택된 점을 확인합니다.

그림 24.16 line*age 검정이 선택된 FDR LogWorth 대 효과 크기 그림



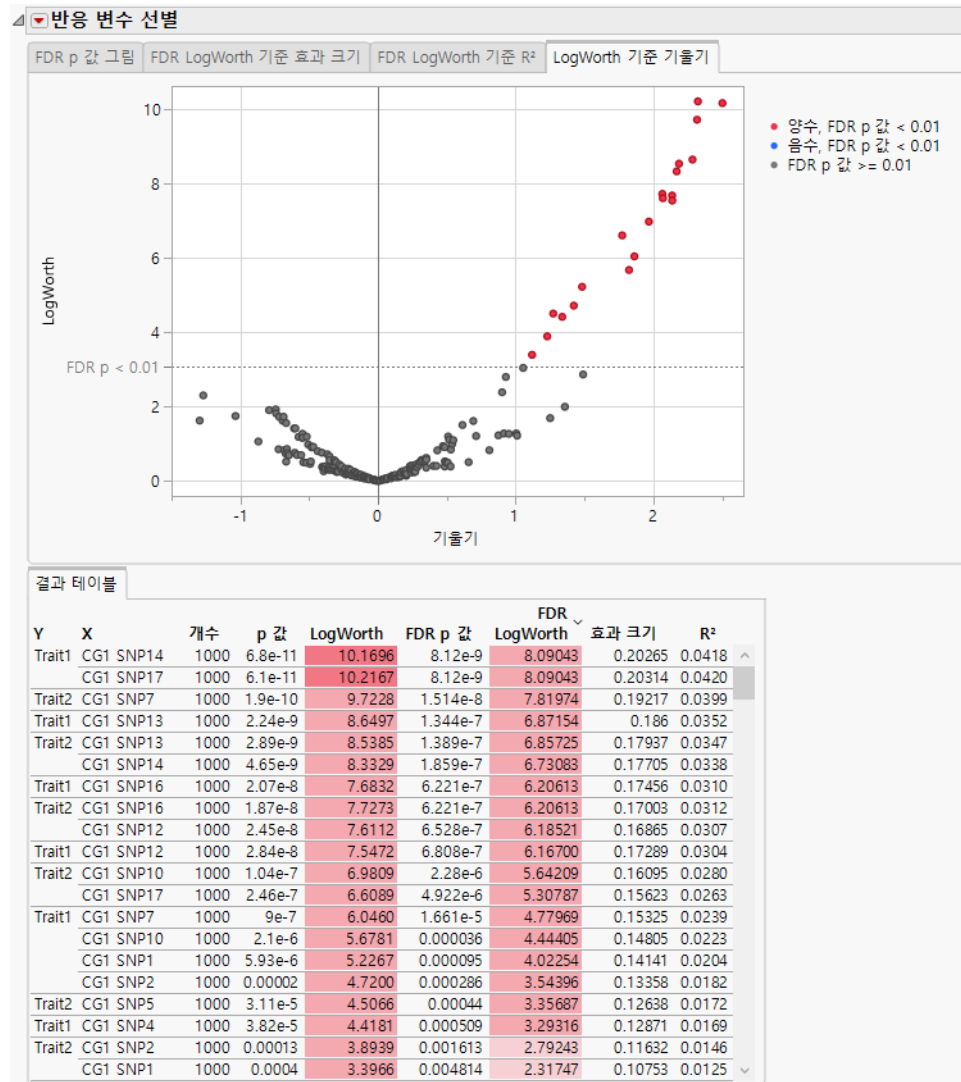
2 를 초과하는 LogWorth 값은 0.01 수준에서 유의합니다 . 데이터 필터를 사용하면 sex 와 channel 을 제외하고 모형 효과가 0.01 수준에서 거의 유의하지 않음을 알 수 있습니다 . line*age 교호작용 효과 검정을 위한 점이 선택되어 있습니다 . 이러한 점은 0.01 수준에서 모두 유의하지 않습니다 .

화산 (Volcano) 그림의 예

반응 변수 선별 플랫폼을 사용하여 네 가지 특성과 연관된 SNP를 확인합니다. SNP는 유전적 변이의 한 유형입니다. 화산(Volcano) 그림은 유형이 동일한 많은 가설 검정의 결과를 시각화하는데 유용합니다. 이 경우 모든 반응과 변수가 연속형이므로 모든 검정이 단순 선형 회귀를 통해 수행됩니다.

1. **도움말 > 샘플 데이터 폴더 > Life Sciences** 를 선택하고 Genotypes Pedigree.jmp 를 엽니다 .
2. **분석 > 선별 > 반응 변수 선별** 을 선택합니다 .
3. Trait1 ~ Trait4 를 선택하고 **Y, 반응** 을 클릭합니다 .
4. Markers 열 그룹을 선택하고 **X** 를 클릭합니다 .
5. **공통 Y 척도** 및 **공통 X 척도** 를 선택합니다 .
6. **확인** 을 클릭합니다 .
7. "LogWorth 대 기울기" 탭을 선택합니다 .

그림 24.17 LogWorth 대 기울기 그림



"LogWorth 대 기울기" 그림(화산 그림이라고도 함)에서 색상이 적용된 점은 False Discovery Rate를 제어하는 동시에 0.01 수준에서 유의한 검정을 나타냅니다. 이 예에서는 그림에 빨간색 점만 있습니다. 이는 단순 선형 회귀에서 유의한 모든 X 변수의 기울기가 양수임을 의미합니다. 따라서 특성과 유의하게 연관된 모든 SNP의 경우 SNP 값이 클수록 Trait 값이 더 큼니다. "결과 테이블"에는 개별 검정의 결과가 표시되며 가장 유의한 검정이 맨 위에 나옵니다. 모든 상위 검정에는 Trait 1과 Trait 2가 포함됩니다. 즉, 이 SNP 집합은 다른 두 특성보다 Trait 1 및 Trait 2와 더 높은 연관성이 있다는 결론을 내릴 수 있습니다.

반응 변수 선별 플랫폼에 대한 통계 상세 정보

모든 반응 변수 선별 그림에는 Benjamini and Hochberg(1995)에 설명된 False Discovery Rate 기법을 사용하여 수행되는 검정의 p 값이 포함됩니다. Westfall et al. (2011)의 내용도 참조하십시오. 이 방법에서는 p 값이 독립적이고 균등 분포를 따른다고 가정합니다.

반응 변수 선별 플랫폼에서는 다음 절차를 사용하여 α 수준에서 False Discovery Rate 를 제어합니다.

1. 관심 있는 가설 검정 m 개를 수행하여 p 값 ($p_1, p_2, \dots, p_m$) 을 구합니다.
2. p 값의 순위를 가장 작은 것부터 순서대로 정렬합니다. 이러한 순위화를 $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$ 으로 나타냅니다.
3. $p_{(i)} \leq (i/m)\alpha$ 에 대해 가장 큰 p 값을 찾습니다. 이 첫 번째 p 값을 k 번째로 큰 $p_{(k)}$ 라고 가정합니다.
4. $p_{(k)}$ 보다 작거나 같은 p 값과 관련된 k 개의 가설을 기각합니다.

이 절차를 사용하면 False Discovery Rate가 α 를 초과하지 않습니다.

False Discovery Rate 에 대해 조정된 p 값은 $p_{(i), FDR}$ 로 표시하고 다음과 같이 계산됩니다.

$$p_{(i), FDR} = \begin{cases} p_{(m)} & \text{for } i = m \\ \min\left[p_{(i+1), FDR}, \left(\frac{m}{i}\right)p_{(i)}\right] & \text{for } i = m-1, \dots, 1 \end{cases}$$

가설의 FDR 조정 p 값이 α 미만이면 절차에 의해 가설이 기각됩니다.

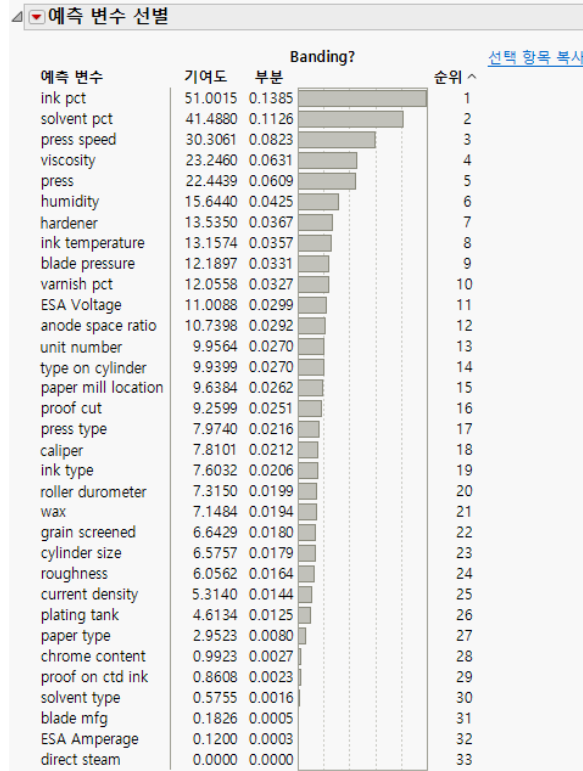
예측 변수 선별

유의한 효과를 위한 여러 예측 변수 선별

예측 변수 선별 플랫폼에서는 결과 예측 능력을 위해 많은 예측 변수를 선별하는 방법을 제공합니다. 이는 부품, 공정 또는 표본에 대해 수백 또는 수천 번의 측정을 수행하는 대규모 데이터 집합 분석에 유용하며, 혁신적인 접근 방식이 필요합니다. 예를 들어 예측 변수 선별을 사용하면 상태 예측을 위한 조건이 있는 환자와 없는 환자의 표본에서 테스트된 수천 개의 생체 지표를 식별할 수 있습니다.

예측 변수 선별은 반응 변수 선별과 다릅니다. 반응 변수 선별은 요인을 한 번에 하나씩 반응의 예측 변수로 검정합니다. 예측 변수 선별은 붓스트랩 포레스트 분할을 사용하여 반응에 대한 예측 변수의 기여도를 평가합니다. 파티션 모형은 여러 예측 변수를 기반으로 합니다. 예측 변수 선별은 단독으로는 약하지만 다른 예측 변수와 함께 사용하면 강력한 예측 변수를 식별할 수 있습니다. 자세한 내용은 "[반응 변수 선별](#)"에서 확인하십시오.

그림 25.1 예측 변수 선별 보고서의 예



목차

예측 변수 선별 플랫폼 개요 485

예측 변수 선별의 예 485

예측 변수 선별 플랫폼 시작 487

예측 변수 선별 보고서 488

예측 변수 선별 플랫폼 옵션 488

예측 변수 선별 플랫폼 개요

예측 변수 선별은 다수의 후보에서 유의한 예측 변수를 식별하는 데 유용합니다. 수백 개의 X 가 있고 이 중 결과의 예측 변수로 가장 유의한 X 를 결정해야 한다고 가정해 보겠습니다.

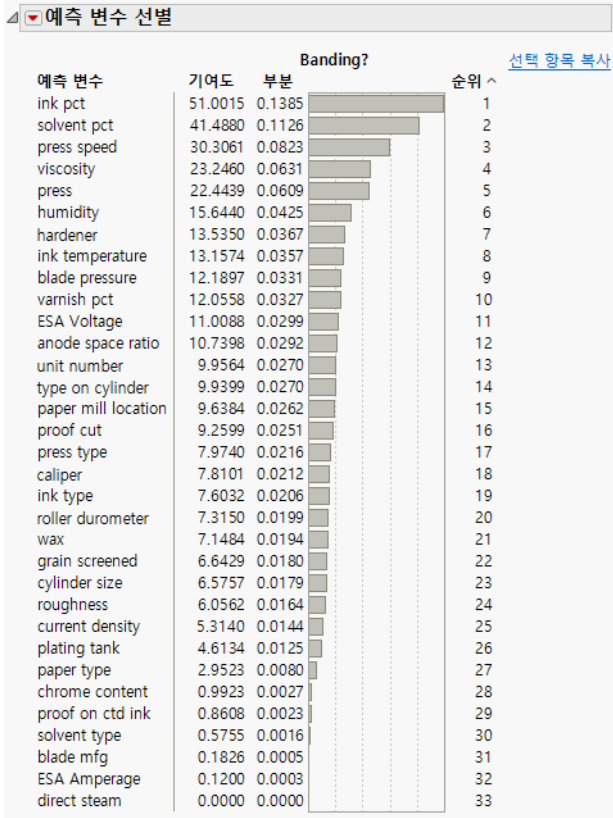
예측 변수 선별 플랫폼에서는 붓스트랩 포레스트를 사용하여 반응의 잠재적 예측 변수를 식별합니다. 각 반응에 대해 100 개의 결정 트리를 사용하는 붓스트랩 포레스트 모형이 생성됩니다. 각 예측 변수의 붓스트랩 포레스트 모형에 대한 열 기여도가 순위화됩니다. 붓스트랩 포레스트 방법은 랜덤 성분을 포함하므로 보고서를 다시 실행하면 열 기여도가 다를 수 있습니다. 결정 트리에 대한 자세한 내용은 "[파티션 모형](#)"에서 확인하십시오.

예측 변수 선별의 예

반응을 예측할 확률이 가장 높은 요인 또는 피처를 알아보려고 합니다. 그라비아 인쇄 회사의 기계에서 측정한 값이 포함된 데이터 테이블을 예로 들겠습니다. 데이터 집합에는 539 개의 레코드와 38 개의 변수가 있습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Bands Data.jmp 를 엽니다.
2. **분석 > 선별 > 예측 변수 선별**을 선택합니다.
3. Banding? 을 **Y, 반응**으로 선택합니다.
4. 그룹화된 grain screened ~ chrome content 열을 선택하고 **X** 를 클릭합니다.
5. **확인**을 클릭합니다.

그림 25.2 순위화된 열 기여도



참고: 이 분석은 랜덤 선택 성분이 있는 붓스트랩 포레스트 방법을 기반으로 하므로 결과가 [그림 25.2](#)에 표시된 것과 약간 다를 수 있습니다. 자세한 내용은 "[붓스트랩 포레스트](#)"에서 확인하십시오.

열은 붓스트랩 포레스트 모형에서 기여도에 따라 정렬되고 순위가 매겨집니다. 기여도가 가장 높은 예측 변수가 관심 반응에 대한 강력한 후보 예측 변수입니다.

팁: 테이블에서 열 정렬 방식을 변경하려면 열 머리글을 클릭합니다.

예측 변수 선별 플랫폼 시작

분석 > 선별 > 예측 변수 선별을 선택하여 예측 변수 선별 플랫폼을 시작합니다.

그림 25.3 예측 변수 선별 시작 창

많은 예측 변수의 결과 예측 능력을 선별합니다.

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

Y, 반응 반응 열입니다.

X 예측 변수 열입니다.

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.

트리 수 붓스트랩 포레스트 모형의 결정 트리 수입니다.

난수 시드값 설정 적합 절차에 사용되는 시작 값의 시드를 설정합니다. "난수 시드값 설정"은 분석을 재현하려는 경우에 유용합니다. 난수 시드값을 설정하고 스크립트를 저장하면 시드 값이 스크립트에 자동으로 저장됩니다.

예측 변수 선별 보고서

"예측 변수 선별" 보고서에는 각 기여도 및 순위가 포함된 예측 변수 목록이 표시됩니다. 기여도가 가장 높은 예측 변수가 Y 예측에 중요할 수 있습니다.

"기여도" 열에는 붓스트랩 포레스트 모형에 대한 각 예측 변수의 기여도가 표시됩니다. 보고서의 "부분" 열에는 각 변수의 기여도가 백분율로 표시됩니다.

"예측 변수 선별" 보고서에서 중요한 예측 변수를 선택할 수 있습니다. 중요한 예측 변수를 선택하면 데이터 테이블에서 해당 열이 선택되므로 모델링 플랫폼의 시작 창에 열을 쉽게 입력할 수 있습니다. "선택 항목 복사" 버튼을 사용하면 선택한 열을 클립보드에 복사하여 모형에도 붙여넣을 수 있습니다. 이런 방식으로 예측 변수 선별은 모델링 프로세스를 향상시킵니다.

예측 변수 선별 플랫폼 옵션

"예측 변수 선별"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

결과 저장 보고서의 예측 변수 순위 테이블을 새 데이터 테이블에 저장합니다.

다음 옵션에 대한 자세한 내용은 **JMP 사용**의 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.

26 장

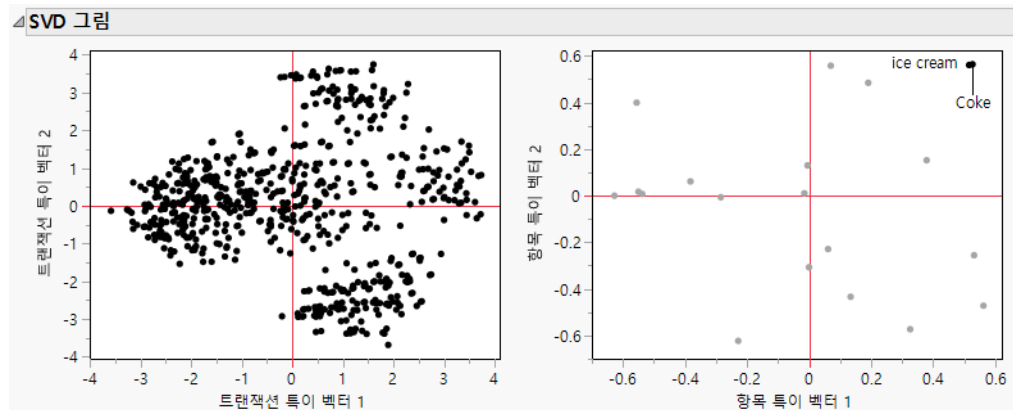
JMP PRO 연관성 분석 장바구니 분석 수행

연관성 분석 플랫폼은 JMP Pro 에서만 사용할 수 있습니다.

연관성 분석 플랫폼을 사용하면 서로 선호도가 있는 항목을 식별할 수 있습니다. 트랜잭션에서 자주 함께 나타나는 항목을 식별하기 위해 트랜잭션 데이터(장바구니라고도 함)를 분석하는 데 흔히 사용됩니다. 예를 들어 식료품점과 온라인 판매자는 연관성 분석을 사용하여 함께 구매하는 경향이 있는 상품을 전략적으로 구성하고 추천합니다.

연관성 분석은 종속 사건 또는 연관된 사건을 식별할 때도 사용됩니다. 예를 들어 비슷한 시기에 고장 난 것 같은 자동차 부품을 식별할 수 있습니다. 이 경우 차량 검사를 장바구니로 간주하고 각 검사에서 발견된 결함 부품 그룹 간의 연관성을 분석합니다.

그림 26.1 특이값 분해 그림의 예



목차

연관성 분석 플랫폼 개요.....	491
연관성 분석 플랫폼의 예.....	492
연관성 분석 플랫폼 시작.....	494
데이터 형식.....	495
연관성 분석 보고서.....	495
빈발 항목 집합.....	496
규칙.....	496
연관성 분석 플랫폼 옵션.....	497
특이값 분해 개요.....	497
SVD 보고서.....	498
SVD 보고서 옵션.....	499
주제 분석 개요.....	500
주제 분석 보고서.....	500
주제 분석 보고서 옵션.....	501
연관성 분석 플랫폼의 추가 예.....	501
연관성 분석 플랫폼에 대한 통계 상세 정보.....	504
빈발 항목 집합 생성에 대한 통계 상세 정보.....	504
성능 측도에 대한 통계 상세 정보.....	504

JMP^{PRO} 연관성 분석 플랫폼 개요

연관성 분석 플랫폼에서는 독립 사건 또는 트랜잭션에서 항목 그룹 간의 연관성을 식별합니다. 연관성 분석에서 항목은 기본 관심 대상입니다. 예를 들어 항목은 제품, 웹 페이지 또는 서비스일 수 있습니다. 항목 집합은 하나 이상의 항목 목록입니다.

두 항목 집합 간의 관계는 연관 규칙으로 정의됩니다. 연관 규칙은 조건 항목 집합과 결과 항목 집합으로 구성됩니다. 선행 항목은 조건 항목 집합의 개별 항목입니다. 연관성 분석은 조건 항목 집합이 이미 트랜잭션에 있는 경우 결과 항목 집합이 트랜잭션에 있을 것으로 예측하는 연관 규칙을 식별합니다. 일부 연관 규칙은 다른 규칙보다 더 강력하므로 더 유용합니다. 다음과 같은 세 가지 성능 측도로 연관 규칙 강도를 설명합니다.

- 지지도는 항목 집합이 나타나는 트랜잭션의 비율입니다. 지지도 값이 높으면 해당 항목 집합이 자주 발생한다는 것을 나타냅니다.
- 신뢰도는 트랜잭션에 조건 항목 집합이 있을 때 결과 항목 집합을 포함하는 트랜잭션의 비율입니다. 신뢰도는 연관 규칙의 영향력 또는 예측 검정력을 측정합니다. 연관성 분석의 신뢰도는 신뢰 구간의 개념과 관련이 없습니다.
- 향상도는 조건 항목 집합과 결과 항목 집합이 트랜잭션에서 독립적으로 발생한다고 가정할 때 기대 신뢰도에 대한 연관 규칙 신뢰도의 비율입니다. 향상도는 조건 항목 집합이 있는지 여부에 따라 결과 항목 집합이 얼마나 달라지는지 측정합니다. 향상도 최소값은 0입니다.
 - 향상도 비율이 1 보다 작으면 조건과 결과가 함께 발생하는 빈도가 우연히 발생할 경우보다 낮으므로 서로 밀어낸다는 것을 나타냅니다.
 - 향상도 비율이 1 에 가까우면 결과가 조건이 포함된 트랜잭션에서 발생하는 비율이 우연히 발생할 비율과 같음을 나타냅니다.
 - 향상도 비율이 1 보다 크면 결과 항목 집합에 조건 항목 집합에 대한 선호도가 있음을 나타냅니다. 결과 항목 집합은 우연히 발생하는 것보다 조건 항목 집합과 함께 발생할 빈도가 더 높습니다.

이러한 성능 측도에 대한 자세한 내용은 "[성능 측도에 대한 통계 상세 정보](#)"에서 확인하십시오.

연관성 분석 플랫폼을 사용하면 특이값 분해도 수행할 수 있습니다. SVD(특이값 분해)는 연관성 분석에 사용되는 것과 다른 행렬 축소 방법론을 사용하여 유사한 트랜잭션 및 유사한 항목을 그룹화합니다. 연관성 분석에서 얻은 결과를 보완하는 정보를 제공하려면 SVD 방법론을 사용합니다.

연관성 분석에 대한 자세한 내용은 Hastie et al. (2009) 및 Shmueli et al. (2010) 연구 자료에서 확인하십시오. 특이값 분해에 대한 자세한 내용은 Jolliffe(2002) 연구 자료에서 확인하십시오.

JMP PRO 연관성 분석 플랫폼의 예

이 예에서는 식료품점에서 보고한 트랜잭션 데이터를 사용합니다. 1001 명의 고객이 구매한 품목에 대해 소비자 행동 패턴을 식별하기 위해 항목 간의 연관성을 탐색하려고 합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Grocery Purchases.jmp** 를 엽니다.
2. **분석 > 선별 > 연관성 분석**을 선택합니다.
3. **Product** 를 선택하고 **항목**을 클릭합니다.
4. **Customer ID** 를 선택하고 **ID** 를 클릭합니다.

참고 : 1001 명의 각 고객에 대해 고유한 사용자 ID 가 할당됩니다.

5. **확인**을 클릭합니다.

기본적으로 "규칙" 보고서는 신뢰도에 따라 내림차순으로 정렬됩니다. 그러나 신뢰도가 매우 높은 연관 규칙은 조건 집합의 항목 수가 많은 더 경향도 있습니다. 더 작은 조건 집합의 연관 규칙을 검토할 예정이므로 보고서를 신뢰도에 따라 오름차순으로 정렬합니다.

6. "규칙" 보고서에서 마우스 오른쪽 버튼을 클릭하고 **열별 정렬**을 선택합니다.
"열 선택" 창이 나타납니다.
7. 신뢰도를 선택한 후 "오름차순" 옵션을 선택합니다.
8. **확인**을 클릭합니다.

그림 26.2 연관성 분석 보고서

규칙				
조건	결과	신뢰도	항상도	
apples	sardines	40%	1.357	^
Coke	turkey	40%	1.422	
sardines	chicken, Heineken	40%	2.002	
herring, olives	ham, turkey	40%	3.006	
ice cream	turkey	40%	1.424	
artichoke	baguette, Heineken	40%	1.547	
ham	corned beef, herring	40%	1.648	
ham	corned beef, olives	40%	1.703	
peppers	sardines	41%	1.371	
sardines	peppers	41%	1.371	
Coke	Heineken, sardines	41%	2.23	
sardines	Coke, Heineken	41%	2.332	
avocado, Heineken	crackers, ham	41%	2.707	
soda	baguette, herring	41%	1.631	
corned beef	ham	41%	1.343	
Heineken, soda	bourbon, crackers, olives	41%	3.152	
artichoke, Heineken	avocado, baguette	41%	1.921	
Heineken, herring	avocado, baguette	41%	1.924	

9. 신뢰도 값 "58%" 가 나올 때까지 보고서를 아래로 스크롤합니다.

"규칙" 보고서 테이블에 avocado(아보카도)를 구입한 고객의 58%가 artichoke(아티초크)도 구입했음을 나타내는 항목이 있습니다. "항상도" 값은 1.908로, 종속성 가능성이 있음을

나타냅니다 . 트랜잭션에서 아보카도와 아티초크가 발생하는 비율이 유의한지 확인하려고 합니다 .

10. " 빈발 항목 집합 " 옆의 표시 아이콘을 클릭합니다 .

그림 26.3 빈발 항목 집합 보고서

빈발 항목 집합		
항목 집합	지지도 ~ 항목 수	
{Heineken}	60%	1
{crackers}	49%	1
{herring}	49%	1
{olives}	47%	1
{bourbon}	40%	1
{baguette}	39%	1
{corned beef}	39%	1
{crackers, Heineken}	37%	2
{avocado}	36%	1
{soda}	32%	1
{chicken}	31%	1
{apples}	31%	1
{ice cream}	31%	1
{artichoke}	30%	1
{ham}	30%	1
{Coke}	30%	1
{peppers}	30%	1
{sardines}	30%	1
{Heineken, herring}	29%	2
{turkey}	28%	1
{baguette, Heineken}	26%	2
{Heineken, soda}	26%	2
{herring, olives}	26%	2
{artichoke, Heineken}	25%	2

" 빈발 항목 집합 " 보고서에 따르면 고객의 36% 가 아보카도를 구입했습니다 . 그림 26.2 의 " 규칙 " 보고서에서는 이러한 고객의 58% 가 아티초크도 구입했음을 보여 줍니다 . 이 행동을 보이는 고객의 비율이 높기 때문에 식료품점 관리자는 이 정보를 사용하여 아보카도와 아티초크를 전략적으로 배치할 수 있습니다 .

향상도가 가장 높은 연관 규칙도 살펴보려고 합니다 .

11. " 규칙 " 보고서 테이블에서 마우스 오른쪽 버튼을 클릭하고 **열별 정렬**을 선택합니다 .

" 열 선택 " 창이 나타납니다 .

12. 향상도를 선택하고 **확인**을 클릭합니다 .

" 규칙 " 테이블이 향상도 값에 따라 내림차순으로 정렬됩니다 . 두 번째 연관 규칙은 향상도가 5.642 이고 신뢰도가 83% 입니다 . 조건 집합 {chicken, ice cream} 과 결과 항목 집합 {Coke, sardines} 모두 지지도가 적절한지 확인하려고 합니다 .

13. " 빈발 항목 집합 " 보고서에서 마우스 오른쪽 버튼을 클릭하고 **열별 정렬**을 선택합니다 .

" 열 선택 " 창이 나타납니다 .

14. 항목 집합을 선택하고 " 오름차순 " 옵션을 선택합니다 .

15. **확인**을 클릭합니다 .

" 빈발 항목 집합 " 테이블이 항목 집합에 따라 사전순으로 정렬됩니다 . 목록을 스크롤하여 조건 항목 집합 {chicken, ice cream} 의 지지도가 14% 인지 확인합니다 . 결과 항목 집합

{Coke, sardines}의 지지도는 15%입니다. 이 연관 규칙은 항상도 값이 높지만 첫 번째로 검토한 연관 규칙보다 트랜잭션 수가 적음을 나타냅니다.

JMP PRO 연관성 분석 플랫폼 시작

분석 > 선별 > 연관성 분석을 선택하여 연관성 분석 플랫폼을 시작합니다.

그림 26.4 연관성 분석 시작 창

함께 발생하는 항목에 대한 데이터를 탐색합니다.

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

항목 분석할 항목 데이터가 포함된 하나 이상의 범주형 열입니다. 열의 모델링 유형이 "다중 반응"인 경우 각 행의 각 다중 반응이 하나의 항목으로 처리됩니다. 여러 열을 지정한 경우 각 행의 각 열이 하나의 항목으로 처리됩니다.

ID 항목이 속한 트랜잭션을 식별하는 열입니다. "항목" 역할에 할당된 열의 모델링 유형이 "다중 반응"이거나 "항목" 역할에 여러 열이 지정된 경우를 제외하고 "ID" 역할은 필수입니다.

참고 : "ID" 역할이 지정되면 ID가 동일한 모든 행이 하나의 트랜잭션으로 결합됩니다.

빈도 "항목" 열에서 트랜잭션 빈도를 식별하는 열입니다. 예를 들어 빈도가 3인 트랜잭션은 "항목" 열에서 항목이 포함된 세 개의 동일한 트랜잭션으로 간주됩니다.

참고 : "항목" 역할에 할당된 열의 모델링 유형이 "다중 반응"이 아니면 "빈도" 역할이 무시됩니다.

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

최소 지지도 항목 집합 발생 비율에 대한 최소값을 지정합니다. 이 값은 0에서 1 사이여야 합니다. 지지도가 이 값보다 크거나 같은 항목 집합만 분석에 포함됩니다.

최소 신뢰도 조건 항목 집합이 포함된 트랜잭션 내에서 결과 항목 집합이 발생할 비율에 대한 최소값을 지정합니다. 이 값은 0에서 1 사이여야 합니다. 신뢰도가 이 값보다 크거나 같은 연관 규칙만 보고서에 나타냅니다.

최소 향상도 최소 종속성 비율을 지정합니다. 향상도 값은 0 이상이어야 합니다. 향상도가 이 값보다 크거나 같은 연관 규칙만 보고서에 나타납니다.

최대 선행 항목 수 조건 항목 집합의 최대 항목 수를 지정합니다. 조건 집합의 항목 수가 이 값보다 큰 연관 규칙은 분석에 포함되지 않습니다.

최대 규칙 크기 조건 항목 집합과 결과 항목 집합의 결합에 나타나는 최대 항목 수를 지정합니다. 결합 항목 수가 이 값보다 큰 연관 규칙은 분석에 포함되지 않습니다. 기본값은 4입니다.

참고 : 시작 창에서 최소 지지도, 최대 선행 항목 수 및 최대 규칙 크기 옵션을 사용하여 대규모 데이터 집합의 계산 시간을 줄일 수 있습니다. 이러한 측도에 대한 자세한 내용은 "[연관성 분석 플랫폼에 대한 통계 상세 정보](#)"에서 확인하십시오.

데이터 형식

연관성 분석 플랫폼에는 다음과 같은 세 가지 데이터 형식이 허용됩니다.

- 각 행의 단일 항목 반응을 "항목" 역할에 지정하고 "ID" 역할을 사용하여 각 트랜잭션에 포함되는 항목을 식별할 수 있습니다. 이 형식의 경우 "항목" 및 "ID" 역할은 필수이고 "빈도" 역할은 무시됩니다.
- 모델링 유형이 "다중 반응"인 열을 "항목" 역할에 지정할 수 있습니다. 이 형식의 경우 "항목" 역할은 필수이고 "ID" 및 "빈도" 역할은 선택적입니다. "ID" 역할이 지정되면 ID가 동일한 모든 행이 하나의 트랜잭션으로 결합됩니다. "빈도" 역할이 지정되면 트랜잭션의 특정 항목이 아니라 트랜잭션에 빈도가 적용됩니다.
- 다중 항목 반응 열을 "항목" 역할에 지정할 수 있습니다. 이 형식의 경우 "ID" 및 "빈도" 역할은 선택적입니다. "ID" 역할이 지정되면 ID가 동일한 모든 행이 하나의 트랜잭션으로 결합됩니다. "빈도" 역할이 지정되면 트랜잭션의 특정 항목이 아니라 트랜잭션에 빈도가 적용됩니다.

연관성 분석 보고서

기본적으로 "연관성 분석" 보고서에는 다음 보고서가 포함됩니다.

- "[빈발 항목 집합](#)"
- "[규칙](#)"

팁 : 보고서의 테이블 내용을 특정 열을 기준으로 정렬하려면 테이블에서 마우스 오른쪽 버튼을 클릭하고 **열별 정렬**을 선택합니다.

JMP[®] PRO 빈발 항목 집합

연관성 분석 플랫폼의 "빈발 항목 집합" 보고서에는 항목 집합이 지지도에 따라 내림차순으로 나열됩니다. 나열된 항목 집합은 시작 창에 지정한 최소 지지도 값을 충족합니다. 각 항목 집합은 조건 항목 집합 및 연관 규칙을 구성하기 위한 결과 항목 집합으로 간주됩니다. 이 테이블에는 다음 열이 포함됩니다.

항목 집합 조건 집합 또는 연관 규칙을 위한 결과 집합으로 간주되는 항목 집합입니다.

지지도 항목 집합의 모든 항목이 발생하는 트랜잭션의 비율입니다.

항목 수 항목 집합의 항목 수입니다.

JMP[®] PRO 규칙

연관성 분석 플랫폼의 "규칙" 보고서에는 신뢰도에 따라 내림차순으로 정렬된 연관 규칙 테이블이 표시됩니다. 시작 창에서 지정한 최소 지지도, 최소 신뢰도, 최소 향상도, 최대 선행 항목 수 및 최대 규칙 크기 요구 사항을 충족하는 연관 규칙만 이 보고서에 나타납니다.

"규칙" 보고서 테이블에는 다음 열이 포함됩니다.

규칙 조건 항목 집합과 결과 항목 집합을 결합하여 구성된 연관 규칙입니다.

조건 트랜잭션 내에서 결과 항목 집합의 존재 여부에 영향을 미치는 것으로 간주되는 항목 집합입니다.

결과 존재 여부가 조건 항목 집합의 존재 여부에 영향을 받는 것으로 간주되는 항목 집합입니다.

신뢰도 조건 항목 집합이 트랜잭션에 있을 때 결과 항목 집합을 포함하는 트랜잭션의 비율입니다. 신뢰도는 연관 규칙의 영향력 또는 예측 검정력을 측정합니다.

향상도 조건 항목 집합과 결과 항목 집합이 트랜잭션에서 독립적으로 발생한다고 가정할 때 기대 신뢰도에 대한 연관 규칙 신뢰도의 비율입니다. 향상도는 조건 항목 집합이 있는지 여부에 따라 결과 항목 집합이 얼마나 달라지는지 측정합니다. 향상도 최소값은 0입니다.

- 향상도 비율이 1 보다 작으면 조건 항목 집합과 결과 항목 집합이 함께 발생하는 빈도가 우연히 발생할 경우보다 낮으므로 서로 밀어낸다는 것을 나타냅니다.
- 향상도 비율이 1 에 가까우면 결과 항목 집합이 조건 항목 집합이 포함된 트랜잭션에서 발생하는 비율이 우연히 발생할 비율과 같음을 나타냅니다.
- 향상도 비율이 1 보다 크면 결과 항목 집합에 조건 항목 집합에 대한 선호도가 있음을 나타냅니다. 결과 항목 집합은 우연히 발생하는 것보다 조건 항목 집합과 함께 발생할 빈도가 더 높습니다.

JMP[®] PRO 연관성 분석 플랫폼 옵션

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

트랜잭션 목록 각 트랜잭션 ID 값 및 해당 트랜잭션에 포함된 항목을 나열하는 테이블을 표시하거나 숨깁니다. 테이블은 "트랜잭션 ID" 열을 기준으로 정렬됩니다.

빈발 항목 집합 지지도가 시작 창에 지정된 최소 지지도 값을 초과하는 항목 집합 목록을 표시하거나 숨깁니다. 자세한 내용은 "**빈발 항목 집합**"에서 확인하십시오.

규칙 시작 창에 지정된 최소 지지도, 최소 신뢰도, 최소 향상도, 최대 선행 항목 수 및 최대 규칙 크기 요구 사항을 충족하는 연관 규칙 테이블을 표시하거나 숨깁니다. 자세한 내용은 "**규칙**"에서 확인하십시오.

SVD 항목에 대한 따름 행렬의 부분 SVD(특이값 분해) 보고서를 표시하거나 숨깁니다. 이 분해는 따름 행렬을 분석을 위해 사용자가 지정한 차원 수로 축소합니다. SVD 보고서에는 특이 벡터, 군집 트랜잭션 및 군집 항목을 시각화하고, 특이 벡터를 저장하거나, 회전된 SVD 분석을 수행하는 옵션도 포함되어 있습니다. 자세한 내용은 "**특이값 분해 개요**"에서 확인하십시오.

다음 옵션에 대한 자세한 내용은 **JMP 사용**의 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.

JMP[®] PRO 특이값 분해 개요

SVD(특이값 분해)는 서로 선호도가 있는 항목을 식별할 수 있는 다른 방법을 제공하여 연관성 분석을 보완합니다. 트랜잭션 항목 행렬의 특이값 분해는 행렬을 관리 가능한 차원 수로 축소하여 유사한 트랜잭션 및 유사한 항목을 그룹화할 수 있도록 합니다. SVD 분석은 상관 행렬에 대해 PCA(주성분 분석)를 수행하는 것과 동일합니다.

트랜잭션 항목 행렬은 각 행이 트랜잭션에 해당하고 각 열이 항목에 해당하는 행렬입니다. 행렬의 항목은 0과 1입니다. 트랜잭션에서 항목이 발생하면 해당 행과 열 항목이 1입니다. 그렇지 않

으면 행과 열 항목이 0입니다. 일반적으로 트랜잭션 항목 행렬에는 1보다 0 값이 더 많으므로 희소 행렬이라고 합니다.

부분 특이값 분해는 세 개의 행렬 $\mathbf{U}$, $\mathbf{S}$ 및 $\mathbf{V}'$ 를 사용하여 열 표준화 트랜잭션 항목 행렬에 근사한 값을 산출합니다. 이러한 행렬 간의 관계는 다음과 같이 정의됩니다.

$$\text{트랜잭션 항목 행렬} \approx \mathbf{U} * \mathbf{S} * \mathbf{V}'$$

$nTran$ 을 트랜잭션 항목 행렬의 트랜잭션(행) 수, $nItem$ 을 트랜잭션 항목 행렬의 항목(열) 수, $nVec$ 를 지정된 특이 벡터 수로 각각 정의합니다. $nVec$ 는 $\min(nTran, nItem)$ 보다 작거나 같아야 합니다. 따라서 $\mathbf{U}$ 는 트랜잭션 항목 행렬의 왼쪽 특이 벡터를 포함하는 $nTran \times nVec$ 행렬입니다. $\mathbf{S}$ 는 $nVec$ 차 대각 행렬입니다. $\mathbf{S}$ 의 대각 항목은 트랜잭션 항목 행렬의 특이값입니다. $\mathbf{V}'$ 는 $nVec \times nItem$ 행렬입니다. $\mathbf{V}'$ 의 행(또는 $\mathbf{V}$ 의 열)은 오른쪽 특이 벡터입니다.

오른쪽 특이 벡터는 함수 또는 주제 영역이 유사한 서로 다른 항목 간의 연관성을 포착합니다. 세 개의 항목이 동일한 트랜잭션에 나타나는 경향이 있는 경우 SVD에서는 $\mathbf{V}'$ 에 이 세 개 항목에 대해 큰 값을 갖는 특이 벡터를 생성할 가능성이 높습니다. $\mathbf{U}$ 특이 벡터는 이 새로운 항목 공간에 투영된 트랜잭션을 나타냅니다.

SVD는 간접 관계도 포착합니다. 두 개의 항목이 동일한 트랜잭션에 함께 나타나지는 않지만 일반적으로 다른 세 번째 항목이 있는 트랜잭션에 나타나는 경우 SVD는 이러한 연관성을 포착할 수 있습니다. 두 개의 트랜잭션에 공통된 항목은 없지만 차원 축소 공간에서 연결되는 항목이 포함된 경우 이 두 트랜잭션은 SVD 그림에서 유사 벡터에 매핑됩니다.

SVD는 트랜잭션 데이터를 고정 차원 벡터 공간으로 변환하여 군집화, 분류 및 회귀 기법에 적용할 수 있도록 합니다. 저장 옵션을 사용하면 이 벡터 공간을 다른 JMP 플랫폼에서 분석할 수 있는 형태로 내보낼 수 있습니다.

특이값 분해가 수행되기 전에 트랜잭션 항목 행렬을 중심화하고 척도화한 후 $nTran - 1$ 로 나눕니다. 이 분석은 트랜잭션 항목 행렬의 상관 행렬에 대한 PCA와 동일합니다. SVD 구현은 트랜잭션 항목 행렬의 희소성을 활용합니다.

JMP[®] PRO SVD 보고서

연관성 분석 플랫폼의 SVD 옵션은 특이값 분해를 통해 두 개의 SVD 그림과 한 개의 특이값 테이블을 생성합니다.

JMP[®] PRO SVD 그림

첫 번째 그림에는 각 트랜잭션에 대한 점이 포함되어 있습니다. 특정 트랜잭션에 대해 표시되는 점은 처음 두 개의 특이 벡터($\mathbf{U}$ 행렬의 처음 두 열)에 있는 트랜잭션 값에 대각 특이값 행렬($\mathbf{S}$)을 곱한 값으로 정의됩니다. 이 그림은 주성분 플랫폼의 스코어 그림과 동일합니다. 이 그림의 각 점은 트랜잭션을 나타냅니다. 트랜잭션 SVD 그림에서 눈에 띄게 그룹화된 점은 구성이 유사한 트랜잭션을 나타냅니다.

두 번째 그림에는 각 항목에 대한 점이 포함되어 있습니다. 특정 항목에 대해 표시되는 점은 처음 두 개의 특이 벡터($\mathbf{V}'$ 행렬의 처음 두 행)에 있는 항목 값에 대각 특이값 행렬($\mathbf{S}$)을 곱한 값으로

정의됩니다. 이 그림은 주성분 플랫폼의 적재 그림과 동일합니다. 이 그림의 각 점은 항목을 나타냅니다. 항목 SVD 그림에서 눈에 띄게 그룹화된 항목은 함수 또는 주제 영역이 유사한 항목을 나타냅니다.

JMP PRO 특이값

트랜잭션 및 항목 SVD 그림 아래에 특이값 테이블이 나타납니다. 이 특이값은 트랜잭션 항목 행렬의 특이값 분해에서 **S** 행렬의 대각 항목입니다. 특이값 테이블에는 동등한 주성분 분석의 해당 고유값이 들어 있는 열이 포함되어 있습니다. 주성분 플랫폼에서와 마찬가지로 각 고유값(또는 특이값)이 설명하는 변동의 백분율 및 누적 백분율을 나타내는 열이 있습니다. "누적 백분율" 열을 사용하여 트랜잭션 항목 행렬에서 유지할 분산의 백분율을 결정한 후 해당하는 수의 특이값 벡터를 사용할 수 있습니다.

JMP PRO SVD 보고서 옵션

연관성 분석 플랫폼의 SVD 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

SVD 산점도 행렬 항목 및 트랜잭션 특이값 분해 벡터에 대한 산점도 행렬을 표시하거나 숨깁니다. 이 옵션을 선택할 경우 산점도 행렬의 크기를 선택해야 합니다. 이 산점도 행렬을 사용하면 특이값 분해의 처음 두 차원보다 더 많은 차원을 시각화할 수 있습니다.

주제 분석, 회전된 SVD 항목 트랜잭션 행렬에 대한 Varimax 회전 부분 특이값 분해를 수행하여 주제라는 항목 그룹을 생성합니다. 이 옵션을 여러 번 선택하여 서로 다른 개수의 주제를 찾을 수 있습니다. 자세한 내용은 "[주제 분석 개요](#)"에서 확인하십시오.

군집 항목 데이터에 있는 항목의 계층적 군집화 분석을 표시하거나 숨깁니다. 덴드로그램 오른쪽에 다음과 같은 옵션이 있습니다.

군집 설정 덴드로그램에서 색상이 적용되고 "군집 저장" 옵션에 사용되는 군집 수를 지정합니다. 항목 SVD 그림의 점도 군집별로 색상이 적용됩니다.

군집 저장 항목 및 해당 항목의 군집 할당을 포함하는 데이터 테이블을 저장합니다. "군집 저장"을 클릭하기 전에 군집 수를 설정해야 합니다.

군집화 및 덴드로그램에 대한 자세한 내용은 [다변량 방법](#)에서 확인하십시오.

군집 트랜잭션 데이터에 있는 트랜잭션의 계층적 군집화 분석을 표시하거나 숨깁니다. 덴드로그램 오른쪽에 다음과 같은 옵션이 있습니다.

군집 설정 덴드로그램에서 색상이 적용되고 "군집 저장" 옵션에 사용되는 군집 수를 지정합니다. 트랜잭션 SVD 그림의 점도 군집별로 색상이 적용됩니다.

군집 저장 트랜잭션 및 해당 트랜잭션의 군집 할당을 포함하는 데이터 테이블을 저장합니다. "군집 저장"을 클릭하기 전에 군집 수를 설정해야 합니다.

군집화 및 덴드로그램에 대한 자세한 내용은 [다변량 방법](#)에서 확인하십시오.

트랜잭션 특이 벡터 저장 트랜잭션 특이값 분해에서 사용자가 지정한 수의 특이 벡터를 새 데이터 테이블의 열로 저장합니다. 이 데이터 테이블에서 각 행은 하나의 트랜잭션에 해당합니다. 저

장된 처음 두 열은 트랜잭션 SVD 그림에 표시된 점을 나타냅니다. 자세한 내용은 "[특이값 분해 개요](#)"에서 확인하십시오.

항목 특이 벡터 저장 항목 특이값 분해에서 사용자가 지정한 수의 특이 벡터를 새 데이터 테이블의 열로 저장합니다. 이 데이터 테이블에서 각 행은 하나의 항목에 해당합니다. 저장된 처음 두 열은 항목 SVD 그림에 표시된 점을 나타냅니다. 자세한 내용은 "[특이값 분해 개요](#)"에서 확인하십시오.

JMP PRO 주제 분석 개요

연관성 분석 플랫폼의 "주제 분석, 회전된 SVD" 옵션은 트랜잭션 항목 행렬의 부분 SVD(특이값 분해)에 대해 Varimax 회전을 수행합니다. 자세한 내용은 "[특이값 분해 개요](#)"에서 확인하십시오. 트랜잭션 항목 행렬에서 유지할 주제 수에 해당하는 회전된 특이 벡터의 수를 지정해야 합니다. 주제 수를 지정하면 주제 분석 보고서가 나타납니다.

주제 분석은 회전 PCA(주성분 분석)와 동일합니다. Varimax 회전은 특이 벡터 집합을 선택하고 이를 회전시켜 더 직접적으로 좌표 방향(항목 방향)을 향하게 합니다. 이렇게 회전하면 각 회전 벡터가 항목 집합을 향하므로 벡터가 트랜잭션의 구성을 설명하는 데 도움이 됩니다. 음수 값은 반발력을 나타냅니다. 음수 값을 갖는 항목은 양수 값을 갖는 항목에 비해 주제에 나타나는 빈도가 낮습니다.

JMP PRO 주제 분석 보고서

연관성 분석 플랫폼의 "주제 분석" 보고서에서는 회전 후 각 주제에서 적재량이 가장 큰 항목을 보여 줍니다. 추가 보고서에서는 회전된 특이값 분해의 성분을 보여 줍니다.

"주제별 상위 적재" 보고서에서는 각 주제에 대한 테이블을 보여 줍니다. 각 테이블의 항목은 각 주제에 대한 적재량 절대값이 가장 큰 항목입니다. 각 테이블은 적재량 절대값을 기준으로 내림차순 정렬되어 있습니다. 이러한 테이블을 사용하여 각 주제에 해당하는 개념 테마를 판별할 수 있습니다.

주제 분석 보고서에는 다음 보고서도 포함됩니다.

주제 적재 각 항목의 주제별 적재량 행렬이 포함됩니다. 이 행렬은 회전 PCA의 요인 적재 행렬과 동일합니다.

주제별 단어 클라우드 각 주제당 하나씩의 단어 클라우드가 포함됩니다.

주제 스코어 각 주제의 트랜잭션 스코어 행렬이 포함됩니다. 주제에서 스코어가 높은 트랜잭션은 해당 주제와 연관성이 있을 가능성이 높습니다.

주제 스코어 그림 모든 트랜잭션에 대한 주제 스코어를 1차원 산점도에 표시합니다. 음수 값은 주제와 음의 관계를 갖는 트랜잭션을 나타냅니다. 이 그림을 사용하여 각 주제 내에서 트랜잭션의 분포를 살펴볼 수 있습니다.

각 주제에 의해 설명된 분산 각 주제에 의해 설명되는 분산의 테이블이 포함됩니다. 이 테이블에는 각 주제에 의해 설명되는 변동이 백분율 또는 누적 백분율에 대한 열도 포함됩니다.

회전 행렬 Varimax 회전에 대한 회전 행렬이 포함됩니다.

JMP PRO 주제 분석 보고서 옵션

연관성 분석 플랫폼의 "주제 분석" 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

주제 산점도 행렬 회전된 특이값 분해 벡터에 대한 산점도 행렬을 표시하거나 숨깁니다.

표시 옵션 주제 분석 보고서에 나타나는 내용을 표시하거나 숨기기 위한 옵션이 포함되어 있습니다. 자세한 내용은 "[주제 분석 보고서](#)"에서 확인하십시오.

주제 이름 바꾸기 하나 이상의 주제에 대해 설명이 포함된 이름을 추가할 수 있는 대화상자를 표시합니다.

트랜잭션 주제 벡터 저장 회전 특이값 분해에서 구한 특이 벡터 중 사용자가 지정한 수만큼을 데이터 테이블에 열로 저장합니다.

항목 주제 벡터 저장 주제 벡터를 "항목 주제 스코어"라는 새 데이터 테이블의 열로 저장합니다.

제거 SVD 보고서에서 주제 분석 보고서를 제거합니다.

JMP PRO 연관성 분석 플랫폼의 추가 예

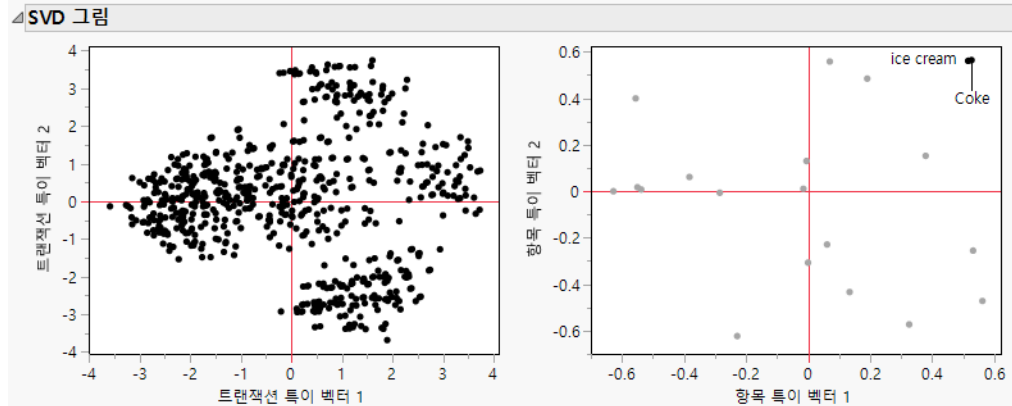
이 예에서는 트랜잭션 항목 행렬의 특이값 분해를 사용하여 식료품점에서 보고한 트랜잭션 데이터를 분석합니다. 데이터에는 1001 명의 고객이 구매한 품목에 대한 정보가 포함되어 있습니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 Grocery Purchases.jmp 를 엽니다.
2. **분석 > 선별 > 연관성 분석**을 선택합니다.
3. Product 를 선택하고 **항목**을 클릭합니다.
4. Customer ID 를 선택하고 **ID** 를 클릭합니다.

참고 : 1001 명의 각 고객에 대해 고유한 사용자 ID 가 할당됩니다.

5. **확인**을 클릭합니다.
6. "연관성 분석"의 빨간색 삼각형을 클릭하고 **SVD** 를 선택합니다.
7. **확인**을 클릭합니다.

그림 26.5 SVD 그림



트랜잭션 SVD 그림은 두 개 또는 세 개의 트랜잭션 그룹이 있을 수 있음을 나타냅니다. 항목 SVD 그림의 오른쪽 위에 Coke(콜라) 와 ice cream(아이스크림) 을 나타내는 점이 중첩되어 있습니다. 이 두 항목의 근접성은 두 항목 간에 강한 선호도가 있음을 나타냅니다.

8. "SVD" 의 빨간색 삼각형을 클릭하고 **주제 분석, 회전된 SVD** 를 선택합니다.
9. "주제 수 (회전된 특이 벡터)" 옆에 "3" 을 입력하고 **확인** 을 클릭합니다.

주제 항목 및 주제 스코어 보고서가 나타납니다.

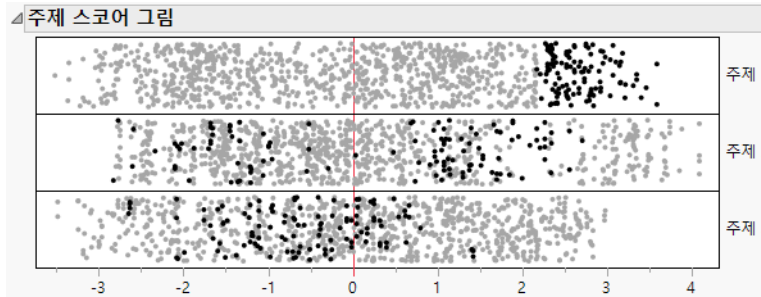
그림 26.6 주제 항목 보고서

▲ 주제별 상위 적재

주제 1		주제 2		주제 3	
항목	적재	항목	적재	항목	적재
avocado	-0.7191	Coke	0.7017	crackers	0.7050
olives	0.7142	ice cream	0.6945	apples	-0.6437
baguette	-0.6089	herring	-0.6550	soda	0.6411
turkey	0.5697	sardines	0.6122	Heineken	0.6409
artichoke	-0.4889	chicken	0.5158	sardines	-0.3618
bourbon	0.4860	corned beef	-0.4257	steak	-0.3468
Heineken	-0.4329	steak	-0.3395	corned beef	-0.3409
corned beef	0.3836	olives	-0.3035	bourbon	0.3340
Coke	0.3096	ham	-0.3022		

세 개의 그룹 또는 주제가 생성되어 주제 항목 보고서에 표시됩니다. 주제 항목 테이블에 나열된 첫 번째 항목은 해당 그룹의 기본 항목을 나타냅니다. 예를 들어 "주제 1" 은 avocado(아보카도) 를 포함하지 않지만 olives(올리브) 를 포함하는 트랜잭션으로 주로 식별되는 그룹입니다.

그림 26.7 주제 스코어



1001 개의 각 트랜잭션에 할당된 주제 스코어가 " 주제 스코어 " 보고서에 표시됩니다 . 해당 트랜잭션이 다른 주제와 어떤 관련이 있는지 보려면 주제에 대한 점 그룹을 선택합니다 . 예를 들어 주제 1 에서 값이 매우 높은 트랜잭션이 주제 2 와 3 에서는 값이 낮은 경향이 있습니다 .

10. " 특이값 " 보고서를 엽니다 .

그림 26.8 특이값 테이블

순자	특이값	고유값	백분율	누적 백분율
1	1.7399	3.0272	15.1362	15.1362
2	1.6494	2.7204	13.6021	28.7383
3	1.5093	2.2780	11.3901	40.1284
4	1.4077	1.9816	9.9079	50.0363
5	1.2565	1.5788	7.8938	57.9301
6	1.1331	1.2839	6.4196	64.3497
7	1.0333	1.0676	5.3381	69.6877
8	0.9927	0.9854	4.9270	74.6148
9	0.9265	0.8583	4.2916	78.9063
10	0.7888	0.6223	3.1114	82.0177
11	0.7247	0.5253	2.6263	84.6440
12	0.6647	0.4419	2.2094	86.8534
13	0.6465	0.4179	2.0897	88.9431
14	0.6379	0.4069	2.0346	90.9777
15	0.6266	0.3927	1.9634	92.9411
16	0.6140	0.3770	1.8848	94.8259
17	0.6077	0.3693	1.8465	96.6724
18	0.5817	0.3383	1.6917	98.3641
19	0.5590	0.3125	1.5623	99.9264
20	0.1213	0.0147	0.0736	100.000

그림 26.8 에서 알 수 있듯이 처음 두 특이값은 식료품점 데이터의 변동을 약 30% 만 설명합니다 . 변동을 충분히 설명하려면 다른 차원이 더 필요할 수 있습니다 .

JMP^{PRO} 연관성 분석 플랫폼에 대한 통계 상세 정보

이 섹션에는 다음 항목에 대한 통계 상세 정보가 포함되어 있습니다.

- "빈발 항목 집합 생성에 대한 통계 상세 정보"
- "성능 측도에 대한 통계 상세 정보"

JMP^{PRO} 빈발 항목 집합 생성에 대한 통계 상세 정보

연관성 분석 플랫폼에서는 빈발 항목 집합을 생성할 때 **Apriori 알고리즘**을 사용하여 계산 시간을 줄입니다. Apriori 알고리즘은 항목 집합의 지지도가 부분집합의 지지도보다 크지 않다는 사실을 활용합니다. 플랫폼에서는 최소 지지도 수준을 충족하는 더 작은 항목 집합의 조합에서 더 큰 항목 집합을 생성합니다. 또한 지정된 최대 선행 항목 수 또는 최대 규칙 크기를 초과하는 항목 집합을 생성하지 않습니다. 가능한 총 규칙 수는 항목 수에 따라 기하급수적으로 증가하므로 대규모 데이터 집합으로 작업할 때 이러한 옵션이 유용합니다. Apriori 알고리즘에 대한 자세한 내용은 Agrawal and Srikant(1994) 연구 자료에서 확인하십시오.

JMP^{PRO} 성능 측도에 대한 통계 상세 정보

이 섹션에서는 연관성 분석 플랫폼에 사용되는 성능 측도를 정의합니다. 조건 항목 집합 X 와 결과 항목 집합 Y 를 사용하는 연관 규칙을 $X \Rightarrow Y$ 로 나타냅니다. Hahsler(2015)에는 지지도, 신뢰도 및 향상도를 비롯하여 연관성 분석에 사용되는 성능 측도 모음이 포함되어 있습니다.

JMP^{PRO} 지지도

지지도는 항목 집합이 발생하는 트랜잭션의 비율입니다. 지지도를 트랜잭션에 항목 집합이 포함될 확률로 볼 수도 있습니다.

조건 항목 집합 X 의 지지도 S 는 다음과 같이 정의됩니다.

$$\text{supp}(X) = \frac{N_X}{N} = \text{Pr}(X)$$

다음은 각 요소에 대한 설명입니다.

N_X = 항목 집합 X 를 포함하는 트랜잭션의 수

N = 총 트랜잭션 수

연관 규칙의 지지도는 다음과 같이 정의됩니다.

$$\text{supp}(X \Rightarrow Y) = \text{supp}(X \cup Y) = \text{Pr}(X \cap Y)$$

이 경우 지지도는 트랜잭션에 항목 집합 X 와 Y 가 모두 포함될 확률과 동일합니다.

항목 집합과 연관 규칙의 경우 모두 지지도 범위는 0% ~ 100% 사이입니다.

JMP[®] PRO 신뢰도

신뢰도는 트랜잭션에 조건 항목 집합이 있을 때 결과 항목 집합을 포함하는 트랜잭션의 비율입니다. 신뢰도를 트랜잭션에 조건 항목 집합이 있을 때 해당 트랜잭션에 결과 항목 집합이 포함될 조건부 확률로 볼 수도 있습니다.

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} = \frac{\text{Pr}(X \cap Y)}{\text{Pr}(X)} = \text{Pr}(Y|X)$$

신뢰도 범위는 0% ~ 100% 사이입니다. 신뢰도가 0%인 연관 규칙의 결과 항목 집합은 조건 항목 집합이 있는 모든 트랜잭션에 나타나지 않습니다. 신뢰도가 100%이면 조건 항목 집합을 포함하는 모든 트랜잭션에 결과 항목 집합도 포함된다는 것을 나타냅니다.

참고 : 연관성 분석의 신뢰도는 신뢰 구간의 개념과 관련이 없습니다.

JMP[®] PRO 향상도

향상도는 X와 Y 간의 종속성을 측정합니다.

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}(Y)} = \frac{\text{Pr}(X \cap Y)}{\text{Pr}(X)\text{Pr}(Y)}$$

향상도의 분자는 X와 Y가 공동으로 발생하는 트랜잭션의 비율입니다. 분모는 X와 Y가 독립적으로 발생한다고 가정할 때 예상되는 결합 발생의 추정값입니다.

향상도 범위는 0 ~ ∞ 사이입니다. 향상도 값이 1이면 X와 Y가 트랜잭션에서 공동으로 발생하는 빈도가 단지 우연히 발생할 가능성과 같음을 나타냅니다. 향상도 값이 증가하면 X가 있을 때 Y가 예상보다 자주 발생한다는 것을 나타냅니다.

참고 : $X \Rightarrow Y$ 연관 규칙의 향상도는 $Y \Rightarrow X$ 연관 규칙의 향상도와 같습니다.

공정 기록 탐색기

복잡한 공정 기록에서 문제 성분 식별

공정 기록 탐색기는 반도체 공장과 같이 도구 또는 경로 조합에 결함이 있으면 수율과 품질이 저하될 수 있는 복잡한 제조 환경에서 유용합니다. 하나의 유닛에 서로 다른 여러 도구를 포함하는 수백 개의 단계를 거쳐야 할 수 있습니다. 낮은 수율의 원인이 되는 정확한 요인 조합을 판별하기 어려울 수 있습니다.

공정 기록 탐색기에서는 모든 생산 추적 데이터를 정리하고 낮은 수율과 관련된 것으로 보이는 요인을 식별할 수 있습니다. 그런 다음 단계별 회귀를 수행할 수 있습니다. 공정에서 단계 또는 도구 사이의 유닛 대기 시간을 분석하고 이러한 대기 시간이 수율과 어떤 관련이 있는지 확인할 수 있습니다. 또한 다른 도구로의 전환이 수율에 미치는 영향을 분석할 수도 있습니다.

참고 : 공정 기록 탐색기 플랫폼은 시험용입니다. 향후 릴리스에서 변경될 수 있습니다.

그림 27.1 공정 기록 탐색기 보고서

공정 기록 탐색기					
1250개의 유닛에서 다음 ID 열 사용: Lot, Wafer					
125개의 단계에서 다음 열 사용: Layer, Operation					
156250번의 작업에서 다음 시간 열 사용: TimeIn, TimeOut					
195개의 수준이 다음 X 열에 있음: Tool, Route					
1 반응: Yield					
반응 Yield					
최저 수율 수준					
항	수준	개수	평균	표준 오차	표준편차
Tool	PlanarTool008	1000	0.75299	0.00188	0.05952
Tool	DepTool003	1375	0.75577	0.00168	0.06245
Tool	EtchTool019	750	0.76192	0.0021	0.05749
Tool	EtchTool012	500	0.76203	0.0026	0.05807
Tool	EtchTool034	850	0.76279	0.00199	0.0581
Tool	PlanarTool021	1050	0.7637	0.00186	0.06028
Tool	PlanarTool030	1025	0.76539	0.00184	0.05894
Tool	EtchTool016	900	0.76566	0.00191	0.05734
Tool	EtchTool009	925	0.76598	0.00197	0.05994
Tool	Dep2Tool018	1525	0.76766	0.00156	0.06088
Tool	Dep2Tool014	1825	0.76813	0.00138	0.05893
Tool	DepTool019	1650	0.76839	0.00143	0.05794
Tool	EtchTool020	1000	0.76909	0.00177	0.05586
Tool	Dep2Tool015	1475	0.76948	0.0015	0.05775
Tool	PhotoTool010	3525	0.76996	0.00098	0.05822
Tool	Dep2Tool013	1825	0.7705	0.00141	0.06025
Tool	DepTool012	1750	0.77063	0.00145	0.06051
Tool	Dep2Tool006	1400	0.77063	0.00158	0.05905
Tool	PlanarTool028	900	0.77103	0.00202	0.06071
Tool	Dep2Tool011	1600	0.77124	0.00145	0.05797
Tool	EtchTool010	850	0.77125	0.00213	0.06208

목차

공정 기록 탐색기의 예.....	509
공정 기록 탐색기 플랫폼 시작.....	511
데이터 형식.....	512
공정 기록 탐색기 보고서.....	512
공정 기록 탐색기 플랫폼 옵션.....	512

공정 기록 탐색기의 예

공정 기록 탐색기를 사용하여 복잡한 제조 공정을 거친 웨이퍼의 수율을 조사합니다.

1. **도움말 > 샘플 데이터 폴더**를 선택하고 **Quality Control > Lot Wafer History.jsp** 및 **Quality Control > Lot Wafer Yield.jsp** 를 엽니다.

Lot Wafer History.jsp에는 50 개 로트의 25 개 웨이퍼에 대한 단계별 기록이 포함되어 있습니다. Lot Wafer Yield.jsp에는 각 로트의 각 웨이퍼에 대한 수율이 포함되어 있습니다.

2. Lot Wafer History.jsp 데이터 테이블에서 **분석 > 선별 > 공정 기록 탐색기**를 선택합니다.
3. Lot 와 Wafer 를 선택하고 **ID** 를 클릭합니다.
4. Tool 과 Route 를 선택하고 **X, 공정**을 클릭합니다.
5. Layer 와 Operation 을 선택하고 **단계**를 클릭합니다.
6. TimeIn 과 TimeOut 을 선택하고 **타임스탬프**를 클릭합니다.
7. **확인**을 클릭합니다.

해당하는 수율 테이블을 선택할 수 있는 창이 나타납니다.

8. Lot Wafer Yield 를 선택하고 **확인**을 클릭합니다.

수율 열을 선택할 수 있는 창이 나타납니다.

9. Yield 를 선택하고 **확인**을 클릭합니다.
10. " 공정 기록 탐색기 " 의 빨간색 삼각형을 클릭하고 **최저 수율 수준**을 선택합니다.

그림 27.2 공정 기록 탐색기 보고서

공정 기록 탐색기					
1250개의 유닛에서 다음 ID 열 사용: Lot, Wafer					
125개의 단계에서 다음 열 사용: Layer, Operation					
156250번의 작업에서 다음 시간 열 사용: TimeIn, TimeOut					
195개의 수준이 다음 X 열에 있음: Tool, Route					
1 반응: Yield					
반응 Yield					
최저 수율 수준					
항	수준	개수	평균	표준 오차	표준편차
Tool	PlanarTool008	1000	0.75299	0.00188	0.05952
Tool	DepTool003	1375	0.75577	0.00168	0.06245
Tool	EtchTool019	750	0.76192	0.0021	0.05749
Tool	EtchTool012	500	0.76203	0.0026	0.05807
Tool	EtchTool034	850	0.76279	0.00199	0.0581
Tool	PlanarTool021	1050	0.7637	0.00186	0.06028
Tool	PlanarTool030	1025	0.76539	0.00184	0.05894
Tool	EtchTool016	900	0.76566	0.00191	0.05734
Tool	EtchTool009	925	0.76598	0.00197	0.05994
Tool	Dep2Tool018	1525	0.76766	0.00156	0.06088
Tool	Dep2Tool014	1825	0.76813	0.00138	0.05893
Tool	DepTool019	1650	0.76839	0.00143	0.05794
Tool	EtchTool020	1000	0.76909	0.00177	0.05586
Tool	Dep2Tool015	1475	0.76948	0.0015	0.05775
Tool	PhotoTool010	3525	0.76996	0.00098	0.05822
Tool	Dep2Tool013	1825	0.7705	0.00141	0.06025
Tool	DepTool012	1750	0.77063	0.00145	0.06051
Tool	Dep2Tool006	1400	0.77063	0.00158	0.05905
Tool	PlanarTool028	900	0.77103	0.00202	0.06071
Tool	Dep2Tool011	1600	0.77124	0.00145	0.05797
Tool	EtchTool010	850	0.77125	0.00213	0.06208

이 보고서에는 유닛, 단계 및 작업 수와 X 열의 수준 수를 비롯하여 공정에 대한 정보가 포함됩니다. "최저 수율 수준" 테이블에는 X 열의 195개 수준에 대한 개수 및 수율 정보가 포함되어 있습니다. PlanarTool008 도구의 수율이 가장 낮고 1,000개의 유닛이 해당 도구를 거쳤습니다.

공정 기록 탐색기 플랫폼 시작

분석 > 선별 > 공정 기록 탐색기를 선택하여 공정 기록 탐색기 플랫폼을 시작합니다.

그림 27.3 공정 기록 탐색기 시작

낮은 수율과 연관된 공정 단계를 파악합니다.

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

시작 창 역할

ID 공정의 유닛을 식별하는 하나 이상의 열을 지정합니다.

참고 : ID 열은 수율 테이블에 해당 열이 있어야 합니다.

X, 공정 공정 단계를 구성하는 성분 또는 특정 상세 정보를 식별하는 하나 이상의 열을 지정합니다.

단계 유닛이 거쳐 가는 단계 및 하위 단계를 지정하는 선택적 열입니다. 예를 들어 Layer 와 해당 Layer 내의 Operation 유형입니다.

타임스탬프 작업 시간을 지정합니다. 시간은 단일 시점으로 지정하거나, 작업 시작 및 종료 시간을 나타내는 한 쌍의 시점으로 지정할 수 있습니다.

순서 유닛 배치가 "X, 공정" 역할에 지정된 성분을 거쳐 가는 순서를 지정하는 선택적 열입니다.

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

시작 창 옵션

목표 : Y 최소화 분석 목표가 반응을 최소화하는 것임을 지정합니다. 기본적으로 분석은 반응을 최대화하려고 합니다.

참고 : 반응은 플랫폼을 시작할 때 사용된 데이터 테이블이 아니라 두 번째 데이터 테이블에 지정됩니다.

데이터 형식

공정 기록 탐색기에 사용할 데이터는 두 개의 데이터 테이블에 지정되어야 합니다. 첫 번째 테이블은 플랫폼을 시작하는 데 사용되며 공정의 단계 및 상세 정보를 포함합니다. 두 번째 테이블은 ID 열의 각 수준에 대한 반응 또는 수율을 포함합니다.

Quality Control > Lot Wafer History 및 Quality Control > Lot Wafer Yield 샘플 데이터 테이블은 데이터가 두 데이터 테이블에 지정되는 방식을 보여 주는 예입니다.

공정 기록 탐색기 보고서

초기 "공정 기록 탐색기" 보고서에서는 데이터 테이블의 정보를 요약하여 보여 줍니다.

- ID 열에 의해 결정되는 유닛 수. ID 열이 두 개 이상인 경우 고유 조합의 수입니다.
- 단계 열에 의해 결정되는 단계 수. 단계 열이 두 개 이상인 경우 고유 조합의 수입니다.
- 타임스탬프 열에 의해 결정되는 작업 수. 이 값은 데이터 테이블의 행 수와 같습니다.
- X, 공정 열에 의해 결정되는 수준 수. X, 공정 열이 두 개 이상인 경우 이 값은 각 열에 있는 수준의 수의 합입니다.
- Yield 열에 의해 결정되는 반응 수

"공정 기록 탐색기"의 빨간색 삼각형 메뉴에서 옵션을 선택하여 분석을 계속합니다. 자세한 내용은 "공정 기록 탐색기 플랫폼 옵션"에서 확인하십시오.

공정 기록 탐색기 플랫폼 옵션

"공정 기록 탐색기"의 빨간색 삼각형 메뉴에는 다음 옵션이 포함되어 있습니다.

최저 수율 수준 각 X, 공정 변수의 각 수준에 대한 수율 요약 통계량 테이블을 표시하거나 숨깁니다. 각 수준에 대해 제공되는 요약 통계량은 작업 수와 수율의 평균, 표준 오차 및 표준편차입니다. 기본적으로 이 테이블은 평균 수율에 따라 오름차순으로 정렬됩니다.

시간 필터를 사용한 최저 수율 수준 수율이 가장 낮은 상위 50 개 수준의 시간 구간에 대한 요약 통계량 테이블을 표시하거나 숨깁니다. 이 루틴은 구간 내 평균 수율을 최소화하는 "최저 수율 수준" 테이블에서 상위 50 개의 각 수준에 대한 시간 구간을 검색합니다. 각 구간은 해당 수

준의 작업을 25% 이상 포함해야 합니다. 각 수준에 대해 제공되는 요약 통계량은 식별된 구간의 작업 수, 식별된 구간의 평균 수율, 수준의 첫 번째 작업과 마지막 작업에 대한 타임스탬프 (첫 번째 시간, 마지막 시간), 식별된 구간의 첫 번째 작업과 마지막 작업에 대한 타임스탬프 (런시작 시간, 런종료 시간) 입니다.

단계별 회귀 단계별 회귀 분석을 위한 옵션을 지정할 수 있는 "단계별 회귀" 창을 표시합니다. 단계별 회귀를 사용하여 공정에서 수율 감소의 원인이 되는 문제를 찾을 수 있습니다. 각 회귀 단계에서 음수 수율이 가장 큰 향에 따라 향이 입력됩니다. 이렇게 하면 수율에 가장 큰 영향을 미치는 공정 단계를 식별하는 데 도움이 됩니다. 단계별 회귀는 각 회귀 단계에서 향이 모형에 입력된 순서를 보여 주는 테이블을 제공합니다.

참고: 시작 창에서 "목표: Y 최소화" 옵션을 선택한 경우에는 단계별 회귀가 수율 증가의 원인이 되는 향을 감지하도록 조정됩니다.

목표 단계별 회귀 모형에 사용되는 기준을 지정합니다. "최대 개별 차이" 옵션은 개별적으로 수율을 가장 많이 변경시키는 향을 식별합니다. "최대 전체 차이" 옵션은 전체 차이 (차이에 영향을 받는 유닛 수를 곱한 값) 를 가장 크게 만드는 향을 식별합니다. "일반적 예측" 옵션은 기존의 단계별 회귀 모형을 생성합니다.

시간 필터링 단계별 회귀 모형에서 성분의 시작 시간을 사용할 수 있습니다. "없음" 옵션을 선택하면 전체 기간에 대한 평균이 시간으로 사용됩니다. "시작 시간" 옵션을 선택하면 특정 시작 시간에 성분이 실패할 수 있습니다. 시작 시간은 음수 수율 효과를 최대화하는 시간입니다. "시작 시간" 옵션을 선택하면 "단계별 회귀" 보고서 테이블에 "시작 시간" 열이 포함됩니다.

X 변환 단계별 회귀에서 X, 공정 변수에 대해 일반 개수 데이터 또는 로그 (개수 +1) 데이터를 사용할 수 있습니다. "로그 (개수 +1)" 옵션은 빈도 수가 높으면 회귀 분석에서 선형 영향력이 적도록 회귀 분석의 변수를 변환합니다. "없음" 옵션을 선택하면 높은 빈도 수가 회귀 분석에서 선형 영향력을 가집니다.

회귀 단계 수 단계별 회귀 분석의 단계 수입니다. 기본값은 10 입니다.

그림 27.4 단계별 회귀 창

단계별 옵션 지정

목표

☒ 최대 개별 차이
☐ 최대 전체 차이
☐ 일반적 예측

시간 필터링

☒ 없음
☐ 시작 시간

X 변환

☒ 없음
☐ 로그(개수+1)

회귀 단계 수

10

대기 시간 분석 (타임스탬프 변수를 두 개 지정한 경우에만 사용 가능) 각 고유 ID 는 한 번에 하나의 단계 수준만 진행할 수 있습니다. 또한 각 단계는 한 번에 하나의 ID 만 처리할 수 있습니다. 대기 시간 분석은 각 ID 에 대한 대기 시간 요약 통계량 테이블을 제공합니다. 이 테이블에서 대기 시간은 ID 가 한 단계에서 완료된 후 다른 단계를 시작할 때까지의 시간입니다. 다른 대기 시간 분석 테이블도 제공됩니다. 이러한 테이블에는 각 단계 변수의 수준 내에서 각 ID 에 대한 대기 시간 요약 통계량이 포함됩니다. 각 테이블에서 대기 시간은 ID 가 단계 변수의 특정 수준에서 완료된 후 동일한 수준을 다시 진행할 때까지의 시간입니다. 이 테이블에는 다음 통계량이 제공됩니다.

대기 시간 합계 주어진 ID 및 단계 변수 수준 조합에 대한 대기 시간의 합입니다.

대기 시간 개수 주어진 ID 및 단계 변수 수준 조합에 대한 대기 시간의 수입니다.

평균 대기 시간 주어진 ID 및 단계 변수 수준 조합에 대한 대기 시간의 평균입니다.

최대 대기 시간 주어진 ID 및 단계 변수 수준 조합에 대한 최대 대기 시간입니다.

전환 분석 결과 수열에 문제가 되는 성분 간 전환을 식별하는 데 유용한 보고서를 표시합니다. 이 보고서는 공정에서 정렬 및 등록 문제를 식별하는 데 도움이 될 수 있습니다. 예를 들어 공정 열의 일부 수준이 해당 공정 열의 다른 수준에 대해 정렬되지 않았을 수 있습니다. 전환 문제가 있으면 특정 수준을 반복하는 항목이 공정 열의 여러 수준을 거치는 항목보다 수율이 더 높을 것으로 기대할 수 있습니다.

이 옵션을 선택하면 분석에 포함할 하나 이상의 X, 공정 열과 각 열에서 하나 이상의 수준을 선택해야 합니다. 분석을 위해 지정한 각 열에 대해 네 개의 테이블이 포함된 보고서가 나타납니다. 첫 번째 테이블에는 열의 특정 수준에서 열의 다른 특정 수준으로 전환하는 경우의 수율과 개수가 표시됩니다. 두 번째 테이블에는 열의 특정 수준에서 외부로 전환하는 경우의 수율과 개수가 표시됩니다. 세 번째 테이블에는 열의 특정 수준으로 전환하는 경우의 수율과

개수가 표시됩니다. 네 번째 테이블에는 유닛이 열의 특정 수준을 반복하는 전환에 대한 수
율과 개수가 표시됩니다.

순서 교호작용 (시작 창에서 순서 변수를 지정한 경우에만 사용 가능) 순서 열을 기준으로 수율에
대한 일원 분석 보고서를 표시합니다. 각 X_i 공정 열에 대한 테이블이 보고서에 포함됩니다. 각
테이블에서 공정 열의 수준이 해당 분산 분석 모형의 LogWorth 값에 따라 내림차순으로 나열
되어 있습니다. 이 테이블에는 공정 열의 각 수준에 대한 p 값과 히트맵도 포함됩니다.

개수 테이블 저장 X_i 공정 변수의 각 수준을 거친 유닛 수가 포함된 새 데이터 테이블을 저장합니
다. 개수 테이블에는 ID 값에 의해 결정되는 각 고유 유닛에 대한 행이 있습니다.

로그 개수 테이블 저장 로그 (개수 +1) 항목을 제외하고 개수 테이블과 동일한 형식의 새 데이터
테이블을 저장합니다.

대기 시간 저장 (타임스탬프 변수를 두 개 지정한 경우에만 사용 가능) 대기 시간 분석 보고서의
각 테이블을 별도의 데이터 테이블에 저장합니다.

부록 A

통계 상세 정보 예측 및 전문 모델링

이 부록에서는 각 플랫폼에서 사용할 수 있는 적합 측도, 교차 검증 유형 및 트리 방법의 예측 확률에 대해 설명합니다.

목차

엔트로피 R^2 519

JMP 모델링의 검증 519

 검증 열 역할..... 520

 K 폴드 및 홀드백 검증..... 522

트리 방법의 예측 확률..... 523

엔트로피 R^2

엔트로피 R^2 통계량은 훈련 데이터 집합과 검증 및 테스트 데이터 집합 (검증이 사용되는 경우)에 대해 계산됩니다. 훈련 데이터 집합에 대한 엔트로피 R^2 은 다음과 같이 계산됩니다.

- 훈련 데이터 집합을 사용하여 지정된 모형이 적합됩니다.
- 모형을 기반으로 반응의 각 수준에 대한 예측 확률이 계산됩니다.
- 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 반응 수준에 대한 예측 확률을 사용하여 훈련 데이터 집합의 관측값에 대한 가능도가 계산됩니다. 이를 $Likelihood\_Reduced_{Training}$ 이라고 합니다.
- 훈련 데이터 집합에 대한 엔트로피 R^2 은 다음과 같습니다.

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

검증 데이터 집합 및 테스트 데이터 집합의 엔트로피 R^2 은 훈련 데이터 집합의 엔트로피 R^2 과 유사한 방식으로 계산됩니다.

JMP 모델링의 검증

과대적합을 방지하고 양호한 예측 모형을 선택하기 위해 모델링 전에 데이터를 여러 데이터 집합으로 분할할 수 있습니다. 이 프로세스는 원래 데이터의 일부를 사용하여 모수를 추정하고 나머지 데이터를 사용하여 모수를 조정 또는 평가하거나 둘 다 수행합니다. JMP Pro에서는 다음과 같은 방법을 사용하여 데이터를 두 개 또는 세 개의 집합으로 분할할 수 있습니다.

훈련 및 평가 데이터를 훈련 데이터 집합과 검증 데이터 집합이라는 두 집합으로 분할합니다. 훈련 데이터 집합은 모형 모수를 추정하는 데 사용됩니다. 검증 데이터 집합은 적합 모형의 성능을 독립적으로 평가하는 데 사용됩니다.

훈련 및 조정 데이터를 훈련 데이터 집합과 검증 데이터 집합이라는 두 집합으로 분할합니다. 훈련 데이터 집합은 모형 모수를 추정하는 데 사용됩니다. 검증 데이터 집합은 모형 적합 알고리즘에서 모형 모수를 조정하고 궁극적으로 예측 능력이 양호한 모형을 선택하는 데 사용됩니다. 이 경우에는 독립적인 모형 평가가 수행되지 않습니다.

훈련, 조정 및 평가 데이터를 훈련 데이터 집합, 검증 데이터 집합, 테스트 데이터 집합이라는 세 집합으로 분할합니다. 훈련 데이터 집합은 모형 모수를 추정하는 데 사용됩니다. 검증 데이터 집합은 모형 적합 알고리즘에서 모형 모수를 조정하고 궁극적으로 예측 능력이 양호한 모형을 선택하는 데 사용됩니다. 그런 다음 테스트 데이터 집합을 사용하여 적합 모형의 성능을 독립적으로 평가합니다.

JMP
PRO

검증 열 역할

데이터 파티션을 생성하는 한 가지 방법은 검증 열 역할을 사용하는 것입니다. 검증 열 역할은 해당 열의 값을 사용하여 데이터를 여러 부분으로 나눕니다. 이 열은 플랫폼 시작 창에서 "검증" 역할을 사용하여 할당합니다. 검증 열을 생성하는 방법에 대한 자세한 내용은 “[검증 열 생성](#)”에서 확인하십시오.

검증 열 역할은 JMP Pro에서만 사용할 수 있습니다. JMP의 경우 “[제외된 행을 검증 홀드백으로 사용](#)”의 내용을 참조하십시오.

주의: 검증 열 사용은 플랫폼에 따라 다릅니다. 플랫폼마다 검증 열의 수준을 다르게 사용합니다. [표 A.1](#)의 "참고" 내용을 참조하십시오.

표 A.1 플랫폼별 검증 열

플랫폼	훈련 및 평가	훈련 및 조정	훈련, 조정 및 평가	참고
모형 적합				
최소 제곱 적합	예	아니요	아니요	수준 수가 3 개를 초과하면 검증 열이 무시됩니다.
단계별 회귀	아니요	예	예	수준 수가 3 개를 초과하면 K 폴드 교차 검증이 사용됩니다.
로지스틱 회귀	예	아니요	아니요	수준 수가 3 개를 초과하면 검증 열이 무시됩니다.
일반화 회귀	아니요	예	예	수준 수가 3 개를 초과하면 K 폴드 교차 검증이 사용됩니다.
부분 최소 제곱	아니요	예	예	수준 수가 3 개를 초과하면 K 폴드 교차 검증이 사용됩니다.
예측 모형				
신경망	아니요	예	예	수준 수가 3 개를 초과하면 K 폴드 교차 검증이 사용됩니다.
파티션	아니요	예	예	수준 수가 3 개를 초과하면 가장 작은 3 개의 값을 포함하는 행만 사용됩니다.

플랫폼	훈련 및 평가	훈련 및 조정	훈련, 조정 및 평가	참고
붓스트랩 포레스트	아니요	예	예	수준 수가 3 개를 초과하면 가장 작은 3 개의 값을 포함하는 행만 사용 됩니다.
부스티드 트리	아니요	예	예	수준 수가 3 개를 초과하면 가장 작은 3 개의 값을 포함하는 행만 사용 됩니다.
K 최근접 이웃	아니요	예	예	수준 수가 3 개를 초과하면 가장 작은 3 개의 값을 포함하는 행만 사용 됩니다.
Naive Bayes	아니요	예	예	수준 수가 3 개를 초과하면 가장 작은 3 개의 값을 포함하는 행만 사용 됩니다.
서포트 벡터 머신	예	예	예	수준 수가 3 개를 초과하면 K 폴드 교차 검증이 사용됩니다.
전문 모형				
함수 데이터 탐색기	예	아니요	아니요	그룹별 랜덤 검증 열로 생성해야 합니다. 수준 수가 2 개를 초과하면 가장 작은 값으로 훈련 데이터 집합이 정의 되고 다른 모든 값으로 검증 데이터 집합이 정의 됩니다.
다변량 모형				
판별	예	예	예	수준 수가 3 개를 초과하면 가장 작은 3 개의 값을 포함하는 행만 사용 됩니다.
부분 최소 제곱	아니요	예	예	수준 수가 3 개를 초과하면 K 폴드 교차 검증이 사용됩니다.
Uplift	아니요	예	예	수준 수가 3 개를 초과하면 가장 작은 3 개의 값을 포함하는 행만 사용 됩니다.

K 폴드 및 홀드백 검증

검증 열 역할은 데이터를 교차 검증 데이터 집합으로 분할하기 위한 프레임워크를 제공합니다. 또한 일부 JMP 플랫폼에서는 K 폴드 및 다양한 유형의 홀드백 검증을 지원합니다.

K 폴드 교차 검증 원래 데이터를 K 개의 부분집합으로 나눕니다. 그러면 K 개의 각 집합이 나머지 데이터에 대한 모형 적합을 검증하는 데 사용되며 총 K 개 모형을 적합시킵니다. 최량 검증 통계량을 생성하는 모형이 최종 모형으로 선택되고, 해당 모형을 생성하는 데 사용되지 않은 폴드는 보고서에 표시된 검증 데이터 집합 성능 통계량을 제공합니다.

참고: 일부 플랫폼의 경우 모형 제어판에서 K 폴드 교차 검증을 지정해야 합니다. 다른 플랫폼의 경우 플랫폼 시작 창에서 K 폴드 교차 검증을 지정해야 합니다. 수준이 네 개 이상 포함된 검증 열을 통해 K 폴드 교차 검증을 지정해야 하는 플랫폼도 있습니다.

랜덤 검증 홀드백 (특정 플랫폼의 시작 옵션으로 사용 가능) 원래 데이터를 무작위로 훈련 데이터 집합과 검증 데이터 집합으로 나눕니다. 테스트 데이터 집합을 포함할 수도 있습니다. 원래 데이터 중 각 데이터 집합에 사용할 데이터의 비율을 지정할 수 있습니다.

Leave-One-Out 검증 홀드백 (특정 플랫폼의 옵션으로 사용 가능) 관측값을 한 번에 하나씩 제외하면서 모형을 반복적으로 적합시킵니다. Leave-One-Out 검증은 잭나이프 절차라고도 합니다.

제외된 행을 검증 홀드백으로 사용 데이터 테이블에서 제외된 행을 검증 홀드백 집합으로 사용합니다. JMP Pro의 경우 플랫폼 환경 설정에서 이 옵션을 선택하여 사용할 수 있습니다.

참고: 제외된 행을 검증 홀드백 집합으로 사용할 수 있는 플랫폼에서는 시작 창에서 검증 열 또는 검증 비율을 지정하지 않은 경우에만 제외된 행이 사용됩니다.

표 A.2 플랫폼별 K 폴드 및 홀드백 검증

플랫폼	제외된 행을 검증 홀드백으로 사용	랜덤 검증 홀드백	Leave-One-Out 홀드백	K 폴드 교차 검증
모형 적합				
최소 제곱 적합	아니요	아니요	아니요	아니요
단계별 회귀	아니요	아니요	아니요	예 (연속형 반응 모형의 경우만 해당)
로지스틱 회귀	아니요	아니요	아니요	아니요
JMP PRO 일반화 회귀	아니요	예	예	예 (모형 제어판을 통해)
JMP PRO 부분 최소 제곱	아니요	예	예	예 (모형 제어판을 통해)

플랫폼	제외된 행을 검증 홀드백으로 사용	랜덤 검증 홀드백	Leave-One-Out 홀드백	K 폴드 교차 검증
예측 모형				
신경망	예	예	아니요	예(모형 시작 창 또는 검증 열을 통해)
파티션	예	예	아니요	예(플랫폼 환경 설정에서 옵션 선택)
 붓스트랩 포레스트	예	예	아니요	아니요
 부스티드 트리	예	예	아니요	아니요
 K 최근접 이웃	예	예	아니요	아니요
 Naive Bayes	예	예	아니요	아니요
 서포트 벡터 머신	아니요	예	아니요	예(모형 시작 창을 통해)
전문 모형				
 함수 데이터 탐색기	아니요	아니요	아니요	아니요
다변량 모형				
판별	선택적	아니요	아니요	아니요
부분 최소 제곱	아니요	예	예	예(모형 시작 창 또는 검증 열을 통해)
 Uplift	아니요	예		아니요

트리 방법의 예측 확률

결정 트리 및 붓스트랩 포레스트 방법에 대한 예측 확률은 아래 설명과 같이 확률 통계량에 의해 계산됩니다.

결정 트리의 범주형 반응에 대해 "분할 확률 표시" 명령은 다음 통계량을 표시합니다.

비율 각 반응 수준에 대한 노드의 관측값 비율입니다.

확률 트리의 해당 노드에 대한 예측 확률입니다. 주어진 노드에서 i 번째 반응 수준에 대한 확률을 계산하는 방법은 다음과 같이 정의됩니다.

$$\text{Prob}_i = \frac{n_i + \text{prior}_i}{\sum (n_i + \text{prior}_i)}$$

여기서 총합은 모든 반응 수준을 대상으로 합니다. n_i 는 i 번째 반응 수준에 대한 노드의 관측값 수이고, prior_i 는 i 번째 반응 수준에 대한 사전 확률이며 다음과 같이 계산됩니다.

$$\text{prior}_i = \lambda p_i + (1-\lambda)P_i$$

여기서 p_i 는 상위 노드의 prior_i , P_i 는 상위 노드의 Prob_i , λ 는 현재 0.9 로 설정된 가중치 요인입니다.

확률 (Prob) 계산 방법은 예측 확률이 항상 0 이 아님을 보장합니다.

부록 B

참조 자료

다음 자료가 예측 및 전문 모델링에서 참조됩니다.

- Agrawal, R., and Srikant, R. (1994). "Fast Algorithms for Mining Association Rules." In *Proceedings of the 20th VLDB Conference*. Santiago, Chile: IBM Almaden Research Center. Accessed July 5, 2016. <https://web.stanford.edu/class/cs345d-01/rl/ar-mining.pdf>.
- Bates, D. M., and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. New York: John Wiley & Sons.
- Benford, F. (1938). "The law of anomalous numbers." *Proceedings of the American philosophical society*, 551–572.
- Benjamini, Y., and Hochberg, Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society, Series B* 57:289–300.
- Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*. 3rd ed. Englewood Cliffs, NJ: Prentice-Hall.
- Candes, E. J., Li, X., Ma, Y., and Wright, J. (2009). "Robust Principal Component Analysis?" *Journal of the ACM*, 58:1–37.
- Cleveland, W. S. (1994). *Visualizing Data*, Summit, NJ: Hobart Press.
- Conover, W. J. (1999). *Practical Nonparametric Statistics*. 3rd ed. New York: John Wiley & Sons.
- Cureton, E. E. (1967). "The Normal Approximation to the Signed-Rank Sampling Distribution when Zero Differences are Present." *Journal of the American Statistical Association* 62: 319, 1068–1069.
- Donoho, D. L. (1995). "De-noising by soft-thresholding." *IEEE transactions on information theory*, 41(3), 613–627.
- Efron, B. (1981). "Nonparametric Standard Errors and Confidence Intervals." *Canadian Journal of Statistics*, 9:139–172.
- Geladi, P., MacDougall, D., & Martens, H. (1985). "Linearization and scatter-correction for near-infrared reflectance spectra of meat." *Applied spectroscopy*, 39(3), 491–500.
- Hahsler, M. (2015). "A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules." Accessed September 17, 2018. http://michael.hahsler.net/research/association_rules/measures.html.
- Haldane, J. B. (1956). "The estimation and significance of the logarithm of a ratio of frequencies." *Ann Hum Genet*, 20, 309–311.
- Hand, D. J., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. Cambridge, MA: MIT Press.

- Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer-Verlag.
- Hawkins D. M., and Kass G. V. (1982). "Automatic Interaction Detection." In *Topics in Applied Multivariate Analysis*, edited by D. M. Hawkins, 267–300. Cambridge: Cambridge University Press.
- Huber, P. J., and Ronchetti, E. M. (2009). *Robust Statistics*. 2nd ed. New York: John Wiley & Sons.
- Hyndman, R. J., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Berlin: Springer-Verlag.
- Jolliffe, I. T. (2002). *Principal Component Analysis*. New York: Springer-Verlag.
- Kass, G. V. (1980). "An Exploratory Technique for Investigating Large Quantities of Categorical Data." *Journal of the Royal Statistical Society, Series C* 29:119–127
- Lehman, E. L. (2006). *Nonparametrics: Statistical Methods Based on Ranks*. 2nd ed. New York: Springer.
- Lin, Z., Chen, M., and Ma, Y. (2013). The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv preprint arXiv:1009.5055.
- Mason, R. L., and Young, J. C. (2002). *Multivariate Statistical Process Control with Industrial Applications*. Philadelphia: SIAM.
- McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. London: Chapman & Hall.
- Nagelkerke, N. J. D. (1991). "A Note on a General Definition of the Coefficient of Determination." *Biometrika* 78:691–692.
- Nason, G. P. (Ed.). (2008). *Wavelet methods in statistics with R*. New York, NY: Springer New York.
- Nelder, J. A., and Wedderburn, R. W. M. (1972). "Generalized Linear Models." *Journal of the Royal Statistical Society, Series A* 135:370–384.
- Paixão, P., Gouveia, L. F., Silva, N., & Morais, J. A. (2017). "Evaluation of dissolution profile similarity—Comparison between the f_2 , the multivariate statistical distance and the f_2 bootstrapping methods." *European Journal of Pharmaceutics and Biopharmaceutics*, 112, 67–74.
- Parker, R. J. (2015). *Efficient Computational Methods for Large Spatial Data*. Ph.D. diss., Department of Statistics, North Carolina State University. Accessed June 30, 2016. <https://repository.lib.ncsu.edu/ir/bitstream/1840.16/10572/1/etd.pdf>.
- Platt, J. (1998). *Sequential minimal optimization: A fast algorithm for training support vector machines*. Technical Report MST-TR-98-14, Microsoft Research. <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-98-14.pdf>.
- Qian, P. Z., Huaiqing, W., and Wu, C. F. (2012). "Gaussian process models for computer experiments with qualitative and quantitative factors." *Technometrics* 50:383–396.
- Ramsay, J. O., and Silverman, B. W. (2005). *Functional Data Analysis*. 2nd ed. New York: Springer.
- Ratkowsky, D. A. (1990). *Handbook of Nonlinear Regression Models*. New York: Marcel Dekker.

- Sall, J. (2002). "Monte Carlo Calibration of Distributions of Partition Statistics." SAS Institute Inc., Cary, NC. Accessed July 29, 2015.
<https://www.jmp.com/content/dam/jmp/documents/en/white-papers/montecarlocal.pdf>.
- Santer, T., Williams, B., and Notz, W. (2003). *The Design and Analysis of Computer Experiments*. New York: Springer-Verlag.
- SAS Institute Inc. (2020). *SAS/ETS 15.2 User's Guide*. Cary, NC: SAS Institute Inc.
<https://go.documentation.sas.com/api/docsets/etsug/15.2/content/etsug.pdf>.
- Savitzky, A., and Golay, M. J. (1964). "Smoothing and differentiation of data by simplified least squares procedures." *Analytical Chemistry* 36(8), 1627–1639.
- Schäfer, 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.
- Schuirman, D. J. (1987). "A Comparison of the Two One-sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability." *Journal of Pharmacokinetics and Biopharmaceutics* 15:657–680.
- Shiskin, J., Young, A. H., and Musgrave, J. C. (1967). *The X-11 Variant of the Census Method II Seasonal Adjustment Program*. Technical Report 15, US Department of Commerce, Bureau of the Census.
- Shmueli, G., Patel, N. R., and Bruce, P. C. (2010). *Data Mining For Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*. 2nd ed. Hoboken, NJ: John Wiley & Sons.
- Shmueli, G., Bruce, P. C., Stephens M. L., and Patel, N. R. (2017). *Data Mining For Business Intelligence: Concepts, Techniques, and Applications with JMP Pro*. Hoboken, NJ: John Wiley & Sons.
- Smyth, Gordon K. (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." *Statistical applications in genetics and molecular biology* 3(1).
- Westfall, P. H., Tobias, R. D., and Wolfinger, R. D. (2011). *Multiple Comparisons and Multiple Tests Using SAS*. 2nd ed. Cary, NC: SAS Institute Inc.

기술 라이선스 고지 사항

- Scintilla is Copyright © 1998–2017 by Neil Hodgson <neilh@scintilla.org>.

All Rights Reserved.

Permission to use, copy, modify, and distribute this software and its documentation for any purpose and without fee is hereby granted, provided that the above copyright notice appear in all copies and that both that copyright notice and this permission notice appear in supporting documentation.

NEIL HODGSON DISCLAIMS ALL WARRANTIES WITH REGARD TO THIS SOFTWARE, INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS, IN NO EVENT SHALL NEIL HODGSON BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.

- Progress® Telerik® UI for WPF: Copyright © 2008-2019 Progress Software Corporation. All rights reserved. Usage of the included Progress® Telerik® UI for WPF outside of JMP is not permitted.
- ZLIB Compression Library is Copyright © 1995-2005, Jean-Loup Gailly and Mark Adler.
- Made with Natural Earth. Free vector and raster map data @ naturalearthdata.com.
- Packages is Copyright © 2009–2010, Stéphane Sudre (s.sudre.free.fr). All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

Neither the name of the WhiteBox nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES

(INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- iODBC software is Copyright © 1995–2006, OpenLink Software Inc and Ke Jin (www.iodbc.org). All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
- Neither the name of OpenLink Software Inc. nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL OPENLINK OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- This program, “bzip2”, the associated library “libbzip2”, and all documentation, are Copyright © 1996–2019 Julian R Seward. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. The origin of this software must not be misrepresented; you must not claim that you wrote the original software. If you use this software in a product, an acknowledgment in the product documentation would be appreciated but is not required.
3. Altered source versions must be plainly marked as such, and must not be misrepresented as being the original software.

4. The name of the author may not be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE AUTHOR "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE AUTHOR BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Julian Seward, jseward@acm.org

bzip2/libbzip2 version 1.0.8 of 13 July 2019

- R software is Copyright © 1999–2012, R Foundation for Statistical Computing.
- MATLAB software is Copyright © 1984-2012, The MathWorks, Inc. Protected by U.S. and international patents. See www.mathworks.com/patents. MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See www.mathworks.com/trademarks for a list of additional trademarks. Other product or brand names may be trademarks or registered trademarks of their respective holders.
- libopc is Copyright © 2011, Florian Reuter. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and / or other materials provided with the distribution.
- Neither the name of Florian Reuter nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF

USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- libxml2 - Except where otherwise noted in the source code (e.g. the files hash.c, list.c and the trio files, which are covered by a similar license but with different Copyright notices) all the files are:

Copyright © 1998–2003 Daniel Veillard. All Rights Reserved.

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the “Software”), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED “AS IS”, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL DANIEL VEILLARD BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Except as contained in this notice, the name of Daniel Veillard shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization from him.

- Regarding the decompression algorithm used for UNIX files:

Copyright © 1985, 1986, 1992, 1993

The Regents of the University of California. All rights reserved.

THIS SOFTWARE IS PROVIDED BY THE REGENTS AND CONTRIBUTORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE REGENTS OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

3. Neither the name of the University nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

- Snowball is Copyright © 2001, Dr Martin Porter, Copyright © 2002, Richard Boulton. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

3. Neither the name of the copyright holder nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- Pako is Copyright © 2014–2017 by Vitaly Puzrin and Andrei Tuputcyn.

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED “AS IS”, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

- HDF5 (Hierarchical Data Format 5) Software Library and Utilities are Copyright 2006–2015 by The HDF Group. NCSA HDF5 (Hierarchical Data Format 5) Software Library and Utilities Copyright 1998-2006 by the Board of Trustees of the University of Illinois. All rights reserved. DISCLAIMER: THIS SOFTWARE IS PROVIDED BY THE HDF GROUP AND THE CONTRIBUTORS “AS IS” WITH NO WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED. In no event shall The HDF Group or the Contributors be liable for any damages suffered by the users arising out of the use of this software, even if advised of the possibility of such damage.
- agl-aglfn technology is Copyright © 2002, 2010, 2015 by Adobe Systems Incorporated. All Rights Reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
- Neither the name of Adobe Systems Incorporated nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- dmlc/xgboost is Copyright © 2019 SAS Institute.

Licensed under the Apache License, Version 2.0 (the “License”); you may not use this file except in compliance with the License. You may obtain a copy of the License at

<http://www.apache.org/licenses/LICENSE-2.0>

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

- libzip is Copyright © 1999–2019 Dieter Baron and Thomas Klausner.

This file is part of libzip, a library to manipulate ZIP archives. The authors can be contacted at <libzip@nih.at>.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
3. The names of the authors may not be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE AUTHORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE AUTHORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- OpenNLP 1.5.3, the pre-trained model (version 1.5 of en-parser-chunking.bin), and dmlc/xgboost Version .90 are licensed under the Apache License 2.0 are Copyright © January 2004 by Apache.org.

You may reproduce and distribute copies of the Work or Derivative Works thereof in any medium, with or without modifications, and in Source or Object form, provided that You meet the following conditions:

- You must give any other recipients of the Work or Derivative Works a copy of this License; and
 - You must cause any modified files to carry prominent notices stating that You changed the files; and
 - You must retain, in the Source form of any Derivative Works that You distribute, all copyright, patent, trademark, and attribution notices from the Source form of the Work, excluding those notices that do not pertain to any part of the Derivative Works; and
 - If the Work includes a “NOTICE” text file as part of its distribution, then any Derivative Works that You distribute must include a readable copy of the attribution notices contained within such NOTICE file, excluding those notices that do not pertain to any part of the Derivative Works, in at least one of the following places: within a NOTICE text file distributed as part of the Derivative Works; within the Source form or documentation, if provided along with the Derivative Works; or, within a display generated by the Derivative Works, if and wherever such third-party notices normally appear. The contents of the NOTICE file are for informational purposes only and do not modify the License. You may add Your own attribution notices within Derivative Works that You distribute, alongside or as an addendum to the NOTICE text from the Work, provided that such additional attribution notices cannot be construed as modifying the License.
 - You may add Your own copyright statement to Your modifications and may provide additional or different license terms and conditions for use, reproduction, or distribution of Your modifications, or for any such Derivative Works as a whole, provided Your use, reproduction, and distribution of the Work otherwise complies with the conditions stated in this License.
- LLVM is Copyright © 2003–2019 by the University of Illinois at Urbana-Champaign. Licensed under the Apache License, Version 2.0 (the “License”); you may not use this file except in compliance with the License. You may obtain a copy of the License at:
<http://www.apache.org/licenses/LICENSE-2.0>
Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.
 - clang is Copyright © 2007–2019 by the University of Illinois at Urbana-Champaign. Licensed under the Apache License, Version 2.0 (the “License”); you may not use this file except in compliance with the License. You may obtain a copy of the License at:
<http://www.apache.org/licenses/LICENSE-2.0>
Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS”, WITHOUT WARRANTIES OR CONDITIONS OF

ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

- lld is Copyright © 2011–2019 by the University of Illinois at Urbana-Champaign.

Licensed under the Apache License, Version 2.0 (the “License”); you may not use this file except in compliance with the License. You may obtain a copy of the License at:

<http://www.apache.org/licenses/LICENSE-2.0>

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS”, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

- libcurl is Copyright © 1996–2021, Daniel Stenberg, daniel@haxx.se, and many contributors, see the THANKS file. All rights reserved.

Permission to use, copy, modify, and distribute this software for any purpose with or without fee is hereby granted, provided that the above copyright notice and this permission notice appear in all copies.

THE SOFTWARE IS PROVIDED “AS IS”, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OF THIRD PARTY RIGHTS. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Except as contained in this notice, the name of a copyright holder shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization of the copyright holder.

- On the Windows operating system, JMP utilizes the OpenBLAS library. OpenBLAS is licensed under the 3-clause BSD license. Full license text follows: Copyright © 2011-2015, The OpenBLAS Project

All rights reserved. Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
3. Neither the name of the OpenBLAS project nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE OPENBLAS PROJECT OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.