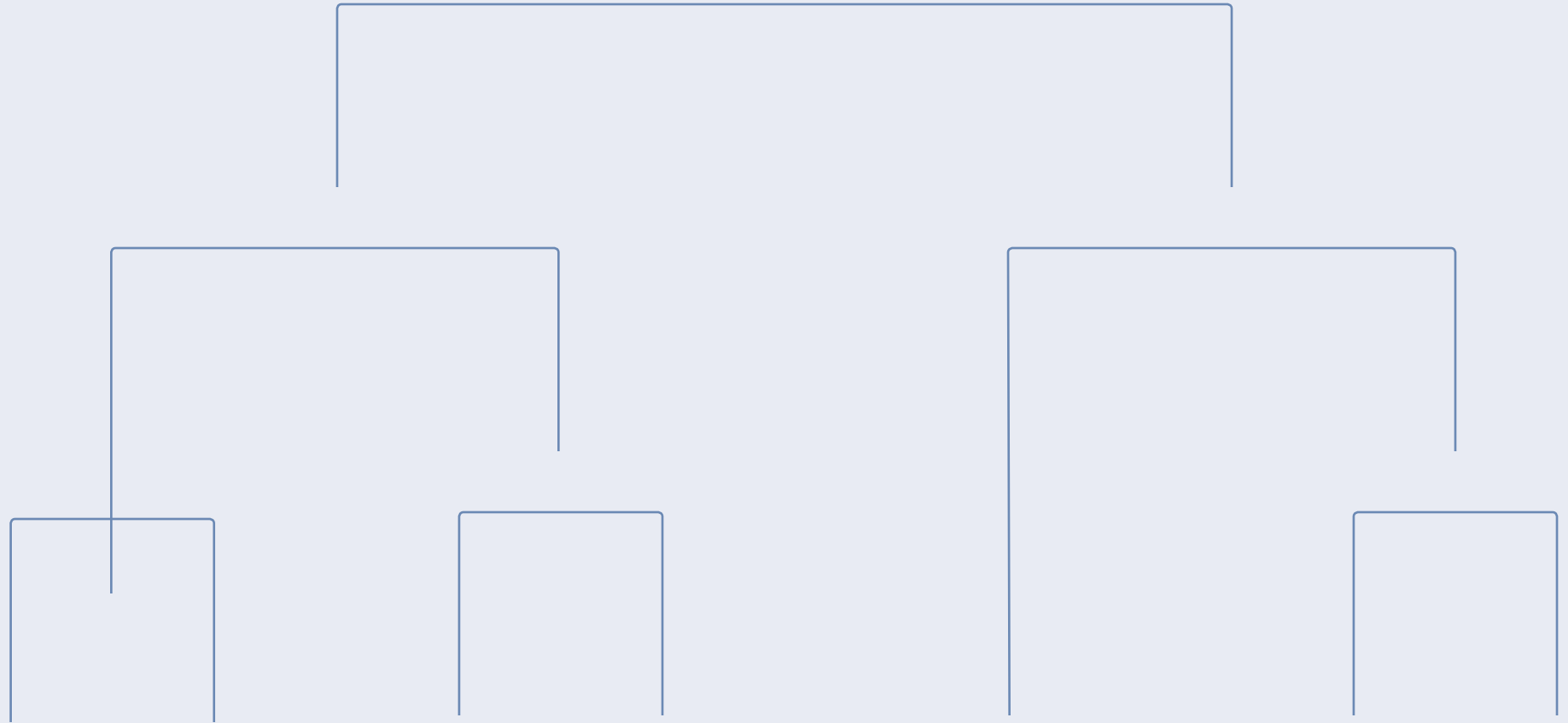
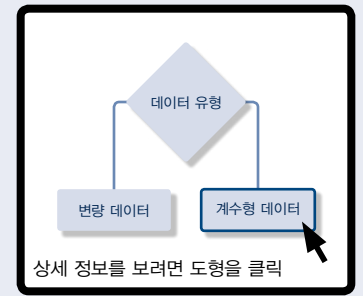


METHOD CHOOSER
(통계 방법 길잡이)

Minitab 15 ™
Statistical Software

회귀 & 분산 분석 편
(Regression and ANOVA)

회귀 & 분산 분석

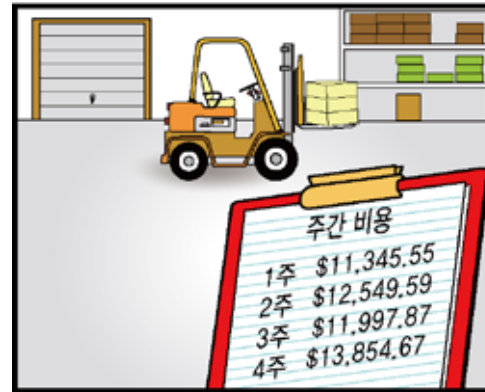


연속형 반응값 혹은 범주형 반응값인가?

반응값(Y)의 데이터 유형

부품 혹은 프로세스의 특성을 측정한다.
 하나 혹은 둘 이상의 X 변수를 이용하여 평균 반응값을 추정한다.

예시
 재무 분석가는 대형 할인점의 총 주간 비용을 조사한다.
 분석가는 주간 비용이 선적 건수, 근무 시간 그리고 전력 사용량과 얼마나 많은 관련이 있는지를 파악해 보고자 한다.

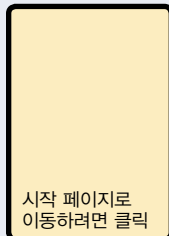
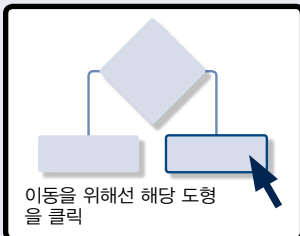


반응값을 나쁨, 좋음, 우수와 같은 범주로 구분한다.
 하나 혹은 둘 이상의 X 변수를 이용하여 반응값의 각 수준에 대한 확률을 평가한다.

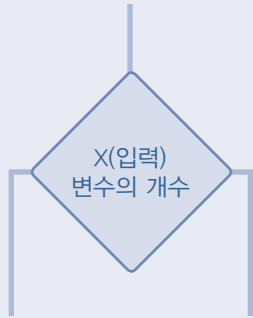
예시
 호텔 매니저는 고객에게 서비스 만족도를 1에서 5점 척도로 평가하게 한다.
 매니저는 객실 타입에 대한 고객 불만족(예시: 1,2점이면 불만족) 확률을 알아보고자 한다.



반응값은 X 변수에 의해 설명되거나 예측하고자 하는 변수이다. 반응값은 흔히 Y 혹은 결과 변수로 불린다.
 연속형 반응 변수는 길이, 무게, 온도와 같은 측정치이다. 이것은 보통 소수 값을 포함한다. 범주형 반응 변수는 원자재 등급과 방법의 종류와 같은 특성이나 상, 중, 하와 같은 비율을 나타낸다.
 만약에 반응 변수가 순서를 매길 수 있는 여러 개의 범주를 가지고 서수로 나타낼 수 있다면 그 변수는 연속형 혹은 범주형으로써 취급될 수 있다. 예를 들어, 신발 브랜드들에 대한 품질 수준을 1에서 10까지 점수를 매긴다. 그 데이터를 연속형 모형으로 분석하면, 'A 브랜드 평균값은 4.4이다'와 같은 결과를 얻을 수 있다. 그러나 범주형(계수형) 모델로 분석하게 되면, 'A 브랜드 품질 수준이 3이하의 확률은 40%이다'와 같은 결과를 얻는다.



모델에 포함하기를 원하는 X 변수의 개수는?



반응값을 설명하기 위해 다른 잠재 X 변수들이 중요하지 않거나, 혹은 가장 단순한 모델을 사용하기를 원하는 경우에, 하나의 X만을 사용한다.

예시

선생님은 표준화된 시험 점수로 미래의 학생의 등급을 예측할 수 있는지 알기를 원한다. 다른 변수들이 학생의 등급에 영향을 미칠 것이라 예상은 되지만 시험 점수와 학생의 등급만으로 관계성을 조사하기를 원한다.



반응값을 적당히 설명하기 위해 하나 이상의 X 변수가 필요하거나 다른 X 변수들의 변동성이 없다는 가정 하에 하나의 X 변수의 효과를 연구하기를 원하는 경우에 모델에서 둘 혹은 셋 이상의 X 변수를 사용한다.

예시

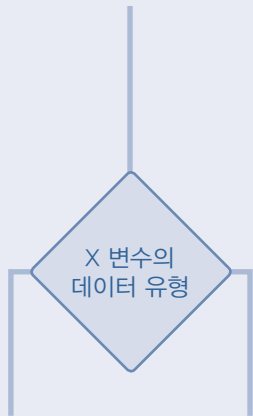
부동산 감정사는 도시 외곽의 콘도 시세에 대하여 다음 요인들이 중대한 영향을 미칠 것으로 판단한다; 평형, 도심으로부터 거리, 인근 주택들의 평균 재산세. 그래서 감정사는 콘도 가격을 예측하기 위한 모델에서 세 개의 X 변수를 사용한다.



X는 반응값을 설명하고 예측하기 위해 사용되는 입력값이다. X가 바뀌면 반응값도 변화게 된다. X는 예측변수, 입력변수 혹은 설명변수라고 불린다. X가 범주형일 때, 이것을 종종 요인(factor)이라고 한다.

만약에 X 변수가 모델에 포함하기에 충분히 중요한지를 모를 경우, 우선은 포함시키는 것이 낫다. 어떤 X 변수가 반응값에 가장 큰 영향을 미치는 지를 파악하여 중대한 X 변수를 파악할 수 있기 때문이다. 그러나 그 프로세스 지식에 근거하여 각 X 변수가 실무적으로 중요한지를 동시에 고려해야 한다.

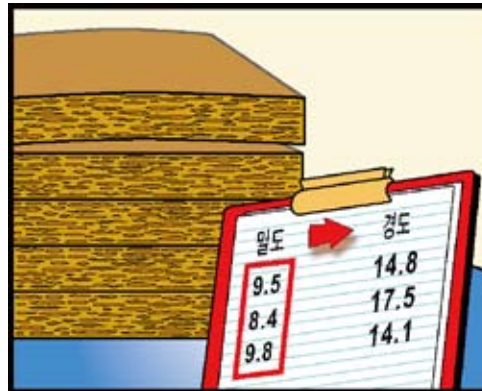
X 변수가 연속형 혹은 범주형인가?



길이, 무게, 온도와 같은 제품 혹은 프로세스의 특성을 측정한다. 데이터는 종종 소수값을 포함한다.

예시

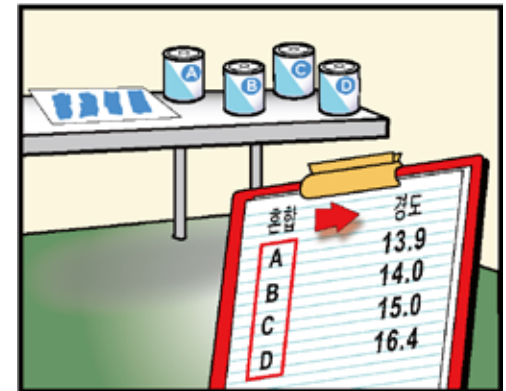
스티로폼 제조사는 재료 입자가 제품 강도에 얼마나 영향을 미치는지 알기를 원한다. 담당자는 입자의 밀도를 증가 혹은 약화시켰을 때 강도가 어떻게 변화 하는지 측정한다.



원자재 등급 혹은 실패, 합격과 같은 조건 혹은 특성 하에서 범주를 구분한다.

예시

화학회사 담당자는 A, B, C, D 타입의 혼합 페인트가 건조된 상태 후 경도에 얼마나 영향을 미칠지 평가하기를 원한다. 검사자는 각 페인트를 금속 조각에 바른 후 건조된 후에 경도를 측정한다.



연속형 X는 종종 예측치(Predictor)라고 불린다. 범주형 X는 종종 요인(factor)이라고 불리고 그것의 범주는 수준이라고 한다.

만약 X가 범주형이고 서열이 있거나, 순서를 갖는 숫자(1부터 10까지의 척도와 같이)으로 재표현 될 수 있는 많은 수준을 포함하고 있다면 그 변수는 연속형 혹은 범주형으로 처리할 수 있다. 만약 연속형으로 처리한다면 X의 모든 연속 값에 대한 반응값을 예측할 수 있다. 만약 범주형으로 처리한다면, X의 각 수준에 대한 평균 반응값을 계산할 수 있다.

만약 X가 선적된 물품 수량 혹은 콜 센터에 걸려온 전화 수와 같은 빈도라면, 비록 그 값이 정수이더라도 연속형처럼 빈도를 다룰 수 있다.

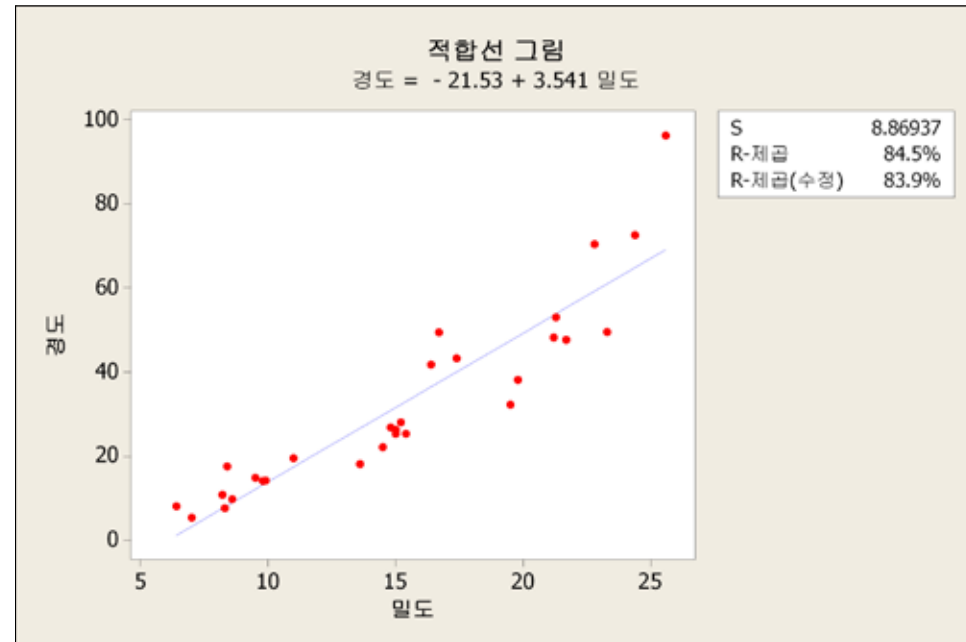
▶ 적합선 그림 (Fitted Line Plot)

적합선 그림은 예측치(연속형 X)와 반응값(연속형 Y) 간의 관계를 조사한다.

예시

스티로폼 제조사의 품질 담당자는 제품의 밀도와 강도 간의 관계를 측정하기를 원한다. 그는 둘 사이의 관계를 파악하기 위해 적합선 그림을 사용한다.

Minitab에서 적합선 그림 메뉴는 통계분석 > 회귀분석 > 적합선 그림 이다



일반적으로, 처음에는 데이터에 직선을 적합 시킨다. 직선은 X, Y 관계에 대해 가장 단순한 모형을 제공한다. 그러나 만약 데이터에 직선이 잘 적합 되지 않는다면, 다음을 시도해 보라.

- 2차, 3차 곡선 적합
- 반응값(Y) 혹은 예측 변수(X)에 로그 변환 적용

▶ 일원 분산 분석 (One-Way ANOVA)

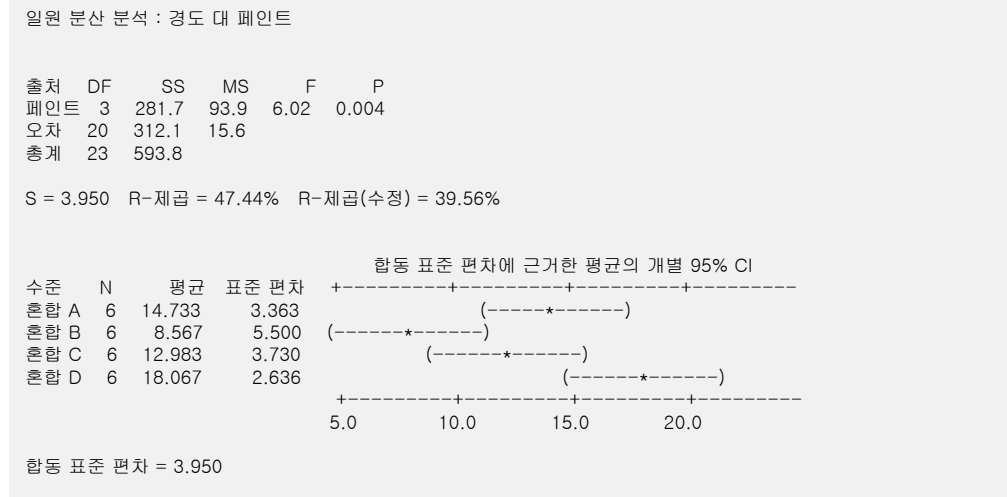
일원 분산 분석은 요인(범주형 X)과 반응값(연속형 Y)간의 관계에 대해 조사한다.

예시

화학회사 품질 담당자는 각 혼합 페인트가 건조된 후 경도를 측정하기를 원한다. 담당자는 페인트의 혼합 유형(X)과 그 경도(Y)를 조사하기 위해 일원 분산 분석을 사용한다.

Minitab에서 일원 분산 분석을 수행하기 위해선 다음과 같다

- 만약 반응값이 첫번째 컬럼에 있고, 요인의 수준이 두번째 컬럼에 입력되어 있다면, 통계분석 > 분산 분석 > 일원 분산 분석을 선택한다
- 만약 반응값이 각 요인 수준별로 각각 다른 열에 입력되어 있다면, 통계분석 > 분산 분석 > 일원 분산 분석(분할된 데이터)를 선택한다.



일원 분산 분석에서 정확한 결과를 얻기 위해선 요인(X)의 각 수준별로 반응값이 정규분포를 따라야 한다. 분석에 대한 가정을 만족하는지를 보기 위해, 통계분석 > 일원 분산 분석 > 그래프 > 잔차 그림 > 네 개 모두를 체크한 후 분석한다.

반응값 데이터가 비정규성이고 요인(X)의 각 수준에서 많은 이상치를 포함한다면, 통계분석 > 비모수 통계학 > Kruskal-Wallis 검정을 고려해 보라.

그룹간에 중요한 차이를 검출하기 위해 얼마나 많은 데이터가 필요한지를 결정하기 위해 통계분석 > 검정력 및 표본 크기 > 일원 분산 분석을 사용하라.

관심 대상인 대표 X 변수가 연속형 혹은 범주형인가?

대표 X 변수의
데이터 유형

길이, 무게, 온도와 같은 부품 혹은 프로세스의 특성을 측정된 데이터이다 그 값은 소수를 종종 포함한다. 방정식으로 X값이 연속된 범위에 대한 반응값을 예측한다.

예시

의사는 환자의 나이와 신체질량지수(BMI)가 입원 기간에 얼마나 연관되어 있는지를 조사하기를 원한다.

나이	BMI	입원기간
17	17.2	3
44	22.6	4
37	24.3	7

원자재 등급, 방법의 종류와 같은 특성 또는 상태, 합격/실패와 같은 비율처럼 범주로 분류된 데이터이다. 각 X 수준에 대한 반응값을 비교한다.

예시

대형 소매점의 담당자는 대금 지급 방법과 지급 요일이 거래 비용에 어떻게 연관되는지 조사하기를 원한다.

요일	지급방법	합계
월	신용카드	\$113.55
화	현금	\$13.99
수	수표	\$66.40

연속형 X는 종종 예측치(Predictor)라고 불린다. 범주형 X는 종종 요인(factor)이라고 불리고 그것의 범주는 수준이라고 한다.

어떤 프로세스에서, 연속형 변수와 범주형 변수 둘 다 반응값을 설명하기 위한 X 변수로 될 수 있다. 그렇다고 하면, 주요한 X 변수가 연속형 혹은 범주형인지를 결정하라.

- 만약 주요한 X 변수가 연속형이면, X의 변화가 반응값에 어떻게 연관되는지 평가하기 위해 예측 모형을 적합 시킬 수 있다. 예를 들어, 혈압과 콜레스테롤 변화가 심장 발작에 얼마나 위험한지 예측 할 수 있다.
- 만약 주요한 X 변수가 범주형이면, 반응값에 대해 각 요인의 수준의 효과와 요인들 간의 상호효과를 평가할 수 있다. 예를 들어, 성별(여자, 남자)과 인종(아프리카-아메리칸, 코카시안, 히스패닉)이 심장 발작에 어떻게 영향을 미치는지 그리고 성별과 인종의 상호효과가 심장발작에 영향을 미치는지를 조사한다.

회귀 분석 (Regression)

회귀분석은 반응값(연속형 Y)과 하나 이상의 예측치(연속형 X 변수들)간의 관계를 조사한다.

예시

병원 담당자는 나이와 신체질량지수(BMI)에 기초해 입원기간을 예측하기를 원한다. 하나의 반응값과 여러 개의 예측치들 간의 관계를 조사하기 위하여 회귀분석을 수행한다. Minitab에서 회귀분석을 수행하기 위해선, 통계분석 > 회귀 분석 > 회귀 분석을 선택한다.

```
회귀 분석: 입원 기간 대 BMI, 나이

회귀 방정식
입원 기간 = - 0.666 + 0.182 BMI + 0.0630 나이

예측
변수      계수   계수 SE      T      P
상수     -0.6660  0.7084     -0.94  0.350
BMI       0.18248  0.02418    7.55  0.000
나이     0.063043  0.005510   11.44  0.000

S = 1.15395  R-제곱 = 62.8%  R-제곱(수정) = 62.1%

분산 분석

출처      DF      SS      MS      F      P
회귀       2     218.22  109.11  81.94  0.000
잔차 오차  97     129.17   1.33
총계      99     347.39
```



회귀 모형은 선형 그리고 다항을 포함할 수 있다. 예를 들어, 입원기간과 신체지수간의 곡선 관계를 모형화하기 위해 BMI² 로써 다항을 포함할 수 있다. 또한 신체지수와 나이간의 교호작용을 설명하기 위한 항을 포함할 수 있다.

회귀분석에서 범주형 예측변수를 포함하기 위해선, 각 X 수준을 의미하는 숫자 값인 지시변수(indicator variables)를 만들어라.

미니탭에서, 계산 > 지시 변수 만들기를 선택한다. 회귀 분석에서 정확한 결과를 얻기 위해선, 데이터는 몇 개의 가정을 만족해야 한다. 분석에 대한 가정이 만족되는지 체크하기 위해선, 통계분석 > 회귀 분석 > 회귀 분석 > 그래프 > 잔차 그림 '네 개 모두' 를 선택한다.

▶ 일반 선형 모형 (General Linear Model)

일반 선형 모형은 하나의 반응값 (연속형 Y)과 하나 이상의 요인 (범주형 X 변수)간의 관계를 조사한다. 또한 모형에서 연속형 X 변수를 공변량으로 포함할 수 있다.

예시

대형 소매점 회계 담당자는 대금 지급 방법, 대금 지급 요일 그리고 각 거래 비용을 조사한다. 담당자는 대금 지급 방법과 지급 요일이 거래 비용과 관련이 있는지 조사하기 위해 일반 선형 모형을 사용한다. Minitab에서 통계분석 > 분산분석 > 일반 선형 모형을 선택한다.

일반 선형 모형: 거래가 대 지불 방법, 요일

요인	유형	수준	값
지불 방법	고정됨	4	수표, 신용카드, 직불카드, 현금
요일	고정됨	7	금요일, 목요일, 수요일, 월요일, 일요일, 토요일, 화요일

거래가에 대한 분산 분석(검정을 위해 수정된 제곱합 사용)

출처	DF	Seq SS	Adj SS	Adj MS	F	P
지불 방법	3	59745.9	57783.8	19261.3	69.90	0.000
요일	6	3331.3	3331.3	555.2	2.01	0.068
오류	130	35821.8	35821.8	275.6		
총계	139	98898.9				

S = 16.5998 R-제곱 = 63.78% R-제곱(수정) = 61.27%

거래가에 대한 비정상적 관측치

관측치	거래가	적합치	SE 적합치	잔차	표준화
					잔차
25	8.1000	39.7587	6.0686	-31.6587	-2.05 R
140	31.5800	63.9319	4.0800	-32.3519	-2.01 R

R은 표준화 잔차가 큰 관측치를 나타냅니다.



일반 선형 모형의 정확한 결과를 얻기 위해선, 데이터는 몇 개의 가정을 만족해야 한다. 분석에 대한 가정이 만족하는지 조사하기 위해선, 통계분석 > 분산분석 > 일반 선형 모형 > 그래프 > 잔차 그림 '네 개 모두' 를 선택한다.

반응값이 몇 개의 범주인가?



반응이 정확히 두 개의 범주로 분류된다 ; 실패/성공, 예/아니오.

예시

시리얼 제조사는 신상품에 대한 광고가 구매에 영향을 미치는지를 조사하고자 한다. 마케팅 담당자는 제품을 구매한 집단을 대상으로 광고를 보았는지, 그래서 구매를 했는지를 문의한다.



반응이 두 개 이상의 범주로 분류된다 ; 나쁨, 보통, 우수, 또는 동, 서, 남, 북.

예시

농업 연구관은 암탉의 몸무게가 달걀의 크기와 관련이 있는지 알기를 원한다. 암탉을 무작위로 선택하여 몸무게를 측정한 후 그 달걀의 크기를 소, 중, 대로 분류한다.



반응값은 X 변수로 설명하거나 예측하려는 변수이다. 반응값은 흔히 Y 혹은 결과변수라고 한다.

▶ 이항 로지스틱 회귀 분석 (Binary Logistic Regression)

이항 로지스틱 회귀분석은 하나 이상의 X 변수와 두 개의 범주를 가지는 범주형 반응값 간의 관계를 조사한다.

예시

시리얼 제조사의 마케팅 담당자는 고객에게 광고를 본 후에 제품을 구매했는지를 질문한다. 광고가 구매에 영향을 주었는지 파악하기 위해서 이항 로지스틱 회귀분석을 이용한다.

미니탭에서, 통계분석 > 회귀 분석 > 이항 로지스틱 회귀분석 을 선택한다.

이항 로지스틱 회귀 분석: 구입 대 광고 시청

연결 함수: 로짓

반응 정보

변수	값	카운트
구입	1	22 (사건)
	0	49
총계		71

로지스틱 회귀 분석 표

예측 변수	계수	계수 SE	Z	P	승산비	95% CI	
						하한	상한
상수	-1.45529	0.419750	-3.47	0.001			
광고 시청							
예	1.21890	0.543589	2.24	0.025	3.38	1.17	9.82

로그 우도 = -41.278

모든 기울기가 0인지 검정: G = 5.341, DF = 1, P-값 = 0.021

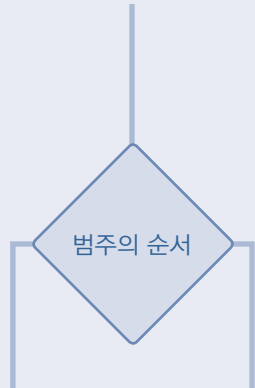
* 참고 * 적합도 검정이 수행되지 않았습니다.

* 참고 * 모형이 모든 자유도를 사용합니다.

연관성 측도:
(반응 변수와 예측 확률 사이)

쌍	번호	백분율	요약 측도	
일치	450	41.7	Somers의 D	0.29
불일치	133	12.3	Goodman-Kruskal 감마	0.54
같은 값	495	45.9	Kendall의 타우-a	0.13
총계	1078	100.0		

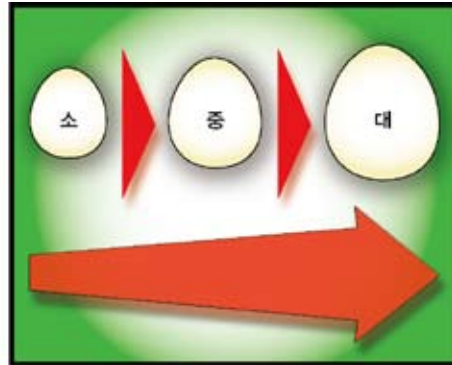
반응 범주가 순서형을 따르는가?



반응값 범주가 최소에서 최대까지 정렬될 수 있다.

예시

농업 연구관은 암탉의 몸무게가 달걀의 크기와 연관이 있는지 조사하기를 원한다. 각 암탉의 몸무게를 기록하고 각 달걀을 소, 중, 대인지를 본다.



반응값 범주가 최소에서 최대까지 정렬될 수 없다.

예시

자동차 제조사의 마케팅 담당자는 구매자의 성별 혹은 나이가 구매 차량의 색상과 관련이 있는지 알기를 원한다. 차량의 색상은 최소에서 최대까지 정렬할 수 없기 때문에 반응값 범주는 순서가 없다.



▶ 순서형 로지스틱 회귀분석 (Ordinal Logistic Regression)

순서형 로지스틱 회귀분석은 하나 이상의 X 변수와 순서가 있는 세 개 이상의 범주를 가지는 범주형 반응값 간의 관계를 조사한다.

예시

농업 연구관은 암탉의 몸무게와 그것의 달걀의 크기(소, 중, 대)를 기록한다. 암탉의 몸무게와 달걀의 크기가 관련이 있는지 파악하기 위해 순서형 로지스틱 회귀분석을 사용한다.

미니탭에서, 통계분석 > 회귀 분석 > 순서형 회귀 분석 을 선택한다.

순서형 로지스틱 회귀 분석: 계란 크기 대 무게

연결 함수: 로짓

반응 정보

변수	값	카운트
계란 크기	1	25
	2	9
	3	26
총계		60

로지스틱 회귀 분석 표

예측 변수	계수	계수 SE	Z	P	승산비	95% CI	
						하한	상한
상수(1)	4.62840	1.32875	3.48	0.000			
상수(2)	5.53347	1.38379	4.00	0.000			
무게	-0.780097	0.212678	-3.67	0.000	0.46	0.30	0.70

로그 우도 = -47.834

모든 기울기가 0인지 검정: G = 25.739, DF = 1, P-값 = 0.000

적합도 검정

방법	카이-제곱	DF	P
Pearson	85.8181	61	0.020
이탈도	63.4791	61	0.389

연관성 측도:

(반응 변수와 예측 확률 사이)

쌍	번호	백분율	요약 측도	
일치	943	85.0	Somers의 D	0.72
불일치	148	13.3	Goodman-Kruskal 감마	0.73
같은 값	18	1.6	Kendall의 타우-a	0.45
총계	1109	100.0		

▶ 명목형 로지스틱 회귀분석 (Nominal Logistic Regression)

명목형 로지스틱 회귀분석은 하나 혹은 그 이상의 X 변수와 순서가 없는 세 개 이상의 범주를 가지는 범주형 반응값 간의 관계를 조사한다.

예시

자동차 제조사의 마케팅 담당자는 구매자의 나이와 성별 그리고 그들이 구매한 차량 색상을 기록한다. 차량의 색상이 구매자의 나이와 성별과 관련이 있는지를 결정하기 위해 명목형 로지스틱 회귀 분석을 사용한다.

미니탭에서, 통계분석 > 회귀분석 > 명목형 로지스틱 회귀분석 을 선택한다.

명목형 로지스틱 회귀 분석: 색상선호도 대 성별, 나이

반응 정보

변수	값	카운트
색상선호도	흰색	108 (기준 사건)
	회색	107
질은 청색	빨강	43
	노랑	71
밝은 청색	노란색	35
	검정	51
	총계	40
		455

로지스틱 회귀 분석 표

예측 변수	계수	계수 SE	Z	P	승산비	95% CI 하한
로짓 1: (회색/흰색)						
상수	-0.355370	0.413894	-0.86	0.391		
성별						
여성	-0.527942	0.309343	-1.71	0.088	0.59	0.32
나이	0.0161152	0.0100709	1.60	0.110	1.02	1.00
로짓 2: (질은 청색/흰색)						
상수	-1.21441	0.545298	-2.23	0.026		
성별						
여성	-0.307114	0.408562	-0.75	0.452	0.74	0.33
나이	0.0117568	0.0131747	0.89	0.372	1.01	0.99
로짓 3: (빨강/흰색)						
상수	1.14244	0.463385	2.47	0.014		
성별						
여성	-0.379432	0.327794	-1.16	0.247	0.68	0.36
나이	-0.0361957	0.0124037	-2.92	0.004	0.96	0.94
로짓 4: (밝은 청색/흰색)						
상수	1.29152	0.666117	1.94	0.053		
성별						
여성	-1.93182	0.504388	-3.83	0.000	0.14	0.05
나이	-0.0472456	0.0199261	-2.37	0.018	0.95	0.92

연락처

Global Statistical Software
for Quality improvement worldwide



성공을 위한 열쇠 Minitab이 같이 합니다.

- 모든 분야의 강력한 통계 분석 지원
- Six Sigma를 위한 성과 창출 지원
- 품질 관리를 위한 개선 활동 지원
- 사업 평가를 위한 정량적 분석 지원



(주)이레테크

(주)이레테크 소프트웨어사업부

경기도 군포시 당정동 522번지 SK벤티움 101동 703호

TEL : 031-436-1101 FAX : 031-436-1110 Email : minitab@minitab.co.kr

* 통계 방법 길잡이는 Minitab Inc.(www.minitab.com)의 Method Chooser의 한글 번역판이며, 콘텐츠에 대한 일체의 권리는 Minitab Inc.에 있으며 무단 사용을 금합니다.