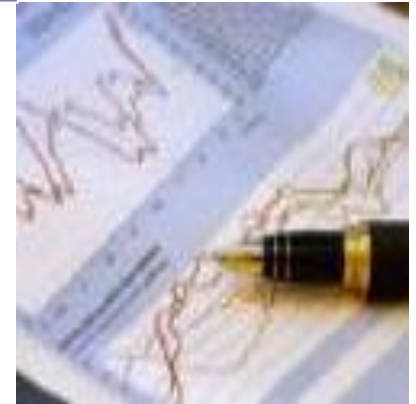


SPSS를 이용한 회귀분석(regression analysis) 〈단순회귀분석을 중심으로〉

2011년 11월 29일



임 찬 수

Table of Contents

1**회귀분석이란?**

2**단순회귀분석**

3**단순회귀분석의 예제**

■ 회귀분석(Regression analysis)

- 회귀분석이란 독립변수에 대한 종속변수의 결과를 예측하는 선형 회귀식을 추정하는 것이다. 독립변수와 종속변수의 이해와 함께 추정된 회귀직선의 적합도와 추정된 회귀직선은 어느 정도 자료를 설명할 수 있는 지를 어떻게 파악할 수 있는 지 알아보자.
- 회귀분석과 상관관계의 차이점
 - 어떤 개체들에 대하여 두 변수를 측정하였다고 할 때, 두 변수의 관계는 상관관계로 표현되며, 한 변수가 다른 변수에 어느 정도 영향을 미치는 변수인지를 파악하고자 할 때에는 예측을 생각한다. 관련성을 생각할 때는 두 변수가 대등한 관계이지만 예측에서는 두 변수가 대등한 관계가 아니고 영향을 주는 변수와 영향을 받는 변수로 나뉘어진다.
 - 상관관계와 회귀분석의 차이 : 두 변수 X, Y 에 대하여 X 가 독립변수, Y 가 종속변수라고 하면

상관관계 : $X \leftrightarrow Y$

회귀분석 : $X \rightarrow Y$

앞에서 배운 상관이 관련성에 목적을 둔 것이라면 회귀분석은 예측을 목적으로 한다. 예측을 위해서는 두 변수간의 관계식 혹은 여러 변수와의 관계식을 구해야 하는데 이렇게 구하는 관계식을 회귀모형 또는 회귀방정식이라고 한다.

■ 회귀모형(Regression model)

● 회귀모형이 무엇인가?

- ▶ 변수의 관계를 선형의 방정식으로 표현하는 것을 회귀모형이라고 한다.
- ▶ 한 변수(종속변수)가 다른 변수들(독립변수)에 의해 어떻게 설명되는지 구체적인 형태로 표현하는 것 : $X \rightarrow Y$
- ▶ 독립변수 : 영향을 주는 변수, 설명변수라고도 한다.
- ▶ 종속변수 : 영향을 받는 변수, 반응변수라고도 한다.

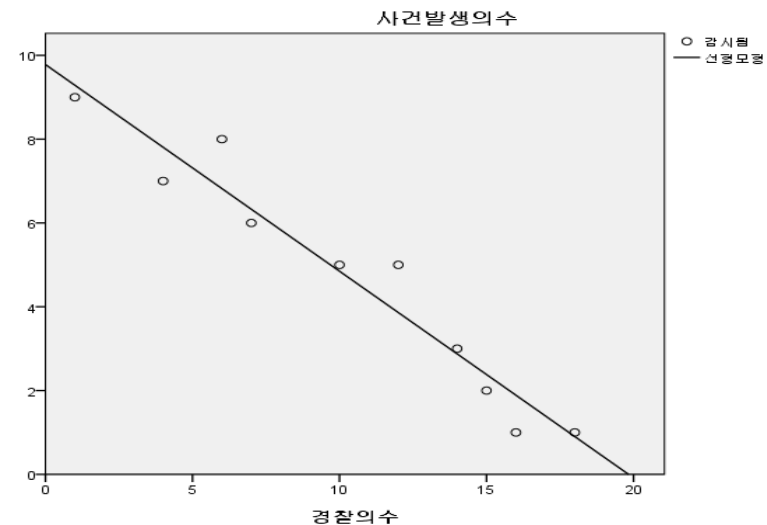
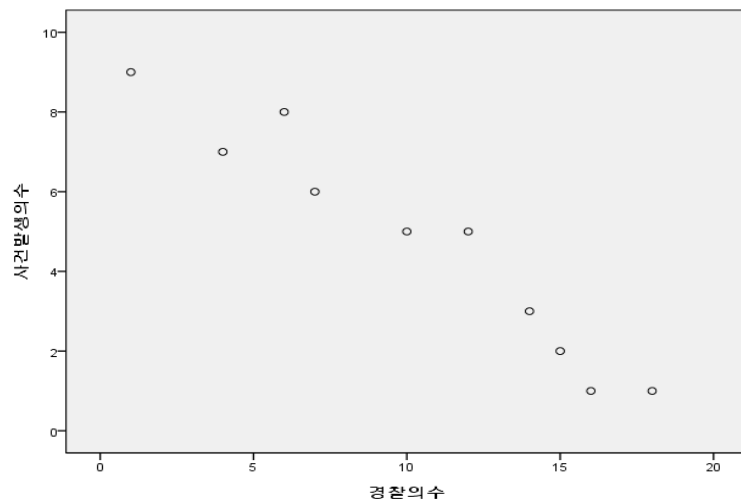
● 독립변수는 반드시 하나인가? ➔ 아닙니다. 독립변수는 여러 개가 가능. 독립변수의 수에 따라 회귀분석은 아래와 같이 나눌 수 있다.

- ▶ Simple linear regression(단순선형회귀분석)
- ▶ Multiple linear regression(다중선형회귀분석)

● 회귀모형 중에서 가장 기본적인 회귀분석은 설명변수가 한 개 뿐인 단순회귀(simple regression)이다. 만약 이 때 설명변수와 종속변수의 관계가 1차식으로 표현될 경우를 단순선형회귀(Simple linear regression)이라 한다.

■ 단순 회귀방정식의 기본적 형태

- 단순 회귀방정식의 기본적 형태 - 산점도로 나타난 그래프가 아래 좌측과 같다면, 이 자료를 가장 잘 설명할 수 있는 일차함수는 우측과 같이 나타낼 수 있다.

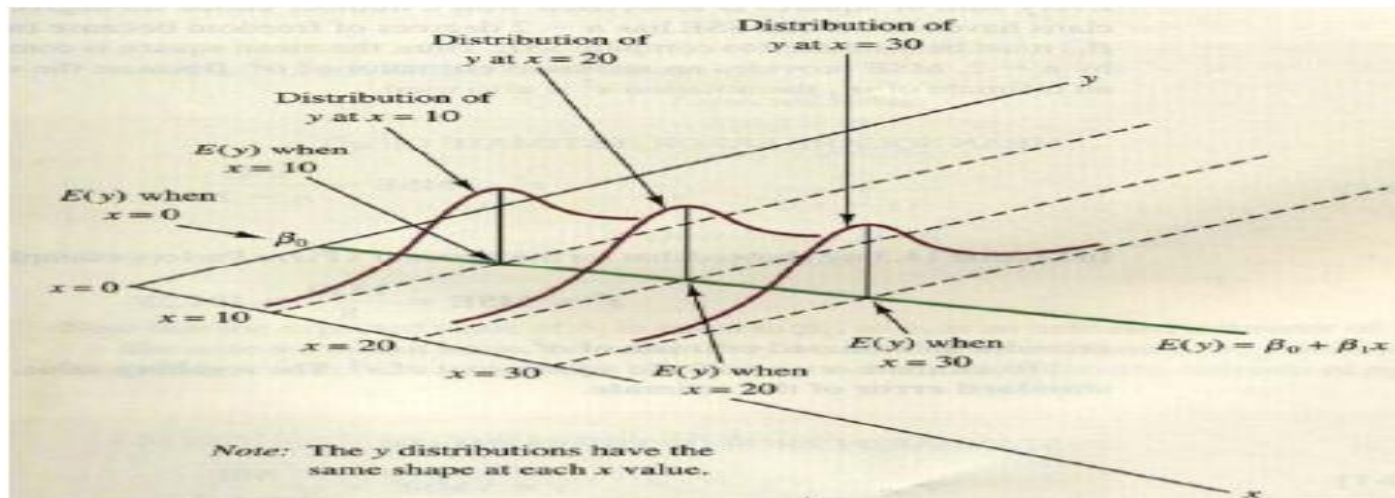


■ 단순 선형 회귀 모형의 정의

- 단순선형모형이란 독립변수가 하나인 일차 선형함수의 형태이다. 간단히 생각하면 $Y=ax+b$ 의 형태를 가지게 될 것이다. 일반적으로 단순 회귀모형은 다음과 같은 형태를 취한다.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

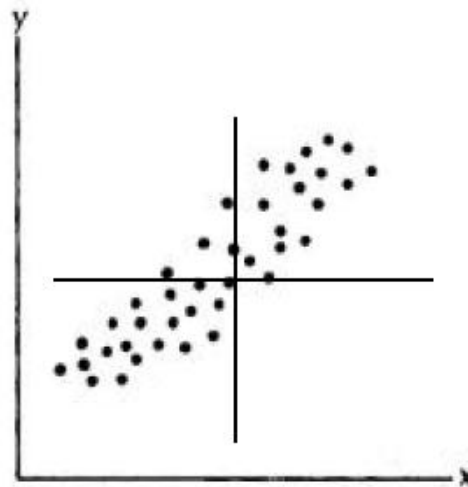
- ✓ β_0 가 상수항 b ($y=ax+b$ 에서), β_1 이 직선의 기울기 a ($y=ax+b$ 에서), β_0, β_1 를 회귀계수(regression coefficient), $\varepsilon \sim N(0, \sigma^2)$
- 그러면 ε 은 무엇일까? 실제 자료값들을 보면 여러 요인들(환경의 변화, 측정의 부정확 등)에 의하여 그 값이 직선을 벗어나는 경우가 많다. 이러한 요인을 표현한 것이 오차항 ε 이다. 일반적으로 ε 은 $N(0, \sigma^2)$ 을 따른다.



회귀방정식의 추정

산점도를 통한 추정

- ✓ 어떤 변수 X 와 Y 에 대한 산점도를 그렸을 때 아래와 같이 나왔다고 하자. 산점도를 보면 두 변수가 선형의 관계를 가짐을 짐작할 수 있다.
- 자료를 처음 접했을 때 우선 산점도를 그려 두 변수간의 관계가 선형적인 확인한 후 회귀방정식을 추정하는 것이 바람직하다. 산점도에서 선형의 관계가 보이지 않는데 선형회귀를 구하는 것은 무의미한 일이다.



■ 최소제곱법

- 주어진 자료에 가장 적합한 회귀방정식은 어떻게 구할 수 있을까? 쉽게 생각해보면 회귀방정식을 만들었을 때 실제 자료값과 오차, 즉 ε 이 가장 작은 회귀방정식을 구하면 될 것이다.

$$\varepsilon = Y_i - \beta_0 - \beta_1 X_i$$

- 앞의 회귀식을 통해 위와 같은 식을 찾을 수 있다. $\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$ 이 최소가 되는 β_0, β_1 을 구하면 된다. 이러한 추정방법을 최소제곱추정법(method of least squares)이라 하며, 이렇게 구해진 β_0, β_1 을 최소제곱추정량(least squares estimator)이라 한다. 위의 식을 편미분하면 최소제곱추정량을 구할 수 있으며 그 결과는 아래와 같다.

$$\hat{\beta}_0 = \bar{y} - b\bar{x}, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- 여기서 $\hat{\beta}_0, \hat{\beta}_1$ 는 β_0, β_1 의 추정량을 뜻한다.

■ 최소제곱법

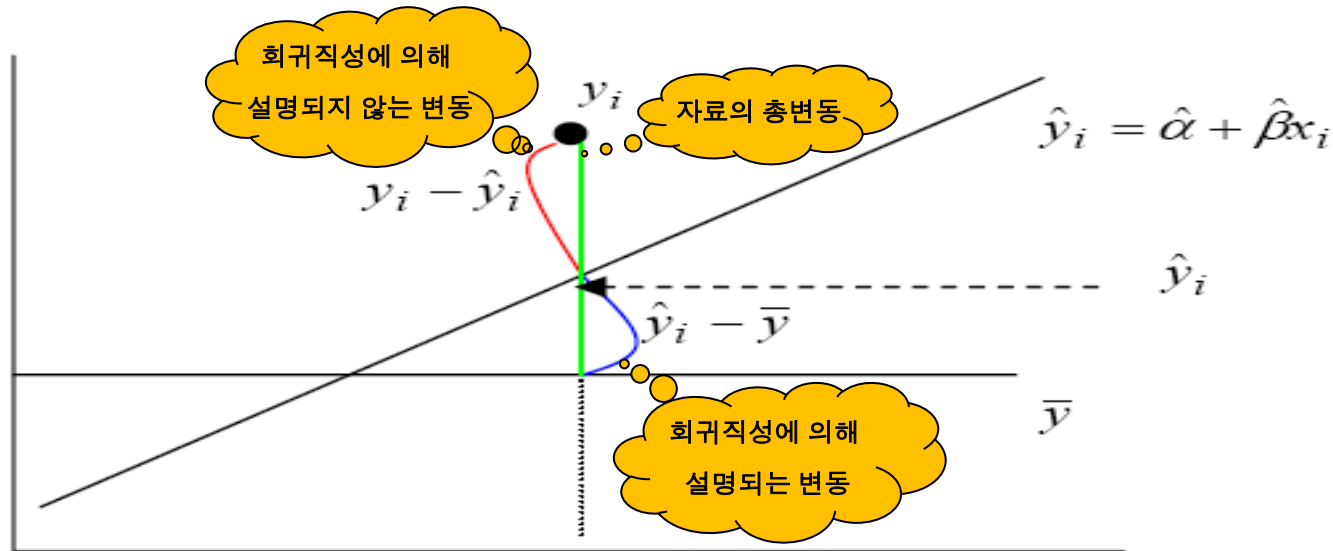
- 앞에서 구한 $\hat{\beta}_0, \hat{\beta}_1$ 에 의해 추정된 회귀직선은 다음과 같다.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

- 예측값 : $x=x_i$ 일 때 $\hat{y}$ 의 값, 즉 $\hat{y}_i$
- 잔차(오차) : 실제 관측값 y_1 와 적합값 $\hat{y}_i$ 과의 차이

- 그러면 위에서 추정된 회귀 직선이 실제 자료를 어느 정도 설명할 수 있는 지를 알아볼 수 있는 측도는 없을까? 이러한 방법을 살펴볼 수 있는 것이 R^2 이다. 다음 장에서 R^2 를 살펴보자.

■ 결정계수(R^2)



- 총변동=회귀직선에 의해서 설명되는 변동+회귀직선에 의해서 설명되지 않는 변동. 이것을 수식으로 살펴보자.

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

- 위 식은 직선상의 어느 한점에서만 계산한 것이다. 이것을 직선상의 모든 점에서 계산한다면 식은 다음과 같이 된다.

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

SST
 (총변동 혹은 총제곱합)

SSE
 (잔차제곱합)

SSR
 (회귀제곱합)

■ 결정계수(R^2)

- 여기에서 결정계수 R^2 은 다음과 같이 정의된다.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

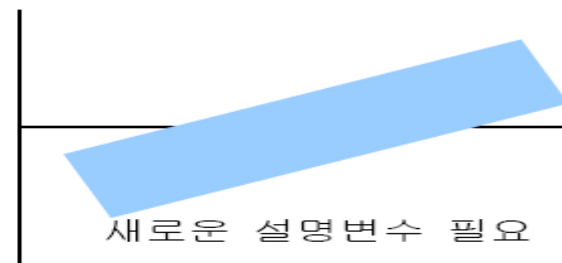
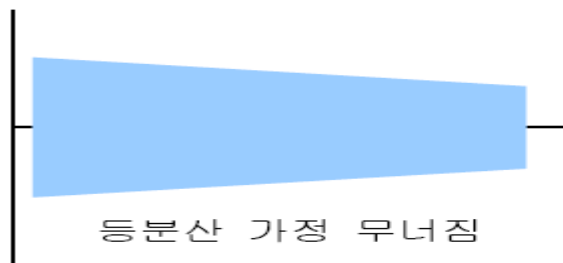
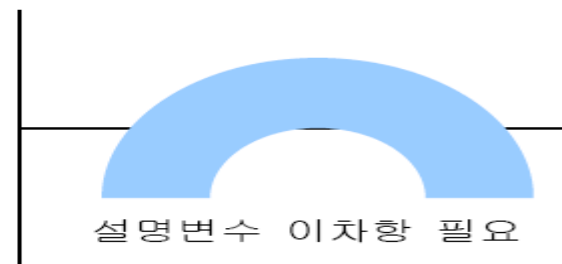
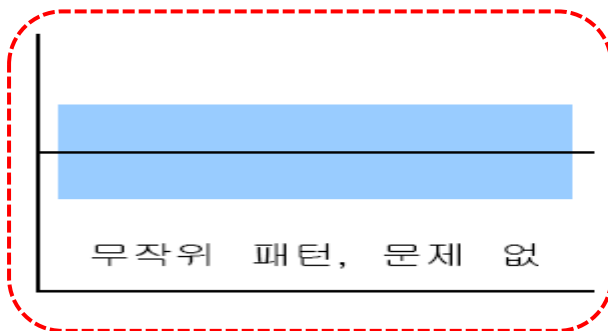
- 즉 R^2 는 총변동 중 회귀직선에 의해 설명되는 변동이 차지하는 비율을 뜻한다. 즉, 추정된 회귀직선에 의해 변동이 설명되는 변동이 차지하는 비율을 뜻한다. 즉, 추정된 회귀직선에 의해 변동이 설명되는 정도를 나타낸다. R^2 가 클수록 회귀직선이 잘 추정되었다고 생각할 수 있다. 반대로 R^2 가 작을수록 회귀직선은 잘 추정되지 않았고 자료를 잘 설명할 수 없다고 할 수 있다. ➔ 그러면 R^2 (결정계수)가 높으면 무조건 좋은가?? 그렇지 않다. R^2 는 변수가 많아지거나 혹은 관측치가 많아질수록 무조건 커지는 현상이 있다. 또한 R^2 을 통계적으로 검정할 방법이 없다. 그러므로 무조건 R^2 이 높으면 좋은 것은 아니다. 단지 설명력만을 의미할 뿐이다. R^2 가 0.52이면 회귀모형은 전체자료의 약 52%를 설명한다고 판단하면 된다.

■ 단순회귀분석의 과정

- 회귀분석은 단순히 회귀모형을 구하는 것으로 끝나는 것이 아니고 자료로부터 추정된 회귀모형이 어느 정도로 잘 구해진 회귀직선인지(적합한지) 확인하기 위해서는 여러 가지 검정과정을 거쳐야 한다.
- 일반적인 회귀분석의 과정은 다음과 같은 순서로 이루어진다.
 - ✓ 산점도를 그려서 대략적인 직선관계 파악
 - ✓ 잔차(residual)검정을 통해 이상치 확인 및 제거
 - ✓ 회귀직선의 적합도 검정 : F -검정을 통해 적어도 하나의 이상의 유의한 변수가 있는지 확인(회귀직선의 적합도)
 - ✓ 독립변수의 유의성 검정 : t -검정을 통해 검정. 단순회귀분석에서는 독립변수가 하나밖에 없으므로 변수의 선택과정이 필요없다.
 - ✓ 회귀모형이 유의하다면, R^2 을 통해 설명력을 확인하고 최종 회귀모형을 정리하고 해석.
- ✓ 산점도는 앞에서 확인했으므로 산점도를 통한 확인은 생략.

잔차분석

- 우선 산점도를 통해 선형관계가 있는지 확인을 해보고 잔차분석을 실시한다. 잔차분석을 하는 이유는 첫째로 우리는 잔차가 정규분포이고 독립을 가정한다. 이에 잔차들이 동일한 분포를 가지는 확인해 봐야 한다. 두 번째로는 잔차를 통해 이상치를 확인하고 제거한 후 회귀분석을 실시해야 한다. 이상치는 회귀식에 영향을 주므로 우리는 이상치를 제거(± 2 이상)하고 회귀분석을 실시해야 한다. 영향치는 Cook's distance를 이용(1이상이면 영향치)하여 확인한다.



■ 모형의 유의성과 독립변수의 유의성 확인

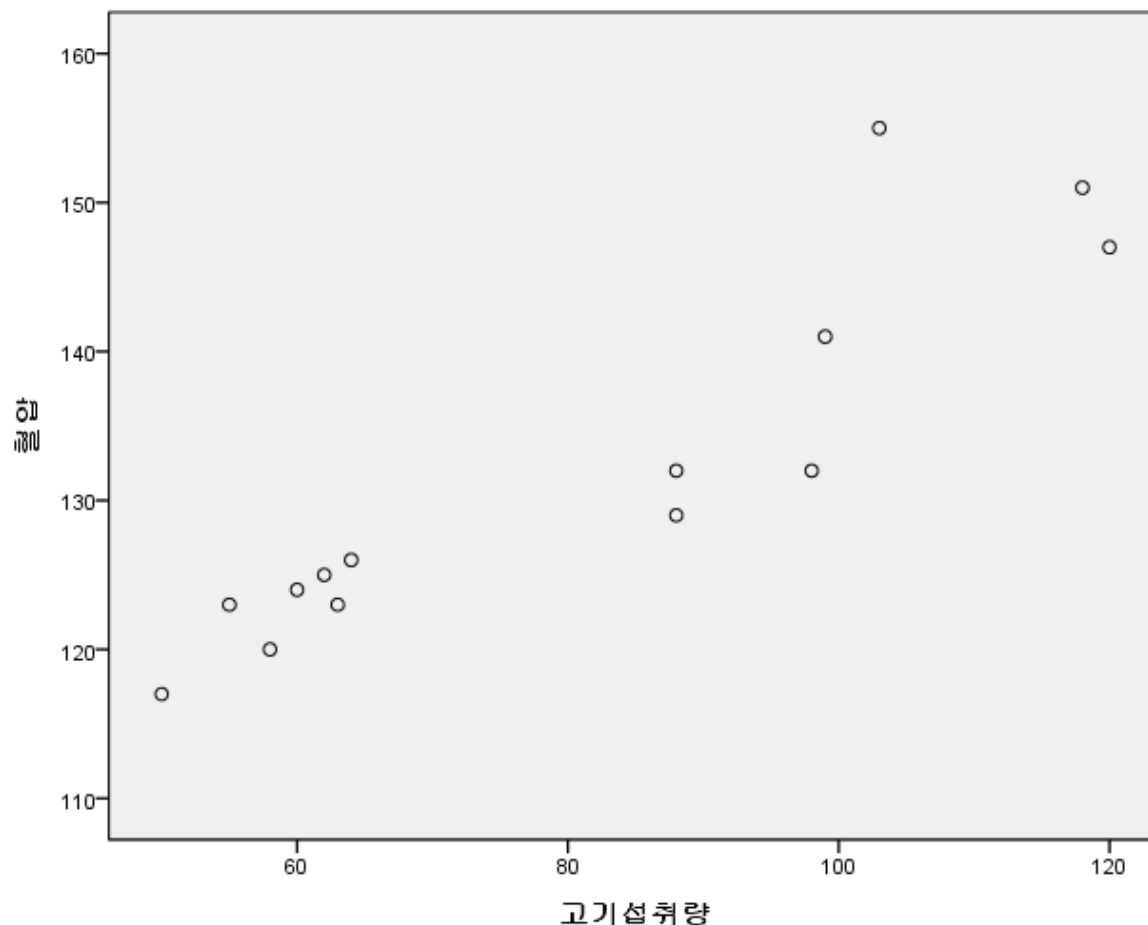
- 잔차분석을 통해 동일한 분포와 형태가 없는 것을 확인하고, 이상치를 제거한 후 이제 본격적인 회귀분석을 실시한다.
- 다음으로 회귀모형의 적합성을 판단한다.
 - ✓ 귀무가설(H_0) : $\beta_0=\beta_1=\dots=\beta_n=0$ (회귀모형이 적합하지 않다) vs 대립가설(H_1) : $\text{not } H_0$ (회귀모형이 적합하다.)
 - ✓ 위 가설에 대하여 F 검정을 통하여 판단. 귀무가설이 선택되면 회귀분석을 더 이상 진행하지 않음. 만약 대립가설이 선택되면 회귀모형이 적합하다고 판단하고 다음으로 각 독립변수들에 대한 회귀계수를 검정.
- 이제 다음으로 각 회귀계수에 대하여 검정을 한다.
 - ✓ 귀무가설(H_0) : $\beta_i=0$ (개별 독립변수의 영향력은 없다.) vs 대립가설(H_1) : $\beta_i \neq 0$ (개별 독립변수는 영향력은 있다.)
 - ✓ 유의수준(α)와 p -value를 비교하여 귀무가설 혹은 대립가설을 선택한다.
- 이제 각 개별 독립변수의 유의성을 판단하고, 추정되는 회귀계수를 이용하여 최종 회귀모형을 작성하고 해석한다.

■ 단순회귀분석 예제

- 혈압에 영향을 주는 요인으로 다양한 변수들이 존재한다. 이를 실제 확인해보기 위해 14명의 남자들에 대하여 혈압과 체중, 나이, 고기섭취량을 조사하였다. 이중 고기 섭취량은 혈압에 영향을 주는지 살펴보기 위하여 회귀분석을 실시하고 영향을 준다면 어떠한 영향을 주는지 살펴보자.
- 자료는 SPSS에서 확인. 고기섭취량이 혈압에 어떠한 영향을 주는지를 살펴보는 것이므로 인과관계를 포함. 그러므로 회귀분석을 실시해야 한다. 독립변수가 하나이므로 단순회귀분석 실시.
- 우선 선형관계가 있는지 산점도를 통해 선형관계를 확인해보고, 이상치를 확인하고 제거하고, 잔차분석을 통해 확인하고 회귀분석을 실시한다.

■ 단순회귀분석 예제

● 산점도

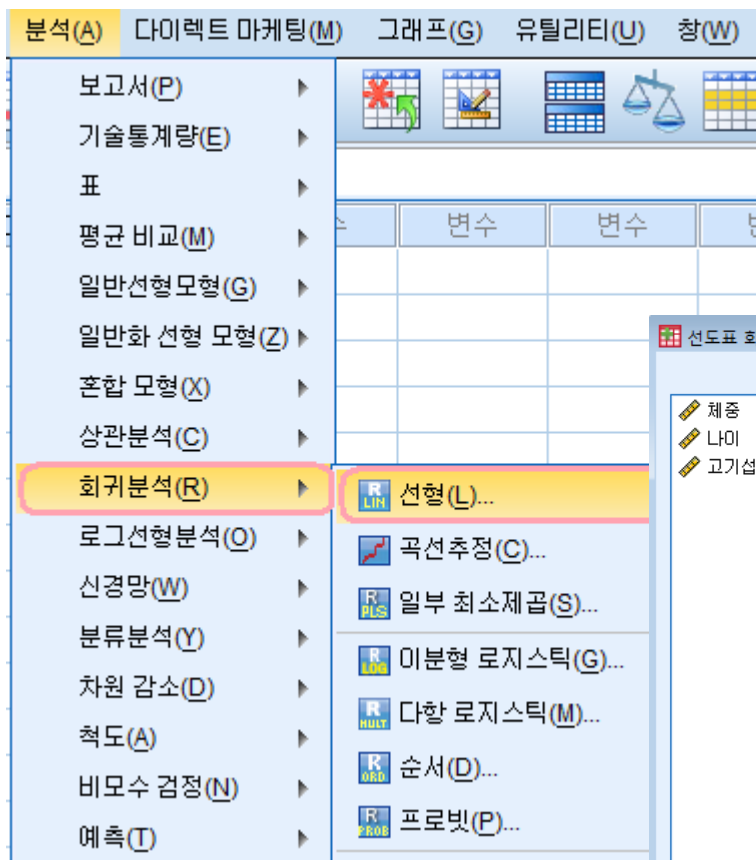


- 그래프 - 산점도를 선택
- 단순산점도를 선택. 만약에 우리가 그리고 싶은 산점도가 많다면 행렬로 선택하면 각 조합별로 그려준다.
- Y축과 X축을 각각 변수로 지정
- 일반적으로 원인이 될 수 있는 변수를 X축, 결과가 될 수 있는 변수를 Y축으로 지정한다.
- 고기섭취량이 증가하면 혈압이 대략적으로 증가하는 것으로 보임.
- 선형관계가 있는 것으로 판단됨

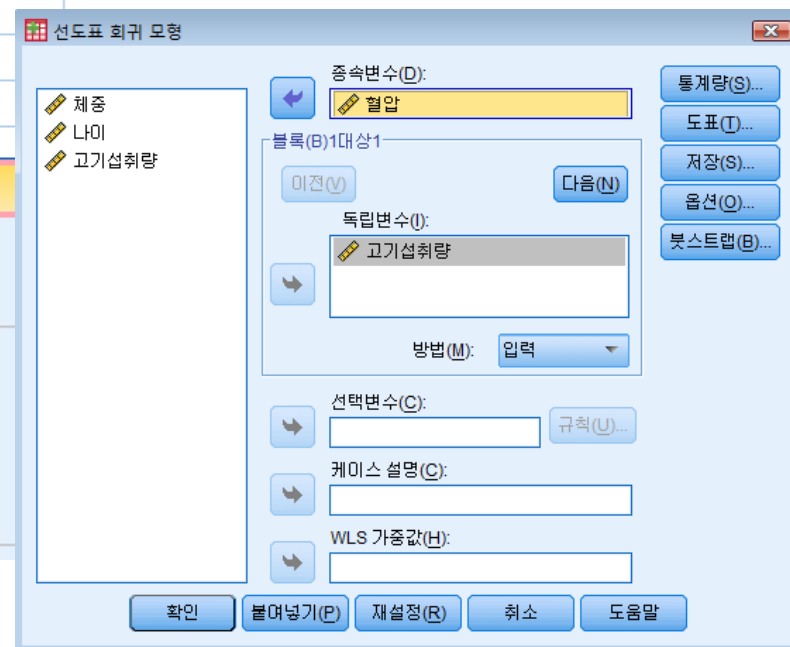
■ 단순회귀분석 예제

● 자료의 입력과 회귀분석

	원인	체중	나이	고기섭취량
1	151	89	35	118
2	120	76	51	58
3	141	91	22	99
4	124	85	20	60
5	126	83	33	64
6	117	79	25	50
7	129	80	52	88
8	123	74	63	55
9	125	79	55	62
10	132	85	41	98
11	123	77	54	63
12	132	82	43	88
13	155	95	42	103
14	147	92	21	120



- 자료를 입력
- 분석-회귀분석-선형을 선택
- 독립변수(원인)와 종속변수(결과)의 지정



■ 단순회귀분석 예제

● 잔차분석(옵션 : 도표)

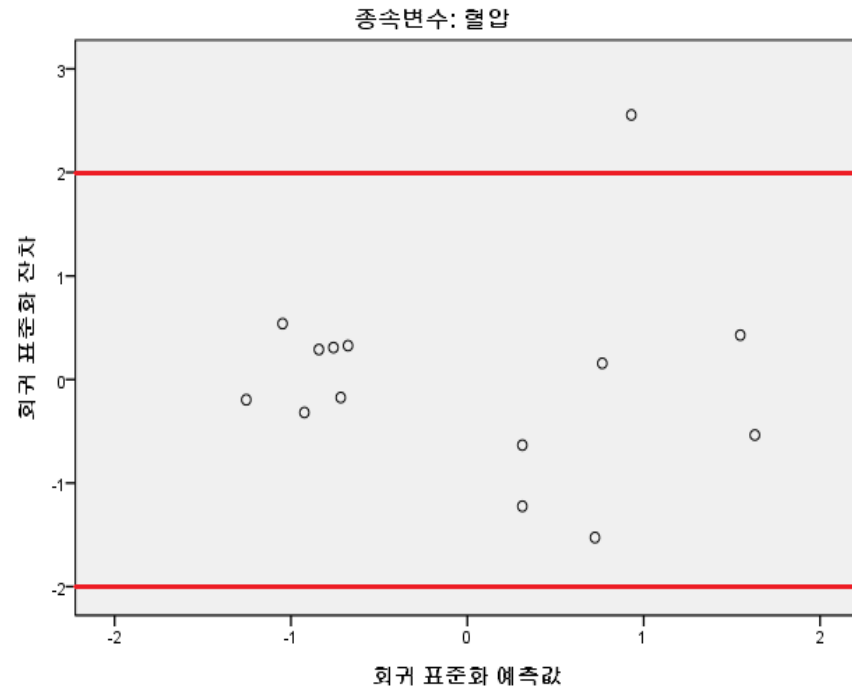
- ▶ 잔차도표를 그리기 위해 도표에서 Y축과 X축을 지정.
- ▶ Y축: ZRESID(표준화된 잔차), X축: ZPRED(표준화된 예측치)
- ▶ 정규확률도표선택 후 계속

● 이상치, 영향치 확인(옵션 : 저장)

- ▶ 이상치 및 영향치 확인을 위해 회귀분석 옵션에서 저장을 클릭
- ▶ 거리에서 Cook의 거리를 체크, 잔차에서 표준화를 체크.

■ 단순회귀분석 예제

● 잔차분석



- 잔차도표를 보면 2이상의 값이 있음.
- 이상치가 있으므로 제거해야 한다.
- 데이터를 보면 잔차의 값과 Cook's distance를 확인

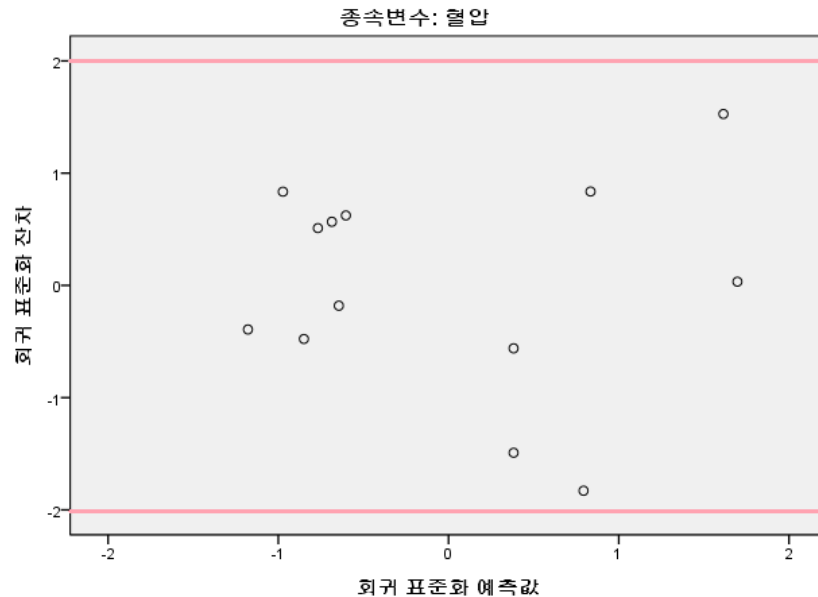
● 이상치, 영향치 확인

	혈압	체중	나이	고기섭취량	ZRE_1	COO_1
1	151	89	35	118	.42935	.04254
2	120	76	51	58	-.31842	.00933
3	141	91	22	99	.15647	.00183
4	124	85	20	60	.29042	.00695
5	126	83	33	64	.32715	.00715
6	117	79	25	50	-.19506	.00561
7	129	80	52	88	-1.22390	.06966
8	123	74	63	55	.53975	.03185
9	125	79	55	62	.30879	.00706
10	132	85	41	98	-1.52573	.16482
11	123	77	54	63	-.17410	.00213
12	132	82	43	88	-.63343	.01866
13	155	95	42	103	2.55512	.60583
14	147	92	21	120	-.53641	.07566

- ZRE_1의 값은 표준화 잔차를 COO_1은 Cook's distance를 확인.
잔차의 기준값은 ± 2 , Cook's 거리는 1이상이다.
- 13번째 관측치만 이 이상치. 제거하고 회귀분석 시행.
- 이상치와 영향치 제거는 변수선택을 통해서 실시(생략)
- 앞의 과정을 반복

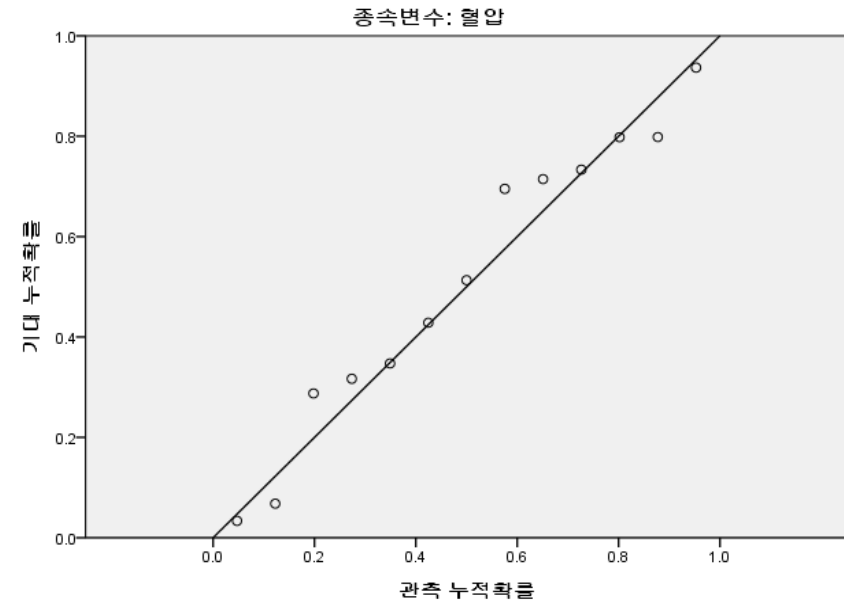
단순회귀분석 예제(이상치 제거 후)

잔차분석(이상치 및 분포)



- ▶ 잔차도표를 보면 2이상의 값 없음.
- ▶ 특정한 패턴이 없음. 잔차들이 동일한 분포의 형태이라고 판단 할 수 있다.

잔차의 정규성검정



- ▶ 잔차의 정규성 검정을 위해 정규성 도표를 확인.
- ▶ 관측치가 직선에 대체적으로 가깝다. 잔차들이 정규분포를 따른다고 할 수 있다
- ▶ 최종적으로 이상치도 없고 잔차의 분포도 동일하고 정규성을 만족함.

■ 단순회귀분석 예제

● 출력결과

분산분석^b

모형	제곱합	자유도	평균 제곱	F	유의확률
1 회귀 모형	1189.731	1	1189.731	114.528	.000 ^a
잔차	114.269	11	10.388		
합계	1304.000	12			

a. 예측값: (상수), 고기섭취량

b. 종속변수: 혈압

계수^a

모형	비표준화 계수		표준화 계수	t	유의확률
	B	표준오차	베타		
1 (상수)	97.816	3.137		31.178	.000
고기섭취량	.409	.038	.955	10.702	.000

a. 종속변수: 혈압

모형 요약^b

모형	R	R 제곱	수정된 R 제곱	조정값의 표준오차
1	.955 ^a	.912	.904	3.223

a. 예측값: (상수), 고기섭취량

b. 종속변수: 혈압

- F-검정 결과 유의확률이 0.0001로 유의수준 0.05보다 적음. 귀무가설을 기각. 대립가설을 채택. 회귀모형이 적합하다고 혹은 적어도 하나 이상의 독립변수는 통계적으로 유의하다고 판단. 이제 개별적인 독립변수가 유의한지 계수값은 얼마인지 확인을 위해 독립변수의 검정.
- 계수라는 곳에 각 독립변수 별로 t-검정을 확인. 고기섭취량의 유의확률이 0.0001로 고기섭취량은 통계적으로 혈압에 영향을 준다고 판단 할 수 있다. 회귀모형의 계수값은 0.409이다.
- 최종회귀모형식 : 혈압 = 97.816+0.409*고기섭취량
- 회귀모형의 해석 : 고기섭취량이 한 단위 증가하면 혈압은 평균적으로 0.409으로 증가함을 알 수 있다.
- 모형의 설명력 : $R^2=0.912$ 로 위의 회귀모형은 전체자료의 91%를 설명. 매우 잘 설명함을 알 수 있다.

질의응답

Question
&
Answer