회귀분석

서론

- 지금까지는 하나의 변수에 관하여 한 모집단에 대한 추론을 하거나, 두 모집단을 비교하는 추론의 방법을 공부하였다.

- 그러나 실제 자료에서 많은 경우에 두 개 이상의 변수에 대하여 관측값을 얻게 되는데 그 변수들에 관하여 다음과 같은 질문을 할 수 있다.

(1) 변수들이 서로 관련이 있는가?
(2) 관련이 있다면 얼마나 밀접하게 관련이 있는가?
(3) 관심이 있는 변수의 값을 그 외 다른 변수의 값으로부터 예측할 수 있을까?

- 위와 같은 질문에 대하여는 각각의 변수를 따로 분석하여서는 답을 얻을 수 없고, 모든 변수를 가지고 함께 분석하여야 한다.

서론

- 기초통계실습1에서 두 변수에 대한 기본적인 분석방법으로 두 범주형 변수의 요약은 분할표, 두 연속형 변수의 요약은 산점도, 두 연속형 변수의 수치적 요약은 상관계수를 배웠습니다.

- 간단하게 table(x1,x2), plot(x1,x2), cor(x1,x2)로 구현을 할 수 있습니다.

회귀분석

- 여러 변수가 주어질 때 한 변수를 나머지 변수들로부터 설명하기 위하여 모형을 설정하고 그 모형이 맞다고 할 수 있는지를 검정하고 그 모형을 통하여 추론과 예측을 하게 된다.

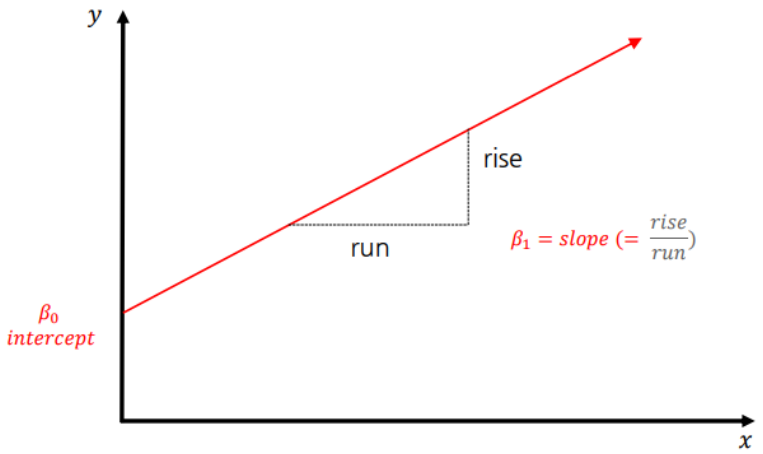
- 단순선형회귀모형: 두 변수 간의 직선 관계를 설명

단순선형회귀모형

- 독립변수 or 설명변수: 보통 실험하는 사람에 의하여 통제되어 독립적으로 주어지는 변수이며, x로 많이 표현한다.

- 종속변수 or 반응변수: 독립변수와 오차에 의하여 결정되는 변수이며, y로 많이 표현한다.

- $y = \beta_0 + \beta_1 x$



단순선형회귀모형

- 오차: 설명될 수 없는 요인으로 관측될 수는 없지만 어떤 확률적 성질을 가짐.

- 단순선형회귀모형

확률변수 Y를 독립변수 x와 오차라는 확률변수 ε에 의해 설명되는 종속변수라 하면 그 직선의 관계는 다음과 같이 표현될 수 있다.

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1,2,\cdots,n$$

조건:

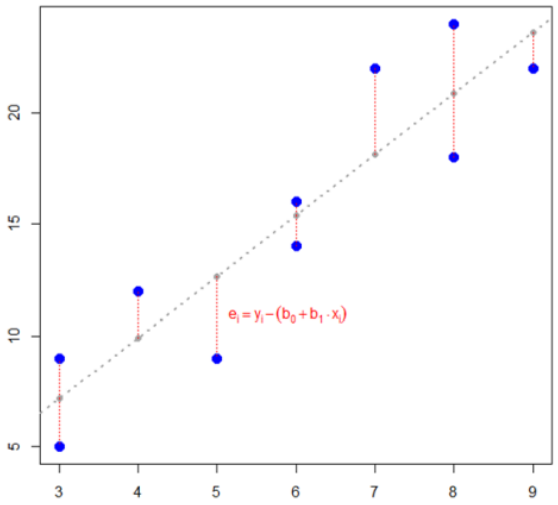
(1) β_0와 β_1은 직선식을 결정하는 미지의 회귀모수이다.
(2) 오차 ε_i들은 서로 독립이며 평균이 0, 분산이 σ^2인 정규분포를 따르는 확률 변수이다.
(3) Y_i는 독립변수를 x_i로 고정시켰을 때의 종속변수의 값이다.

단순선형회귀모형

- $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad where \quad \varepsilon_i \sim i.i.d.N(0, \sigma^2)$
- $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$
- $y = \beta_0 + \beta_1 x$: 모회귀직선(population regression line)을 가정
- 미지의 회귀모수 β_0와 β_1을 두 변수의 자료 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$을 이용하여 추정한 후에 이들을 이용하여 회귀직선을 추정

최소제곱추정법

- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \Rightarrow \quad y = b_0 + b_1 x$
- $\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$을 최소화시키는 직선을 찾는 것(편미분)
- 그 때의 b_0, b_1을 $\widehat{\beta_0}, \widehat{\beta_1}$



- 최소제곱추정량(LSE: Least Squares Estimator)

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} \quad where \quad S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \ and \ S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

- 추정회귀직선(Estimated Regression Line): $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$

최소제곱추정법

- 오차: $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$
- 잔차(residual): $\hat{\varepsilon_i} = e_i = y_i - \hat{y_i} = y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i) \quad \Rightarrow \quad \sum_{i=1}^{n} e_i = 0$
- 잔차제곱합: $SSE = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (e_i - 0)^2$
- 오차의 분산 σ^2의 추정량: $s^2 = MSE = \frac{SSE}{n-2}$

예제

- 예제. 다음 표는 어떤 알레르기 증세에 효과가 있다고 하는 새로 개발된 약품의 복용량(mg)과 효과가 지속되는 기간(일)을 기록한 자료이다.

복용량(x)	3	3	4	5	6	6	7	8	8	9
지속기간(y)	9	5	12	9	14	16	22	18	24	22

수고하셨습니다.

- 과제 X