Linear Regression

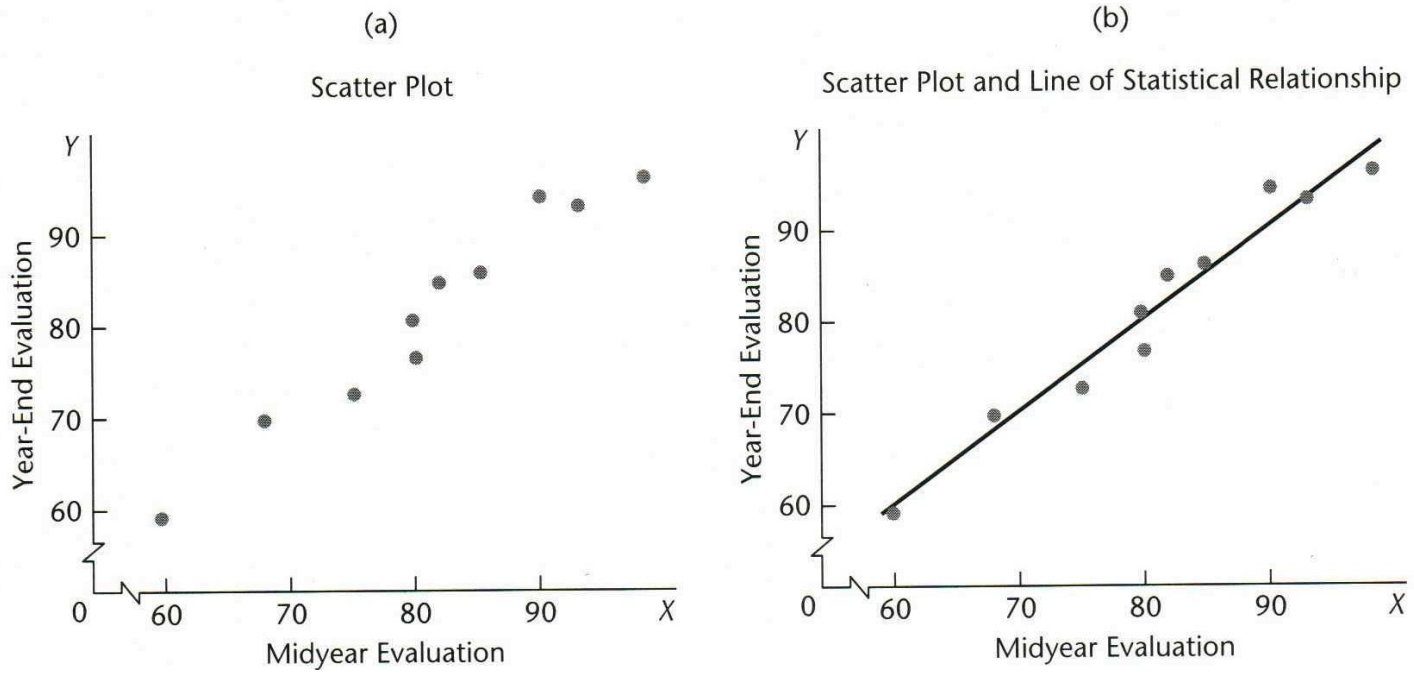
5기 허유진

목차

01 단순 선형 회귀분석

회귀분석의 목적

회귀분석:
결과값(Y)과 입력값(X)간의 함수적 관계를 알아보기 위해 시행하는 분석 기법
X,Y 변수로 회귀분석을 진행하여 도출해 낸 함수 식 f(x)는, 두 변수가 y=f(x)의 관계를 가질 가능성이 높다는 것을 설명해준다.



$$Y = f(X) + \epsilon$$

이때 ϵ 는 오차로, random error component이다. 즉, **오차가 존재하는 통계적 모형**.

회귀분석의 목적

회귀분석의 목적:
데이터가 주어졌을 때 두 변수 간의 관계를 찾아 함수로 나타내기 위한 분석이다.

변수 간의 관계를 perfect하지 않다. 오차가 존재한다.

X변수와 Y변수의 인과관계(causation)를 나타내는 것이 아니라 상관관계(association)를 나타낸다.
(상황에 따라 인과관계로 해석이 가능한 경우가 존재하긴 한다.)

Simple Linear Regression Model

단순 선형 회귀분석(Simple Linear Regression):
설명변수가 1개 있는 선형 회귀 모형에 대한 분석

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad i = 1, 2, \ldots, n}$$

$Y_i =$ 종속변수, 반응변수
$X_i =$ 독립변수, 설명변수
β_1 = 기울기를 나타내는 회귀 계수, 알려져 있지 않은 모수 (우리가 추정해야하는 값)
β_0 = 절편을 나타내는 회귀 계수, 알려져 있지 않은 모수 (우리가 추정해야하는 값)
ϵ_i = 오차항

오차항은 independent and identically distributed (iid) 확률변수이다.

$E(\epsilon_i) = 0$

$Var(\epsilon_i) = \sigma^2$ (unknown parameter)

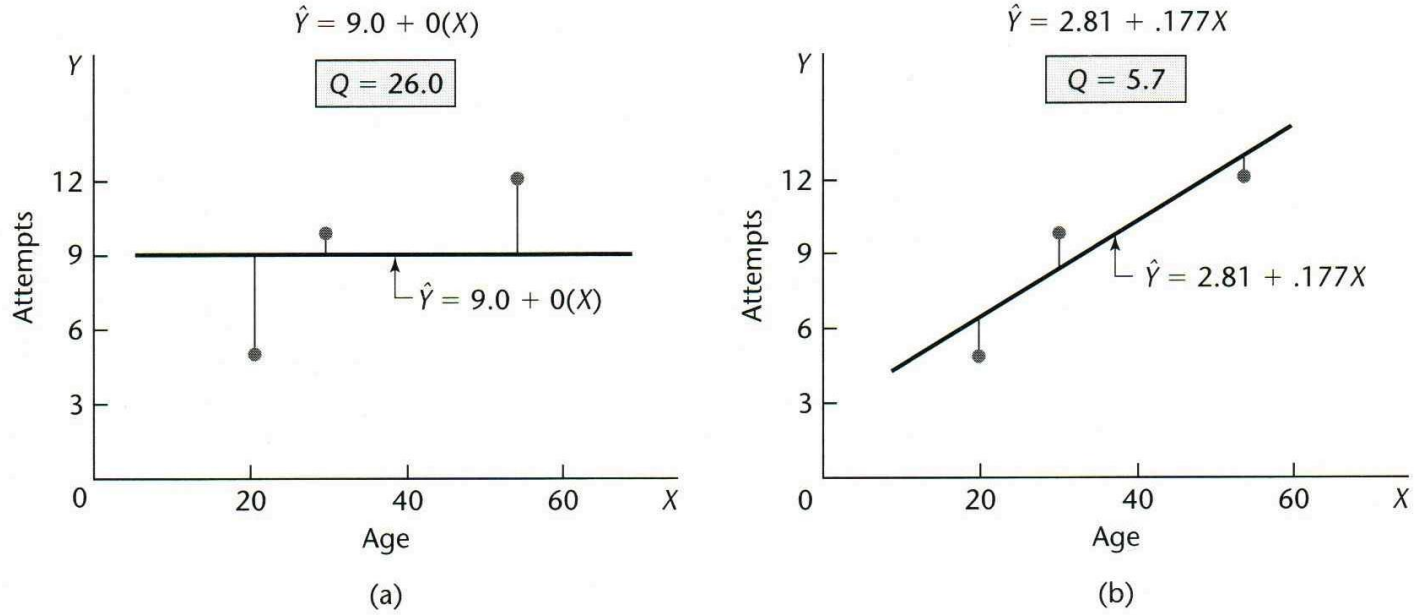
$Cov(\epsilon_i, \epsilon_j) = 0 \; for \; i \neq j$ (uncorrelated)

Least Squares Method

단순 선형 회귀분석(Simple Linear Regression)에서 β_0 과 β_1 값을 추정하는 방법

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad i = 1, 2, \ldots, n$$

최소자승법(Least Squares Method, LSM)
함수관계에 존재하는 오차의 제곱합을 최소로 만드는 식을 찾는 것



Least Squares Method

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad i = 1, 2, \ldots, n$$

최소자승법(Least Squares Method, LSM)

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$(b_0, b_1) = \text{argmin}_{(\beta_0, \beta_1)} Q(\beta_0, \beta_1)$$

$$\left.\frac{\partial Q}{\partial \beta_0}\right|_{\beta_0 = b_0, \beta_1 = b_1} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\left.\frac{\partial Q}{\partial \beta_1}\right|_{\beta_0 = b_0, \beta_1 = b_1} = -2 \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0$$

위의 "정규방정식"에서 다음을 얻는다.

$$b_1 = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) / \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Notation for sum of squares

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

앞 slide의 b_1, b_0 식을 다시 정리하면

$$b_1 = \frac{S_{xy}}{S_{xx}}$$

$$b_0 = \overline{y} - b_1 \overline{x}$$

Estimation of Mean Response

우리가 추정하는 것은 Mean Response. 실제 모수인 $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$값은 알 수 없다.

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad i = 1, 2, \ldots, n}$$

$$\boldsymbol{Mean\ Response}: \boldsymbol{E(Y) = \beta_0 + \beta_1 X}$$
$$\boldsymbol{Estimated\ Mean\ Response}: \boldsymbol{\hat{Y} = b_0 + b_1 X}$$
$$\boldsymbol{Fitted\ value\ for\ the\ ith\ case}: \boldsymbol{\hat{Y}_i = b_0 + b_1 X_i}$$
$$\boldsymbol{Residual}: \boldsymbol{e_i = Y_i - \hat{Y}_i}$$ (잔차는 오차 ϵ_i에 대한 best information 제공)

오차와 잔차

오차(error) = $\boldsymbol{\epsilon_i = Y_i - \beta_0 - \beta_1 X_i}$
잔차(Residual) = $\boldsymbol{e_i = Y_i - b_0 - b_1 X_i}$

잔차는 오차의 근사값으로, 오차의 정보를 가장 잘 반영해주는 숫자이다.
오차는 관측할 수 없고, 잔차는 실제로 계산 가능.

Estimation of σ^2

$\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$ 뿐만 아니라, error의 variance 값인 σ^2 값도 추정해야 한다. 이때 필요한 개념은

Error Sum of Squares

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (\boldsymbol{Y_i} - \boldsymbol{\hat{Y}_i})^2$$

Error Mean Square

$$MSE = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (\boldsymbol{Y_i} - \boldsymbol{\hat{Y}_i})^2}{n-2} = \frac{SSE}{n-2}$$

이때의 n-2는 degree of freedom

MSE는 σ^2 의 unbiased estimator이다. 즉 $E(MSE) = \sigma^2$
따라서 σ^2 의 추정량으로 MSE를 사용

ANOVA Approach to Regression

Total Sum of Squares (추정된 직선에 따라 변하지 않음)

$$SSTO = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Regression Sum of Squares (fit의 퀄리티에 따라 변함)

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

Error Sum of Squares (fit의 퀄리티에 따라 변함)

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$SSTO = SSR + SSE$$

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

Coefficient of Determination R^2

$$SSTO = SSR + SSE$$

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

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

Fit이 좋다 = $\frac{SSE}{SSTO}$ 값이 작다 = $\frac{SSR}{SSTO}$ 값이 크다 = R^2 값이 크다

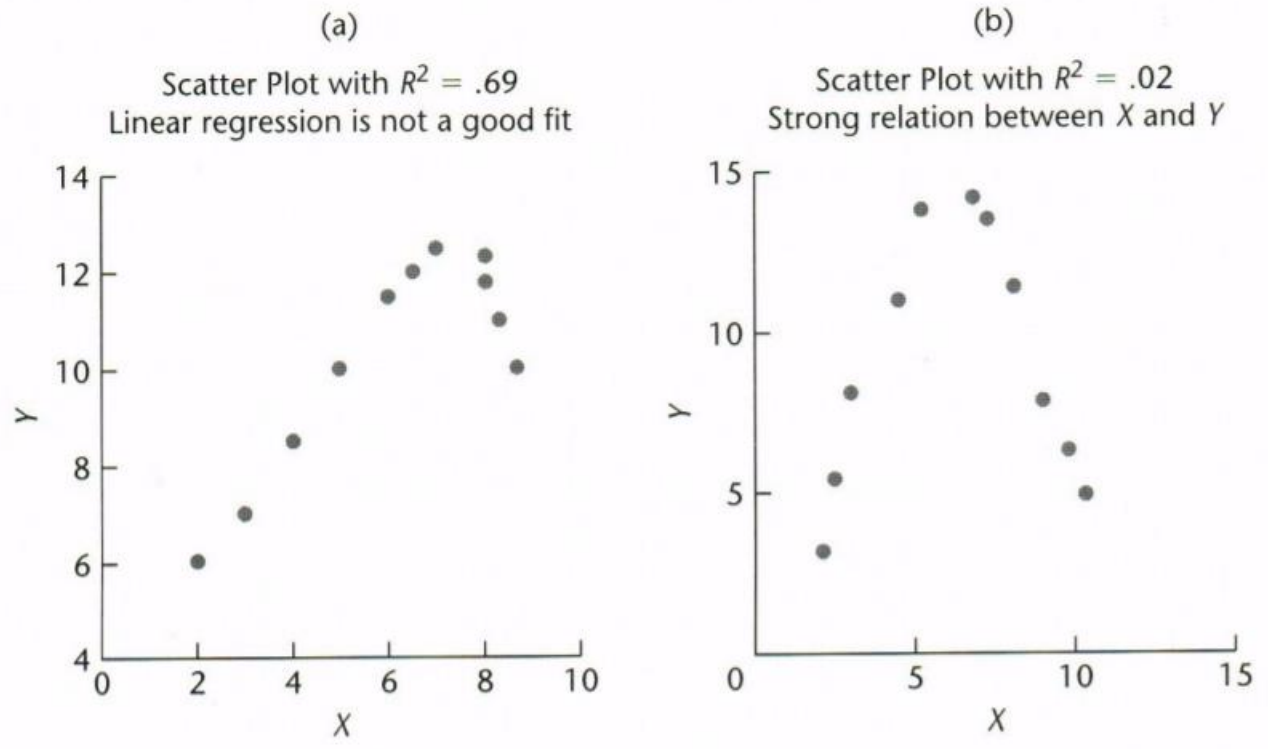
$$0 \leq R^2 \leq 1$$

의미: 전체 변동량 중에서 모형이 설명할 수 있는 변동량의 비율을 의미한다.
결정계수 R square의 값이 클수록, fit의 quality가 높음을 의미한다.

Coefficient of Determination R^2

R^2의 한계점

1. 결정계수의 값이 크다는 것이 반드시 regression 추정이 잘 되었음을 의미하지는 않는다.
2. R^2값이 작다고 해서 x와 y가 아무런 관계가 없다고 볼 수는 없다. (비선형적 연관성을 찾아낼 수 없다.)



회귀분석의 가정사항

오차를 직접 측정할 수 없기 때문에 오차항의 근사값으로 잔차를 사용

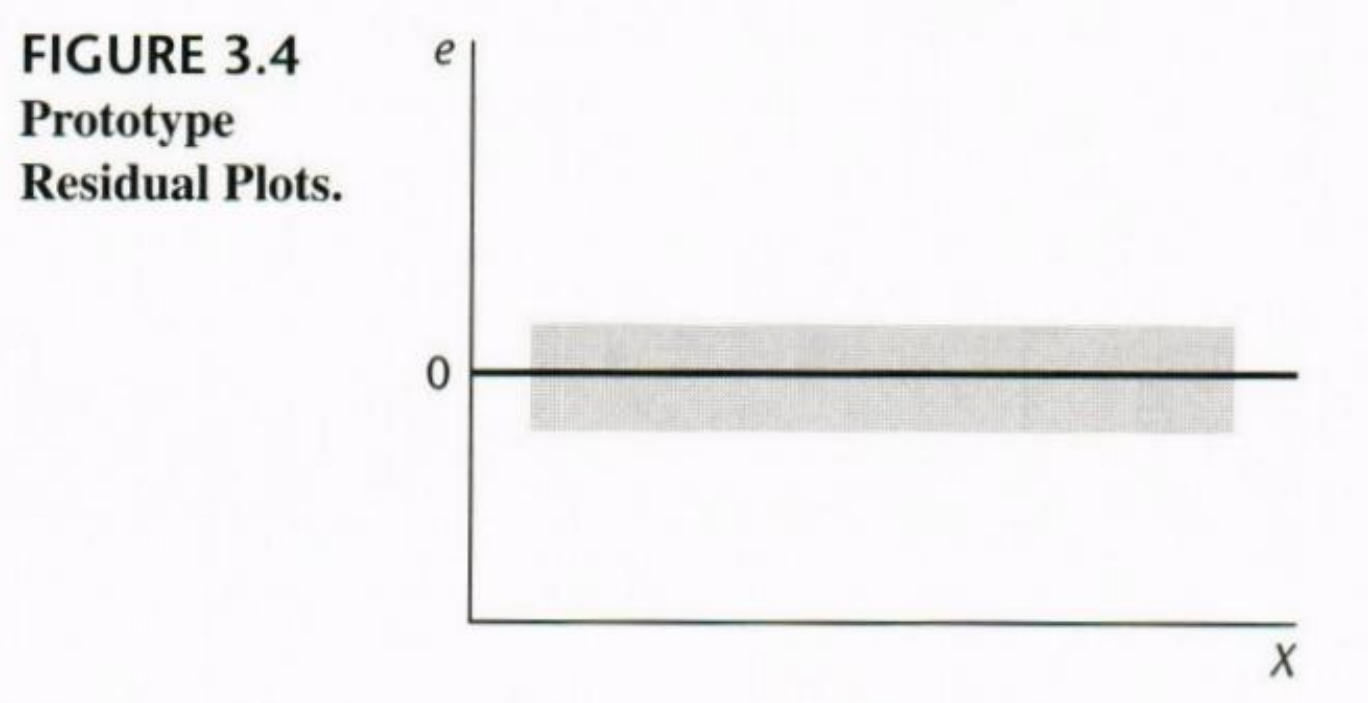
오차(error) = $\boldsymbol{\epsilon_i = Y_i - \beta_0 - \beta_1 X_i = Y_i - E(Y_i)}$ $\quad$ $\boldsymbol{\epsilon_i \sim iid\ N(0, \sigma^2)}$
잔차(Residual) = $\boldsymbol{e_i = Y_i - b_0 - b_1 X_i = Y_i - \hat{Y}_i}$

만약 fitted된 모형이 맞다면, 잔차가 오차의 특성을 반영하여 $\boldsymbol{N(0, \sigma^2)}$의 분포를 따를 것이다.

1. 오차의 등분산성
2. 오차의 독립성
3. 오차의 정규성

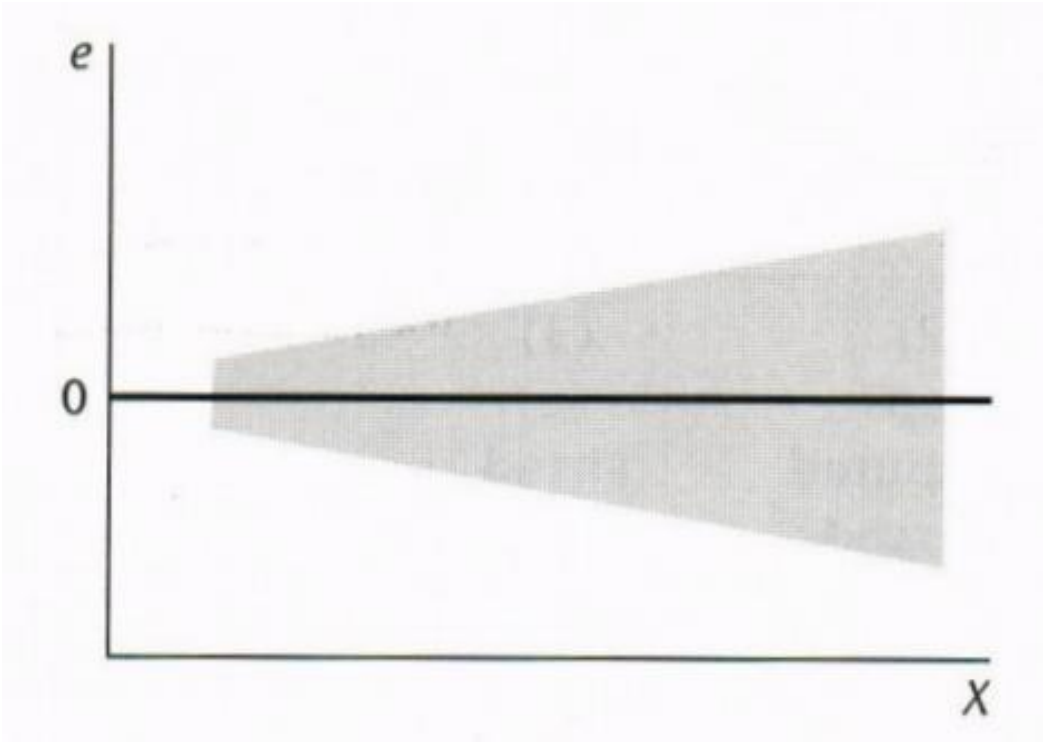
이 세 가지 가정사항은 잔차 그림을 그려 확인할 수 있다.

회귀분석의 가정사항



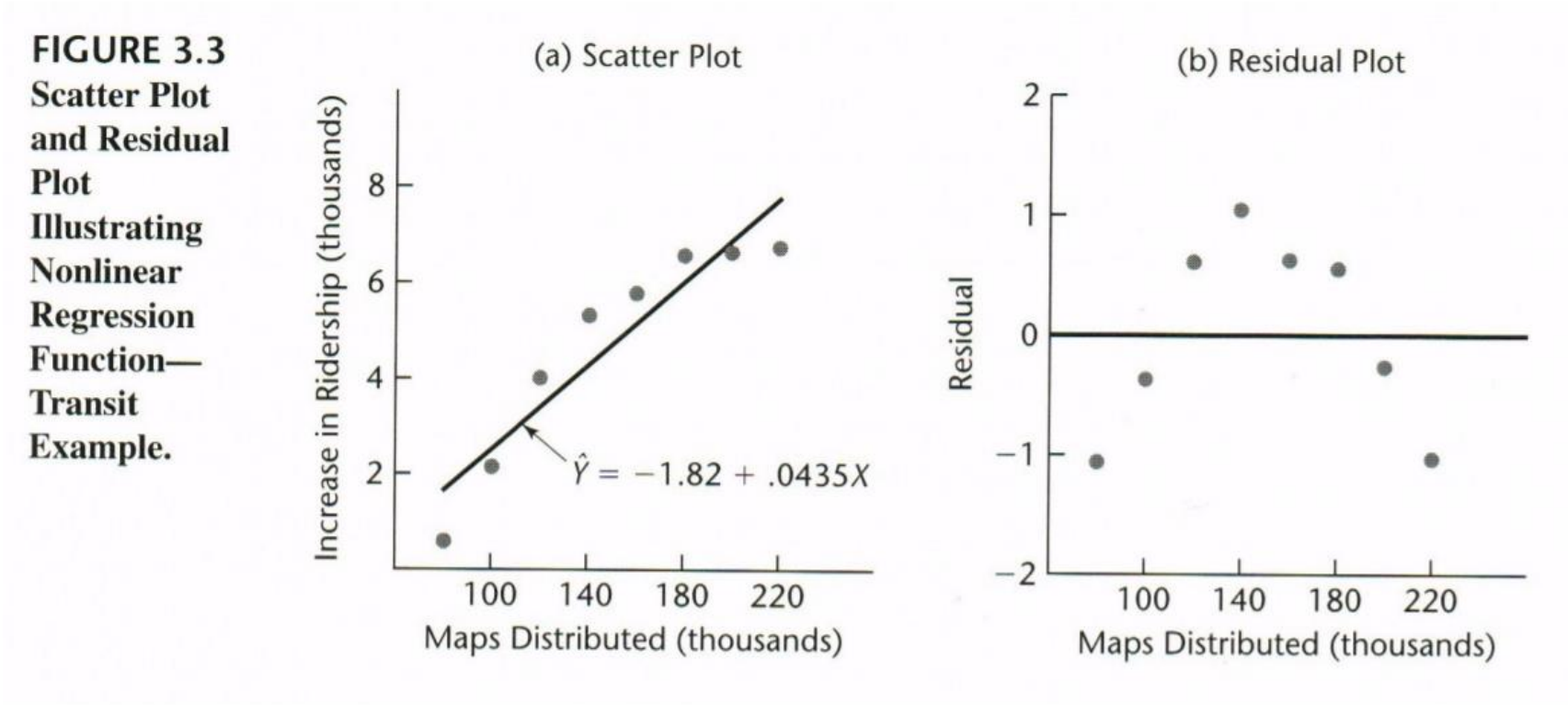
잔차가 0 주위에 적절하게 잘 퍼져 있고, 띠의 폭이 일정하다.
선형회귀가 적절하게 적합 되었을 때에 나타나는 이상적인 잔차 그림

회귀분석의 가정사항



등분산성 가정을 만족하지 못하는 잔차 그림.
분산이 일정하지 않다.

회귀분석의 가정사항



잔차가 일정한 패턴을 보인다. 잔차의 독립성을 만족하지 못한다.
(이러한 경우에는 2차 항이 필요하지 않을까..?라는 의심)

02 다중 선형 회귀분석

Multiple Regression Models

Multiple Regression Models 다중 선형 회귀분석 모델
설명변수가 2개 이상 있는 선형 회귀모형에 대한 분석.

설명변수가 2개라고 가정했을 때, $\boldsymbol{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i}$ $\boldsymbol{\epsilon_i \sim iid\ N(0, \sigma^2)}$

이 경우 회귀 함수는 $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\beta_0 + \beta_1 X_1 + \beta_2 X_2}$

설명변수가 p-1개라고 가정했을 때,

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i} \qquad \boldsymbol{\epsilon_i \sim iid\ N(0, \sigma^2)}$$

$Y_i =$ 종속변수, 반응변수
$X_1, \ldots, X_{p-1} =$ 독립변수, 설명변수
$\beta_0, \ldots, \beta_{p-1} =$ 회귀 계수, 알려져 있지 않은 모수 (우리가 추정해야하는 값)
$\epsilon_i =$ 오차항 (오차항에 대한 가정은 앞과 동일)

이 경우 회귀 함수는 $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}}$

회귀계수의 해석

회귀계수의 해석

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

β_0: Y절편
β_1: X_2는 고정시킨 채로 X_1이 한 단위 증가했을 때 $E(Y)$가 얼마나 증가하는지를 의미
β_2: X_1은 고정시킨 채로 X_2가 한 단위 증가했을 때 $E(Y)$가 얼마나 증가하는지를 의미

이와 같은 해석은 설명변수들이 서로 correlated 되어있지 않다는 가정 하에 완벽한 해석이다.

만약 correlated 되어 있다면,
β_1: X_1과 Y를 나머지 설명변수로 설명한 다음에, 설명하지 못한 나머지 부분에서 Y를 설명하기 위한 X_1의 기여 정도를 의미
β_2: X_2와 Y를 나머지 설명변수로 설명한 다음에, 설명하지 못한 나머지 부분에서 Y를 설명하기 위한 X_2의 기여 정도를 의미

Coefficient of Determination R^2

R^2의 기본적 의미와 식은 단순회귀와 다중회귀에서 동일하다.

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

회귀모형의 fit이 얼마나 좋은지를 보기 위해 많이 사용하는 척도.

R^2 값이 크면 fit이 좋음을 의미한다. 그러나 한계점 또한 존재한다.

모형의 설명변수가 많아질수록 R^2 값은 계속 증가한다.

이런 단점을 보완하기 위해 Adjusted R^2 값을 사용하게 된다.

$\boldsymbol{Adjusted\ R^2}$

$$Adjusted\ R^2 = R_a^2 = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$$

p: 설명변수+절편의 개수

처음에는 중요한 설명변수를 집어넣으면서 R_a^2 값이 커진다.

하지만 unimportant한 설명변수를 집어넣으면 p가 커지고, 그 결과 R_a^2 값이 작아지게 된다.

So $\boldsymbol{R_a^2}$ can penalize excessive number of predictors in the model.

R_a^2를 최고로 만드는 모형을 선택하면 되겠다.

03 다중공선성 (Multicollinearity)

다중공선성이란?

실제 분석에서 설명변수들은 서로 correlated 되어있는 경우가 많다.

서로 다른 변수들이 아주 강하게 correlated 되어 있는 경우에 다중공선성이 발생한다.

두 변수의 선형관계가 매우 강할 경우, 나머지 설명변수들이 특정한 한 변수의 역할을 대신할 수 있다.
이럴 경우에 회귀분석에서 계수의 p-value 값이 매우 크게 나올 수 있다.

- 다중공선성은 VIF(Variance Inflation Factor)로 확인 가능
- 변수 간의 산점도를 그려서 파악 가능

- 해결방법: Principal Component Analysis 또는 Ridge Regression

04 회귀분석에서의 Model Selection

모형 선택 기준

Model Selection

다양한 모델을 만들 수 있다. 어떠한 변수를 사용한 모델을 최적의 모델로 선택하여 사용할 것인가?

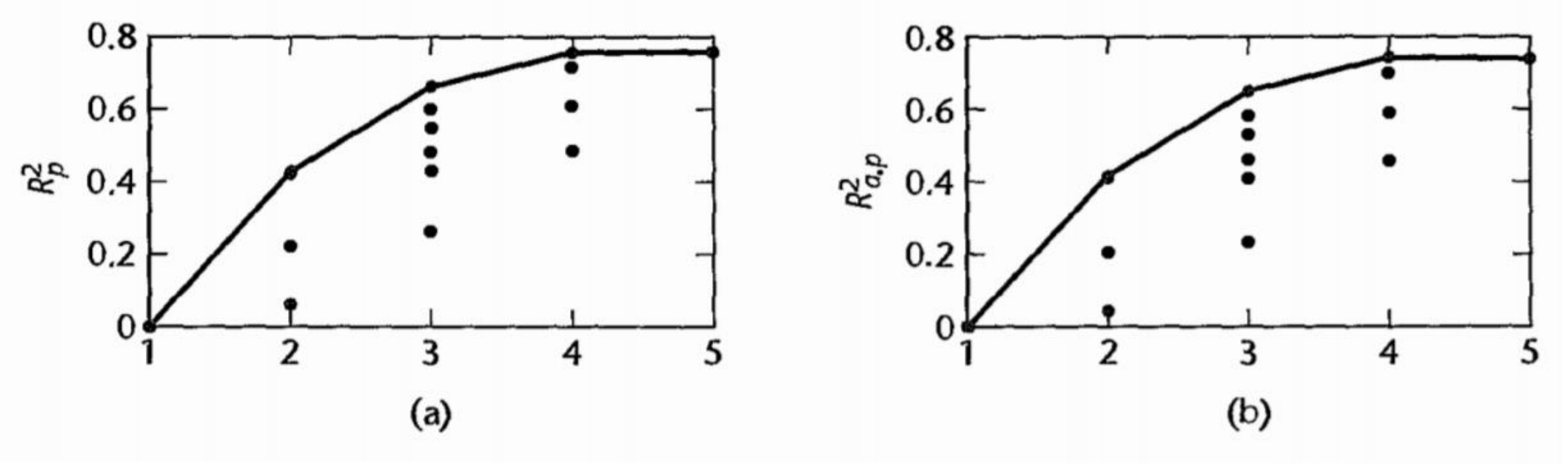
Model Selection의 기준

1. R^2
2. $Adjusted\ R^2 = R_a^2$
3. $Mallows'\ C_p$
4. $AIC_p = n \times lnSSE_p - n \times ln(n) + 2p$ (값이 작을수록 좋은 모형)
5. $BIC_p = n \times lnSSE_p - n \times ln(n) + (ln(n)) \times p$
6. $PRESS_p$

모형 선택 기준

Model Selection의 기준

1. R^2
2. $Adjusted\ R^2 = R_a^2$



모형 선택 기준

Automatic Search Procedures for Model Selection

만약 p-1개의 설명변수가 있다면 2^{p-1}개의 Regression을 모두 수행해야한다.
이런 상황을 방지하기 위한 special algorithm이 존재한다.

1. Forward Selection Procedure

아무런 변수도 없는 모델에서 시작. 각 step에서 1개의 설명변수를 더하면서, SSE를 최소화하는 모델을 찾는다.

2. Backward Elimination Procedure

모든 설명변수를 추가한 모델에서 시작. 각 step에서 1개의 설명변수를 빼면서, SSE를 최소화하는 모델을 찾는다.

3. Forward Stepwise Procedure

아무런 변수도 없는 모델에서 시작. 각 step에서 1개의 설명변수를 더하거나 빼면서, SSE를 최소화하는 모델을 찾는다.

감사합니다

5기 허유진