Linear Regression: Part II

Prediction & Inference

송 준

고려대학교
통계학과 / 융합데이터과학 대학원

Recap: Function Estimation

Optimal predictor:

$$f^* = argmin_f \mathbb{E}[(f(X) - Y)^2]$$

Empirical Risk Minimizer:

$$\widehat{f}_n = argmin_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} ((f(X_i) - Y_i))^2$$

Class of predictors *Empirical mean*

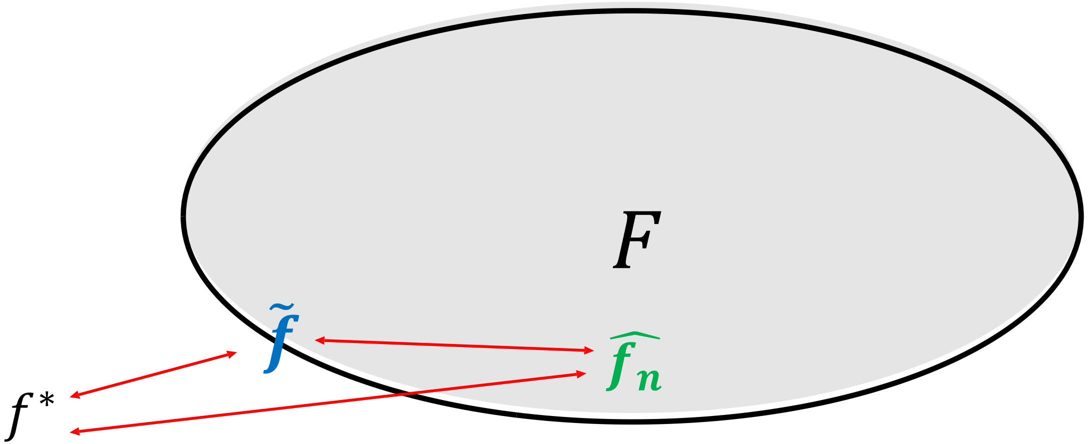
Recap: Function Estimation

Ideal goal: *Construct prediction rule* $f^* : \mathcal{X} \rightarrow \mathcal{Y}$

$$\boldsymbol{f^* = argmin_f \mathbb{E}_{XY}[loss(Y, f(X))]}$$

$$\tilde{\boldsymbol{f}} = arg \min_{f \in F} \mathbb{E}_{XY}[\text{loss}(Y, f(X))]$$

$$\widehat{\boldsymbol{f_n}} = \arg \min_{f \in F} \sum_{i=1}^{n} \text{loss}\big(Y_i, f(X_i)\big)$$



Recap : Function Estimation – one more

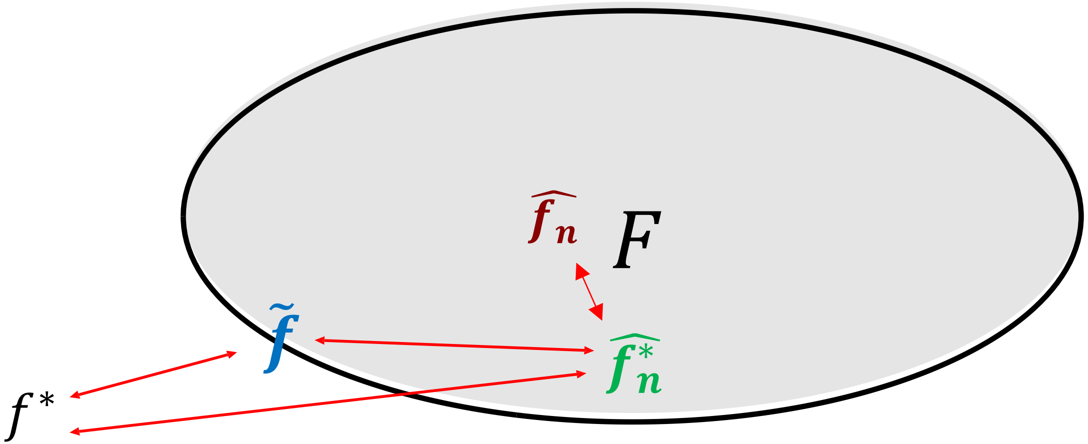
Ideal goal: *Construct prediction rule* $f^*: \mathcal{X} \rightarrow \mathcal{Y}$

$$\boldsymbol{f^* = argmin_f \mathbb{E}_{XY}[loss(Y, f(X))]}$$

$$\tilde{\boldsymbol{f}} = \arg\min_{f \in F} \mathbb{E}_{XY}[\text{loss}(Y, f(X))]$$

$$\widehat{\boldsymbol{f}_{\boldsymbol{n}}^*} = \arg\min_{f \in F} \sum_{i=1}^{n} \text{loss}\big(Y_i, f(X_i)\big)$$

$$\widehat{\boldsymbol{f}_{\boldsymbol{n}}} = \widehat{\arg\min_{f \in F}} \sum_{i=1}^{n} \text{loss}\big(Y_i, f(X_i)\big)$$



함수공간 F가 복잡할 경우 minimizer를 찾지 못할 수 있음

Recap: Introduction to Linear Regression

- **Data Assumption:**
  - $(x_1, y_1), \dots, (x_n, y_n), x_i \in R^p,\ \ y_i \in R$
  - (x_i, y_i): a realization of $(X_i, Y_i) \sim i.i.d.(X, Y)$

- **Model Assumption:** X와 Y는 선형관계를 가짐.

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

- ***f* 의 형태제약:**

$$\mathrm{F} = \{f: f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, for\ some\ \beta_0, \beta_1, \dots, \beta_p\}$$

1차 목표: $\beta_0, \beta = (\beta_1, \dots, \beta_p)^T$ 찾기,

Recap : Solution to the Optimization Problem

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \left(y_i - \beta_0 \cdot 1 \ - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}\right)^2$$

$$= \text{argmin}_{\beta \in \mathbb{R}^{p+1}} (Y - \ \mathbf{X}\beta)^{\mathrm{T}} (Y - \ \mathbf{X}\beta)$$

- **The Analytic Solution**
- *Estimating Equation:* $(\boldsymbol{X^T X})\hat{\beta} = \boldsymbol{X^T Y}$

$$\hat{\beta} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y}$$

$$\hat{\mathbf{Y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y}$$

Goal of Regression Models

- **회귀 모형 (Regression Model)**:

$$\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{X}) + \boldsymbol{\epsilon}$$

- **Goal of Regression Models:**
  - **추정 (Estimation):** 관계를 나타내는 함수 *f*에 대한 추정
  - **예측 (Prediction)**: *X* 값이 주어졌을 때 대응되는 *Y* 값의 예측
  - **추론 (Inference)**: *Further investigation*
    - 예측이 "얼마나" 정확한가?
    - 함수 *f()* 가 얼마나 정확한가?
    - 예측변수가 여러 개 있을 때 모든 변수가 *Y*의 값에 영향을 주나?
    - 모형이 충분히 적합 됐나?

Linear Regression Models

Linear Regression Models

- **선형 회귀 모형 (Linear Regression Model)**:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

  - $X = (X_1, \cdots, X_p)$: p차원 확률변수
  - *Y: 1*차원 확률변수
  - 오차 항: $\mathrm{E}(\epsilon) = 0, var(\epsilon) = \sigma^2, \epsilon \sim N(0, \sigma^2)$,
  - $\beta = (\beta_0, \beta_1, \dots, \beta_p) \in \mathbb{R}^{p+1}$
  - 회귀 모수 (**parameters**) 혹은 회귀계수 (**coefficients**), *unknown, non-random parameters (to be estimated)*

- **표본 Sample**: *n*개의 *data* $(X_1, Y_1), \dots, (X_n, Y_n)$ 는 위 모델을 따르는 *random copy*

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i, \qquad i = 1, \dots, n$$

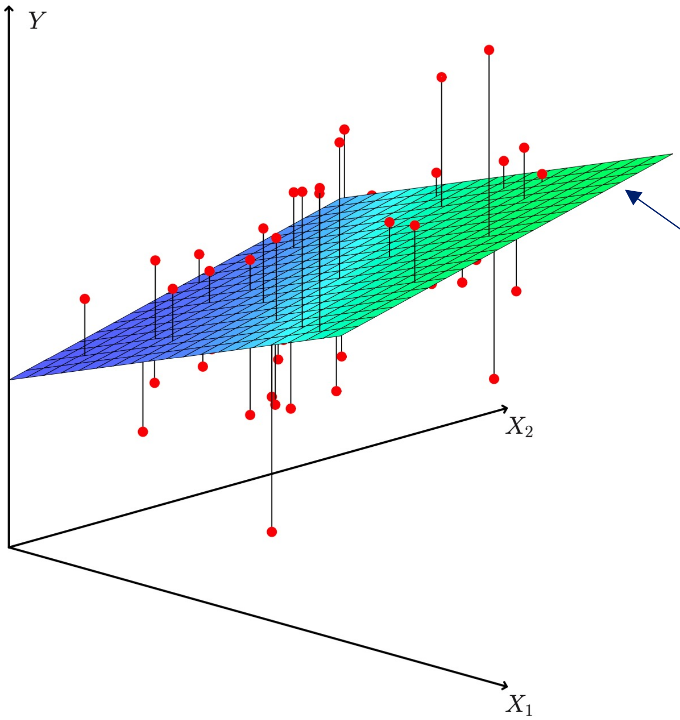
Estimation : Least Square Estimation (LSE)

- **Input** : *Dataset* $Z = \{(x_1, y_1), \cdots, (x_n, y_n)\}$
- Compute

$$\hat{\beta}(Z) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} L(\beta; Z)$$
$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta^T x_i)^2$$

- **Output** : $f_{\hat{\beta}(Z)}(x) = \hat{\beta}(Z)^T x$
- *Discuss algorithm for computing the minimal* β *later*

Least Square Estimation 최소제곱추정



- 예제*:* 두개의 예측변수 *(p=2)*
- *Data Points*

  $(x_{i1}, x_{i2}, y_i) \in \mathbb{R}^3, \qquad i = 1, \dots, n$
- *Surface generated by* $(\beta_0, \beta_1, \beta_2)$

  $\{(X_1, X_2, Y) \in \mathbb{R}^3 : \beta_0 + \beta_1 X_1 + \beta_2 X_2 = Y\}$
- 선*: Data point*와 *surface* 사이 거리

  $|y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})|, \qquad i = 1, \dots, n$

Least Square Estimation 최소제곱추정

- $\hat{\beta} = argmin_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right)^2$

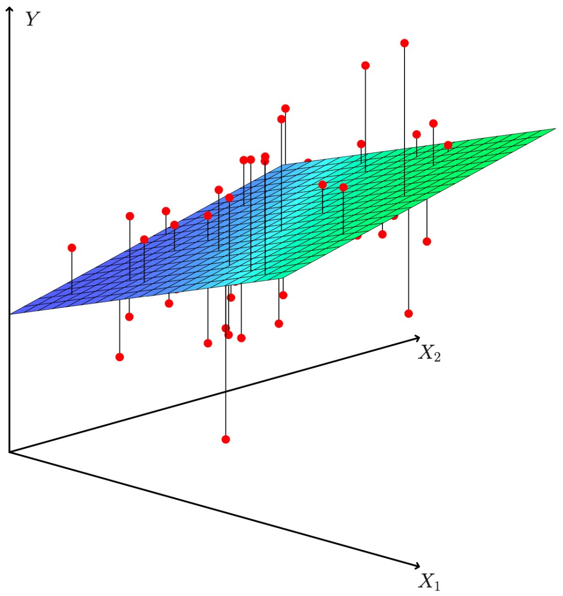
- **평면**: $\{(X_1, X_2, Y) \in \mathbb{R}^3 : \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 = Y\}$

Collection of Predicted Values

- 점: $(x_{i1}, x_{i2}, y_i) \in \mathbb{R}^3,\ i = 1, \ldots, n$

- 선: $y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}),\ i = 1, \ldots, n$

- **잔차 (residual)**: $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2})$



Statistical Inference

예측 (Prediction)

- **Prediction:** $X = x$ 값이 주어질 때 이와 대응되는 Y의 값 예측

- **Idea**: 조건부 **기댓값(평균)**: $X_i = x_i$ 라고 값이 주어졌을 경우, (x_i는 상수)

$$Model: Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

$$E(Y_i | X_i = x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

- $X_i = x_i$ 라고 값이 주어졌을 경우 가능한 Y의 값 중 **평균**으로 예측.

- **평균으로의 회귀(Regression)**

예측 (Prediction)

- **Idea**: 조건부 **기대값**: $X_i = x_i$ 라고 값이 주어졌을 경우, (x_i는 상수)

$$Model:\ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

$$E(Y_i|X_i = x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

조건부 기대값을 취하면 평균이 0인 오차항 제거

- **예측 (Prediction):** 적절한 추정값 $\hat{\beta}_j, j = 0, \dots, p$ 을 구했다면,

새로운 $X = x^* = (x_1^*, \dots, x_p^*)$값에 대응되는 Y의 예측은 조건부 기대값

$$\hat{y} = E(Y|X = x^*) = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \hat{\beta}_2 x_2^* + \cdots + \hat{\beta}_p x_p^*$$

- **적합 값 (Fitted values):** 이미 관측된 $x_i, i = 1, \dots, n$ 에 대응 하는 y값

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}, \qquad i = 1, \dots, n$$

추론

- *X*들이 *Y* 에 어떻게 영향을 미치는가*?* 각 변수별로 *Positive? Negative?* 얼마나*?*
- 해당 데이터가 *Linear Regression* 하는게 적합한가*?*
- *Y* 와 관계가 있는 *X* 변수들이 모두 다 모델에 필요한 변수들인가*?*
- *Linear regression* 결과가 믿을 만 한가*?*

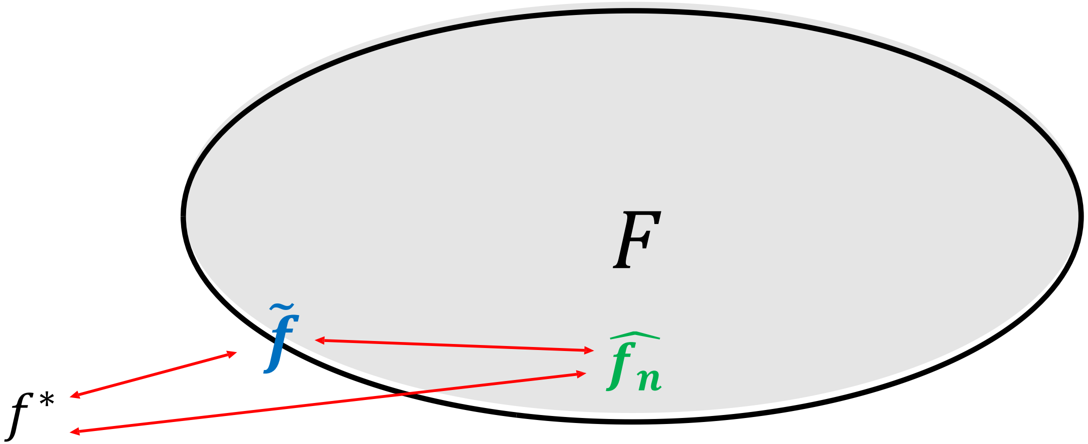
Recap: Function Estimation

Ideal goal: *Construct prediction rule* $f^* : \mathcal{X} \to \mathcal{Y}$

$$\boldsymbol{f^* = argmin_f \mathbb{E}_{XY}[loss(Y, f(X))]}$$

$$\tilde{f} = \arg\min_{f \in F} \mathbb{E}_{XY}[\mathrm{loss}(Y, f(X))]$$

$$\widehat{f_n} = \arg\min_{f \in F} \sum_{i=1}^{n} \mathrm{loss}\big(Y_i, f(X_i)\big)$$



F 의 구조가 단순할 수록 추론이 용이함

Model Interpretation

- **Idea**: 조건부 **기대값**: $X_i = x_i$ 라고 값이 주어졌을 경우, (x는 상수)

$$Model:\ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

$$E(Y_i | X_i = x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

오차항 제거 가능

- $\beta_j, j = 1, \ldots, p$: X의 j번째 변수가 1단위 커질 때마다 **Y의 평균이** β_j만큼 증가한다. (X의 다른 변수들이 통제된(*controlled*) 상태 하에서)

- β_0: X의 모든 값이 0일때 Y의 평균. 예측변수(X)항이 곱해져 있지 않기에 Y와 X의 관계를 보는 회귀분석에서는 β_0의 해석을 생략하는 경우가 많다.

추론: 모형 적합성

- **회귀모형이 적합한지 확인**하기 위해 회귀모형을 *Fit* 했을 때와 하지 않았을 때를 비교

- 만약 X 값에 관계 없이 Y를 예측한다면 $\hat{y}$은?

$$\hat{y} = E(Y\widehat{|X =} x^*) = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \hat{\beta}_2 x_2^* + \cdots + \hat{\beta}_p x_p^*$$

추론: 모형 적합성

- **회귀모형이 적합한지 확인**하기 위해 회귀모형을 *Fit* 했을 때와 하지 않았을 때를 비교

- 만약 X 값에 관계 없이 Y를 예측한다면 $\hat{y}$은?

$$\hat{y} = E(\widehat{Y|X} = x^*) = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \hat{\beta}_2 x_2^* + \cdots + \hat{\beta}_p x_p^*$$

$$E(\widehat{Y|X} = x^*) = \bar{Y}$$

추론: 모형 적합성

- **SST (총 편차제곱합)**

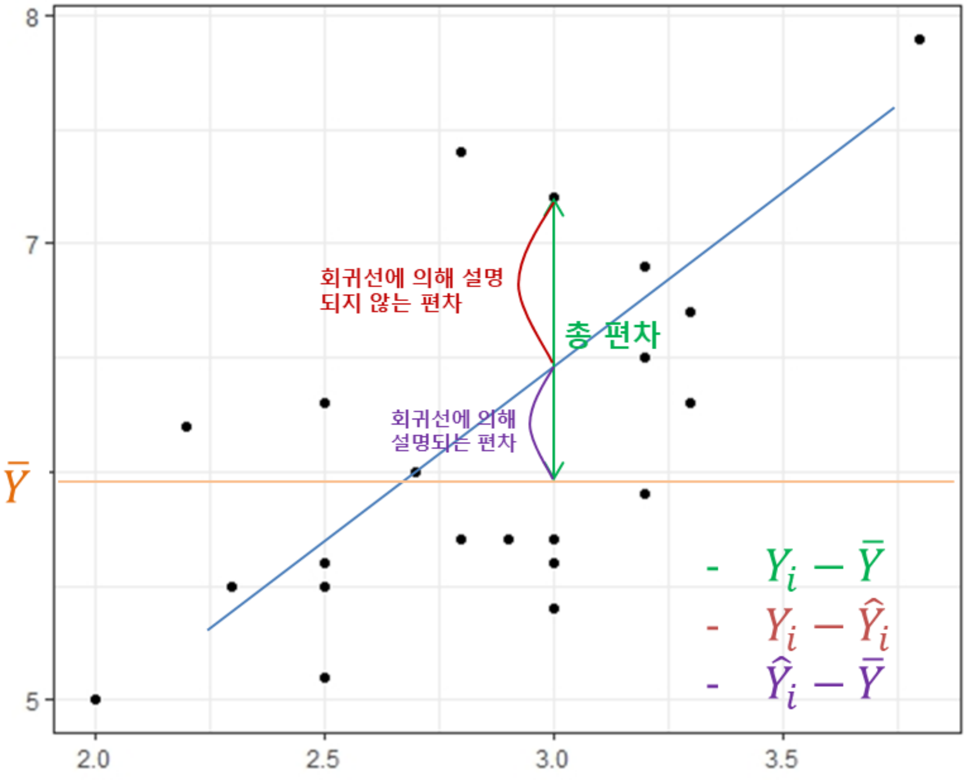
$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

- **SSR (회귀제곱합)**

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

- **SSE (오차제곱합)**

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



추론: R^2

- **SST (총 편차제곱합)**

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

- **SSR (회귀제곱합)**

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

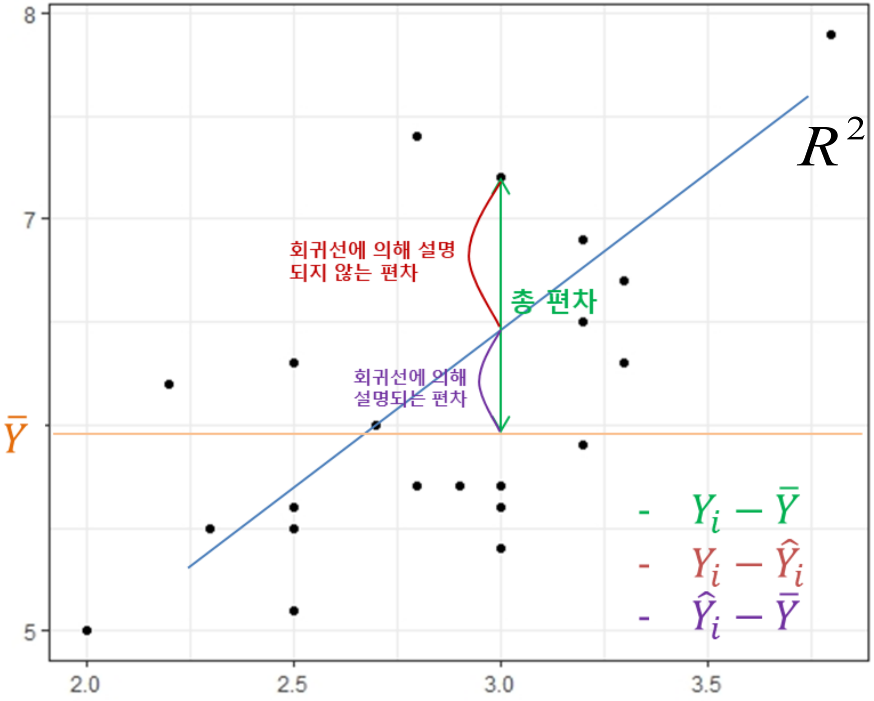
- **SSE (오차제곱합)**

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

결정계수 R^2(coefficient of determination)

회귀직선의 적합도를 평가하는 방법

전체변동에서 회귀로 설명되는 부분이 차지하는 비율



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

추론: 추정된 값의 분포

- *Central Limit Theorem:* 평균/(표본오차)는 정규분포로 수렴
- *Visualization (평균): Streamlit Example (sample mean)*
- *Visualization (Simple Linear Regression): Streamlit Example (simple linear regression)*

추론: 통계적 가설 검정(모델 전체)

- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \quad vs \quad H_1 : not\ H_0$

- $F = \frac{MSR}{MSE} \sim F_{(p,n-p-1)}\ under\ H_0$

- 만약 $F \geq F_\alpha(p, n-p-1)$ 또는 p값이 유의수준 α보다 작으면 H_0기각(*reject* H_0),
  α는 보통 *0.05* 또는 *0.01.*

추론 : 통계적 가설 검정(개별 변수)

- 모델 가정 하에서 *(독립성, 정규성 등) LSE* 를 이용한 $\hat{\beta}_j$의 분포는

$$\frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim t(n-p-1)$$

- 가설검정 *(testing hypothesis):*

$$H_0: \beta_j = 0 \left(X\text{의 } j\text{번째 변수와 } Y\text{는 선형관계가 없다}\right). \quad vs. \quad H_1: \beta_j \neq 0$$

- 검정통계량 *(test statistic):*

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim t(n-p-1) \; under \; H_0$$

만약 $|t| \geq t_{\frac{\alpha}{2}}(n-p-1)$ 또는 p값이 유의수준 α보다 작으면 H_0기각

Things to Consider (Multicollinearity)

Recall

$$\hat{\beta} = \mathrm{argmin}_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \left(y_i - \beta_0 \cdot 1 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}\right)^2$$
$$= \mathrm{argmin}_{\beta \in \mathbb{R}^{p+1}} (Y - \mathbf{X}\beta)^{\mathrm{T}}(Y - \mathbf{X}\beta)$$

- **The Analytic Solution**
- *Estimating Equation:* $(\boldsymbol{X}^T\boldsymbol{X})\hat{\beta} = \boldsymbol{X}^T\boldsymbol{Y}$

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$
$$\hat{\mathbf{Y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

Suppose that there are $c_0, c_1, \ldots, c_n \in \mathbb{R}$, *such that*

$$X_j = c_0 1_n + c_1 \mathrm{X}_1 + \cdots + c_{j-1}\mathrm{X}_{j-1} + c_{j+1}\mathrm{X}_{j+1} + \cdots + c_p \mathrm{X}_p + \delta,$$

If $\delta = 0$*, ?*

Things to Consider (Multicollinearity)

Suppose that there are $c_0, c_1, \dots, c_n \in \mathbb{R}$, such that

$$X_j = c_0 1_n + c_1 \mathrm{X}_1 + \cdots + c_{j-1}\mathrm{X}_{j-1} + c_{j+1}\mathrm{X}_{j+1} + \cdots + c_p \mathrm{X}_p + \delta,$$

If δ = 0,

- 관측된 *X* 변수들이 서로 상관관계가 *1*이면
  - $(\boldsymbol{X}^T\boldsymbol{X})\hat{\beta} = \boldsymbol{X}^T\boldsymbol{Y}$ **:** 해가 존재하지 않음.

- 관측된 *X* 변수들이 서로 상관관계가 매우 높다면 *(δ 가 0에 가까움)*
  - $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ <- 이 값이 매우 불안정함 *(분산이 매우 높음!)*
  - 계산이 가능하더라도 결과값에 대한 신뢰도는 매우 낮게 됨

- 통계모델 학습 자체는 가능. 하지만 결과값이 *j*번째 변수에 의한 것인지 아니면 다른 변수에 의한 것인지 판단이 어려움.

Things to Consider

- *Input* 변수간 상관관계가 높을 경우 어떻게 처리하나*?*
    - *(*키*,* 몸무게*),* 이미지의 첫번째 픽셀과 두번째 픽셀*, …*

- *Input* 변수가 과하게 많은 경우는*? (*해가 존재하지 않음*)*

- *Input* 변수에 *Categorical* 변수가 *mixed* 된 경우는*?*
    - 성별*,* 소속*,* 국적*, etc*

- 변수간 *synergy effect* 는*?*

- *Missing? Outlier?* 데이터 소실*,* 잘못 입력된 데이터셋*, etc*