The Simple Linear Regression Model

GR 5205 / GU 4205
Section 2/ Section 3

Columbia University
Xiaofei Shi

Last Class and Homework 1

- Modeling: Trying to find a function g(X) to minimize mean squared loss

Minimizer:

$$\text{argmin}_{g(x)} \mathbb{E}[(Y - g(X))^2] = \mathbb{E}[Y|X]$$

- Modeling: Linear relationship

$$Y = \beta_0 + X\beta_1 + \epsilon$$

We want to link: $$\mathbb{E}[Y|X] = \beta_0 + X\beta_1$$

Question: does there exists real joint distribution of (X, Y) such that the linear conditional expectation relationship holds?

Answer: yes when (X,Y) follows bivariate normal distribution!

Important takeaways

- Motivation for the least square estimator (LSE):

  why we consider correlation between X and Y

  Expect: $$Y = rX + \text{constant} + \text{noise}$$

- Normalization rocks (and why we usually normalize the numerical variables

  once we get them)

The Simple Linear Regression Model

More general case...

- Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be samples from the same model
- If the SLR model holds, we write $Y_i = \beta_0 + X_i\beta_1 + \epsilon_i$,
- Here, ϵ_i satisfies $\mathbb{E}[\epsilon_i] = 0, \mathrm{Var}[\epsilon_i] = \sigma^2$, and for $i \neq j$, $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0.$
- Observations: predictor : $x_1, x_2, \ldots, x_n$ response : $y_1, y_2, \ldots, y_n$
- Preference: $Q = \sum_{i=1}^{n}(y_i - \beta_0 - x_i\beta_1)^2$
- Model parameters: $\beta_0, \beta_1 (, \sigma^2)$

General Methodology

- Preference + data $\Rightarrow$ Q = Q(model parameters; data)
- Estimation of model parameters $\Leftrightarrow$ Minimizing Q wrt model parameters

$\Rightarrow$ Taking partial derivatives of Q wrt model parameters and set them to 0!

Prediction and residual

$$b_1 = \frac{(x - \bar{x}1_n)^\top (y - \bar{y}1_n)}{\|x - \bar{x}1_n\|^2} \qquad b_0 = \bar{y} - \bar{x}b_1$$

- Prediction: $\hat{y}_i = b_0 + x_i b_1$
- Residual: $e_i = y_i - \hat{y}_i = y_i - b_0 - x_i b_1$
- Residual can be viewed as the estimation of unobservable error terms

$$\hat{\epsilon}_i = e_i = y_i - \hat{y}_i = y_i - b_0 - x_i b_1$$

- Estimation of $\widehat{\sigma}^2 = \mathrm{MSE} = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\|y - \hat{y}\|^2}{n-2}$

n-2?