Linear Regression

1 Derived Distributions of OLS Estimates

1.1 General Matrix Expression

$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \overset{\text{i.i.d}}{\sim} Dist(\boldsymbol{0}_n, \sigma^2 \boldsymbol{I}_{n\times n})$
$\boldsymbol{\beta} = \boldsymbol{(X'X)^{-1}X'Y}$
$cov(\boldsymbol{\beta}) = \sigma^2 \boldsymbol{(X'X)^{-1}}$
$\boldsymbol{\beta} \sim Dist(\boldsymbol{\beta}, \sigma^2 \boldsymbol{(X'X)^{-1}})$

1.2 When Errors are Normally Distributed

$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where $\epsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$

1. $\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum\limits_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum\limits_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum\limits_{i=1}^{n}(X_i - \bar{X})^2} = \sum\limits_{i=1}^{n} W_i Y_i$ (i.e., weighted sum), where

$W_i = \frac{(X_i - \bar{X})}{\sum\limits_{i=1}^{n}(X_i - \bar{X})^2}$, so $\sum\limits_{i=1}^{n} W_i (X_i - \bar{X}) = \sum\limits_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sum\limits_{i=1}^{n}(X_i - \bar{X})^2} = 1$, then

$$\begin{aligned}
\sum_{i=1}^{n} W_i Y_i &= \sum_{i=1}^{n} W_i(\beta_0 + \beta_1 X_i + \epsilon_i) = \sum_{i=1}^{n} W_i(\bar{Y} - \beta_1 \bar{X} - \bar{\epsilon} + \beta_1 X_i + \epsilon_i) \\
&= \bar{Y}\sum_{i=1}^{n} W_i - \bar{\epsilon}\sum_{i=1}^{n} W_i + \beta_1 \sum_{i=1}^{n} W_i (X_i - \bar{X}) + \sum_{i=1}^{n} W_i \epsilon_i \\
&= \bar{Y}\sum_{i=1}^{n} W_i - 0 + \beta_1 + \sum_{i=1}^{n} W_i \epsilon_i = \hat{\beta}_1
\end{aligned}$$

$\hat{\beta}_1 - \beta_1 = \text{constant} + \sum\limits_{i=1}^{n} W_i \epsilon_i$

$$Var(\hat{\beta}_1) = E[(\hat{\beta}_1 - \beta_1)^2] = E[(\sum_{i=1}^{n} W_i \epsilon_i)^2] = E[\sum_{i=1}^{n} W_i^2 \epsilon_i^2 + \sum_{i=1}^{n} \sum_{j \neq i}^{n} W_i W_j \epsilon_i \epsilon_j]$$

$$= \sum_{i=1}^{n} W_i^2 E(\epsilon_i^2) + \sum_{i=1}^{n} \sum_{j \neq i}^{n} W_i W_j E(\epsilon_i \epsilon_j) = \sum_{i=1}^{n} W_i^2 E(\epsilon_i^2)$$

$E(\epsilon_i \epsilon_j) = 0$ due to the uncorrelation assumption between error pair

$$= \sigma^2 \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^4} = \frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\boxed{\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2})}$$

2. $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

$$Var(\hat{\beta}_0) = Var(\bar{Y} - \hat{\beta}_1 \bar{X}) = Var(\frac{1}{n} \sum_{i=1}^{n} Y_i) + (\bar{X})^2 Var(\hat{\beta}_1)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} Var(\epsilon_i) + \frac{\sigma^2 \bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\boxed{\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}))}$$

3. $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

$$Var(\hat{Y}) = Var(\hat{\beta}_0 + \hat{\beta}_1 X) = Var(\hat{\beta}_0) + X^2 Var(\hat{\beta}_1)$$

$$= \sigma^2(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}) + X^2 \frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$= \sigma^2(\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2})$$

$$\boxed{\hat{Y} \sim N(\beta_0 + \beta_1 X, \sigma^2(\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}))}$$