Properties of Least Square Estimator

Pandega Abyan Zumarsyah

November 2023

Known General Properties

Expected value has linearity property:

$$E[aX + bY] = aE[X] + bE[Y]$$
$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

From expected value, variance and covariance are defined as:

$$Var(X) = E[(X - E[X])^2]$$
$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Known Regression Properties

Relation between regressor/response and the true regression line can be defined as:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$B_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad B_0 = \bar{Y} - B_1\bar{x}$$

with x is regressor, Y is response, ϵ is error, β_0 and B_0 are related to the intercept, while β_1 and B_1 are related to the slope. Note that Y, B_0, and B_1 are random variables while the others are deterministic

The error ϵ is a random variable that has mean and variance:

$$E[\epsilon_i] = 0, \quad Var(\epsilon_i) = \sigma^2$$

It can be assumed that values of response Y are independent to each other, thus the covariance:

$$Cov(Y_i, Y_j) = \begin{cases} Var(Y_i) \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

Useful General Properties

Variance has symmetric property:

$$\begin{aligned}
Var(-X) &= E[(-X - E[-X])^2] \\
&= E[((-1) - X - (-1)E[X])^2] \\
&= E[(-1)^2(X - E[X])^2] \\
&= E[(X - E[X])^2] \\
&= Var(X)
\end{aligned}$$

Variance also has addition rules:

$$\begin{aligned}
Var(X+Y) &= E[((X+Y) - E[X+Y])^2] \\
&= E[((X - E[X]) + (Y - E[Y]))^2] \\
&= E[(X - E[X])^2 + (Y - E[Y])^2 + 2(X - E[X])(Y - E[Y])] \\
&= E[(X - E[X])^2] + E[(Y - E[Y])^2] + 2E[(X - E[X])(Y - E[Y])] \\
&= Var(X) + Var(Y) + 2\,Cov(X, Y)
\end{aligned}$$

Note that the covariance is zero when X and Y are independent

For independent case, it can be simply extended to summation:

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i)$$

Covariance has properties related to multiplication with constants:

$$\begin{aligned}
Cov(aX, bY) &= E[(aX - E[aX])(bY - E[bY])] \\
&= E[(aX - aE[X])(bY - bE[Y])] \\
&= E[ab(X - E[X])(Y - E[Y])] \\
&= ab\,E[(X - E[X])(Y - E[Y])] \\
&= ab\,Cov(X, Y)
\end{aligned}$$

Covariance also has properties related to summation:

$$\begin{aligned}
Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} Y_j\right) &= E\left[\left(\sum_{i=1}^{n} X_i - E\left[\sum_{i=1}^{n} X_i\right]\right)\left(\sum_{j=1}^{n} Y_j - E\left[\sum_{j=1}^{n} Y_j\right]\right)\right] \\
&= E\left[\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} E[X_i]\right)\left(\sum_{j=1}^{n} Y_j - \sum_{j=1}^{n} E[Y_j]\right)\right] \\
&= E\left[\sum_{i=1}^{n} (X_i - E[X_i]) \sum_{j=1}^{n} (Y_j - E[Y_j])\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} E[(X_i - E[X_i])(Y_j - E[Y_j])] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, Y_j)
\end{aligned}$$

Useful Regression Properties

Expected value of response:

$$\begin{aligned} E[Y_i] &= E[\beta_0 + \beta_1 x_i + \epsilon_i] \\ &= E[\beta_0] + E[\beta_1 x_i] + E[\epsilon_i] \\ &= \beta_0 + \beta_1 x_i \end{aligned}$$

Variance of response:

$$\begin{aligned} Var(Y_i) &= E[(Y_i - E[Y_i])^2] \\ &= E[((\beta_0 + \beta_1 x_i + \epsilon_i) - (\beta_0 + \beta_1 x_i))^2] \\ &= E[(\epsilon_i)^2] \\ &= E[(\epsilon_i)^2] + E[\epsilon]^2 \\ &= Var(\epsilon) \\ &= \sigma^2 \end{aligned}$$

Simplification of slope random variable B_1:

$$\begin{aligned} B_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i - (x_i - \bar{x})\bar{Y}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i - \bar{Y} \sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i - \bar{Y} \cdot 0}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

Mean of B_1

$$
\begin{aligned}
E[B_1] &= E\left[\frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})E[Y_i]}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})\beta_0}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\sum_{i=1}^n (x_i - \bar{x})\beta_1 x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\beta_0 \sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\beta_1 \sum_{i=1}^n (x_i - \bar{x})x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\beta_0 \cdot 0}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\beta_1 \sum_{i=1}^n (x_i^2 - \bar{x}x_i)}{\sum_{i=1}^n (x_i^2 + \bar{x}^2 - 2x_i\bar{x})} \\
&= \beta_1 \frac{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n \bar{x}x_i}{\sum_{i=1}^n x_i^2 + \sum_{i=1}^n \bar{x}^2 - \sum_{i=1}^n 2x_i\bar{x}} \\
&= \beta_1 \frac{\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 + \bar{x}^2 \sum_{i=1}^n 1 - 2\bar{x}\sum_{i=1}^n x_i} \\
&= \beta_1 \frac{\sum_{i=1}^n x_i^2 - \bar{x}(n\bar{x})}{\sum_{i=1}^n x_i^2 + \bar{x}^2 n - 2\bar{x}(n\bar{x})} \\
&= \beta_1 \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\
&= \beta_1
\end{aligned}
$$

Mean of B_0

$$
\begin{aligned}
E[B_0] &= E[\bar{Y} - B_1\bar{x}] \\
&= E[\bar{Y}] - E[B_1\bar{x}] \\
&= E\left[\frac{1}{n}\sum_{i=1}^n Y_i\right] - E[B_1]\bar{x} \\
&= \frac{1}{n}\sum_{i=1}^n E[Y_i] - E[B_1]\frac{1}{n}\sum_{i=1}^n x_i \\
&= \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 x_i) - \beta_1 \frac{1}{n}\sum_{i=1}^n x_i \\
&= \frac{1}{n}\sum_{i=1}^n \beta_0 + \beta_1 \frac{1}{n}\sum_{i=1}^n x_i - \beta_1 \frac{1}{n}\sum_{i=1}^n x_i \\
&= \frac{1}{n}\sum_{i=1}^n \beta_0 \\
&= \beta_0
\end{aligned}
$$

Variance of B_1

$$
\begin{aligned}
Var(B_1) &= Var\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) \\
&= Var\left(\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{S_{xx}}Y_i\right) \\
&= \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{S_{xx}}\right)^2 Var(Y_i) \\
&= \sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{S_{xx}^2}\sigma^2 \\
&= \frac{\sigma^2}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \frac{\sigma^2}{S_{xx}^2}S_{xx} \\
&= \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}
\end{aligned}
$$

Variance of B_0

$$
\begin{aligned}
Var(B_0) &= Var(\bar{Y}-B_1\bar{x}) \\
&= Var(\bar{Y}) + Var(-B_1\bar{x}) + 2Cov(\bar{Y},-B_1\bar{x}) \\
&= Var(\bar{Y}) + \bar{x}^2 Var(B_1) - 2\bar{x}\,Cov(\bar{Y},B_1)
\end{aligned}
$$

Before continuing, $Var(\bar{Y})$ and $Cov(\bar{Y},B_1)$ need to be solved:

$$
\begin{aligned}
Var(\bar{Y}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right) \\
&= \left(\frac{1}{n}\right)^2 Var\left(\sum_{i=1}^{n}Y_i\right) \\
&= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n}Var(Y_i) \\
&= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n}\sigma^2 \\
&= \left(\frac{1}{n}\right)^2 n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

Note that the values of Y are independent to each other so that variance of summation can simply become summation of variance

$$
\begin{aligned}
Cov(\bar{Y}, B_1) &= Cov\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \frac{\sum_{j=1}^{n}(x_j - \bar{x})Y_j}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right) \\
&= Cov\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \sum_{j=1}^{n}\frac{(x_j - \bar{x})}{S_{xx}}Y_j\right) \\
&= Cov\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \frac{1}{S_{xx}}\sum_{=1}^{n}(x_j - \bar{x})Y_j\right) \\
&= \frac{1}{nS_{xx}} Cov\left(\sum_{i=1}^{n} Y_i, \sum_{j=1}^{n}(x_j - \bar{x})Y_j\right) \\
&= \frac{1}{nS_{xx}}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_j - \bar{x})\, Cov(Y_i, Y_j) \\
&= \frac{1}{nS_{xx}}\sum_{j=1}^{n}(x_j - \bar{x})\sum_{i=1}^{n} Cov(Y_i, Y_j) \\
&= \frac{1}{nS_{xx}}\sum_{j=1}^{n}(x_j - \bar{x})\left(Cov(Y_j, Y_j) + \sum_{i=1, i\neq j}^{n} Cov(Y_i, Y_j)\right) \\
&= \frac{1}{nS_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})(Var(Y_j) + 0) \\
&= \frac{1}{nS_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})\sigma^2 \\
&= \frac{\sigma^2\sum_{k=1}^{n}(x_k - \bar{x})}{nS_{xx}} \\
&= \frac{\sigma^2 \cdot 0}{nS_{xx}} \\
&= 0
\end{aligned}
$$

With $Var(\bar{Y})$ and $Cov(\bar{Y}, B_1)$ in hand, the variance of B_0:

$$\begin{aligned}
Var(B_0) &= Var(\bar{Y}) + \bar{x}^2 Var(B_1) - 2\bar{x}\, Cov(\bar{Y}, B_1) \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} - 0 \\
&= \sigma^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2 + n\bar{x}^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \sigma^2 \frac{\sum_{i=1}^n (x_i^2 + \bar{x}^2 - 2x_i\bar{x}) + n\bar{x}^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \sigma^2 \frac{\sum_{i=1}^n x_i^2 + \sum_{i=1}^n \bar{x}^2 - \sum_{i=1}^n 2x_i\bar{x} + n\bar{x}^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \sigma^2 \frac{\sum_{i=1}^n x_i^2 + \bar{x}^2 \sum_{i=1}^n 1 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \sigma^2 \frac{\sum_{i=1}^n x_i^2 + n\bar{x}^2 - 2n\bar{x}^2 + n\bar{x}^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \sigma^2 \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}
\end{aligned}$$