Extra Proofs

Regression, F24 - CC

2024-08-29

(Variance of Intercept Est) For the simple linear regression model discussed in lecture, show that

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right).$$

Proof:

Recall

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{j=1}^n (X_j - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{j=1}^n (X_j - \bar{X})^2}.$$

Note: Here I'm using j as the index dummy variable in the denominator just to prevent confusing it with the index in the numerator. It's also acceptable to just call the entire denominator $SSX = \sum_{j=1}^n (X_j - \bar{X})^2$.

Let $k_i = \frac{(X_i - \bar{X})}{\sum_{j=1}^n (X_j - \bar{X})^2}$ (think of them as weights), so that $\hat{\beta}_1 = \sum_{i=1}^n k_i Y_i$.

$\hat{\beta}_0$ can be written as a linear combination of Y_i's:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{1}{n} \sum_{i=1}^n Y_i - \sum_{i=1}^n k_i Y_i \bar{X} = \sum_{i=1}^n \left(\frac{1}{n} - k_i \bar{X} \right) Y_i$$

So,

$$\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var} \left(\sum_{i=1}^n \left(\frac{1}{n} - k_i \bar{X} \right) Y_i \right) = \sum_{i=1}^n \mathrm{Var} \left(\left(\frac{1}{n} - k_i \bar{X} \right) Y_i \right) = \sum_{i=1}^n \left(\frac{1}{n} - k_i \bar{X} \right)^2 \sigma^2$$

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

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

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

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

$$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{j=1}^n (X_j - \bar{X})^2} \right) \text{ since } \sum_{i=1}^n (X_i - \bar{X}) = n\bar{X} - n\bar{X} = 0 \text{ in the middle term.}$$

(Gauss-Markov) For the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, \ldots, n,$$

where ε_i are iid with mean 0 and variance σ^2. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the least squares estimators for β_0 and β_1, respectively, and $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ be the fitted value. Prove the Gauss-Markov theorem for the least squares estimator of the slope.

In other words: For any other linear unbiased estimator for β_1, show that $\hat{\beta}_1$ has the smallest variance, i.e., let $\tilde{\beta}_1 = \sum_{i=1}^n c_i Y_i$, such that $\mathrm{E}(\tilde{\beta}_1) = \beta_1$, then $\mathrm{Var}(\tilde{\beta}_1) \geq \mathrm{Var}(\hat{\beta}_1)$.

Proof:

We want to show that our OLS estimator is BLUE (the best linear unbiased estimate). To do this, we define a new linear unbiased estimator $\tilde{\beta}_1$ and show that its variance must be larger than $\hat{\beta}_1$'s variance is.

Proof: Recall we can express $\hat{\beta}_1$ as $\hat{\beta}_1 = \sum_{i=1}^n k_i Y_i$ where $k_i = \dfrac{x_i - \bar{x}}{SSX}$.

Then write

$$\begin{aligned}
\tilde{\beta}_1 &= \sum_{i=1}^n \tilde{k}_i Y_i \\
&= \sum_{i=1}^n (\tilde{k}_i - k_i + k_i) Y_i \\
&= \sum_{i=1}^n (\tilde{k}_i - k_i) Y_i + \sum_{i=1}^n k_i Y_i \\
&= \sum_{i=1}^n d_i Y_i + \hat{\beta}_1 \qquad \text{(where } d_i = \tilde{k}_i - k_i\text{)}
\end{aligned}$$

Now we attempt to find out some information about these d_i's with the information that $\tilde{\beta}_1$ is unbiased, i.e. $E(\tilde{\beta}_1) = \beta_1$.

$$\begin{aligned}
E(\tilde{\beta}_1) = \beta_1 &\implies E(\sum_{i=1}^n d_i Y_i + \hat{\beta}_1) = \beta_1 \\
&\implies \sum_{i=1}^n d_i E(Y_i) + E(\hat{\beta}_1) = \beta_1 \\
&\implies \sum_{i=1}^n d_i (\beta_0 + \beta_1 x_i) + \beta_1 = \beta_1 \\
&\implies \beta_0 \sum_{i=1}^n d_i + \beta_1 \sum_{i=1}^n d_i x_i + \beta_1 = \beta_1 \\
&\implies \beta_0 \sum_{i=1}^n d_i + \beta_1 (\sum_{i=1}^n d_i x_i + 1) = \beta_1
\end{aligned}$$

We need to match coefficients. By this I mean the coefficients of β terms on the right hand side and left hand side must agree. For β_0, the coefficient on the left ($\sum_{i=1}^n d_i$) must equal the coefficient on the right (0). So:

$$\sum_{i=1}^n d_i = 0. \tag{A}$$

Similarly, the coefficient of β_1 on the left ($\sum_{i=1}^n d_i x_i + 1$) must equal the coefficient of β_1 on the right (1). So:

$$\sum_{i=1}^n d_i x_i + 1 = 1 \implies \sum_{i=1}^n d_i x_i = 0. \tag{B}$$

We will use these results later, so I'll label them (A) and (B) for reference. Next, we calculate the variance of $\tilde{\beta}$, which we express as:

$$\tilde{\beta} = \sum_{i=1}^n d_i Y_i + \hat{\beta}_1 = \sum_{i=1}^n d_i Y_i + \sum_{i=1}^n k_i Y_i = \sum_{i=1}^n (d_i + k_i) Y_i$$

, so that,

$$\begin{aligned}
Var(\tilde{\beta}) &= Var(\sum_{i=1}^n (d_i + k_i) Y_i) \\
&= \sum_{i=1}^n (d_i + k_i)^2 Var(Y_i) \\
&= \sigma^2 \sum_{i=1}^n (d_i + k_i)^2 \\
&= \sigma^2 \sum_{i=1}^n (d_i^2 + 2 d_i k_i + k_i^2) \\
&= \sigma^2 (\sum_{i=1}^n k_i^2 + 2 \sum_{i=1}^n d_i k_i + \sum_{i=1}^n d_i^2)
\end{aligned}$$

Let's focus on the middle term and plug in k_i:

$$\begin{aligned}
\sum_{i=1}^n d_i k_i &= \sum_{i=1}^n d_i \frac{x_i - \bar{x}}{SSX} \\
&= \frac{1}{SSX} \Big[\sum_{i=1}^n d_i x_i - \bar{x} \sum_{i=1}^n d_i \Big] \\
&= \frac{1}{SSX} (0 - 0) = 0 \qquad \text{using (A) and (B)}
\end{aligned}$$

Then we're done since:

$$\begin{aligned}
Var(\tilde{\beta}) &= \sigma^2 (\sum_{i=1}^n k_i^2 + \sum_{i=1}^n d_i^2) \\
&\geq \sigma^2 \sum_{i=1}^n k_i^2 = Var(\hat{\beta}_1)
\end{aligned}$$

because $\sum_{i=1}^n d_i^2 \geq 0$ always as it's the sum of squared numbers (which are always nonnegative).