HANDOUT ON THE OLS ESTIMATOR

Contents

1. Derivation of the OLS Estimator

(1) Consider a linear regression model with an intercept and only one explanatory variable (for n observations):

$$Y_i = \beta_0 + \beta_1 X_i + u_i.$$

Show that the least squares estimator for the slope coefficient β_1 is given by

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}$$

and the least squares estimator for the intercept is given by

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

Solution: The least-squares estimator is the solution for

$$(\widehat{\beta}_0, \widehat{\beta}_1)' = \underset{\beta_0, \beta_1}{\arg\min} \sum_{i=1}^n \left(Y_i - \beta_0 - \beta_1 X_i\right)^2.$$

We can find the analytical solution for this minimization problem using tools from calculus. At the critical points of a function, the first derivatives must be zero. Hence, we need to take the first (partial) derivative of the objective function with respect to β_0 and β_1, and set them to zero. This yields two equations for two unknowns for which the solution exists.

$$\frac{\partial \sum_{i=1}^n \left(Y_i - \beta_0 - \beta_1 X_i\right)^2}{\partial \beta_0} = -2\sum_{i=1}^n \left(Y_i - \beta_0 - \beta_1 X_i\right) \overset{set}{=} 0$$

$$\frac{\partial \sum_{i=1}^n \left(Y_i - \beta_0 - \beta_1 X_i\right)^2}{\partial \beta_1} = -2\sum_{i=1}^n X_i\left(Y_i - \beta_0 - \beta_1 X_i\right) \overset{set}{=} 0$$

From the first equation, we can solve for β_0.

$$\begin{aligned}
-2\sum_{i=1}^n \left(Y_i - \beta_0 - \beta_1 X_i\right) = 0 &\Rightarrow \sum_{i=1}^n \left(Y_i - \beta_0 - \beta_1 X_i\right) = 0 \\
&\Rightarrow \sum_{i=1}^n \left(Y_i - \beta_1 X_i\right) - \sum_{i=1}^n \beta_0 = 0 \\
&\Rightarrow \sum_{i=1}^n \left(Y_i - \beta_1 X_i\right) = n\beta_0
\end{aligned}$$

Solution for β_0 is

$$\widehat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \beta_1 X_i\big) = \frac{1}{n}\sum_{i=1}^{n} Y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{Y} - \beta_1 \bar{X}.$$

Substituting the solution for β_0 into the second equation yields

$$\begin{aligned}
-2\sum_{i=1}^{n} X_i\big(Y_i - [\bar{Y} - \beta_1\bar{X}] - \beta_1 X_i\big) = 0 &\Rightarrow \sum_{i=1}^{n} X_i\big(Y_i - [\bar{Y} - \beta_1\bar{X}] - \beta_1 X_i\big) = 0 \\
&\Rightarrow \sum_{i=1}^{n} X_i\big(Y_i - \bar{Y}\big) - \beta_1 \sum_{i=1}^{n} X_i\big(X_i - \bar{X}\big) = 0.
\end{aligned}$$

Hence, solution for β_1 is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i\big(Y_i - \bar{Y}\big)}{\sum_{i=1}^{n} X_i\big(X_i - \bar{X}\big)}.$$

Let's take a closer look at the numerator and the denominator. For the numerator, we have

$$\begin{aligned}
\sum_{i=1}^{n} X_i\big(Y_i - \bar{Y}\big) &= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \bar{Y} \\
&= \sum_{i=1}^{n} X_i Y_i - \bar{Y}\sum_{i=1}^{n} X_i - n\bar{X}\bar{Y} + n\bar{X}\bar{Y} \\
&= \sum_{i=1}^{n} X_i Y_i - \bar{Y}\sum_{i=1}^{n} X_i - \bar{X}\sum_{i=1}^{n} Y_i + n\bar{X}\bar{Y} \\
&= \sum_{i=1}^{n} X_i Y_i - \bar{Y}\sum_{i=1}^{n} X_i - \bar{X}\sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} \bar{X}\bar{Y} \\
&= \sum_{i=1}^{n} \big(X_i Y_i - \bar{Y} X_i - \bar{X} Y_i + \bar{X}\bar{Y}\big) \\
&= \sum_{i=1}^{n} \big(X_i - \bar{X}\big)\big(Y_i - \bar{Y}\big)
\end{aligned}$$

which is a measure of sample covariance between X and Y. For the denominator, we have

$$\begin{aligned}
\sum_{i=1}^{n} X_i\big(X_i - \bar{X}\big) &= \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i \bar{X} \\
&= \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i \bar{X} - n\bar{X}^2 + n\bar{X}^2 \\
&= \sum_{i=1}^{n} X_i^2 - \bar{X}\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \bar{X}\bar{X} + \sum_{i=1}^{n} \bar{X}^2 \\
&= \sum_{i=1}^{n} X_i^2 - \bar{X}\sum_{i=1}^{n} X_i - \bar{X}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \bar{X}^2 \\
&= \sum_{i=1}^{n} \big(X_i^2 - 2\bar{X}X_i + \bar{X}^2\big) \\
&= \sum_{i=1}^{n} \big(X_i - \bar{X}\big)^2.
\end{aligned}$$

Now we can substitute $\widehat{\beta}_1$ back to $\widehat{\beta}_0$ to obtain the solution for β_0

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

2. Algebraic Properties of the OLS Estimator

(a) The least squares residuals sum to zero: $\sum_{i=1}^{n} \widehat{u}_i = 0$. Note that,

$$\begin{aligned}\sum_{i=1}^{n} \widehat{u}_i &= \sum_{i=1}^{n} \big(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i\big) \\ &= \sum_{i=1}^{n} \big(Y_i - [\bar{Y} - \widehat{\beta}_1 \bar{X}] - \widehat{\beta}_1 X_i\big) \\ &= \sum_{i=1}^{n} \big(Y_i - \bar{Y}\big) - \widehat{\beta}_1 \sum_{i=1}^{n} \big(X_i - \bar{X}\big) \\ &= \sum_{i=1}^{n} \big(Y_i - \bar{Y}\big) - \left(\frac{\sum_{i=1}^{n} X_i\big(Y_i - \bar{Y}\big)}{\sum_{i=1}^{n} X_i\big(X_i - \bar{X}\big)}\right) \sum_{i=1}^{n} \big(X_i - \bar{X}\big) \\ &= 0,\end{aligned}$$

because $\sum_{i=1}^{n} \big(Y_i - \bar{Y}\big) = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \bar{Y} = n\bar{Y} - n\bar{Y} = 0$ and $\sum_{i=1}^{n} \big(X_i - \bar{X}\big) = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \bar{X} = n\bar{X} - n\bar{X} = 0$.

(b) The correlation between residuals and regressors is zero: $\sum_{i=1}^{n} X_i \widehat{u}_i = 0$. Because,

$$\begin{aligned}\sum_{i=1}^{n} X_i \widehat{u}_i &= \sum_{i=1}^{n} X_i\big(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i\big) \\ &= \sum_{i=1}^{n} X_i\big(Y_i - [\bar{Y} - \widehat{\beta}_1 \bar{X}] - \widehat{\beta}_1 X_i\big) \\ &= \sum_{i=1}^{n} X_i\big(Y_i - \bar{Y}\big) - \widehat{\beta}_1 \sum_{i=1}^{n} X_i\big(X_i - \bar{X}\big) \\ &= \sum_{i=1}^{n} X_i\big(Y_i - \bar{Y}\big) - \left(\frac{\sum_{i=1}^{n} X_i\big(Y_i - \bar{Y}\big)}{\sum_{i=1}^{n} X_i\big(X_i - \bar{X}\big)}\right) \sum_{i=1}^{n} X_i\big(X_i - \bar{X}\big) \\ &= \sum_{i=1}^{n} X_i\big(Y_i - \bar{Y}\big) - \sum_{i=1}^{n} X_i\big(Y_i - \bar{Y}\big) = 0.\end{aligned}$$

(c) The regression line passes through the sample mean: $\bar{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{X} \equiv \bar{\widehat{Y}}$. Because,

$$\begin{aligned}\bar{Y} &= \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} (\widehat{Y}_i + \widehat{u}_i) = \frac{1}{n}\sum_{i=1}^{n} \widehat{Y}_i + \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i \\ &= \frac{1}{n}\sum_{i=1}^{n} (\widehat{\beta}_0 + \widehat{\beta}_1 X_i) + \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i = \frac{1}{n}\sum_{i=1}^{n} \widehat{\beta}_0 + \frac{1}{n}\sum_{i=1}^{n} \widehat{\beta}_1 X_i + \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i \\ &= \frac{1}{n}\sum_{i=1}^{n} \widehat{\beta}_0 + \widehat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{X} + 0 \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 \bar{X}.\end{aligned}$$