Introduction to Econometrics: Recitation 2

Seung-hun Lee

1 Least Squares Estimation

1.1 Main Assumptions of the OLS Estimators

For the ordinary least squares to have desired properties of unbiasednesss, consistency, efficiency, and asymptotic normality, the following assumptions must be made

Assumption 1.1. *Here are the assumptions for the classical linear regression model*

A1 *Linearity: The regression is assumed to be linear in parameters.*

$$\text{Okay: } Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

$$\text{Not: } Y_i = \beta_0 + \beta_1 X_i + \beta_2^2 X_i + u_i,$$

A2 $E(u_i|X_i) = 0$*: This means that conditional on letting* X_i *take a certain value, we are not making any systematical error in the linear regression. This is required for the OLS to be unbiased.*

A3 *Homoskedasticity:* $var(u_i) = \sigma_2$*, or variance of* u_i *does not depend on* X_i*. If this condition is broken, there exists a heteroskedasticity*

A4 *No Autocorrelation (Serial Correlation): For* $i \neq j$*,* $cov(u_i, u_j) = 0$*. In other words, error at the previous period does not have any impact on the current* *period. This is usually broken in time series settings, where the error in the previous period carries over to the next period.*

A5 *Orthogonality:* $cov(X_i, u_i) = 0$. *Or, the error term and the independent variable is uncorrelated. If otherwise, there exists an endogeneity.*

A6 $n > K$: *There should be more observations than independent variables.*

A7 *Variability in X: Independent variables should take somewhat different values for each observation.*

A8 *Correct Specification: The model has all the necessary independent variables in a correct functional form.*

A9 *No perfect multicollinearity: If one of the* X_i *variable is a linear combination of other variables, some of these variables are not estimated.*

A10 *i.i.d.:* (X_i, Y_i) *is assumed to be from independent, identical distribution*

A11 *No Outliers: Outlier has no impact on the regression results.*

1.2 Measure of Fitness

These numbers tell us how informative the sample linear regression we used is in telling us about the population data. We discussed two types of measure

- $\mathbf{R^2}$: It is defined as a fraction of total variation which is explained by the model. Mathematically, this is

$$Y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 X_i}_{\hat{Y}_i} + \hat{u}_i, \ \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} + \bar{\hat{u}},$$

$$\implies Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + \hat{u}_i$$

$$\implies \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} \hat{u}_i^2 + 2\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})\hat{u}_i$$

Note that

$$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})\hat{u}_i = \sum_{i=1}^{n}\hat{Y}_i\hat{u}_i - \bar{Y}\sum_{i=1}^{n}\hat{u}_i$$

Since the conditional mean of error u_i is assumed to be 0 (A2), and the covariance between X_i and u_i is 0 (A5), the above equation becomes zero. So we are left with

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{TSS} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{ESS} + \underbrace{\sum_{i=1}^{n}\hat{u}_i}_{RSS} \implies 1 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} + \frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Thus, the R^2 can be found as

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Intuitively, higher R^2 implies that the model explains more of the total variance, which implies that the regression fits the data well.

- **SER**: Standard Error of Regression. It estimate the standard deviation of the error term in Y_i, or mathematically

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

where $u_i = y_i - \hat{y}_i$ and we use $n-2$ since there is loss of d.f. by two due to $\hat{\beta}_0, \hat{\beta}_1$. If SER turns out to be large, this implies that our model might be missing a key variable.

- **RMSE**: Root mean squared error. It is similar to SER in terms of how it looks,

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}$$

this is used to assess the accuracy of the predictions.

1.3 Sampling distribution of OLS

Note that the OLS estimate that we are getting is a random variable - the estimate we get is different depending on which sample we work with. This is why we can discuss the distributional properties - mean and variance, in particular - of the OLS.

- $\hat{\beta}_1$: Recall that we can write

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

Now, replace Y_i an $\bar{Y}$ with

$$Y_i = \beta_0 + \beta X_i + u_i,\ \bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{u},$$

which allows us to write

$$(Y_i - \bar{Y}) = (\beta_1 (X_i - \bar{X}) + (u_i - \bar{u}))$$

and get

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

We are now ready to discuss the distributional properties

  - $E[\hat{\beta}_1]$: It can be written as

$$\begin{aligned} E[\hat{\beta}_1] &= E\left[\beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^n (X_i - \bar{X})^2}\right] \\ &= \beta_1 + E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^n (X_i - \bar{X})^2}\right] \end{aligned}$$

Note that the $\sum_{i=1}^n (X_i - \bar{X})(u_i - \bar{u})$ can be written to something simpler. This is equal to

$$\sum_{i=1}^n (X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^n X_i u_i - \bar{u} \sum_{i=1}^n X_i + \bar{X} \sum_{i=1}^n u_i + n\bar{X}\bar{u}$$

Since $\bar{X}$ is a sample mean of X, $\sum_{i=1}^{n} X_i = n\bar{X}$. The assumption that conditional mean is zero and (X_i, u_i) are uncorrelated means that the term on the left hand side is zero. Therefore, IF THE CLASSICAL ASSUMPTIONS ARE VALID, $E[\hat{\beta}_1] = \beta_1$.

– $var[\hat{\beta}_1]$: We use the definition of the variances and the fact that the expected value of $\hat{\beta}_1$ is unbiased (at least for now) to get

$$\begin{aligned}
var(\hat{\beta}_1) &= E\left[\left(\hat{\beta}_1 - E[\hat{\beta}_1]\right)^2\right] \\
&= E\left[\left(\hat{\beta}_1 - \beta_1\right)^2\right] \\
&= E\left[\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)^2\right] \\
&= E\left[\left(\frac{(X_1 - \bar{X})(u_1 - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + ... + \frac{(X_n - \bar{X})(u_n - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)^2\right]
\end{aligned}$$

At the moment, we are assuming homoskedasticity and no autocorrelation (A3, A4). Since X_i is from the data and u_i is a random error term, we can take all the X_i terms in and keep the u_i terms in the expectation to get (i.i.d assumption is also useful here)

$$\begin{aligned}
var(\hat{\beta}_1) &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 E[(u_i - \bar{u})^2]}{[\sum_{i=1}^{n}(X_i - \bar{X})^2]^2} \\
&= \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sigma_u^2}{[\sum_{i=1}^{n}(X_i - \bar{X})^2]^2} \ (\because E[(u_i - \bar{u})^2 = var(u_i)) \\
&= \sigma_u^2 \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{[\sum_{i=1}^{n}(X_i - \bar{X})^2]^2} = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}
\end{aligned}$$

Note that to decrease the variance in the estimates, the variance of the error should be small relative to the variation in the X_i.

At the end of the day, we can say the following about the distribution of our $\hat{\beta}_1$ estimator and use this to test our hypothesis

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

- $\hat{\beta}_0$: The formula for $\hat{\beta}_0$ is $\bar{Y} - \hat{\beta}_1 \bar{X}$. By changing $\bar{Y}$, we can get

$$\begin{aligned}\hat{\beta}_0 &= (\beta_0 + \beta_1 \bar{X} + \bar{u}) - \hat{\beta}_1 \bar{X} \\ &= \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{X} + \bar{u}\end{aligned}$$

  Then we can say the following about the sampling distribution

  - $E[\hat{\beta}_0]$: We can write

  $$E[\hat{\beta}_0] = \beta_0 + E[(\beta_1 - \hat{\beta}_1)\bar{X}] + E[\bar{u}] = \beta_0$$

    since $\hat{\beta}_1$ is unbiased and conditional expectation of u_i is zero. Thus, under our current assumptions, $\hat{\beta}_0$ is unbiased.

  - $var[\hat{\beta}_0]$: Using the definition of the variance, we can write

  $$\begin{aligned}var(\hat{\beta}_0) &= E\left[\left(\hat{\beta}_0 - E[\hat{\beta}_0]\right)^2\right] \\ &= E\left[\left(\hat{\beta}_0 - \beta_0\right)^2\right] \\ &= E\left[\left((\beta_1 - \hat{\beta}_1)\bar{X} + \bar{u}\right)^2\right] \\ &= \bar{X}^2 E\left[\left(\beta_1 - \hat{\beta}_1\right)^2\right] + 2\bar{X} E\left[\left(\beta_1 - \hat{\beta}_1\right)\bar{u}\right] + E[\bar{u}^2]\end{aligned}$$

    Under the assumption (A2), we can ignore the middle term as this is zero. The rest of the terms are $\bar{X}^2 var(\hat{\beta}_1)$ and $\frac{\sigma_u^2}{n}$. the final result is

  $$var(\hat{\beta}_0) = \frac{\sigma_u^2 \bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} + \frac{\sigma_u^2}{n} = \frac{\sigma_u^2}{n} \frac{\sum_{i=1}^n X_i^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

  So we can write

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