Recitation 3

Seung-hun Lee

Columbia University

Ordinary Least Squares

Assumptions

- For OLS to be unbiased, consistent, efficient, and asymptotic normal, the following assumptions must be made

Assumptions

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

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

A2 $E(u_i|X_i) = 0$: Conditional on X_i taking a certain value, we are not making any systematical error in the linear regression, required condition for unbiasedness.

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 $i \neq j$, $cov(u_i, u_j) = 0$. 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.

Ordinary Least Squares

Assumptions

Assumptions (Continued)

A5 Orthogonality: $cov(X_i, u_i) = 0$. 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.

Ordinary Least Squares

Measure of fitness

- These numbers tell us how informative the sample linear regression we used is in telling us about the population data
- **$\mathbf{R^2}$**: It is defined as a fraction of total variation which is explained by the model. Mathematically, this is

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

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

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

Ordinary Least Squares

Measure of fitness

- Note that

$$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})\hat{u}_i = \sum_{i=1}^{n} \hat{Y}_i\hat{u}_i - \bar{Y}\sum_{i=1}^{n} \hat{u}_i = 0$$

  This is because

  - The conditional mean of error u_i is assumed to be 0 (A2),
  - The covariance between X_i and u_i is 0 (A5),

- So we are left with

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

Ordinary Least Squares

Measure of fitness

- The above is equivalent to

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

- 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.

Ordinary Least Squares

Measure of fitness

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

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

  we use $n-2$ since there is loss of d.f. by two due to $\hat{\beta}_0, \hat{\beta}_1$.
  If SER is large, our model might be missing a key variable.

- **RMSE**: Root mean squared error. Similar to SER in looks,

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

  this is used to assess the accuracy of the predictions.

Ordinary Least Squares

Sampling distribution

- OLS estimate that we are getting is a random variable - getting different estimates depending on sample we work with.
- $\hat{\beta}_1$: 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

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

Ordinary Least Squares

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

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

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

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

→ 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, UNDER CLASSICAL ASSUMPTIONS, $E[\hat{\beta}_1] = \beta_1$.

Ordinary Least Squares

- $var[\hat{\beta}_1]$: 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

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

Ordinary Least Squares

→ We assume homoskedasticity and no autocorrelation

→ 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.

Ordinary Least Squares

- $\hat{\beta}_0$: 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.

$\rightarrow$ Thus, under our current assumptions, $\hat{\beta}_0$ is unbiased.

Ordinary Least Squares

- $var[\hat{\beta}_0]$: 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

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

Ordinary Least Squares

So what do we take away?

- At the end of the day, we can say

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

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

- The importance of this is that now we can conduct a hypothesis test and create a test statistic based on this distribution