Recitation 4

Seung-hun Lee

Columbia University

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

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

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

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

$\rightarrow$ Since $\bar{X}$ is a sample mean of X, $\sum_{i=1}^{n} X_i = n\bar{X}$.

$\rightarrow$ 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.

$\rightarrow$ 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

- $\rightarrow$ We assume homoskedasticity and no autocorrelation
- $\rightarrow$ 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.

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

Ordinary Least Squares

Hypothesis test

- From the sample distribution of $\hat{\beta}_1$, we can break down into two cases
- **Know** σ_u: Since the $\hat{\beta}_1$ takes a normal distribution, we can "standardize" it to get the test statistic and the distribution for it

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{var(\hat{\beta}_1)}} \sim N(0, 1)$$

  and compare against the critical values (depending on significance level, two vs one-sided test)

Ordinary Least Squares

Hypothesis test

- **Don't know** σ_u; need to have an estimate for $var(\hat{\beta}_1)$ due to not knowing σ_u. The test statistics and its distribution is

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{var(\hat{\beta}_1)}}} \sim t_{n-2}$$

  where $\widehat{var(\hat{\beta}_1)}$ is the estimate for the variance and t_{n-2} is a t-distribution with $n-2$ degrees of freedom.
- The d.f. is determined by the number of observations, where 2 is subtracted because we are estimating β_0 and β_1 in the process.
- When n is large, t-distribution becomes similar to the normal distribution

Confidence interval

- **Confidence interval**: A 95% confidence interval is a range of numbers that form a random interval that has a 95% chance of including a (nonrandom) true value of a parameter.
- This can be obtained by inverting the rejection region that we have used in the critical value approach.

$$\Pr\left(-1.96 \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{var(\hat{\beta}_1)}} \leq 1.96\right) = 0.95$$

$$\implies \Pr\left(\hat{\beta}_1 - 1.96 \times \sqrt{var(\hat{\beta}_1)} \leq \beta_1 \leq \hat{\beta}_1 + 1.96 \times \sqrt{var(\hat{\beta}_1)}\right) = 0.95$$

- If they encompass the null test value, then we cannot reject the null hypothesis. Otherwise, we can reject the null.

Ordinary Least Squares

Binary X_i

- We may be interested in whether there is a difference in outcome due to affiliation to a certain group, which is usually binary
- **Dummy variable**:

$$X_i = \begin{cases} 1 & \text{if } i \text{ belongs in group } X \\ 0 & \text{if otherwise} \end{cases}$$

- OLS method can be applied, with different interpretation

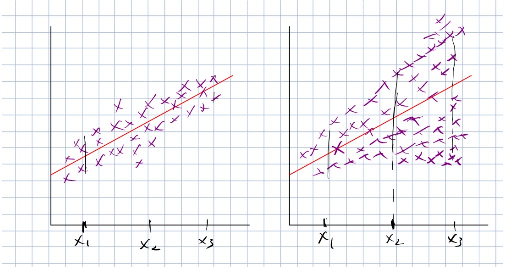
$$\begin{aligned} E[Y_i|X_i = 0] =& \beta_0 \\ E[Y_i|X_i = 1] =& \beta_1 + \beta_0 \end{aligned}$$

β_1 is then the difference in mean between $X_i = 1$ and $X_i = 0$ groups

Ordinary Least Squares

Heteroskedasticity

- The assumption that $var(u_i)$ is constant may not hold.
- If we stick to homoskedasticity in this case, the standard errors are incorrectly estimated (usually underestimated)



- In such case, standard errors of our estimators must take this into account.

Ordinary Least Squares

Consequences of heteroskedasticity

```
. regress testscr str

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.58
       Model |  7794.11919         1  7794.11919   Prob > F        =    0.0000
    Residual |  144315.475       418  345.252333   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0490
       Total |  152109.594       419  363.030058   Root MSE        =    18.581

------------------------------------------------------------------------------
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |   -2.27981   .4798255    -4.75   0.000    -3.222981   -1.336638
       _cons |    698.933   9.467491    73.82   0.000     680.3232    717.5428
------------------------------------------------------------------------------
```

```
. regress testscr str, vce(robust)

Linear regression                               Number of obs     =        420
                                                F(1, 418)         =      19.26
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0512
                                                Root MSE          =     18.581

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |   -2.27981   .5194894    -4.39   0.000    -3.300947   -1.258672
       _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

- The variance rises (usually) in the heteroskedastic regression, so we may make a wrong hypothesis test
- The coefficients are unchanged, since estimation of OLS estimates did not rely on homoskedasticity

Multivariate regression

Omitted variable bias

- Suppose that there are more than one possible independent variable
- The set of models are

$$\begin{aligned} \text{True: } Y_i &= \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i \\ \text{Mistake: } Y_i &= \beta_0 + \beta_1 X_i + u_i^* \\ \text{Sample: } Y_i &= \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i \end{aligned}$$

- Suppose you run an OLS regression without Z_i. $\hat{\beta}_1$ can be calculated as $\frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$. Replacing this with the true model gives

$$\begin{aligned} \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} &= \frac{\sum_{i=1}^n (X_i - \bar{X})(\beta_1 (X_i - \bar{X}) + \beta_2 (Z_i - \bar{Z}) + (u_i - \bar{u}))}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \beta_1 + \beta_2 \frac{\sum_{i=1}^n (X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^n (X_i - \bar{X})^2} + \frac{\sum_{i=1}^n (X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^n (X_i - \bar{X})^2} \end{aligned}$$

Multivariate regression

Omitted variable bias

- If $\beta_2 \neq 0$ and $\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \neq 0$, then the mean of $\hat{\beta}_1$ is not guaranteed to be β_1. This leads to the **omitted variable bias** problem
- This happens when both of the following cases hold
  - Z should explain Y: If the slope coefficient of Z (β_2) is nonzero, then the Z variable is part of the error term if we forget to include them
  - Z is correlated with X: If $cov(X, Z) \neq 0$ and the regression residual $\hat{u}$ is correlated with Z, the independent variable is now correlated with $\hat{u}$, which leads to violation of the assumption that independent variable and the residual are not correlated.
- We can even determine the direction of the bias
  - $\hat{\beta}_1$ is overestimated if $\beta_2 \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} > 0$
  - $\hat{\beta}_1$ is underestiated if $\beta_2 \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} < 0$

Multivariate regression

Omitted variable bias: What to do about it?

- We can simply include the Z variable if we have the data for it.
- Another way is to conduct an ideal randomized controlled experiment (or randomized control trial) that randomly assigns value of X to all students.
- If none of the two are feasible, we should find another variable that can be a proxy to Z - they have to be related to the X variable and is uncorrelated with the errors - which is the Instrumental Variable method.

Multivariate regression

Interpretation

- The technicalities involved do not change drastically compared to the univariate regression.
- However, one should interpret the coefficients cautiously. Suppose that the regression is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

- To see the impact of X_i and Y_i, one needs to take (partial) derivatives on Y_i with respect to X_i. This leads to

$$\beta_1 = \frac{\partial Y_i}{\partial X_i}$$

- In words, β_1 captures how much Y_i changes with respect to X_i *holding other variables constant* (ceteris paribus).
- If you do not hold other variables (Z_i in this case) fixed, the change will not exactly be β_1 (it could be more or less)