

OH : 4:30-6:30 pm (tomorrow only)

Recitation 3: OLS properties: Sampling distribution and fitness

Seung-hun Lee

Columbia University
Undergraduate Introduction to Econometrics Recitation

September 29th, 2022

Ordinary least squares

Ordinary Least Squares: Population vs sample linear models

- Suppose that the **population linear regression model** (also known as data generating process in some books) is

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

- However, we do not know the true values of the population parameters - β_0 and β_1
- An alternative way to approach the problem is to use the **sample linear regression model** (or just model)

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$$

where $\hat{\beta}_0, \hat{\beta}_1$ are estimates of β_0, β_1

Ordinary Least Squares: Definition

- The ideal estimator minimizes the squared sum of residuals.
- Mathematically, this can be obtained by solving the following minimization problem and the first order conditions

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

$$[\hat{\beta}_0] : -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$[\hat{\beta}_1] : -2 \sum_{i=1}^n X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

—

The resulting **least squares estimators** are

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

<Proof>

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$[\hat{\beta}_0]: -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\rightarrow \sum_{i=1}^n y_i - \hat{\beta}_0 n - \hat{\beta}_1 \sum_{i=1}^n x_i = 0$$

$$\rightarrow \underbrace{\frac{1}{n} \sum_{i=1}^n y_i}_{\bar{y}} - \hat{\beta}_0 - \hat{\beta}_1 \underbrace{\frac{1}{n} \sum_{i=1}^n x_i}_{\bar{x}} = 0 \rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$[\hat{\beta}_1]: -2 \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\rightarrow \sum_{i=1}^n x_i y_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0$$

$$\rightarrow \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0$$

$$\rightarrow \hat{\beta}_1 (\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i) = \sum x_i y_i - \bar{y} \sum x_i$$

$$\rightarrow \hat{\beta}_1 = \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum_{i=1}^n x_i^2 - \bar{x} \sum x_i} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Note: $\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \bar{y} \sum x_i - \bar{x} \sum y_i + n \bar{x} \bar{y}$

and $\sum y_i = n \bar{y}$, so $-\bar{x} \sum y_i = -n \bar{x} \bar{y}$

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - 2 \bar{x} \sum x_i + n(\bar{x})^2 = \sum x_i^2 - \bar{x} \sum x_i$$

(since $n(\bar{x})^2 = \bar{x} \sum x_i$)

Ordinary Least Squares: Main assumptions

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

Assumptions

A0 Linearity: The regression is assumed to be linear in parameters.

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

need for Sampling distributions

A1 $E(u_i|X_i) = 0$: 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. (or $\text{cov}(X_i, u_i) = 0$)

A2 i.i.d. (random sampling): (X_i, Y_i) is assumed to be from independent, identical distribution

A3 No Outliers: Outlier has no impact on the regression results. ($E(X_i^4), E(Y_i^4) < \infty$)

A4 Homoskedasticity: $\text{var}(u_i) = \sigma_u$ (variance of u_i does not depend on X_i). \leftrightarrow heteroskedasticity

A5 No Autocorrelation (Serial Correlation): For $i \neq j$, $\text{cov}(u_i, u_j) = 0$. Error at the previous period does not have any impact on the current period. This is usually broken in time series settings

Ordinary Least Squares: Useful alternative expression for $\hat{\beta}_1$

- 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

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

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: Unbiasedness of $\hat{\beta}_1$

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

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

$$\begin{aligned} E(X\epsilon) &\stackrel{!}{=} \\ E(X) + E(\epsilon) & \end{aligned}$$

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

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

- 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, $E[\hat{\beta}_1] = \beta_1$.

$$E(Xu) - E(X)E(u) = 0$$

Ordinary Least Squares: Unbiasedness of $\hat{\beta}_0$

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

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

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

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.

$$E(\bar{X}) - E(\hat{\beta}_1)$$

~~$E(\bar{X}) - E(\hat{\beta}_1)$~~

Ordinary Least Squares: Variances of $\hat{\beta}_0$ and $\hat{\beta}_1$

for if I can
play around with
this

- Might take bit of a work, but when you follow the notes, you get

$$\text{var}(\hat{\beta}_0) = \frac{\sigma_u^2}{n} \frac{\sum_{i=1}^n X_i^2}{\sum_{i=1}^n (X_i - \bar{X})^2}, \text{var}(\hat{\beta}_1) = \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

- At the end of the day, we can say

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

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

Here, I assume
 X_i is non-random
(classnotes: genral)
cafe

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

$n \uparrow \rightarrow \text{Var} \downarrow$, small T^2 relative
to σ_{ϵ}^2

Ordinary Least Squares: How well does the model capture the data?

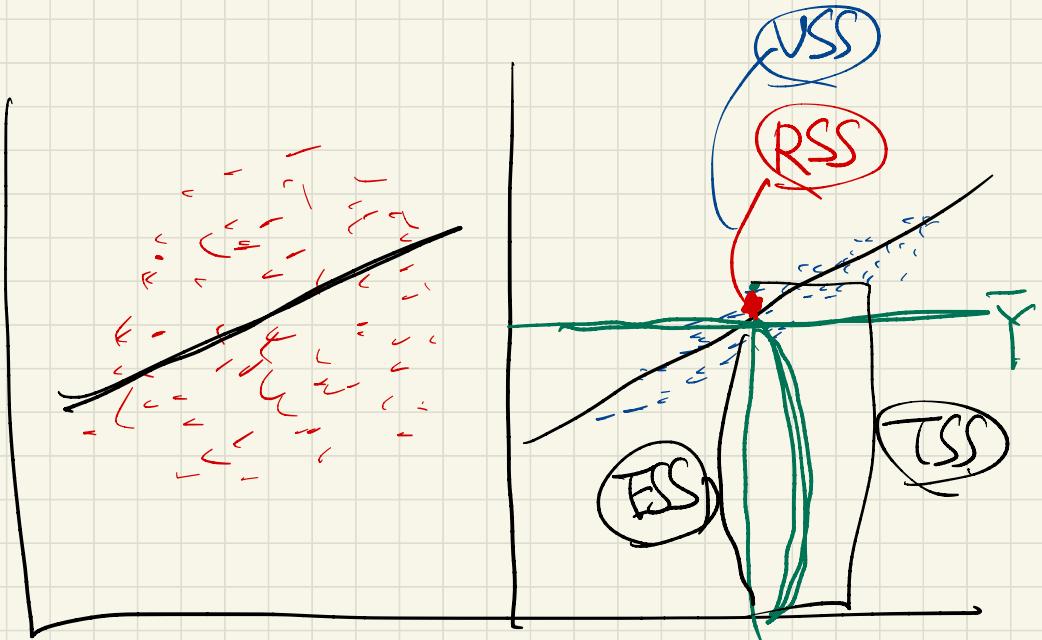
Measure of fitness

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

$$\begin{aligned} Y_i &= \underbrace{\hat{\beta}_0 + \hat{\beta}_1 X_i}_{\hat{Y}_i} + u_i, \quad \bar{Y} = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 \bar{X}}_{\bar{Y}} + \bar{u}, \\ \Rightarrow Y_i - \bar{Y} &= (\hat{Y}_i - \bar{Y}) - (u_i - \bar{u}) \end{aligned}$$
$$\Rightarrow \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (u_i - \bar{u})^2 - \boxed{2 \sum_{i=1}^n (\hat{Y}_i - \bar{Y})(u_i - \bar{u})}$$

Not nice

drop it - o



$$R^2 \downarrow$$

$$\uparrow$$

$$TSS = ESS + RSS$$

$$I = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

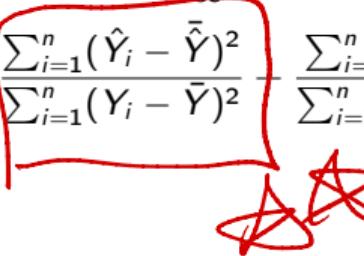
$$R^2 = I - \frac{RSS}{TSS}$$

Ordinary Least Squares: Getting to R^2

- Note that

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

- Since $\sum_{i=1}^n u_i = n\bar{u}$, $\sum_{i=1}^n \hat{Y}_i = n\bar{\hat{Y}}$ and $\sum_{i=1}^n \hat{Y}_i u_i = n\bar{u}\bar{\hat{Y}}$, all terms cancel each other out.
- So we are left with

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


Ordinary Least Squares: Getting to R^2

- Thus, the R^2 can be found as

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

- 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: Setting up the hypothesis test

$$\hat{\beta}_1 \sim N(\beta_1, \text{Var}(\hat{\beta}_1))$$

- 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

test statistic

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

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

$$H_0: \beta_1 = 0 \quad H_A: \beta_1 \neq 0$$

test stat = $\frac{3}{\sqrt{10}} = 0.95$

critical value = ± 1.96

Ordinary Least Squares: Hypothesis test methods

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

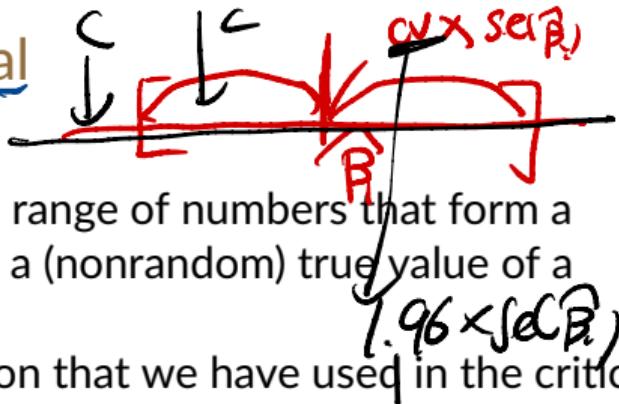
$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{\text{var}}(\hat{\beta}_1)}} \sim t_{n-2}$$

where $\widehat{\text{var}}(\hat{\beta}_1)$ 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

Ordinary Least Squares: Confidence interval

$$H_0: \beta_1 = C$$



- **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.

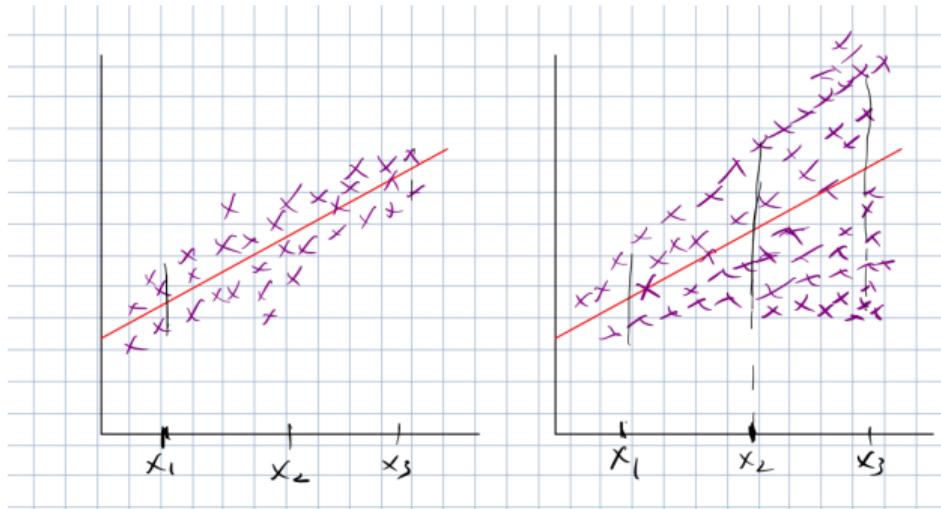
$$\Pr\left(-1.96 \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{var(\hat{\beta}_1)}} \leq 1.96\right) = 0.95$$

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

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

Errors may have different distribution across observations

- The assumption that $\text{var}(u_i)$ is constant may not hold. Thus, be open for heteroskedasticity
- 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.

...but what does heteroskedasticity change?

```
. regress testscr str
```

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

```
. 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	

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

testscr	Robust	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