Recitation 2

Seung-hun Lee

Columbia University

October 4th, 2021

Ordinary Least Squares: Population vs sample linear regression models

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

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

- 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)

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

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

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

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

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

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

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

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

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

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: There are others..

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

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} u_i^2}$$

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. Numerically, the difference between SER and RMSE is minimal and even approximate to identical figure in large sample.

Ordinary Least Squares: Main 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.

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

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

A3 $E(u_i|X_i) = 0$: 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 $cov(X_i, u_i) = 0$)

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

A5 No Autocorrelation (Serial Correlation): For $i \neq j$, $cov(u_i, u_j) = 0$. 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

A6 No Outliers: Outlier has no impact on the regression results. $(E(X_i^4), E(Y_i^4) < \infty)$

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

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

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

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

$\sum_{i=1}^n (X_i - \bar{X})(u_i - \bar{u})$ 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: Tricks for getting $var[\hat{\beta}_1]$

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

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

Ordinary Least Squares: Tricks for getting $var[\hat{\beta}_1]$

- $\rightarrow$ We assume homoskedasticity and no autocorrelation
- $\rightarrow$ Since X_i is from the data[1] 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. This is also achieved with more observations (increase in denominator)

[1] Slightly different angle from class but the key takeaway is same

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

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

  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: Getting to $var[\hat{\beta}_0]$

- $var[\hat{\beta}_0]$: 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 all that for what?

- 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: Setting up the 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 methods

- **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

Ordinary Least Squares: 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.