Recitation 2

Seung-hun Lee

Columbia University

Review: Statistical Inference

- **Statistical Inference** refers to any process of using data analysis to make guesses on some parameters of a population using a randomly sampled observation from the larger population.

- To conduct a statistical inference, one must first identify a statistically testable question (hypothesis), collect and organize the data, carry out an estimation, test the hypothesis, and come to a conclusion using confidence intervals or other methods.
  - **Estimation**: process of guessing the statistic of interest (sample mean and sample variance).
  - **Hypothesis testing**: You test the null hypothesis (H_0) is tested against an alternative hypothesis (H_1).
  - **Confidence interval**: How accurately your statistic of interest is calculated. Typically, researchers use 95% or 99% confidence interval.

Review: Statistical Inference

- A typical type of question answered through statistical inference
  - Do exposure to radiation during pregnancy affect long-run health and educational outcomes? (Almond et al. 2009, Black et al. 2019)
  - Do mask regulations reduce Coronavirus infections?
  - Do migrants affect labor market wages in their new societies? Are they different depending on the occupational sector? (Peri et al. 2020)
- The questions are typically composed of two parts
  - A preceding factor (X), whether accidental, policy-related, or individual's choice based on will
  - An outcome of interest (Y).
- Once you know your X and Y's you need to collect data on these variables. (Not an easy task in reality)

Review: Statistical Inference (example)

- Estimation: Use the data gathered, you derive the sample version of the parameter that you are interested in
  - It could be a sample mean (or differences of the means)
  - In econometrics, you will see $\hat{\beta}$, an estimate of the true β
- Hypothesis test: You set a hypothesis test to claim something about the parameter of interest
  - Null (H_0): Claims about status quo
  - Alternative (H_1): Claim you are trying to make
- Test: To see if estimation from the data supports your claim
  - Compare critical values, construct confidence interval, or use p-value

Review: Statistical Inference (example)

- Suppose you are interested in the effect of class sizes on test scores.
- Define a small classroom is (say, any class with fewer than 20 students)
- Then, calculate the sample mean and the sample variance of test scores of each type of classroom.
- You now calculate the test statistic:

$$\frac{\bar{Y}_b - \bar{Y}_s}{\sqrt{\frac{S_b^2}{n_b} + \frac{S_s^2}{n_s}}}$$

- Your hypothesis would be ”the mean test score of small classroom is different from others”. We can write

$$H_0 : E(Y_b) - E(Y_s) = 0 \text{ vs. } H_1 : E(Y_b) - E(Y_s) \neq 0$$

Review: Statistical Inference (example)

- To carry out the test, we can do as follows
  - Calculate the test statistics and directly compare with the **critical value** derived from assuming that the null hypothesis distribution is correct
    - If we have a standard normal and use 5% significance level, we compare the test statistic against the critical value of 1.96
  - You get the **confidence interval** (usually 95%) to see if this interval includes 0, the value claimed by the null hypothesis.
    - If the confidence interval includes 0, then null hypothesis cannot be rejected. Otherwise, null hypothesis is rejected.
  - Other way is to see the **p-value**, which is roughly defined as the probability of finding a more extreme result than the observed data.
    - Typically, we want to see if the p-value is less than 0.05
    - Even better if less than 0.01

Review: Statistical Inference Properties

Desirable properties

- **Unbiasedness:** $E(\bar{Y}) = \mu_y$, where μ_y is the true parameter value.
- **Efficiency:** $\bar{Y}$ is the efficient estimator if compared against any other estimator $\hat{Y}$, it is the case that $var(\bar{Y}) \leq var(\hat{Y})$
- **Consistency:** $\bar{Y}$ is consistent if $\bar{Y}$ converges to μ_y in probability.
- **Asymptotic Normality:** The estimator is asymptotically normal if it becomes normally distributed as the number of observation increases (central limit theorem)

Ordinary Least Squares

Setup

- 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

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 $E(u_i|X_i) = 0$: 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.

A4 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*

Assumptions

Assumptions (Continued)

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, where the error in the previous period carries over to the next period.

A6 Orthogonality: $cov(X_i, u_i) = 0$. Or, the error term and the independent variable is uncorrelated. If otherwise, there exists an *endogeneity*.

A7 No Outliers: Outlier has no impact on the regression results.

A8 $n > K$: There should be more observations than independent variables.

A9 Variability in X: Independent variables should take somewhat different values for each observation. Otherwise, multicollinearity problem exists.

A10 Correct Specification: The model has all the necessary independent variables. (Otherwise, omitted variable or too much variable)

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
- **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

Measure of fitness

- 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 the mean of error u_i is assumed to be 0, $\bar{u} = 0$ and $\sum_{i=1}^{n} u_i = n\bar{u}$, we only need to take care of $\sum_{i=1}^{n} \hat{Y}_i u_i$.
- Note that it is equal to $\hat{\beta}_0 \sum_{i=1}^{n} u_i + \hat{\beta}_1 \sum_{i=1}^{n} u_i X_i$ (Try replaceing $\hat{Y}_i$)
- The fact that mean of u_i is 0 and (X_i, u_i) are uncorrelated from A6 makes both term 0. So we are left with

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

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

Ordinary Least Squares

Measure of fitness

- 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

Measure of fitness

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

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

  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.