Introduction to Econometrics: Recitation 2

Seung-hun Lee

1 Review of Key Concepts from Statistics

1.1 Statistical Inference

Statistical Inference refers to any process of using data analysis to make judgements 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.

A typical type of question answered through statistical inference are as follows

- 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)

In sum, you can see that the questions are typically composed of two parts - one where there is a preceding factor (X), whether accidental, policy-related, or individual's choice based on will and the other being the 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)

Once the question is identified and the data is obtained, **estimation** of the statistic of interest should follow. Usually, this means obtaining sample mean, sample variance, or an estimate of a coefficient of interest (you will see $\hat{\beta}$ throughout this course). The next step is **hypothesis testing**, where the null hypothesis

(H_0) is tested against an alternative hypothesis (H_1). Typically, claims related to status quo is put as null hypothesis and the new discovery one is trying to sell is classified as alternative hypothesis. Lastly, one constructs a **confidence interval** either support the null hypothesis or reject it. Typically, researchers use 95% or 99% confidence interval. Confidence interval is used to gauge how accurately your test statistic is calculated. If the confidence interval includes the null hypothesis value (zero, in most cases), then your test statistic is calculated inaccurately and you fail to reject the null hypothesis.

For instance, suppose that the question of interest is the effect of class sizes on test scores. Then, one needs to define what a small classroom is. Let's suppose that any class with fewer than 20 students is a small classroom. Then, calculate the sample mean and the sample variance of test scores of each type of classroom. Let $\bar{Y}_s, S_s^2$ denote sample mean and sample variance of test scores from the small classrooms. ($\bar{Y}_b, S_b^2$ are sample mean and sample variance from large classrooms). Once those are calculated, a test statistic is required to test the following null and alternative hypothesis:

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

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

Once test statistic is obtained, it should be compared with a critical value, which differs depending on the significance level and the distribution of the null hypothesis. If the null hypothesis is standard normal and we are using a 5% significance level, we would be using a critical value 1.96. If the test statistics is larger than 1.96, we reject the null. Otherwise, we accept the null.

Another way to conduct this test is to construct a confidence interval of 95% to see if this interval includes the value of $\bar{Y}_b - \bar{Y}_s$ claimed by the null hypothesis, 0. If the confidence interval includes 0, then null hypothesis cannot be rejected. Otherwise, null hypothesis is rejected.

Yet another way to conclude whether null hypothesis should be rejected or not is to see the **p-value**, which is roughly defined as the probability of finding a more extreme result than the observed data, assuming that H_0 is true. The calculation of the p-value depends on how H_1 is defined. For instance, a p-value of 0.061 would imply that there is a 0.061 probability that the next draw would become more extreme than the current draw. If the significance level is chosen to be 5%, the null hypothesis would not be rejected. If it is selected to be 10%, the null hypothesis would be rejected.

Note that the above is a **two-sided test**. One can always have a **one-sided test** by setting the alternative hypothesis as

$$H_1 : E(Y_b) - E(Y_s) > 0 \text{ or } E(Y_b) - E(Y_s) < 0$$

Before discussing the desirable properties of the estimators, some concepts need to be clarified

- **i.i.d.**: Independent and Identical Distribution. In a random sampling, If $Y_1, .., Y_n$ are from the same population and are selected at random, they are identically distributed. Moreover, if a selection of Y_1 has no impact on the selection of Y_2 and so on, they are independently distributed.
- **Sampling Distribution**: It is a probability distribution of a test statistic derived from the random sample you are working on

To ensure that the estimators obtained from sampling has good qualities, the following standards can be used:

- **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 n increases (central limit theorem)

2 Least Squares Estimation

2.1 Derivation of Least Squares Estimator

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 what the true values of the population parameters - β_0 and β_1 in this case- are, an alternative way to approach the problem is to use the **sample linear regression model** (or just model)

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

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

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

The resulting **least squares estimators** are

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

Proof for deriving $\hat{\beta}_0, \hat{\beta}_1$

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

$$\begin{gathered}-2\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \implies \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \\ \implies \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} X_i = 0 \implies \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i \\ \implies \sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i\end{gathered}$$

Using the fact that $\sum_{i=1}^{n} Y_i = n\bar{Y} \iff \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$, I get

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

As for $[\hat{\beta}_1]$: $-2\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$ divide both sides by -2 and rearrange

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

$$\implies \hat{\beta}_1 \left(\sum_{i=1}^{n} X_i^2 - \bar{X} \sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} X_i Y_i - \bar{Y} \sum_{i=1}^{n} X_i \implies \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

Note that

$$\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}_i) = \sum_{i=1}^{n} X_i Y_i - \bar{X} \sum_{i=1}^{n} Y_i - \bar{Y} \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \bar{X}\bar{Y}$$

$$= \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y} - n\bar{X}\bar{Y} + n\bar{X}\bar{Y} = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$$

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

2.2 Main Assumptions of the OLS Estimators

For the ordinary least squares to have desired properties of unbiasednesss, consistency, efficiency, and asymptotic normality, the following assumptions must be made

Assumption 2.1. *Here are the assumptions for the classical linear regression model*

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

$$\textit{Okay: } Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

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

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*

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, there might be an omitted variable or too much variable problem.*

2.3 Measure of Fitness

These numbers tell us how informative the sample linear regression we used is in telling us about the population data. We discussed two types of measure

- $\mathbf{R^2}$: It is defined as a fraction of total variation which is explained by the model. Mathematically, this is

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

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

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

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 $\hat{\beta}_0 \sum_{i=1}^{n} u_i + \hat{\beta}_1 \sum_{i=1}^{n} u_i X_i$, 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

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

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.

- **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.