ØAMET4100 · Spring 2019
Lecture Note 2A

Instructor: Fenella Carpena

January 17, 2019

This lecture note provides a review of linear regressions with one regressor (Stock & Watson, Chapter 4). This lecture note is not intended to be a comprehensive review of lecture or the textbook, since there is a lot more material than we have time to cover. However, I have tried to focus on the concepts which I believe are necessary to be successful in our class.

1 Population Regression vs. Sample Regression, Ordinary Least Squares (OLS) Estimation

- The **population regression line** is the relationship that holds between X and Y on average over the population. We write the population linear regression model as

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

  where Y_i is the dependent variable, X_i is the independent variable, $\beta_0 + \beta_1 X$ is the population regression line, β_0 is the intercept of the population regression line, β_1 is the slope of the population regression line, and u_i is the population error term.

  - Important: β_0 and β_1 are constants; these population parameters are unknown (recall that we usually denote Greek letters as population parameters/unknowns).
  - Important: u_i contains all factors affecting Y that is not captured by X.

- When we collect data on X and Y, we can estimate the population regression function using our sample data. Specifically, with a scatterplot of sample data on X and Y, intuitively, what we are trying to do is to draw a line that "best" fits our sample data. This gives us the **sample regression line**, which we write as

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

- **Some important features to note about the sample regression line:**

  - The "hat" over the Y indicates that it is a **fitted value** (in other words, predicted value) based on the "best-fit" line. That is, for any given X, we have can apply the formula $\widehat{\beta}_0 + \widehat{\beta}_1 X$ to obtain a predicted value $\widehat{Y}$.
  - $\widehat{\beta}_0$ is the intercept and $\widehat{\beta}_1$ is the slope of the sample regression line
  - Not all data points will lie exactly on this line, since it is not usually possible to draw a line that passes through *all* the points on the scatter plot. Hence, there is a **residual**, denoted $\widehat{u}_i$, for each data point.
    - The residual is $\widehat{u}_i = Y_i - \widehat{Y}_i$. Stated in words, it is the difference between the observed value Y in the data, and the predicted value $\widehat{Y}$ from the "best-fit" line. We can think of this as a "prediction error."
    - The units of u_i are the same as Y because residuals are vertical deviations (i.e., uses Y-axis).
    - A positive residual means that the data point is above the sample regression line, while a negative residual means it is below the sample regression line.

- **How do we choose the intercept $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that provides the "best" fit to the data?** We use the **ordinary least squares (OLS)** approach, a type of regression that picks the line that minimizes the sum of squared residuals. Mathematically, OLS solves the following minimization problem:

$$\begin{aligned}\min \sum_{i=1}^{N} \widehat{u}_i^2 &= \min \sum_{i=1}^{N} (Y_i - \widehat{Y}_i)^2 \\ &= \min \sum_{i=1}^{N} (Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i))^2\end{aligned}$$

  In other words, **least squares** selects $\widehat{\beta}_0$ and $\widehat{\beta}_1$ such that the line is as close as possible to the observed data points, where "closeness" is measured by the sum of squared mistakes in predicting Y given X.

- **Important formulas:** If we solve the minimization problem above (for the case where there is only one explanatory variable X), we obtain the following equations for $\widehat{\beta}_0$, $\widehat{\beta}_1$.

$$\widehat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

  where s_{xy} is the sample covariance of X and Y, s_y is the sample standard deviation of Y, s_x is the sample standard deviation of X, $\overline{Y}$ is the sample mean of Y, and $\overline{X}$ is the sample mean of X. See Appendix 4.2 of Stock & Watson for the derivation of these formulas.

- **Important properties of the OLS regression line:**
  - The OLS regression line always passes through the point $(\overline{X}, \overline{Y})$. We can see this directly by rearranging the formula for $\widehat{\beta}_0$, i.e. $\overline{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{X}$
  - $\sum_{i=1}^{n} \widehat{u}_i = 0$, so the sample mean of the OLS residuals $\widehat{u}_i$ is zero
  - $\sum_{i=1}^{n} \widehat{u}_i X_i = 0$, so the sample covariance between the OLS residuals and the regressors is zero

2 Interpreting OLS Regression Coefficients

- **How do we interpret $\widehat{\beta}_0$ and $\widehat{\beta}_1$?** Recall that the sample regression line is written as $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ where $\widehat{\beta}_0$ is the intercept, and $\widehat{\beta}_1$ is the slope.

- $\widehat{\beta}_0$ is the value of regression line when $X = 0$; in other words, it is the point at which the sample regression line intersects the Y-axis.
  - $\widehat{\beta}_0$ has the same units of measurement as Y.
  - In some applications, this intercept has a meaningful economic interpretation. For example, consider the case where $\widehat{Price} = 15 + 2697 \cdot Weight$, where the variable $Price$ is the price of a diamond, and $Weight$ is the diamond's weight. The intercept \$15 is the predicted price of the diamond if the weight is zero. Here, we can interpret it as fixed costs, i.e. the portion of costs that is present regardless of weight such, as the cost of a jewelry box.
  - In other applications, the intercept has no real-world meaning. For instance, in the medical field, you may see a regression $\widehat{BloodLoss} = 552.44 - 130 \cdot Height$, where the variable $BloodLoss$ is the amount of blood in milliliters that a person lost during an operation, and $Height$ is the person's height in meters. Strictly speaking, the intercept 552.44 is the predicted blood loss for a person with a height of 0 meters. This meaning is non-sensical, so it is best to just think of the intercept mathematically in this case, as the coefficient that determines the level of the regression line.

- $\widehat{\beta}_1$ is the change in Y *associated with* a one unit change in X.

- The units of $\widehat{\beta}_1$ are those of Y divided by those of X.
- It is incorrect to describe $\widehat{\beta}_1$ as the "change in Y *caused by* the change in X" because of confounding variables. Hence, we say "associated with" instead of "caused by."

- We can only apply the interpretations of the slope and intercept over the range of X values in our data. This means that we should be careful when making **extrapolations** or **out-of-sample predictions**—that is, making predictions or statements beyond the scope of what is observed in the data—since they are less reliable.

3 Measures of Fit

- Having estimated a linear regression, you might wonder how well that regression line describes the data. Does the explanatory variable account for much or for little of the variation in the response variable? Are the observations tightly clustered around the regression line, or are they spread out? Two measures that can help answer these questions are the R^2 and **standard error of the regression**.

- The R^2 of a regression is the fraction of the sample variance of Y that is explained by X. For instance if $R^2 = 0.25$, then we say "25% of the variation in Y is explained by X"
  - $R^2 = \frac{ESS}{TSS}$ where ESS (explained sum of squares) $= \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$ and TSS (total sum of squares) $= \sum_{i=1}^{n}(Y_i - \overline{Y})^2$
  - In the case of a regression where there is only one independent variable X, R^2 is the square of the sample correlation between X and Y, i.e. $R^2 = r_{XY}^2$.
  - R^2 ranges between 0 and 1. An R^2 close to 1 indicates that X is good at predicting Y (that is, the data points Y are closer to the line), while an R^2 near 0 indicates that X is not very good at predicting Y.
  - R^2 is unit-free, just like correlation.

- The **standard error of the regression (SER)**, denoted $s_{\widehat{u}}$, is an estimator of the standard deivation of the population regression error u_i.
  - It measures the spread of the data points around the regression line.
  - The formula for the SER is given by

$$s_{\widehat{u}} = \sqrt{\frac{\widehat{u}_1^2 + \widehat{u}_2^2 + \ldots + \widehat{u}_n^2}{n-2}}$$

    where the denominator is $n-2$ due to a degrees of freedom correction We subtract 2 because we have used the data to estimate two parameters, namely $\widehat{\beta}_0$ and $\widehat{\beta}_1$.
  - The units if $s_{\widehat{u}}$ are the same as those of Y.

- A high R^2 means that SER is low (and vice versa). Intuitively, this is because if R^2 is high, the data points are closer to the regression line, meaning the residuals $\widehat{u}_i$ are small. As you can see in the formula above for SER, if the residuals $\widehat{u}_i$ are low, then SER will be low as well.

- A high R^2 does not mean that the regression indicates a causal relationship between X and Y. Consider the following example: the regressions $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$ and $\widehat{X} = \widehat{\beta}_0 + \widehat{\beta}_1 Y$ (that is, switching the roles of X and Y) would yield the same R^2. This is because $R^2 = r_{XY}^2$ (the sample correlation of X and Y) in both cases. Hence, even if R^2 is high, we cannot know whether X causes Y or Y causes X.

- Note that you cannot compare R^2 between two regressions that do not have the same response variable and/or do not use the same dataset.

4 Least Squares Assumptions

These assumptions are very critical for our course so it is important that we understand what they mean. These are the assumptions under which OLS provides an appropriate estimator for β_0 and β_1. Understanding these assumptions is essential for understanding when OLS will and will not give useful estimates of the regression coefficients.

- **Assumption # 1: The conditional mean of u_i given X_i is zero, $E(u_i|X_i) = 0$.**
  - This assumption means that all "other factors" contained in u_i are unrelated to X_i in the sense that, given a value of X_i, the mean of the distribution of these other factors is zero.
  - If X_i is randomly assigned, then u_i and X_i are independent $\implies E(u_i|X_i) = 0$
  - Note that $E(u_i|X_i) = 0 \implies cov(u_i, X_i) = 0$ (try to show this on your own). Hence, if X_i and u_i are correlated, this assumption is violated.
  - This assumption is necessary for the OLS estimates $\widehat{\beta}_1$ and $\widehat{\beta}_0$ to be unbiased.
  - In practice, this is usually the most important assumption to consider.
- **Assumption # 2: (X_i, Y_i) are independent and identically distributed (i.i.d.)**
  - Independence here means that any pair of X and Y, any pair of X's, and any pair of Y's are all independent of each other. This implies that $cov(X_i, Y_j) = 0$, $cov(X_i, X_j) = 0$, $cov(Y_i, Y_j) = 0$ for all i and j (in words, any pair of X and Y, any pair of X's, and any pair of Y's have zero covariance).
  - Identically distributed here means that any pair of X and Y have the same joint distribution, all X's have the same distribution, and all Y's have the same distribution. This implies that $cov(X_i, Y_j) = \sigma_{XY}$, $var(X_i) = \sigma^2_X$, and $var(Y_i) = \sigma^2_Y$ for all i and j (in words, this says that any pair of X and Y have the same covariance, all X's have the same variance, and all Y's have the same variance).
  - This assumption holds when we take a simple random sample of the population.
  - Conversely, if the sample is not drawn randomly from the population and is set by the researcher, then this assumption could fail. For example, if only rural schools are included in the sample while the population of interest is all schools, then this assumption is violated.
  - This assumption could also fail when we have time-series data, e.g., if X is US dollar exchange rate, and Y is US exports over time, we might think that the exchange last month will affect exports this month, in which case $cov(X_1, Y_2) \neq 0$. This violates independence between any pair of X_i and Y_j.
- **Assumption # 3: Large outliers are unlikely.**
  - Mathematically, this assumption is expressed by saying that X_i and Y_i have non-zero finite fourth moments (i.e., non-zero finite kurtosis, $0 < E(X_i^4) < \infty$, $0 < E(Y_i^4) < \infty$.
  - This generally holds if the X and Y have a finite range (for example, variables such as age, education).
  - However, if there are data entry errors or misreporting in the data, then large outliers are possible and this assumption is violated.
- **Why do we care about these assumptions?** There are at least two reasons.
  - First, if these assumptions hold and the sample size is sufficiently large, we can apply the central limit theorem, which would tell us that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ have sampling distributions that are normal. In turn, this large-sample normal distribution lets us develop methods for hypothesis testing (i.e., we can use the normal distribution when we conduct hypothesis tests of the population coefficients β_0 and β_1.
  - Second, these assumptions helps us understand the circumstances that pose difficulties for OLS estimation.