ØAMET4100 · Spring 2019
Lecture Note 3A

Instructor: Fenella Carpena

January 24, 2019

This lecture note provides a review of linear regressions with multiple regressors (Stock & Watson, Chapter 6). 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 Omitted Variable Bias (OVB)

- So far, we have talked about a regression where we have a single right hand side variable. Specifically, we considered the regression

$$Y_i = \beta_0 + \beta_1 X_i + u_i. \tag{1}$$

  Recall that u_i in this equation contains all factors that affect Y_i but are not contained in X_i.

- If we want to understand the true relationship between Y_i and X_i, our goal is to get an unbiased estimate of β_1. In many cases, however, if we estimate the above simple regression using least squares, we will get a biased estimate since there is correlation between X_i and u_i.

- Note that if $corr(X_i, u_i) \neq 0$, then $E(u_i|X_i) \neq 0$. Hence, the first least squares assumption is violated. Furthermore, recall from last week's lecture that we needed the first least squares assumption to obtain unbiasedness of $\widehat{\beta}_1$. So in short, if there is correlation between u_i and X_i, we would get an unbiased estimate of $\widehat{\beta}_1$.

- In practice, one way in which $corr(X_i, u_i) \neq 0$ is due to **omitted variable bias (OVB)**. OVB happens when we omit a variable that should be included in the regression. For this bias to occur, we need two conditions:

  1. The omitted variable (call it X_2) is correlated with the right-hand side variable (call it X_1)
  2. The omitted variable determines the outcome Y

- Intuitively, if we omit X_2 from the regression but it should have been included there (because it influences the variable Y), then X_2 gets "absorbed" in the error u. Since u now "contains" X_2, and by condition # 1 above X_1 and X_2 are correlated, this means that u and X_1 are also correlated. Hence, the first least squares assumption is violated and we get a biased estimate of β_1.

- **Example:** Suppose that the true population regression mode for returns to schooling is given by the regression:

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 ability_i + v_i \tag{2}$$

  Suppose we are also given that $E(v_i|educ, ability) = 0$. To get an unbiased estimate of β_1, we would collect data on wages, education, and ability for a simple random sample of individuals. However, because ability is in principle an unobservable characteristic, we are unable to measure it, and we don't have any data on ability. As a result, we instead estimate the regression:

$$wage_i = \beta_0 + \beta_1 educ_i + u_i \tag{3}$$

  where $u_i = \beta_2 ability + v_i$. If we estimated Equation (3) using our sample, we would obtain the sample regression

$$\widehat{wage_i} = \widehat{\beta}_0 + \widehat{\beta}_1 educ \tag{4}$$

Here, $\widehat{\beta_1}$ is a biased estimator for β_1, since the error term in equation (3) is correlated with *educ*. Specifically, $cov(educ, u_i) = cov(educ, \beta_2 ability + v_i) = \beta_2 cov(educ, ability)$, where $cov(educ, ability)$ is not equal to zero because it is most likely the case that people who have higher ability also tend to get more schooling (e.g., they get master's degrees).

- Once we have established whether our estimates are biased, we would like to know how much bias there is. Is there only a little bit of bias or a lot of bias? To help quantify the bias, it is useful to consider the formula for the omitted variable bias.
- Going back to Equation (1), let the correlation between X_i and u_i be ρ_{Xu}. Suppose that the second and third least squares assumptions hold, but the first does not because ρ_{Xu} is non-zero. Then, the OLS estimator has the limit

$$\widehat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu}\frac{\sigma_u}{\sigma_X}$$

The above means that as the sample size increases, $\widehat{\beta}_1$ is close to $\beta_1 + \rho_{Xu} \cdot \frac{\sigma_u}{\sigma_X}$ with increasingly high probability.

- The term $\rho_{Xu}\frac{\sigma_u}{\sigma_X}$ represents how much OVB there is. If the first least squares assumption holds, this term would go away because the first least squares assumption implies that $\rho_{Xu} = 0$. Some important things to note about the bias:
  - The OVB doesn't go away even if we increase the sample size.
  - The size of the bias depends on how large $\rho_{Xu}\frac{\sigma_u}{\sigma_X}$ is.
  - The direction of the bias depends on the sign of ρ_{Xu}.
- The direction of the bias is a very important concept, so it's important to understand it.
  - If $\rho_{Xu} < 0$, then $E(\widehat{\beta}_1) < \beta_1$. Hence we say that $\widehat{\beta}_1$ is **downward biased**. This means that if we took many different random samples and calculated many different $\widehat{\beta}_1$'s using those different samples, on average, we would be getting an estimate that is lower than the truth.
  - If $\rho_{Xu} > 0$, then $E(\widehat{\beta}_1) > \beta_1$. Hence we say that $\widehat{\beta}_1$ is **upward biased**. This means that if we took many different random samples and calculated many different $\widehat{\beta}_1$'s using those different samples, on average, we would be getting an estimate that is higher than the truth.
- How do we know if our estimate estimate is upward or downward biased? This is the same question as: how do we know if $\rho_{Xu} > 0$ or $\rho_{Xu} < 0$?
  - The sign of ρ_{Xu} is determined by the product of the signs of β_2 (i.e., whether the omitted variable has a positive or negative effect on Y) and the sign of $corr(X_1, X_2)$ (i.e., the sign of the correlation between the omitted variable and the variable included in the regression)
  - Going back to our example above in Equation (2) to (4), we found that $cov(educ, u_i) = \beta_2 cov(educ, ability)$. Hence, in this case, the sign of the correlation between *educ* and u is determined by the sign of β_2 (the effect of ability on wages) multiplied by the sign of the correlation between *educ* and *ability*.

2 Multiple Regression Model

- One approach to dealing with OVB is to add more X variables to the regression (assuming that data on the X's are available). This gives us the **multiple regression model (MRM)**.
- As before, when we have a regression model, we have a population regression line. Since we have multiple X's now, the population regression is given by

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$$

- Note that in the above equation, we have k right hand side variables.
- How do we interpret β_1 in this population regression equation? It is the effect of a one unit change in X_{1i} on Y, **holding X_2 to X_k constant**. Why is this the case? Suppose we had $k = 3$, i.e., we have $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$. Now considering changing X_{1i} to $X_{1i} + 1$, but holding X_{2i} and X_{3i} fixed.

- Before the change: We have $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$.
- After the change: We have $Y_i = \beta_0 + \beta_1 (X_{1i} + 1) + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$.
- The difference between the two is β_1.

- How do we interpret β_0 in the population regression equation? It is the expected value of Y_i when X_{1i} to X_{ki} are all zero.
- As before, we can estimate the population regression using our sample and obtain OLS estimates for the β's. The sample regression line is:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \ldots + \widehat{\beta}_k X_{ki}$$

where $\widehat{u}_i = Y_i - \widehat{Y}_i$ as before. Again, like in the simple linear regression case, in OLS $\widehat{\beta}_0$, $\widehat{\beta}_1$, ..., $\widehat{\beta}_k$ are chosen to minimize the sum of squared residuals.

3 Adjusted R-squared

- There are three measures of fit that can be applied in multiple regression. These are R-squared, SER, and Adjusted R-squared.
- In multiple regression, R-squared and SER are the same as in simple regression.
- The Adjusted R-squared is denoted as $\overline{R}^2$, and it is equal to

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1} \cdot \frac{SSR}{TSS}$$

where SSR and TSS are (as before) the sum of squared residuals and the total sum of squares.

- One of the "problems" with R-squared as a measure of fit in MRM is that it weakly increases as more regressors are added. So a model with more X variables will appear to have a better fit simply because it has more right-hand side variables, even if those variables offer no explanatory power for Y.
- $\overline{R}^2$ is different from R^2 because $\overline{R}^2$ takes in to account the number of explanatory variables in the regression. This is useful because $\overline{R}^2$ gives us a better sense of the goodness of fit. Whereas R^2 weakly increases when adding a regressor, $\overline{R}^2$ increases if and only if the added regressor has some explanatory power for Y.
- Some useful properties of $\overline{R}^2$ and R^2:
  - $R^2 > \overline{R}^2$ since $\frac{n-1}{n-k-1} > 1$.
  - Adding a new regressor has ambiguous effects on $\overline{R}^2$, since when we add a new regressor, SSR falls but $\frac{n-1}{n-k-1}$ increases.
  - Since $\frac{n-1}{n-k-1} > 1$, $\overline{R}^2$ can be negative if SSR/TSS is sufficiently large.
  - For large n and small k, $\frac{n-1}{n-k-1}$ is about equal to 1, so R^2 and $\overline{R}^2$ will be similar.

4 Perfect and Imperfect Multicollinearity

- In the case of a simple linear regression, recall that we had three least squares assumptions. Now that we have multiple linear regression, the same three assumptions are still necessary, but we need to add a fourth one. That fourth assumption pertains to perfect multicollinearity.
- To summarize, the four least squares assumptions in a multiple regression model are:
  1. $E(u_i|X_{1i}, \ldots, X_{ki}] = 0$. As in simple linear regression, this assumption means that all factors in u_i are "unrelated" to the X's because for any given value of the X's, the mean of u_i is zero.

2. $X_{1i}, \ldots, X_{ki}$ are independent and identically distributed (i.i.d.). This is a statement about how the sample is drawn. Suppose Y is *wage* and X is *educ*. Independence means that the wage and education of the first person in the sample is independent of the wage and education of the second person in the sample, etc. If the persons in our sample are drawn from the same population, then they will be identically distributed.
3. Large outliers are unlikely. Mathematically, this means finite fourth moments.
4. No perfect multicollinearity.

- We say that X_1 and X_2 are perfectly multicollinear if X_1 can be written as a linear function o X_2. That is, we can write $X_1 = a + bX_2$ where a and b are some constants.
- When X_1 and X_2 are perfectly multicollinear, we cannot include both variables in the regression because we will not be able to estimate both the coefficients on X_1 and X_2 at the same time. To understand why, think about the interpretation of β_1: it is the effect on Y of a one-unit increase in X_1, **holding X_2 constant**. If X_1 and X_2 are perfectly multicollinear, then if we hold X_2 constant, then we are also holding X_1 constant. Thus, we cannot have a "one-unit" increase in X_1.
- Two examples of regressions where there is perfect multicollinearity:
  - $wage_i = \beta_0 + \beta_1 male_i + \beta_2 female_i + u_i$ where *male* is a dummy variable for a man and *female* is a dummy variable for a woman. Here, $male = 1 - female$ so *male* and *female* are perfectly multicollinear. This is also an example of the **dummy variable trap**.
  - If we estimate $wage_i = \beta_0 + \beta_1 female_i + u_i$ using a sample consisting only of females, there will also be perfect multicollinearity here. Can you see why?
- Why do we care about perfect multicollinearity? We care because if the four assumptions above hold and we have a large sample size, then the OLS estimators $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$ are unbiased, consistent, and jointly normally distributed. We can then take advantage of this normal distribution when we conduct hypothesis tests.
- If X_1 and X_2 are not perfectly correlated, but have correlation that is close to one, then we say that X_1 and X_2 are **imperfectly multicollinear**. Note that correlation is a measure of linear dependence of two variables.
- Unlike perfect multicollinearity, if we have imperfect multicollinearity, we can still estimate the $\widehat{\beta}$'s. However, our estimates of at least one of the coefficients will be imprecise. Why? Consider a case where we have to regressors X_1 and X_2 that are imperfectly multicollinear. Again, the coefficient $\widehat{\beta}_1$ is the effect of a one-unit change in X_1 holding X_2 constant. If we hold X_2 constant, it is as if we are also holding X_1 constant because both are imperfectly collinear. As a result, there will be low variation in X_1, and this low variance leads to less precise estimates of $\widehat{\beta}_1$.