ØAMET4100 · Spring 2019
Lecture Note 5

Instructor: Fenella Carpena

February 7, 2019

This lecture note provides a review of assessing studies based on multiple regression (Stock & Watson, Chapter 9). 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 Introduction

So far in this course, we have used multiple regression to analyze the relationship between variables. We have seen multiple regression applied in the following ways.

- Estimating causal effects: For example, will one additional year of education lead to higher wages?
- Forecasting: For example, what would be the selling price of an apartment that is 50 square meters and has one bedroom?

Now, our goal is to step back and examine when multiple regression provides reliable estimates for these questions. To do so, we consider the concepts of **internal** and **external validity**.

A statistical analysis is said to have **internal validity** if the statistical inferences about causal effects are valid for the population being studied. The analysis is said to have **external validity** if its inferences and conclusions can be generalized from the population and setting studied to other populations and settings.

2 Internal Validity

Suppose we are interested in the causal effect of X on$Y$, and we estimate $Y_i = \beta_0 + \beta_1 X_{1i} + u_i$. Internal validity requires that the estimate $\widehat{\beta}_1$, which measures the causal effect of interest, satisfies two requirements:

- First, $\widehat{\beta}$ is unbiased and consistent. Recall that unbiasedness means $E(\widehat{\beta}_1) = \beta$, and consistent means $\text{plim}_{n\to\infty} \widehat{\beta}_1 = \beta_1$.
- Second, the hypothesis test should have the desired significance level, and confidence intervals should have the desired confidence level. For example, the calculated confidence interval at the 95% confidence level, $\widehat{\beta} \pm 1.96SE(\widehat{\beta})$, contains the true causal effect β_1 with 95% probability.

We will study the following 7 threats to internal validity: (1) omitted variable bias, (2) functional form misspecification, (3) measurement error, (4) sample selection, (5) simultaneous causality, (6) heteroskedasticity, and (7) correlated error terms. Of these, (1) to (5) violate the first OLS assumption, $E(u_i|X_i) = 0$, while (6) and (7) violate the second OLS assumption, that (X_i, Y_i) are i.i.d.

2.1 Omitted Variable Bias

Omitted variable bias arises when a variable that affects the outcome variable Y_i and that is correlated with X_i is omitted from the regression. Suppose the true population regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + w_i, \qquad E(w_i|X_{1i}, X_{2i}) = 0.$$

However, we omit X_2 from the regression, and instead, the regression we estimate is

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i.$$

In the above regression, note that X_2 gets absorbed in the error term, so that $u_i = \beta_2 X_{2i} + w_i$. Applying the OVB formula, we know that

$$\operatorname*{plim}_{n\to\infty} \widehat{\beta}_1 = \beta_1 + \beta_2 \frac{cov(X_1, X_2)}{var(X_1)}.$$

What does the above mean? It says that omitting X_2 leads to an inconsistent estimate of β_1. This bias persists even in large samples. Further, from the term $\beta_2 \frac{cov(X_1,X_2)}{var(X_1)}$, we see the two conditions that are necessary for OVB to occur.

1. The omitted variable X_2 determines Y, i.e., $\beta_2 \neq 0$.
2. The omitted variable X_2 is correlated with X_1, i.e., $cov(X_1, X_2) \neq 0$.

For example, consider the following example of a regression $wage_i = \beta_0 + \beta_1 educ_i + u_i$. A potential omitted variable bias here is "ability." With this omitted variable,

$$\operatorname*{plim}_{n\to\infty} \widehat{\beta}_1 = \beta_1 + \beta_2 \frac{cov(educ, ability)}{var(educ)}.$$

Note that β_2 is likely to be positive, especially if education is valued in the job market and employers reward workers who have higher education. Furthermore, $cov(educ, ability)$ is likely positive, since people with higher ability are good at school and so are likely to get more education. As a result, our estimate of $\widehat{\beta}_1$ is biased upward, i.e., we are getting an overestimate of the true causal effect of education on wages.

Because of omitted variable bias, the OLS estimates we obtain are not internally valid. In particular, we get biased and inconsistent estimates for the true causal effect of X_1. Potential solutions to this problem will vary depending the circumstances, but here are some possibilites.

- We can include the omitted variable as a control variable in the regression, assuming that data is available for it. Adding too many regressors, however, has a cost mainly in terms of reduced precision of the estimates. The control variables we include in our regression be guided by economic reasoning and knowledge of the study setting.
- If there is no data available on the omitted variable because it is an unobserved, we can try to use a "proxy" variable. For example, the "ability" of a person is a characteristic that is not observable, but we can use GPA or grades in school as proxy for ability.
- We can also make use of panel data, instrumental variables, or randomized experiment. We will discuss these methods in future lecture.

2.2 Misspecification of Functional Form

If the regression model has the wrong form (e.g., variables either on the left or on the right hand side should be in logarithms, polynomial effects are omitted) we say that the **functional form** of the regression model is **misspecified**.

Suppose the true population regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + w_i, \qquad E(w_i|X_{1i}) = 0.$$

However, we omit X_{1i}^2 from the regression, and instead, the regression we estimate is

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i.$$

In the above regression, note that X_{1i}^2 gets absorbed in the error term, so that $u_i = \beta_2 X_{1i}^2 + w_i$. As before, we can apply the OVB formula to find that

$$\operatorname*{plim}_{n\to\infty} \widehat{\beta}_1 = \beta_1 + \beta_2 \frac{cov(X_1, X_1^2)}{var(X_1)}.$$

What did we just show? Note that $cov(X_1, X_1^2)$ is never zero. Hence, if $\beta_2 \neq 0$ (i.e., the X_{1i}^2 should be in the regression), we again have inconsistent estimates, and the bias does not go away even in large samples.

The above shows that functional form misspecification can be though of as a type of omitted variable bias, in which the omitted variables are the terms that capture the non-linearities.

As with the case of OVB discussed previously, if there is misspecification of the functional form, the OLS estimates we obtain are not internally valid. In particular, we get biased and inconsistent estimates for the true causal effect of X_1.

How do we solve this problem? There are no hard and fast rules, but potential approaches we can take include the following.

- We should always use economic intuition to choose the appropriate functional form. For example, if economic reasoning suggests that there are decreasing returns to X_1, then we should add X_1^2 to the regression or use $ln(X_1)$.
- It is important to look at scatter plots to visualize the relationship between X and Y to see if it is non-linear.
- We can also check for possible non-linearities by adding polynomial terms to the regression and testing their significance.

2.3 Measurement Error and Errors-in-Variables Bias

In practice, measurement error may occur because survey respondents give wrong answers or data are entered incorrectly. There are two types of measurement error to be considered.

1. Measurement error in the independent variable X; this usually violates internal validity
2. Measurement error in the dependent variable Y; this usually does not violate internal validity, but it does lead to less precise estimates.

2.3.1 Measurement error in X, assuming classical measurement errors

Suppose that the true regression model is

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad E(u_i|X_i) = 0.$$

However, we don't observe X_i, rather, we observe $\widetilde{X}_i$ which is a "noisy" measure of X_i, where

$$\widetilde{X}_i = X_i + w_i.$$

Hence, the regression we actually estimate is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \widetilde{X}_i + v_i \\ &= \beta_0 + \beta_1 \widetilde{X}_i + \left[u_i - \beta_1 \left(\widetilde{X}_i - X_i\right)\right] \\ &= \beta_0 + \beta_1 \widetilde{X}_i + [u_i - \beta_1 w_i] \end{aligned}$$

We can see that there is a form of omitted variable bias here, because the measurement error w_i is not observed and is by definition correlated with $\widetilde{X}_i$. Under the assumption of **classical measurement error**, w_i is purely random. Hence, $E(w_i) = 0$, $var(w_i) = \sigma_w^2$, $cov(w_i, X_i) = 0$, and $cov(w_i, u_i) = 0$.

As before, using the OVB formula, we can show that

$$\operatorname*{plim}_{n\to\infty} \widehat{\beta}_1 = \beta_1 \cdot \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}.$$

What did we just learn? Notice that the fraction $\frac{\sigma_X^2}{\sigma_X^2+\sigma_w^2}$ is smaller than one. This means that when there is measurement error in X and we assume classical measurement errors, the OLS estimate we

obtain is biased toward zero (e.g., if $\frac{\sigma_X^2}{\sigma_X^2+\sigma_w^2} = 1/2$, we are getting half of the true value β_1). This bias is also known as **attenuation** bias. The bias is larger the ratio $\frac{\sigma_X^2}{\sigma_X^2+\sigma_w^2}$ is small. This ratio is called the **signal-to-noise** ratio. The intuition is that if the noise w_i is large relative to the signal X_i, we will be less able to capture the actual dependence of Y_i on X_i.

What are the potential solutions to this bias? We can always try to do our best to get a more accurate measure of X. In addition, one can develop a mathematical model of the measurement error and use this to adjust the estimates. For example, if you believe there is classical measurement error and you can estimate the signal-to-noise ratio, then you can apply this to correct for the bias. Finally, we can employ instrumental variables, which is a technique we will discuss in future lectures.

2.3.2 Measurement error in Y, assuming classical measurement errors

Suppose that the true regression model is

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad E(u_i|X_i) = 0.$$

However, Y is measured error, and we assume classical measurement errors

$$\widetilde{Y}_i = Y_i + w_i$$

where $E(w_i) = 0$, $var(w_i) = \sigma_w^2$, $cov(w_i, X_i) = 0$, and $cov(w_i, u_i) = 0$. Since we don't observe the true Y_i and we only observe it with error, $\widetilde{Y}_i$, the regression we would actually estimate is

$$\widetilde{Y}_i = \beta_0 + \beta_1 X_1 + v_i.$$

Note that in the above equation, $v_i = w_i + u_i$. What happens to our OLS estimates in this case? Our estimates will still be unbiased and consistent because $E(v_i|X_i) = E(w_i|X_i) + E(u_i|X_i) = 0$, that is, the first OLS assumption holds. But, our OLS estimates are less precise (i.e., the standard errors are higher) because $var(v_i) > var(u_i)$. In other words, $SE(\widehat{\beta}_1)$ is larger when we have measurement error in Y.

2.4 Missing Data and Sample Selection

Whether missing data poses a threat to interval validity depends on why the data are missing. There are three types of missing data that we might encounter.

1. Data are **missing at random**. For example, you survey 100 people but you lost a random subset of 20 questions. Our OLS estimates will still be unbiased and consistent, and internal validity is not affected. However, the sample size is smaller, so our estimates are less precise.

2. Data are **missing based on X**. For example, you estimate a regression of education (X) on wages (Y), but you only have data on those who attended university or beyond. This does not necessarily impose a threat to internal validity. However, it can be a threat to external validity because we may not be able to say anything about the effects of education among those who completed on high school. Furthermore, our OLS estimates may be less precise if there is not enough variation in the independent variables X because it is restricted to certain values.

3. Data are **missing based on Y**. For example, to estimate the effect of education on wages, you conduct a survey of workers, but those workers with higher salary do not want to respond to the survey. This leads to **sample selection bias**, and may introduce correlation between the error term and the regressors (thus violating the first OLS assumption).

2.5 Simultaneous Causality

Simultaneous causality occurs when there is a two-way relationship between X and Y. We usually model causal effects of one variable X_i onto another variable Y_i. What if Y_i also in turn affects X_i? This simulatenous relationship would be represented by a system of equations like the following.

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{1}$$
$$X_i = \gamma_0 + \gamma_1 Y_i + v_i \tag{2}$$

By estimating equation(**??**) separately (or, for that matter, equation(**??**) separately) we are introducing, again, a bias: the error term u_i (or v_i) becomes "mechanically" correlated with the right-hand side variable, because of feedback effects.

Using the OVB formula, we can show that given the above two system of equations and assuming $cov(u_i, v_i) = 0$,

$$\operatorname*{plim}_{n\to\infty} \widehat{\beta}_1 = \beta_1 + \frac{\gamma_i var(u_i)}{(1-\gamma_1\beta_1)var(X_i)}.$$

Since the term $\frac{\gamma_i var(u_i)}{(1-\gamma_1\beta_1)var(X_i)}$ is not equal to zero, we see that simultaneous causality leads to inconsistent estimates, and the bias persists even in large samples.

A classical example of simultaneity is supply and demand. In any market, prices and quantities depend on each other both on the demand and on the supply side. The observations of price and quantities are jointly determined by the market equilibrium. Hence, we are not able to recover the parameters of the supply and demand functions by simpy looking at the data on quantities and prices alone.

What are the potential solutions to simultaneous causality? We can use instrumental variables. Another possibility is to use randomize controlled experiments.

2.6 Heteroskedasticity and Correlated Errors

The threats to internal validity that we have discussed so far involve violations of the first OLS assumption $E(u_i|X_i) = 0$, which subsequently lead to biased and inconsistent OLS estimates.

Now, we consider violations of the second OLS assumption that (X_i, Y_i) are i.i.d. Violations of this assumption do not lead to biased and inconsistent OLS estimates. However, it leads to inconsistent standard errors. As a result, hypothesis tests do not have the desired significance level, and confidence intervals do not have the desired confidence level.

2.6.1 Heteroskedasticity

The threat to internal validity occurs when the true population regression error is homoskedastic, but the standard errors are calculated under the assumption of homoskedasticity. The OLS estimates of the coefficients β will not be affected, however, the standard errors will be inconsistent. The solution to this problem is to use heteroskedasticity-robust standard errors (i.e., the `robust` option in Stata).

2.6.2 Correlated Errors

In some settings, the population regression error can be correlated across observations. For example, this may happen if the data are repeated observations of the same entity over time (i.e., time series data), such as the price of the same stock on different days.

Another situation in which the error terms are correlated is when sampling is based on a geographical unit. If there are omitted variables that reflect geographical influences, these omitted variables could result in correlation of the regression errors for observations that are geographically near to each other.

As with heteroskedasticity, correlation of error terms does not make the OLS estimators biased and inconsisent. However, it violates the standard errors are inconsistent. Generally, the solution to this problem is to adjust standard errors to account for the correlation (i.e., cluster-robust standard errors), which we will discuss more in future lectures.

3 External Validity

External validity is a question of whether statistical inferences can be generalized from the population and setting we are studying to another population and setting. There are two factors that may undermine a study's external validity.

- **Differences in populations**. The results from one study cannot be generalized to another population that has structurally different characteristics. For example, a laboratory study on the effects of chemicals on health that is conducted using animals (e.g., mice) may not generalize to humans. Additionally, a study on effects of education on wages using a sample of men may not apply to the case of women.
- **Differences in settings**. The results from one study cannot be generalized, *even for the same population*, to other settings where entities are subjected to different regulations and incentives. For example, studies of the effect of class size on test scores in California cannot be extended to countries where the school system works differently (e.g. in Europe).

How can we assess the external validity of a study? Sometimes, there are two or more studies on different but related populations. If so, the external validity of both studies can be checked by the comparing their results. In general, similar findings in two or more studies bolster claims to external validity.

4 Forecasting

When regression models are used for forecasting, concerns about external validity are very important but concerns about internal validity (i.e., about unbiased estimation of causal effects) are not. To see this, consider the following regression which estimates the effect of student-teacher ratio (STR) on test scores across different schools.

$$\widehat{testscore} = 698.9 - 2.28STR.$$

Using the above regression, we can ask two types of questions.

First, a representative from the Ministry of Education might ask: if we reduce the student-teacher ratio, will it increase test scores? This question is about the causal effect. In this case, the above regression is not necessarily useful because it suffers from omitted variable bias, and hence, it gives us a biased estimate of the causal effect.

Second, consider a parent who is moving to a new town and would like to pick a school that performs well on standardized tests. The parent doesn't care about the causal effect, but he or she wants a regression that is a reliable predictor of test performance. The parent wants a regression that explains a lot of the variation in test scores (e.g., has a good fit) and that can be applied to the town where he or she is moving to.

How do we assess the validity of regressions that are used for forecasting? To obtain reliable forecasts, the regression must have good explanatory power, its coefficients must be estimated precisely, and it must be stable in the sense that the regression estimated on one set of data can be reliably used to make forecasts using other data. When a regression model is used for forecasting, the main concern is whether the model is externally valid, so that it is applicable to the circumstance in which the forecast is made.