Chapter 7 outline, Econometrics
Instrumental variables and model estimation

1 Correlation between an independent variable and the error

Recall that one of the assumptions that we make when proving the Gauss-Markov theorem is that the independent variables are not correlated with the error term. This is an implicit assumption we make when we assume that the independent variables are nonstochastic. However, it is possible that the error term and independent variable are correlated. When that occurs our estimator for $\hat{\beta}$ becomes biased. To see this, note that $\hat{\beta} = \frac{\beta\sum(X_i-\bar{X})^2+\sum(X_i-\bar{X})\varepsilon}{\sum(X_i-\bar{X})^2} = \beta + \frac{\sum(X_i-\bar{X})\varepsilon}{\sum(X_i-\bar{X})^2}$. If we take the expected value of $\hat{\beta}$, we get: $E[\hat{\beta}] = \beta + \frac{\sum(X_i-\bar{X})E[\varepsilon]}{\sum(X_i-\bar{X})^2} = \beta$ because the $E[\varepsilon] = 0$. However, this only holds IF (A VERY BIG IF – VERY, VERY BIG IF) the independent variables are NOT correlated with the error term. If the independent variables are correlated with the error term, then we CANNOT take the expectations operator through the independent variables (the X's). We would need to find out the $E\left[\sum\left(X_i - \bar{X}\right)\varepsilon\right]$, which will be NOT be equal to zero if X_i and ε are correlated. To see this note that we would have:

$\sum(E[X_i\varepsilon] - E[\bar{X}\varepsilon])$. If X_i and ε are not independent then $E[X_i\varepsilon] \neq E[X_i]E[\varepsilon]$, and this term does not equal zero. Thus we will have a biased estimator of $\hat{\beta}$.

2 Measurement error of variables

We will discuss three cases of measurement errors in variables. First we will discuss what problems arise if only Y is measured with error, then if X is measured with error, and then if both X and Y are measured with error.

2.1 Measuring Y with error

Consider the following model:

$Y = \beta_1 + \beta_2 X_2 + \varepsilon$

We know that if all of our assumptions hold that β_1 and β_2 are best linear unbiased estimators. What happens if Y is measured with error? Suppose we observe the following Y^* instead of the "true" Y, where Y^* is given by:

$Y^* = Y + \mu$

Also suppose that $Cov(\varepsilon, \mu) = 0$.

Substituting into our initial model we get:

$Y^* = \beta_1 + \beta_2 X_2 + \varepsilon + \mu$

When we estimate this model we do not account for the fact that Y is mismeasured and we regress Y^* on a constant and X_2. Luckily for us, if $Cov(\varepsilon, \mu) = 0$, then our estimate for β_2 will still be unbiased and consistent, and we can perform the same statistical tests that we have been performing throughout the course. Thus, measuring Y with error does not pose a serious problem for our estimation of the parameters.

2.2 Measuring the X's with error

Consider the same model from above,

$Y = \beta_1 + \beta_2 X_2 + \varepsilon$

Now suppose that we do not observe X_2, but instead we observe a mismeasured X_2^*, where:

$X_2^* = X_2 + \nu$

Can we still obtain our best linear unbiased estimates of the slope parameter β_2?

The regression model that we will run with the mismeasured X_2^* is:

$Y = \beta_1 + \beta_2 (X_2^* - \nu) + \varepsilon$

This becomes:

$Y = \beta_1 + \beta_2 X_2^* + (\varepsilon - \beta_2 \nu)$, or

$Y = \beta_1 + \beta_2 X_2^* + \varepsilon^*$, where $\varepsilon^* = \varepsilon - \beta_2 \nu$

Recall that one of our assumptions needed for the Gauss-Markov theorem is that X and ε are uncorrelated. Recall that X and ε will be uncorrelated if $Cov(X, \varepsilon) = 0$. Also recall that $Cov(X, \varepsilon) = E\left[(X - E\left[X\right])(\varepsilon - E\left[\varepsilon\right])\right]$. For our example, we want $Cov(X_2^*, \varepsilon^*)$ since this is the covariance we are concerned with in the actual regression model we run. What is $Cov(X_2^*, \varepsilon^*)$? We will make a few assumptions first.

1. Assume $Cov(\nu, \varepsilon) = 0$
2. Assume $\nu \tilde{} N(0, \sigma_\nu^2)$
3. Assume $Cov(\nu, X_2) = 0$
4. Assume (for simplicity) $E\left[X_2\right] = 0$
5. Assume $Cov\left[X_2 \varepsilon\right] = 0$ (this is the classical model assumption)

Then,

$Cov(X_2^*, \varepsilon^*) = E\left[(X_2^* - E\left[X_2^*\right])(\varepsilon^* - E\left[\varepsilon^*\right])\right]$

By substitution

$= E\left[(X_2 + \nu - E\left[X_2 + \nu\right])(\varepsilon - \beta_2 \nu - E\left[\varepsilon - \beta_2 \nu\right])\right]$

By the facts that $E\left[X_2 + \nu\right] = 0$ and $E\left[\varepsilon - \beta_2 \nu\right] = 0$

$= E\left[(X_2 + \nu)(\varepsilon - \beta_2 \nu)\right]$

By expanding the two terms,

$= E\left[X_2\varepsilon + \nu\varepsilon - \beta_2 v X_2 - \beta_2\nu^2\right]$

By the properties of expectations operators

$= E[X_2\varepsilon] + E[\nu\varepsilon] - E[\beta_2 v X_2] - E[\beta_2\nu^2]$

You need to recall that if A and B are random variables with means of zero, then $Cov(A, B) = E[(A - E[A])(B - E[B])] = E[AB]$. Using that result and our assumptions,

- $E[X_2\varepsilon] = Cov[X_2\varepsilon] = 0$
- $E[\nu\varepsilon] = Cov[\nu\varepsilon] = 0$
- $E[\beta_2 v X_2] = \beta_2 E[vX_2] = \beta_2 Cov[vX_2] = 0$

The results in the bullet points give us

$Cov(X_2^*, \varepsilon^*) = E[X_2\varepsilon] + E[\nu\varepsilon] - E[\beta_2 v X_2] - E[\beta_2\nu^2] = 0 + 0 - 0 - E[\beta_2\nu^2] = -E[\beta_2\nu^2]$

Recall the result that if A has zero mean, then $Var(A) = E[(A - E[A])^2] = E[A^2]$

We now have:

$Cov(X_2^*, \varepsilon^*) = -\beta_2 E[\nu^2] = -\beta_2\sigma_\nu^2$

Thus, even WITH all the assumptions we made, X_2^* and ε^* are STILL correlated, and our least squares estimators of the slope coefficient will be biased and inconsistent.

2.3 Measuring Y and X with error

I will appeal to logic rather than rigorous mathematics in this section, and I will not require any derivations to show that coefficient estimates are biased in this case. However, we have seen that if Y is measured with error we have no serious consequences to our coefficient estimates. We have also seen that if X is measured with error than our estimated coefficients will be biased and inconsistent. Is there any logical reason to think that if we measure both X and Y with error that our estimated coefficients will return to being the best linear unbiased estimators of the coefficients? Probably not. As the old saying goes, "Two wrongs don't make a right".

3 The Instrumental Variables (or IV) estimator

In this section I will only briefly describe the concept of the IV estimator so that you are familiar with it. As we will see, one of the problems with the IV estimator is that while it exists in theory it is practically very difficult to find a good "instrument".

3.1 The estimation problem

We have already seen that the estimation problem that we have is that our slope coefficient will be biased and inconsistent if we have errors in measurement of our independent variables because the regressor will be correlated with the error term.

3.2 The solution

Assume that X is mismeasured. A solution to the estimation problem above is to find a variable, Z, that meets two assumptions:

1. Z is highly correlated with X
2. Z is uncorrelated with any measurement error (either ν or μ mentioned above) as well as uncorrelated with the regression error term (ε)

As you can probably tell this is not going to be easy. We need to find some variable that is highly correlated with X and is also NOT correlated with any of the errors in the model. If we can find this Z variable, then we can calculate the slope coefficient (in a simple two-variable regression) as:

$$\beta = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

This particular formula for the slope coefficient will lead to consistent estimates of the "true" slope coefficient. One interesting thing to notice is that ordinary least squares is a special case of this instrumental variables technique. This is because if X is not measured with error and it is uncorrelated with the regression error term, we can replace Z with X. We also know that X will be perfectly correlated with X, which satisfies assumption 1.

There is one other small problem with instrumental variable estimation. If ALL of the independent variables in a multiple regression model are measured with error, then ALL of the independent variables need to be replaced with instrumental variables.

4 Specification Error

Thus far we have been assuming that the models we are estimating are the "correct" or "true" models. ALL of the statistical testing we have done makes this assumption, that the unrestricted model is the "true" model. However, suppose we estimate an incorrect or untrue model – how does this affect our least-squares estimates? We will discuss two types of untrue models – those models where irrelevant variables are included in the regression, and those models where relevant variables are omitted from the regression.

4.1 Including irrelevant variables

Suppose the "true" regression model is given by:

$Y = \beta_1 + \beta_2 X_2 + \varepsilon$

Suppose we estimate:

$Y = \beta_1^* + \beta_2^* X_2 + \beta_3^* X_3 + \varepsilon^*$

If you want to think of this in terms of the models we have run in class, think of X_3 as a dummy variable for whether the owner of a house has brown eyes. This should not affect the selling price of the house, and thus X_3 should NOT be included in our regression. But suppose we include it. Does it harm any of the results?

We can show two results. First, we need to find the estimator for β_2^*. We can show (by either solving the least-squares equations for β_2^* or by recalling that we solved those equations in chapter 4) that:

$\hat{\beta}_2^* = \frac{\left(\sum(X_{2i}-\bar{X}_2)(Y_i-\bar{Y})\right)\left(\sum(X_{3i}-\bar{X}_3)^2\right)-\left(\sum(X_{3i}-\bar{X}_3)(Y_i-\bar{Y})\right)\left(\sum(X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)\right.}{\left(\sum(X_{2i}-\bar{X}_2)^2\right)\left(\sum(X_{3i}-\bar{X}_3)^2\right)-\left(\sum(X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)\right)^2}$

We can also show that:

$\hat{\beta}_2^* = \beta_2 + \frac{\left(\sum(X_{2i}-\bar{X}_2)(\varepsilon_i)\right)\left(\sum(X_{3i}-\bar{X}_3)^2\right)-\left(\sum(X_{3i}-\bar{X}_3)(\varepsilon_i)\right)\left(\sum(X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)\right.}{\left(\sum(X_{2i}-\bar{X}_2)^2\right)\left(\sum(X_{3i}-\bar{X}_3)^2\right)-\left(\sum(X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)\right)^2}$

Now, taking the expected value of β_2^*, we get:

$E[\hat{\beta}_2^*] = E\left[\beta_2 + \frac{\left(\sum(X_{2i}-\bar{X}_2)(\varepsilon_i)\right)\left(\sum(X_{3i}-\bar{X}_3)^2\right)-\left(\sum(X_{3i}-\bar{X}_3)(\varepsilon_i)\right)\left(\sum(X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)\right.}{\left(\sum(X_{2i}-\bar{X}_2)^2\right)\left(\sum(X_{3i}-\bar{X}_3)^2\right)-\left(\sum(X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)\right)^2}\right]$

But, assuming the X's are uncorrelated with ε, we know that the expected value of the second term in the brackets (the really long ugly term) is equal to zero. So we are left with:

$E[\hat{\beta}_2^*] = E[\beta_2] = \beta_2$

Thus, $\hat{\beta}_2^*$ is an unbiased estimator of β_2. This is good news. We can also show that $\hat{\beta}_1^*$ is unbiased and that the expected value of $\hat{\beta}_3^*$ is zero.

What we do lose when an irrelevant variable is included is efficiency, because unless X_2 and X_3 are uncorrelated, $Var(\hat{\beta}_2^*)$ is greater than $Var(\hat{\beta}_2)$. What we could show is that the estimated variance of $\hat{\beta}_2^*$ is an unbiased estimator of the variance of $\hat{\beta}_2$, and all of our statistical tests will be valid.

4.2 Omitting a relevant variable

A larger problem occurs if we omit a relevant variable. Assume the true model is:

$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

Suppose we run the following model:

$Y = \beta_1^* + \beta_2^* X_2 + \varepsilon^*$

We can show that the least squares estimator for β_2^* is:

$\hat{\beta}_2^* = \frac{\left(\sum(X_{2i}-\bar{X}_2)(Y_i-\bar{Y})\right)}{\left(\sum(X_{2i}-\bar{X}_2)^2\right)}$

We can then show:

$\hat{\beta}_2^* = \frac{\beta_2\sum(X_{2i}-\bar{X}_2)^2+\beta_3\left(\sum(X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)+\left(\sum(X_{2i}-\bar{X}_2)(\varepsilon_i)\right)\right.}{\left(\sum(X_{2i}-\bar{X}_2)^2\right)}$

Since the X's are fixed and $E[\varepsilon] = 0$, the last term has expected value of zero and drops from the equation. We can then show:

$E[\hat{\beta}_2^*] = \beta_2 + \beta_3 \frac{\left(\sum(X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)\right)}{\left(\sum(X_{2i}-\bar{X}_2)^2\right)} = \beta_2 + \beta_3 \frac{Cov(X_2,X_3)}{Var(X_2)}$

Unless $Cov(X_2, X_3) = 0$, our estimate for β_2 will be biased if we fail to include all the relevant independent variables. More importantly, this estimator is inconsistent, meaning that if we draw more and more observations the bias will not go away.

The important result of this equation is that the sign of the bias can be determined if we know whether the $Cov(X_2, X_3)$ is positive or negative and whether β_3 is positive or negative. Also, we can see that if the $Cov(X_2, X_3)$ is very small, and the coefficient β_3 is also very small, then the bias of our estimate for β_2 should be small (especially if $Var(X_2)$ is large). This suggests that our estimates will not be badly biased if the variables we fail to include are not that important in determining the dependent variable and/or are not very highly correlated with the other independent variables.

4.3 Efficiency vs. bias

The question now becomes which should we be more concerned about, efficiency or bias. If we include irrelevant variables we lose efficiency, while if we omit relevant variables we obtain biased and inconsistent estimates. On a theoretical level, if the sample size is large this suggests we would favor including irrelevant variables over possibly excluding relevant ones, since the loss of efficiency will decline the larger the sample size. However, this goes AGAINST every thing I have been telling you to do when you build your models – it suggests you include all possible variables and then exclude the ones that are irrelevant, whereas I have been telling you NOT to use the "kitchen sink" approach. It is still a better idea to have your models based on some theory or intuition rather than to include every variable you can find, as this will make interpretation and presentation of the results much clearer if you can appeal to intuition to explain why the result is what it is.

5 Specification tests

Most of the formal tests for specification error involve techniques that we will not discuss in this class. However, I will list two types of tests that are frequently used:

1. Likelihood ratio test
2. Hausman specification test

However, there are a few ways to do "non-statistical" or "eyeball" tests of the data to see if measurement error occurs.

5.1 Regression Diagnostics

The best way to ensure that the estimates you obtain from a regression model are "good" estimates is to make sure that the data you have is very clean. When all is said and done, good, clean data are better than any fancy econometric techniques that an economagician can conjure up. This suggests that one of the MOST important things you can do as an empirical economist is spend time looking at your data. Now, no one will know if every data point is accurate, but you can help yourself out if you can find data points that seem to come out of nowhere. While it is true that some of these data points are accurate, at times you may be able to spot mistakes. There is one example of a researcher who estimated returns to education and obtained a very different estimate than most of the other estimates around that time period. When he looked back over his data he realized that the years to schooling column had been set to -99 if the information was not provided. Including all of these -99's as years of schooling had an influential effect on his results, and when he corrected the problem his results more closely paralleled the other results of the time. Also, I have just recently downloaded data from the National Income and Product Accounts to look at something for one of my own projects (the data has since been removed from my computer due to the fact it was carrying a virus, so be careful when downloading data!). This data had similar numbers for missing observations, and if I had blindly used the data I would have obtained "incorrect" estimates had I run any regressions with the data. Below are two methods you can use to diagnose some of the influential data points in your data set. Be aware that some of them may be influential because they truly ARE influential, while others may be influential because they have been mismeasured.

5.1.1 Regular residuals

Assume that we estimate the following model:

$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$

Recall that the residual, $\hat{\varepsilon}$, is given by:

$\hat{\varepsilon} = Y - \hat{Y}$

If we look at the residuals from a regression model, one method of determining where the "outliers" in the data might be is to look for large residuals. If a particular observation of Y generates a large residual this means that the coefficients and the corresponding X's do not do a good job of predicting Y. This suggests that the regression line is far away from that data point. This also suggests that the observation of Y (or one of its corresponding X's) may have been mismeasured. Again, it is difficult to know if the variable has been mismeasured unless you are very familiar with the data, and I do not suggest removing all data points that have "large" residuals. You could however run another model without the data point that has a large residual and see if your estimates change "significantly". If they do then this suggests the data point is influential – if they don't then this suggests the data point is not that influential even if it has a large residual.

Other uses for residuals will be discussed in chapter 6.

5.2 Studentized residuals and DFBETAs

Looking at the actual residuals is not always helpful however. It could be the case that one observation is so influential that it causes the regression line to run very near it, even though the true regression line does not. Two methods that may be useful in finding influential data points are studentized residuals and DFBETAs. Both of these methods require removing an observation and reestimating the regression model. For studentized residuals, what one does is look at the residual generated for the i^{th} observation of the regression line that is obtained when that observation is omitted. What this means is that we run the regression WITHOUT that observation, and then obtain the residual from the regression without that observation. We do this by calculating the predicted value of Y, or $\hat{Y}$, using the regression coefficients obtained without the variable. We then divide by the standard error of the regression model to obtain a standardized residual. Naturally this could be a very lengthy process – luckily SAS has a very nice method for calculating studentized residuals. I will post a homework that instructs you how to obtain studentized residuals in SAS over spring break. When I obtained studentized residuals for the model $wage = \beta_1 + \beta_2 tenure + \beta_3 age + \beta_4 school + \varepsilon$, I obtained about 380 that were greater (in absolute value) than 1.96 (which is the 5% critical value for 9154 observations). This suggests that there are about 380 observations that may be considered as outliers. Of course, we have very little way of knowing what to do with those outliers, but we do know we should take a closer look at them. We can also use our studentized residuals to check to see if our normality assumption holds. If more than 5% of the observations lie outside the interval $(-1.96, 1.96)$, then our normality assumption may be questionable. In our case, we have $\frac{380}{9154} = .0041512$, which is less than 5%. If we wanted to test at the 1% level, we have about 200 studentized residuals greater (again in absolute value) than 2.57. Since $\frac{200}{9154} = .0021848$ is less than 1%, we can conclude that our data is fairly normally distributed. There are other formal tests that one can perform, but these work as informal tests.

We can also use DFBETAs to test for influential data points. A DFBETA is a concept similar to a studentized residual, only now we are looking at standardized regression coefficients. To find a DFBETA we take the regression coefficient for the model with all of the observations and subtract the regression coefficient obtained from the model with the observation omitted. We then take the difference and divide by the standard error of the regression coefficient obtained from the model with the observation omitted. One method of looking at DFBETAs suggests finding DFBETAs that are greater in absolute value than 1.96; another suggests that as the sample size grows the chance that a particular observation is influential declines, so a better "critical value" to use would be $\frac{2}{\sqrt{N}}$, where N is the number of observations.