Universidade de São Paulo / Faculdade de Filosofia, Letras e Ciências Humanas
Departamento de Ciência Política
FLP-0468 & FLS-6183
2º semestre / 2018

Problem Set# 2: Omitted Variable Bias, Measurement Error and Research Design

Due Date: September 20, 2018

Please submit a write-up and do file at the beginning of class on September 20, 2018.

Answer Key

Part I. Omitted Variable Bias with Simulated Data

1. Based on Stock and Watson Question 6.9 and 6.10. Suppose that Y_i, X_i, Z_i satisfy the key assumptions in Key Concepts 6.4. In other words, we are assuming the zero conditional mean assumption, no outliers, no perfect multicollinearity and that the explanatory variables are i.i.d. You are interested in the causal effect of X on Y. Suppose that X and Z are uncorrelated (corr(X,Z)=0.0) as are the variables that are generated in the simulation in part a of the do file. In other words, we will assume the following:

$$Y \sim (100, 20)$$
$$X \sim (7, 8)$$
$$Z \sim (20, 2)$$
$$Corr(Y, X, Z) = \begin{pmatrix} 1 & 0.7 & 0.3 \\ 0.7 & 1 & 0 \\ 0.3 & 0 & 1 \end{pmatrix}$$

a) Now let's examine these properties in a sample of simulated data. Then, please estimate β_1 by regressing y onto x (without including z in the regression) so that the model you estimate is the following: $y_i = \alpha_i + \beta_1 x_i + u_i$. Does this estimator suffer from omitted variable bias? Please explain.

Since the corr (x,z)=0 and corr (z, y) $\neq$ 0, then the estimate of β_1 does not suffer from Omitted Variable Bias.

In the example, X and Z were supposed to be generated to not be correlated. In the sample, however, the corr (X, Z)<0. However, the null hypothesis that the corr (X,Z)=0 can not be rejected with 10% confidence. Therefore, these are not weakly and negatively correlated.

```
. pwcorr y x z,sig

             |        y        x        z
-------------+---------------------------
           y |   1.0000
             |
             |
           x |   0.7240   1.0000
             |   0.0000
             |
           z |   0.2518  -0.0492   1.0000
             |   0.0000   0.2721
```

```
. esttab m1 m2, ci
```

	(1)	(2)
	y	y
x	1.828***	1.864***
	[1.675,1.982]	[1.724,2.004]
z		2.937***
		[2.373,3.501]
_cons	88.00***	28.28***
	[86.33,89.68]	[16.72,39.85]
N	500	500

95% confidence intervals in brackets
* p<0.05, ** p<0.01, *** p<0.001

b) Now let's estimate β_1 by regressing y onto x and z in the regression so that the model you estimate is the following: $y_i = \alpha_i + \beta_1 x_i + \beta_2 z_i + u_i$. Does this estimator suffer from omitted variable bias? Please explain.

As X and Z are not correlated, the estimate of β_1 does not suffer from Omitted Variable Bias, but it may be less efficient.

c) Please state the formula for β_1 in the bivariate and multivariate cases and explain its interpretation.

In the case of a bivariate regression:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)}$$

In the case of a multivariate regression with two explanatory variables:

$$\beta_1 = \frac{\left(\sum_{i=1}^{n}(z_i - \bar{z})^2\right)\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right) - \left(\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})(z_i - \bar{z})\right)}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(z_i - \bar{z})^2 - \left(\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})\right)^2} =$$

= 1.864.

Note: See Chapter 9.3 FPSR for a simplified version of this formula.

In Mostly Harmless, Angrist and Pischke state (p. 35) the definition with respect to the population parameters:

$\widetilde{\beta_k} = \frac{Cov(y, \widetilde{x_{ki}})}{Var(\widetilde{x_{ki}})}$ where $\widetilde{x_{ki}}$ is the residual from the regression of x_{ki} on all other covariates.

John Fox also provides an explanation on page 93.

For every one-unit change in x, y is predicted to increase by 1.864 *holding Z constant*.

Angrist and Pischke summarize "It shows us that each coefficient in a multivariate regression is the bivariate slope coefficient for the corresponding regressor after partialling out all the other covariates."

d) Please state the formula for the standard error of the regression error u_i and explain its interpretation.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} u_i^{2}}{n-k-1}$$

$$\hat{\sigma}^2 = \frac{85294.1964}{497} = \frac{SSR}{n-k-1} = 171.6181$$

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} u_i^{2}}{n-3}}$$

$$\hat{\sigma} = 13.100309$$

The sum of squared residuals, or SSR, is the sum of the squared OLS residuals:

The **standard error of the regression (*SER*)** is an estimator of the standard deviation of the regression error *ui*. The units of *ui* and *Yi* are the same, so the *SER* is a measure of the spread of the observations around the regression line, measured in the units of the dependent variable.

e) Please state the formula for the variance and standard error of β_1 in the bivariate and multivariate cases. What is the estimated variance and standard error of β_1?

In the two variable regression case,

$$\sigma^2_{\beta_1} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0.0781^2 = .0061$$

$$= \sigma_{\beta_1} = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{\frac{\sum_{i=1}^{n} u_i^{2}}{n-2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 0.078$$

Another way of stating the formula for the standard error of β_1 in the two variable regression case is:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_x(1 - r^2_{xz})}$$

where $R^2_{xz} =$ is the r-squared from regressing x on z, n is the sample size and $S^2_x =$ sample variance of X.

In the multiple regression case,

$\sigma^2_{\widehat{\beta j}} = Var(\hat{\beta}_j) = \dfrac{\sigma^2}{nS^2_j(1 - R^2_j)}$ where $R^2_j =$ is the r-squared from regressing x_j on all other x's,

n is the sample size and $S^2_j =$ sample variance of X.

$\sigma_{\widehat{\beta_1}} =$ 0 .071 and $\sigma^2_{\widehat{\beta j}} = 0.0051$

.

	(1) y	(2) y
x	1.828***	1.864***
	(0.0781)	(0.0711)
z		2.937***
		(0.287)
_cons	88.00***	28.28***
	(0.852)	(5.886)
N	500	500

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

f) Now, let's suppose that X and Z are correlated such that corr(X,Z)=0.75 as are the variables that are generated in the simulation in part b of the do file such that:

$Y \sim (100, 20)$

$X \sim (7, 8)$

$Z \sim (20, 2)$

$$Corr(Y, X, Z) = \begin{pmatrix} 1 & 0.7 & 0.3 \\ 0.7 & 1 & 0.75 \\ 0.3 & 0.75 & 1 \end{pmatrix}$$

Let's estimate β_1 by regressing y onto x and z in the regression so that the model you estimate is the following: $y_i = \alpha_i + \beta_1 x_i + \beta_2 z_i + u_i$. Does this estimator suffer from omitted variable bias? Please explain.

As X and Z are correlated and both are included in the model, the estimate of β_1 does not suffer from Omitted Variable Bias.

```
. corr y x z
(obs=500)

             |        y        x        z
-------------+---------------------------
           y |   1.0000
           x |   0.7240   1.0000
           z |   0.3059   0.7282   1.0000

.
. regress y x

      Source |       SS           df       MS      Number of obs   =       500
-------------+----------------------------------   F(1, 498)       =    548.65
       Model |  113773.299         1  113773.299   Prob > F        =    0.0000
    Residual |  103270.567       498  207.370616   R-squared       =    0.5242
-------------+----------------------------------   Adj R-squared   =    0.5232
       Total |  217043.866       499  434.957647   Root MSE        =      14.4

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    1.82828   .0780542    23.42   0.000     1.674924    1.981636
       _cons |   88.00151   .8524495   103.23   0.000     86.32667    89.67635
------------------------------------------------------------------------------

. estimates store m3

. regress y x z

      Source |       SS           df       MS      Number of obs   =       500
-------------+----------------------------------   F(2, 497)       =    420.37
       Model |  136407.787         2  68203.8934   Prob > F        =    0.0000
    Residual |  80636.0792       497  162.245632   R-squared       =    0.6285
-------------+----------------------------------   Adj R-squared   =    0.6270
       Total |  217043.866       499  434.957647   Root MSE        =    12.738

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   2.694723    .100737    26.75   0.000       2.4968    2.892646
           z |  -5.000178   .4233374   -11.81   0.000     -5.83193   -4.168426
       _cons |   182.4062    8.02821    22.72   0.000     166.6328    198.1796
------------------------------------------------------------------------------

. estimates store m4

. esttab m2 m4, se

--------------------------------------------
                      (1)             (2)   
                        y               y   
--------------------------------------------
x                   1.864***        2.695***
                 (0.0711)         (0.101)   

z                   2.937***       -5.000***
                  (0.287)         (0.423)   

_cons               28.28***        182.4***
                  (5.886)         (8.028)   
--------------------------------------------
N                     500             500   
--------------------------------------------
Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

g) What is the variance of the β_1 estimated in (f)? How does this variance compare to the variance obtained in (b)?

In (e): $\sigma^2_{\beta_1} = 0.0051$

In (f): $\sigma^2_{\beta_1} = 0.0102$

h) Please comment on the following "When X and Z are correlated, the variance of β_1 is larger than it would be if X and Z were uncorrelated. Thus, if you are interested in β_1 it is best to leave Z out the regression if it is correlated with X."

It is true that "When X and Z are correlated, the variance of β_1 is larger than it would be if X and Z were uncorrelated." However, if we are interested in β_1 it is best *not* to leave Z out the regression if it is correlated with X. If X is correlated with Z leaving it out of the regression will cause omitted variable bias.

i) How would the exercise above change if Z is a dummy variable? Please see the hint in the do-file and re-run your analysis. This is a creative exercise and points will be rewarded for students who explore alternative models.

In exercise (h) above, we discovered that when X and Z are correlated, the variance of β_1 is larger than it would be if X and Z were uncorrelated. This question is asking you to think about what is different when Z is a dichotomous variable that is either "0" or "1." The hints in the do file provided examples of how to generate Z variables, but if you used d as was generated in the do file you are working with a variable that is not correlated with Z or Y. The rationale is the same as before. In other words, it is still the case that when X and Z are correlated, the variance of is larger than it would be if X and Z were uncorrelated. However, what changes with a dummy variable is the interpretation. In essence in these simplified models, the dummy variables is changing the intercept.

$$y_i = \alpha_i + \beta_1 x_i + \beta_2 d_i + u_i.$$

When d=0, the intercept α_i.

When d=1, the intercept $\alpha_i + \beta_2$.

Here is an example of a solution of how to generate d to meet the necessary conditions. This solution was proposed by Pedro Castro and Raquel Fernandes:

```
* Part C. Analysis with Dummy Variable

* Repetir análise sem correlação entre X e Z

clear

matrix m = (100,7,20)
matrix sd = (20,8,2)
```

```
matrix C = (1, 0.7, 0.3 \ 0.7, 1, 0 \ 0.3, 0, 1)
drawnorm y x z, n(500) means (m) sds(sd) corr(C) seed(12345)

corr x y z

* Tranformar z em dummy: igual a 0 se abaixo da média, igual a 1 se acima da média

sum(z)
return list
gen zmean = r(mean)

replace z=0 if z <= zmean
replace z=1 if z > zmean

corr x y z

regress y x
estimates store m5

regress y x z
estimates store m6

esttab m1 m2 m5 m6, se

* Repetir análise com correlação entre X e Z
clear

matrix m = (100,7,20)
matrix sd = (20,8,2)
matrix C = (1, 0.7, 0.3 \ 0.7, 1, 0.75 \ 0.3, 0.75, 1)
drawnorm y x z, n(500) means (m) sds(sd) corr(C) seed(12345)

corr x y z

sum(z)
return list
gen zmean = r(mean)

replace z=0 if z <= zmean
replace z=1 if z > zmean

corr x y z // a correlação entre z e as demais variáveis se altera quando a transformo
em dummy

regress y x
estimates store m7
```

```
regress y x z
estimates store m8

esttab m3 m4 m7 m8, se
```

Part II. Research Design

Please critique each of the following research plans.

(a) A professor is interested in determining whether there is gender bias in electing women to Congress. To determine potential bias, the researcher collects data on gender and re-election probabilities for all candidates running for Congress in 2014. The professor plans to conduct a "difference in means" test to determine whether the average rate of election is different for women versus male candidates.

The proposed research is too limited. There might be some potentially important determinants of re-election, such as campaign finance, political parties, previous government experience, and education. With additional data, a multiple regression model could be estimated to examine if the effect of gender is statistically significant after controlling for these additional variables.

(b) A political scientist at USP is interested in determining whether the length of graduate studies has a permanent effect on the earnings of USP political science graduates. She collects data on a random sample of students who were enrolled as graduate students in the program from 1980-1995. The data set includes information on each alumni's current salary, highest degree earned, length of time spent in the political science department, age, ethnicity, gender, time in current job, and whether the person is working in the private or public sector. The professor plans to regress salary on potential determinants of earnings.

In this case, we have to worry about selection bias. Even though the research study has made considerable efforts to address potential sources of omitted variable bias, there might be a problem of selection bias. There may be characteristics associated with pursuing a graduate degree in political science that might be correlated with future salaries. Ideally, the study should attempt to address the potential problem of selection bias.