Universidade de São Paulo / Faculdade de Filosofia, Letras e Ciências Humanas
Departamento de Ciência Política
FLP-406 & FLS-6183
2º semestre / 2017

Answer Key

Exercise 1. Please undertake the following routine in Stata. What do you observe from the results of both regressions for the coefficient estimates? How are these results related? Please explain.

1. set obs 100

2. gen x = invnorm(uniform())

3. gen e1 = invnorm(uniform())

4. gen y1 = 2 + 3*x + e1

5. reg y1 x

6. gen e2 = invnorm(uniform())

7. gen y2 = 2 + 3*x + e2

8. reg y2 x

Answer: This routine is to stress what we understand is behind the results of any regression model. In this example, we are modeling the sampling variability from the error and this is why each time we will get slightly or somewhat different estimates for the **α** and **β**. In empirical analysis, we only observe the results of one draw (of the many possible theoretical draws). In fact, we understand that we could have observed different yi's in a hypothetical replication, where the other causes came out a differently.

.

	(1)	(2)
	y1	y2
x	3.056***	2.986***
	[2.875,3.238]	[2.811,3.161]
_cons	2.106***	1.968***
	[1.902,2.309]	[1.772,2.165]
N	100	100
R-sq	0.919	0.921

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

Exercise 2. Now let us assume that the explanatory variable, X, is measured with error. This measurement error is relatively small (.01). What are the consequences for the regression results reported in Exercise 1?

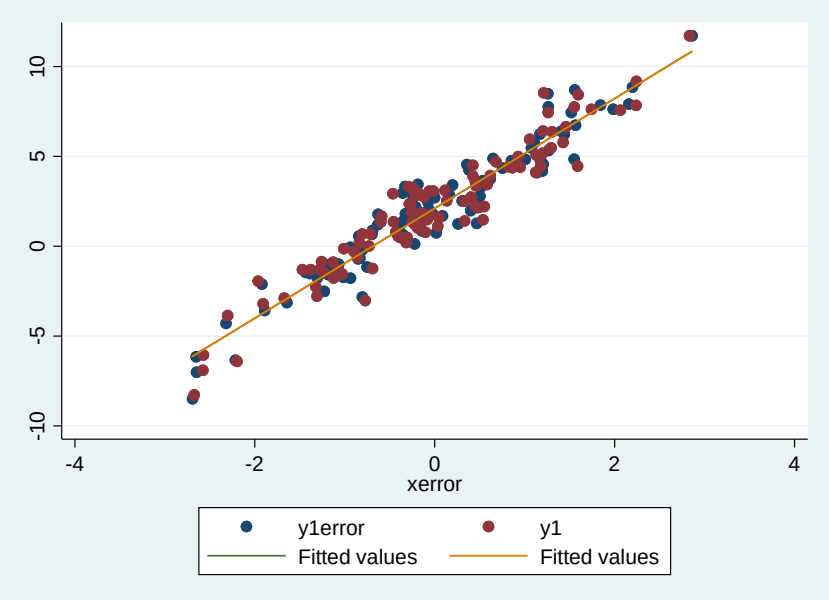
Case 1. Due to how we chose to model the measurement error (as a small constant error), there is only a small effect on the intercept as compared to the regression results reported in Exercise 1.

```
. esttab m1 m1x_error,  se r2
```

	(1) y1	(2) y1_xerror
x	3.056***	
	(0.0915)	
xerror		3.056***
		(0.0915)
_cons	2.106***	2.105***
	(0.103)	(0.103)
N	100	100
R-sq	0.919	0.919

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

Here we see the figure with jittering to make it easier to distinguish between the scatterplot of x and y1 versus the scatterplot of xerror and y1_xerror.



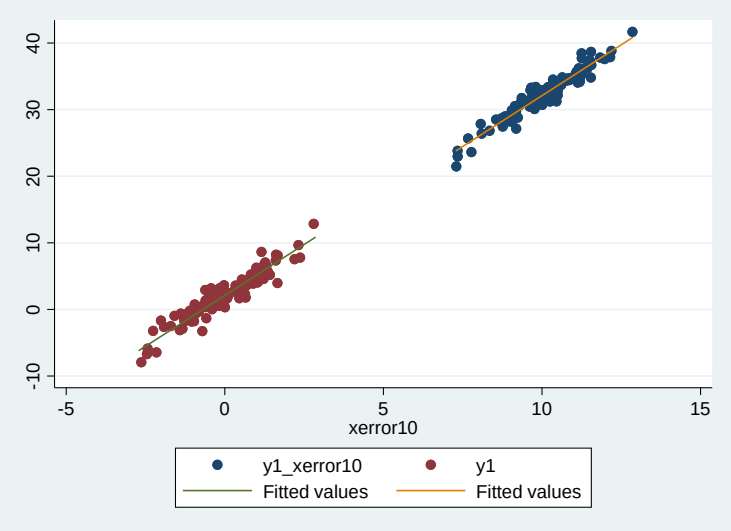
We know that the measurement error modeled in this manner should be reflected in the intercept. We can see this by thinking about our formulas:

$y = \alpha + \beta x + \varepsilon$

we have $x = x + c$ where c is a constant

$y = \alpha + \beta(x - c) + \varepsilon = \alpha + \beta x - \beta c + \varepsilon = (\alpha - \beta c) + \beta x + \varepsilon$

Case 2. We can also see this empirically by increasing the error in X to be a larger number (x+10). The red is the original scatterplot and the blue dots is the measurement error scatterplot.




```
. esttab m1 m1x_error10,  se r2
```

	(1) y1	(2) y1_xerror10
x	3.056***	
	(0.0915)	
xerror10		3.056***
		(0.0915)
_cons	2.106***	1.543
	(0.103)	(0.916)
N	100	100
R-sq	0.919	0.919

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

Case 3. We can show that the implications of measurement error in x are even more serious when we model measurement error as a variable that is measured with variability.

$y = \alpha + \beta x + \varepsilon$

we have $x = x + u$ where u has a mean 0 and standard deviation of 6.

$y = \alpha + \beta(x - u) + \varepsilon = \alpha + \beta x - \beta u + \varepsilon$

In this case, the measurement error in x will create an endogeneity bias as the measurement error will increase the residual error and also bias β.

```
. esttab m1 m1x_error m1x_error10 m3_xerrorcase3, se scalars(r2 rmse)
```

	(1) y1	(2) y1_xerror	(3) y1_xerror10	(4) y1_xerrorc~3
x	3.056***			
	(0.0915)			
xerror		3.056***		
		(0.0915)		
xerror10			3.056***	
			(0.0915)	
xerrorcase3				3.007***
				(0.0147)
_cons	2.106***	2.105***	1.543	2.103***
	(0.103)	(0.103)	(0.916)	(0.103)
N	100	100	100	100
r2	0.919	0.919	0.919	0.998
rmse	1.025	1.025	1.025	1.026

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

Exercise 3. Now let us assume that the dependent variable, Y, is measured with error. This measurement error is relatively small (.01). What are the consequences for the regression results reported in Exercise 1?

Given the assumptions of the CLRM (we will discuss these further), we will still obtain unbiased estimates of the beta coefficient.

However, again have to think about our cases. If the measurement error in y is a constant, it will change the intercept. Consider the case where the error is larger (10).

Case 1.

$y = \alpha + \beta x + \varepsilon$

we have $y = y + c$ where c is a constant

$y = \alpha + \beta x + \varepsilon + c$

```
. esttab m1 mylerror, se scalars (r2 rmse)
```

	(1) y1	(2) y1error
x	3.056***	3.056***
	(0.0915)	(0.0915)
_cons	2.106***	12.11***
	(0.103)	(0.103)
N	100	100
r2	0.919	0.919
rmse	1.025	1.025

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

If the measurement error in y is a variable with a mean 0 and standard deviation of 3, it will change the standard errors and variance.

Case 2.

$y = \alpha + \beta x + \varepsilon$

we have $y = y + u$ where u has a mean 0 and standard deviation of 3.

$(y - u) = \alpha + \beta x + \varepsilon$

In this case, the measurement error in y will increase the residual error and the standard errors of α and β.

Exercise 4. Based on your analysis of the results in Exercises1, 2 and 3, what do you conclude? How do your conclusions related to the regression assumptions outlined by Gujarti?

To review, there are several important assumptions of the Classical Linear Regression Model that are necessary to guarantee that we obtain BLUE estimators in OLS. These are assumptions about the Population Regression Function and not the Sample Regression Function.

1. Linearity in parameters.
2. E(error|x)=0 or E(error)=0 if x is non-stochastic.
3. Error has constant variance.
4. Cov (x, e)=0
5. Cov (u_i, u_j)=0
6. N>k
7. x must vary and there can be no outliers.

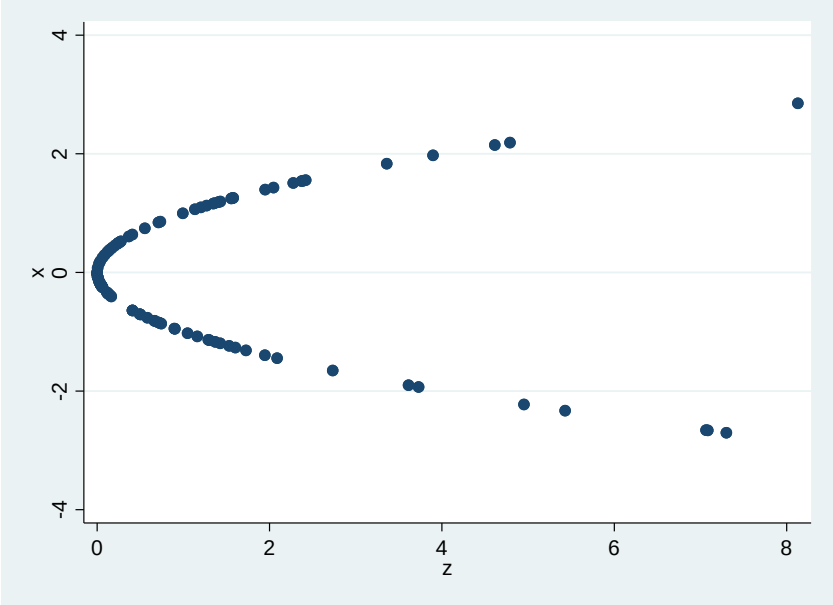
This exercise is calling attention to Assumption #7 and Assumption #4 which are both being violated. When we have measurement error (large measurement error can produce outliers) in the explanatory variables, OLS estimates will be **biased**. See Chapter 13 of Gujarati and Porter for a helpful discussion.

Parte II. The Multiple Regression Model

Exercise 5. Now let's add a second explanatory variable, Z. One of the assumptions of the regression model is that there is no exact collinearity between the explanatory variables in a

multiple regression model. Please create a variable Z that does not violate this assumption. Now, do a scatterplot to analyze the relationship between X and Z.

Let's generate z such that $z=x^2$. Note that this does not violate OLS assumptions.



Exercise 6. Please re-do the analysis reported in Exercise 1 adding a second explanatory variable to the model. What do you observe from the results of both regressions for the coefficient estimates? How is the interpretation of the coefficients different from Exercise 1?

```
. esttab m1 model_z, se scalars(r2 rmse)
```

	(1)	(2)
	y1	y1
x	3.056***	3.044***
	(0.0915)	(0.0919)
z		-0.0692
		(0.0590)
_cons	2.106***	2.192***
	(0.103)	(0.126)
N	100	100
r2	0.919	0.920
rmse	1.025	1.023

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

The second model is now testing the effect of x and z on y. In this model, x continues to have a statistically significant effect on y. However, the interpretation of this effect is different. The effect is now contingent on holding z constant. Similarly, z does not seem to influence Y holding x constant.

Exercise 7. Now let's add a third explanatory variable, V. Please create a variable V that is a dummy variable where precisely 40% of the observations are coded as 1 and 60% are coded as 0. Please re-do the analysis reported in Exercise 1 adding V as a second explanatory variable to the model. What do you observe from the results of both regressions for the coefficient estimates? How is the interpretation of the coefficients different from Exercise 1?

```
. esttab m1 model_z model_zv, se scalars(r2 rmse)
```

	(1)	(2)	(3)
	y1	y1	y1
x	3.056***	3.044***	3.050***
	(0.0915)	(0.0919)	(0.0909)
z		-0.0692	-0.0618
		(0.0590)	(0.0585)
v			1.885
			(1.018)
_cons	2.106***	2.192***	2.164***
	(0.103)	(0.126)	(0.126)
N	100	100	100
r2	0.919	0.920	0.923
rmse	1.025	1.023	1.011

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

The third model is now testing the effect of x, z and v on y. In this model, x continues to have a statistically significant effect on y. However, the interpretation of this effect is different. The effect is now contingent on holding z and v constant. As v is a dummy variable, the concept of holding v constant is less insightful. We will discuss this further in the class on interactions.