**Universidade de São Paulo / Faculdade de Filosofia, Letras e Ciências Humanas**

Departamento de Ciência Política

FLP-0468 & FLS-6183

2º semester / 2019

**Omitted Variable Bias**

**Due Date: October 3, 2019**

**Part I. Omitted Variable Bias with Simulated Data**

1. Based on Stock and Watson Question 6.9 and 6.10. Suppose that  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.
2. Now let’s examine these properties in a sample of simulated data. Then, please estimate  by regressing y onto x (without including z in the regression) so that the model you estimate is the following:$\hat{y}\_{i}=\hat{α}\_{i}+\hat{β}\_{1}x\_{i}+\hat{u}\_{i}.$ Does this estimator suffer from omitted variable bias? Please explain.
3. What assumptions are we violating when we have an omitted variable bias?
4. Now let’s estimate  by regressing y onto x and z in the regression so that the model you estimate is the following:$\hat{y}\_{i}=\hat{α}\_{i}+\hat{β}\_{1}x\_{i}+β\_{2}z\_{i}+\hat{u}\_{i}.$ Does this estimator suffer from omitted variable bias? Please explain.
5. Please state the formula for $\hat{β}\_{1}$ in the bivariate and multivariate cases and explain its interpretation in both cases.
6. Please state the formula for the standard error of the regression error  and explain its interpretation in both cases.
7. Please state the formula for the variance and standard error of $\hat{β}\_{1}$ in the bivariate and multivariate cases. What is the estimated variance and standard error of $\hat{β}\_{1}$?
8. Now, let's suppose that X and Z are correlated such that corr(X,Z)=0.75. Let’s estimate $β\_{1}$ by regressing y onto x and z in the regression so that the model you estimate is the following:$\hat{y}\_{i}=\hat{α}\_{i}+\hat{β}\_{1}x\_{i}+\hat{β}\_{2}z\_{i}+\hat{u}\_{i}.$ Please compare this model to the bivariate regressions where one of the variables is excluded. In otherwords, please compare the multiple regression model to $\hat{y}\_{i}=\hat{α}\_{i}+\hat{β}\_{1}x\_{i}+\hat{u}\_{i}$ and $\hat{y}\_{i}=\hat{α}\_{i}+\hat{β}\_{1}z\_{i}+\hat{u}\_{i}$.
9. What is the variance of the $\hat{β}\_{1}$estimated in the bivariate regression model $\hat{y}\_{i}=\hat{α}\_{i}+\hat{β}\_{1}x\_{i}+\hat{u}\_{i}$ in this case? How does this variance compare to the variance obtained in (c) when the correlation between X and Z was zero?
10. Please comment on the following “When X and Z are correlated, the variance of $\hat{β}\_{1}$is larger than it would be if X and Z were uncorrelated. Thus, if you are interested in it is best to leave Z out the regression if it is correlated with X.”
11. The R-squared is the fraction of the sample variance of Yi explained by the regressors. If we have more factors explaining Y, then we will never see the R2 close to 1. Can we compare the R2 across the models? Why?
12. 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.
13. One researcher wants to analyze the turnout rate by section in the 2016 São Paulo municipal elections. This is his model:  $ turnout = α+β\_{1}income+μ $. In this model, the sections are our units, and income indicates the mean monthly income for the voters on that section. The dataset includes only the aggregated information about the socioeconomic profile of the electorate (education, age, gender, salary range etc.). It also includes geographical information about that districts. Knowing our main explanatory variable is income, can you think of a variable that could be causing omitted variable bias? Can you explain your argument theoretically? What would be a better model that fixes the bias?