Universidade de São Paulo
Faculdade de Filosofia, Letras e Ciências Humanas
Departamento de Ciência Política

FLS-6183 & FLP-468
Métodos Quantitativos de Pesquisa II
2º semestre / 2019

Lorena G. Barberia

Lab 4 // Class 5
Multicollinearity

In this assignment, we will continue to work with simulated data. In the last class, we created and used matrix operations. Today, we will create data with a simulation using the matrix command to generate the correlations between the explanatory variables.

1. **Please fill in the table with the results you obtained after running the do file commands.**

Case 1. Correlation (x1, x2)=0				Case 2. Correlation (x1, x2)=0.4		
Sample size	N=10	N=30	N=100	N=10	N=30	N=100
Coefficient x1	0.834	1.058	0.961	0.837	1.02	0.969
Std Error x1	0.250	0.169	0.098	0.273	0.185	0.107
Vif x1 (1/Vif)	1	1	1	1.19 (0.840)	1.19 (0.840)	1.19 (0.840)
Coefficient x2	0.992	1.082	0.982	0.991	1.09	0.980
Std Error x2	0.250	0.169	0.098	0.273	0.185	0.107
Vif x2 (1/Vif)	1	1	1	1.19 (0.840)	1.19 (0.840)	1.19 (0.840)
Pw Corr x1 x2	Correlation (0.0) Stat sig (1.0)	Correlation (0.0) Stat sig (1.0)	Correlation (0.0) Stat sig (1.0)	Correlation (0.4) Stat sig (0.25)	Correlation (0.4) Stat sig (0.25)	Correlation (0.4) Stat sig (0.0)
Case 3. Correlation (x1, x2)=0.7				Case 4. Correlation (x1, x2)=0.9		
Sample size	N=10	N=30	N=100	N=10	N=30	N=100
Coefficient x1	0.842	0.9770	0.978	0.851	0.887	0.998
Std Error x1	0.350	0.237	0.138	0.574	0.389	0.226

Vif x1 (1/Vif)	1.96(0.51)	1.96(0.51)	1.96(0.51)	5.26(0.19)	5.26(0.19)	5.26(0.19)
Coefficient x2	0.988	1.11	0.974	0.981	1.189	0.958
Std Error x2	0.350	0.237	0.138	0.574	0.389	0.226
Vif x2 (1/Vif)	1.96(0.51)	1.96(0.51)	1.96(0.51)	5.26(0.19)	5.26(0.19)	5.26(0.19)
Pw Corr x1 x2	Correlation (0.7) Stat sig (0.02)	Correlation (0.7) Stat sig (0.0)	Correlation (0.7) Stat sig (0.0)	Correlation (0.9) Stat sig (0. 0004)	Correlation (0.9) Stat sig (0. 0)	Correlation (0.9) Stat sig (0. 0)

2. Let us interpret the results in the table above.

a. In case 1, what did you observe between the samples as the sample size increases?

The accuracy of the coefficients increases as the sample size increases. The estimated coefficients get closer to the value of the "population" parameters.

We generated data such that generate y=.5 + x1 + x2 + r.
Therefore, we expect the regression coefficients to be close to these parameters, which are b1 = 1, b2 = 1, alpha = 0.5.

The standard errors also decrease as we increase the sample size.

b. Case 2 is a case of a positive, but weak correlation between both explanatory variables. When N = 10 what did you observe? How did the standard errors change between N=10 and N= 30? Did you obtain better results when you increased your N to 100 observations? Is there a difference in magnitude of the effects?

There is no big difference in the magnitude of the effect, although as we increase the sample size the coefficients are closer to 1, which is the true or "population" parameters for the betas that we simulated.

The standard errors decrease as we increase the sample size. In the presence of even minor multicollinearity, there are gains from working with a larger sample size. The VIF does not vary from one sample to another because we do not change the correlation between the variables.

c. Case 3 is a case of a positive and slightly stronger correlation between both explanatory variables. When N = 10 what did you observe? How did the standard errors change between N=10 and N= 30? Did you obtain better results when you increased your N to 100 observations? Is there a difference in magnitude of the effects?

When N = 10, the estimated coefficients are farther away from the true "population" parameters as compared to the coefficient estimates when the sample size increases to 30 and 100.

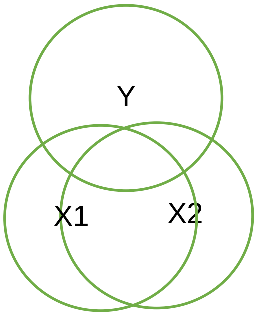
However, the confidence interval of the coefficient estimates are not so different across sample sizes.

The standard errors are smaller when we increase the sample size. This is still a case of multicollinearity that may be tolerable in our model, although it is a bit high.

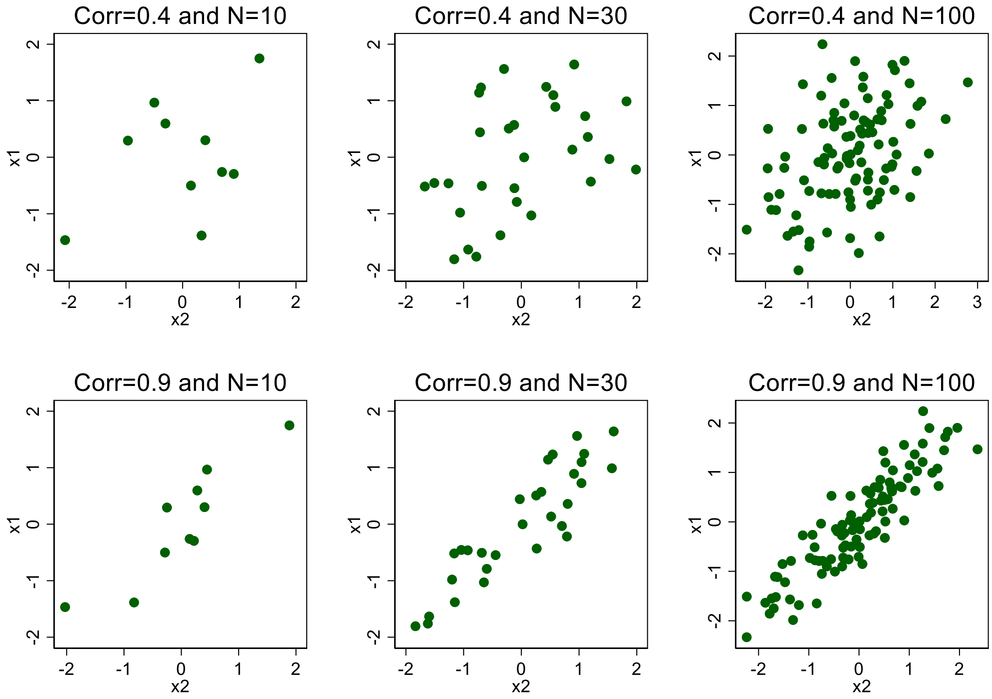
The VIF does not vary from one sample to another because we do not change the correlation between the variables.

d. Case 4 is a case of a positive and strong correlation between both explanatory variables. When N = 10 what did you observe? How did the standard errors change between N=10 and N= 30? Did you obtain better results when you increased your N to 100 observations? Is there a difference in magnitude of the effects?

The standard errors start much higher in the sample with n = 10, but decrease as we increase the sample size. The behavior of betas follows the same, increasing the number of observations brings us closer to population parameters. This is a case of strong multicollinearity, which is illustrated in the figure below. As the figure shows, the covariance of X1 and Y and of X2 and Y is strongly related by the covariance of X1 and X2.



3. Compare the case 2 scatter plots with the case 4 plots, describe what them represent, what do you observe increasing the sample. Please, think in terms of correlation.



Looking at both cases, we can say that increasing the sample makes the behavior of X1 and X2 more visible. In both cases, when n = 100 we see better the correlation between X1 and X2 is stronger as the correlation between these variables increases.

The correlation of X1 and X2 equal to 0.9 poses a more serious problem for making inferences because the relationship between x1 and x2 is almost linear, which violates one of the assumptions for OLS to produce BEST estimators.

4. What does the VIF tell us about the correlation? Compare the simulations for case 2 and case 4 to illustrate your explanation.

According to K&W, the VIF is the variance inflation factor for each independent variable. “This calculation is based on auxiliary regression model in which one independent variable, which we will call Xj, is the dependent variable on all of the other independent variable are the independent variables. The R2 statistic from this auxiliary model, r2j, is then used to calculate the VIF for variable j.”

$$\mathrm{VIF}_j = \frac{1}{(1 - R_j^2)}.$$

To interpret the VIF or the tolerance value (1 / vif) we can say that the higher the vif, or the smaller the tolerance value, the larger the estimated variance of xj in the specified theoretical model. The root square of the vif can be used to assess the impact of multicollinearity on the standard errors.

Case 2: 1.19 (0.840)

$$\sqrt{\mathbf{1.19}} = \mathbf{1,090}$$

Case 4: 5.26(0.19)

$$\sqrt{\mathbf{5.26}} = \mathbf{2,29}$$

We can see in case 4 that the variance for each parameter (X1 and X2) is substantially inflated by multicollinearity, but the inflation value when the correlation between x1 and x2 is less than 5.