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 2 // Class 3

A Deeper Look at Bivariate and Multivariate OLS

In this assignment, we will continue to work with simulated data. Last week, we began with a simple model in which our explanatory variable was a dichotomous variable. In this week, we will expand this analysis moving from a bivariate regression with a continuous explanatory variable and then proceed to estimate a multivariate model with two explanatory variables.

Case 1. The Effect of X on Y

1) In this case, we are estimating a bivariate regression model in the case that our explanatory variable is a continuous variable. Please write the sample regression function we are estimating:

$$\hat{Y} = \hat{\alpha} + \widehat{\beta_1} X + \hat{\mu}$$

2) What is our theoretical expectation of the estimated effect of X on Y given our simulation?

The expected value of the coefficient is 1.5, which is the value we designated for this parameter when we generated the data in the simulation in the do file. We do not expect the coefficient parameter to be exactly 1.5. Why? Given that we are working with data that have an expected mean and variance, we understand that the "true" value of the coefficient parameter will be within the confidence interval.

3) Is this what we observe in the estimated coefficient in the regression? How much does the estimated coefficient vary from its theoretical value? Is the theoretical value of the effect of X on Y within the 95% confidence interval?

The coefficient estimate for β_1 is 1.443804, which differs by 0.06 from the value we established in the simulation. As expected, the value of β_1 is within the confidence interval [1.100528 1.787079].

4) Based on the simulation, what is the expected value of Y? In the do file, we obtain estimates for the predicted or fitted values of Y. What results did we obtain from the estimated regression model?

The expected value of $\hat{y} = \widehat{1.19} + \widehat{1.44} * x + \hat{\mu}$

We use the coefficient estimates to calculate the fitted or predicted values of y when we issue this command in the do file: predict y_hat. In this case, Yhat is 5.46.

sum y_hat

Variable	Obs	Mean	Std. Dev.	Min	Max
y_hat	100	5.460603	.0478125	5.413972	5.509137

	y	y_hat
1.	4.883763	4.18509
2.	5.840159	6.925086
3.	5.091665	5.191964
4.	4.859359	5.289095
5.	5.331532	5.495621
6.	5.04073	5.434401
7.	4.193172	6.668694
8.	3.892908	4.677212
9.	5.705328	4.40942
10.	5.955616	5.694999
11.	6.59692	6.510209
12.	5.453054	5.287431
13.	2.627372	5.363961
14.	4.840024	4.724793
15.	4.969691	5.817614
16.	5.869767	4.187379
17.	6.603791	6.557923
18.	5.606212	5.471128
19.	6.785074	6.421083
20.	6.864299	5.174346
21.	5.687682	6.727281
22.	3.00388	4.325714
23.	6.404084	5.982905
24.	4.869703	4.57487
25.	6.783043	5.945959
26.	4.980854	5.133641
27.	5.076895	4.673073
28.	4.383229	4.416773
29.	6.39509	5.248838
30.	5.434165	6.572775
31.	5.619224	5.462022
32.	6.44482	6.720354
33.	6.508887	6.156434
34.	5.351218	4.649242
35.	6.268963	6.126781
36.	5.097416	5.907262
37.	8.194093	6.50485
38.	8.255773	6.154147
39.	4.28764	5.165313

5) What are the expected values of the residuals? What results did we obtain from the regression model?

As Stock and Watson point out in Chapter 5 and 6, we obtain the coefficient estimates based those coefficients that minimize the sum of squared residuals. Therefore, we would expect the sum of the residuals to be close to zero.

$$E(\hat{\mu}|X) = 3.19e - 10$$

6) Please summarize the estimated RMSE and its interpretation.

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(predicted_i - observed_i)^2}{N}} = 1.0622$$

The RMSE gives us the standard deviation of the unexplained variance (the standard deviation of the residuals). It can help us to understand the accuracy of our model and in the case of this simulation, it seems a bit high. The RMSE is measured in the units of the dependent variable and can be used to compare models with the same dependent variable. Thus, we can use the RMSE in these examples to compare across models.

7) Please summarize the estimated R2 and its interpretation.

0.4155*100 = 41,55. The model explains 41,55% of the sample variance of Y. The R-squared statistic ranges between zero and one, indicating the proportion of the variation in the dependent variable that is accounted for by the model. The value obtained from our model shows a moderate capacity to explain y.

Case 2. The Effect of Z on Y

8) In this case, we are estimating a bivariate regression model in the case that our explanatory variable is a dichotomous variable. Please write the sample regression function we are estimating.

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 z + \hat{\mu}$$

9) What is our theoretical expectation of the estimated effect of Z on Y given our simulation?

Our theoretical expectation is based on the data generating process we created in the simulation:

gen y = 1 + 2*z + r

when z=1, the expected effect is an increase in 2 of y.

10) Is this what we observe in the estimated coefficient in the regression? How much does the estimated coefficient vary from its theoretical value? Is the theoretical value of the effect of Z on Y within the 95% confidence interval?

Our simulation effect of z=1 is 2.034194 on y. This value varies 0.03 from our theoretical expectation and the confidence interval contains the true value of β_1 . CI [1.612371 2.456017]. Note that this variance in CI seems to be high.

11) Based on the simulation, what is the expected value of Y? In the do file today, we obtained estimates for the predicted or fitted values of Y. What results did we obtain from the regression model?

$$\hat{y} = 1.01 + 2.03 * z + \hat{\mu}$$

	y	y_hat
1.	4.883763	5.509137
2.	5.840159	5.509137
3.	5.091665	5.509137
4.	4.859359	5.509137
5.	5.331532	5.413972
6.	5.04073	5.509137
7.	4.193172	5.413972
8.	3.892908	5.509137
9.	5.705328	5.413972
10.	5.955616	5.509137
11.	6.59692	5.413972
12.	5.453054	5.509137
13.	2.627372	5.509137
14.	4.840024	5.509137
15.	4.969691	5.413972
16.	5.869767	5.413972
17.	6.603791	5.413972
18.	5.606212	5.509137
19.	6.785074	5.413972
20.	6.864299	5.413972

```
. sum y_hat
```

Variable	Obs	Mean	Std. Dev.	Min	Max
y_hat	100	5.460603	.8910141	3.178327	7.397727

12) What are the expected values of the residuals? What results did we obtain from the regression model?

As Stock and Watson point out in Chapter 5 and 6, we obtain the coefficient estimates based those coefficients that minimize the sum of squared residuals. Therefore, we would expect the residuals to be close to zero.

$$E(\hat{\mu}|X) = 5.49e - 10$$

13) Please summarize the estimated RMSE and its interpretation.

RMSE= 1.0626

The RMSE gives us the standard deviation of the unexplained variance (the standard deviation of the residuals). It can help us to understand the accuracy of our model and in the case of this simulation it seems a bit high. The RMSE is measured in the units of the dependent variable and cab be used to compare models with the same dependent variable. Thus, we can use the RMSE in these examples to compare across models.

14) Please summarize the estimated R2 and its interpretation.

0.4831*100=48,31

The R-squared statistic ranges between zero and one, indicating the proportion of the variation in the dependent variable that is accounted for by the model. The value obtained from our model shows that the model explains 48 percent of the variation in y.

Case 3. The Effect of X and Z on Y

15) In this case, we are estimating a multivariate regression model with two explanatory variables. Please write the sample regression function we are estimating.

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x + \hat{\beta}_2 z + \hat{\mu}$$

16) What is our theoretical expectation of the estimated effect of X and Z on Y given our simulation? How do we interpret each coefficient?

Based on the simulation command we issued to generate y:

gen y = 1 + 1.5*x + 2*z + r

We are creating a data generating process in which increasing x by 1 unit will result in an increase in 1.5 units of y holding Z constant.

In case of z, when z=1, y is predicted to increase by 2 units holding X constant.

17) Is this what we observe in the estimated coefficients? How much do these coefficients vary from our theoretical values? Are the theoretical values of the effects of X and Z on Y within the 95% confidence intervals?

The model estimates the coefficient for $\hat{\beta}_1$ of 1.443804, which varies 0.06 from its simulated value. The "true" value of $\hat{\beta}_1$ is inside the 95% confidence interval [1.100528 1.787079].

The partial effect when Z equals 1 is estimated to increase y by 2.034194. This value varies 0.03 from the simulated value. The "true value of the coefficient parameter is within the 95% confidence interval: [1.612371 2.456017].

18) What are the expected values of the residuals? What results did we obtain from the regression model?

As Stock and Watson point out in Chapter 5 and 6, we obtain the coefficient estimates based those coefficients that minimize the sum of squared residuals. Therefore, we would expect the residuals to be close to zero.

$$E(\hat{\mu}|X) = 2.61e - 09$$

19) Please summarize the estimated RMSE and its interpretation.

20)

RMSE= 1.0675

The RMSE gives us the standard deviation of the unexplained variance (the standard deviation of the residual). It can help us to understand the accuracy of our model and, in the case of this simulation, it seems a bit high.

21) Please summarize the estimated R2 and its interpretation.

0.6299*100= 62,9%

The R-squared statistic ranges between zero and one, indicating the proportion of the variation in the dependent variable that is accounted for by the model. The value obtained from this model shows a moderated capacity to explain y as we are able to explain 62.9 percent of the variation in Y.