Notes 6: Multivariate regression

ECO 231W - Undergraduate Econometrics

Prof. Carolina Caetano

1 Notation and language

Recall the notation that we discussed in the previous classes.

- We call the outcome variable: ______________________________________
- We call the causal variable of interest:__________________________________
- We call the controls: ____

We say that we want to

____

Alternatively, we sometimes say that we want to

____

____

Examples:

1. Suppose that y = final grade, x_1 = # of classes attended, x_2 = amount of hours spent studying, x_3 = # of office hours attended, x_4 = # of sections attended. We can say that we regress the final grade onto classes, hours studying, office hours, and sections. We can also say that we regress the final grade onto classes, using hours studying, office hours, and sections attended as controls.
2. Suppose that y = birth weight, x_1 = # of cigarettes smoked per day, x_2 = mother's education, x_3 = mother's income, x_4 = mother's race, x_5 = mother's marital status. We can say that we regress the birthweight onto mother's smoking, education, income, race and marital status. We can also say that we regress the birth weight onto smoking, and we use the mother's education, income, race and marital status as controls.

You should also know that people ofter say **run the regression**, as in: "I want to run the regression of y onto x_1, x_2, ..., x_k," or "when you run the regression of y onto x_1, using x_2, ..., x_k as controls, you must..."

In the context of a multivariate regression, sometimes researchers use different nomenclature for the variables. You should get used to the following terms:

- y: ______________________________

  ____

- x_1: ______________________________

  ____

- x_2, ..., x_k: ______________________________

  ____

- Sometimes the entire set of variables x_1, x_2, ..., x_k is called: ______________

  ____

  ____

2 The multivariate regression method

The multivariate regression method is, at the essence, the same as the univariate method we just saw. The whole point is to find what is the line that best predicts the expected outcome y for a given value of x_1, x_2, ..., x_k.

Examples:

1. What is the expected final grade among students that go to a given number of classes, study hours, office hours, and sections?
2. What is the expected birth weight of babies born to mothers that smoked a given amount, and had a given education level, income, race and marital status?

This is no longer a visual method. For univariate regression, we could look at the scatter plot or the graph of averages, which are plots relating the output y to the variable x_1. However, now we would have to draw plots relating the output y to many variables at the same time: x_1, x_2, ..., x_k, which is impossible. We could, of course, plot the tridimensional graph of y as it relates to two variables, say x_1 and x_2, but even that is not too helpful. Our brain can't process tridimensional plots very well, it is hard to see intuitively what is going on in them.

The regression line has the shape

Nomenclature

- a, b_1, b_2, $\ldots$, b_k are called the ________________________
- a is individually known as the ________________________
- b_1, b_2, $\ldots$, b_k are known as the ________________________

How do we find the values of a, b_1, $\ldots$, b_k? The principle is the same as in the univariate case. We minimize the residuals. What are the regression residuals?

Just as in the univariate case, the regression method is called OLS (ordinary least squares) because it is possible to obtain the regression coefficients by minimizing the RMS (residual mean square) errors:

Actually solving this problem by hand can be very difficult unless you know matrix algebra, but you should still remember this point. If you were to do this, you would carefully take the derivative of the summation above with respect to each coefficient $a, b_1, \ldots, b_k$ to arrive at a system of equations. Although not mandatory, you could do this yourself as practice. However, even if you don't derive these equations for yourself, you must know them equations, and the principles that they represent:

1. The regression residuals must average exactly zero. In math terms, this means that the regression line must satisfy (space next page)

Notice: this principle is equivalent to saying that the point of averages $(\bar{y}, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k)$ is always in the regression line. In math terms we would express this to mean that the regression line must satisfy

If you would like to test your skills with the summation notation, try to show that equations (1) and (2) are equivalent.

2. The regression residuals must be uncorrelated with the explanatory variables x_1, x_2, $\ldots$, x_k. In math terms, this means that the regression line must satisfy

If you count, you will see that the two principles above yield $k+1$ equations. We have $k+1$ unknowns: a, b_1, b_2, $\ldots$, b_k. If we solve the system, we will get the coefficients of the regression line.

Of course, solving this system by hand can be very hard. However, solving linear systems of equations is easy for computers. Thus, you won't be expected to do this yourself ever. However, you must understand where the regression results come from, so you don't commit silly (and often embarrassing) mistakes for not understanding how your tools work.

3 The partialling-out formula for multiple regression

I told you that the system above is very hard to solve by hand. It is true, the entire formula of b_1, for example, is very long and complicated. However, it is possible to express it in a very smart way. This formula will turn out to be very useful many times as the course progresses. The formula is:

What is r_{1i}? It is the residual of the regression of x_1 onto $x_2, \ldots, x_k$. Explaining:

1. Write the regression line equation of the regression of x_1 onto $x_2, \ldots, x_k$

$$x_1 = d_1 + d_2 x_2 + \cdots + d_k x_k.$$

2. Run the regression. In other words, calculate $d_1, d_2, \ldots, d_k$ using a computer.

3. For each observation i, calculate the predicted x_1:

$$\hat{x}_{1i} = d_1 + d_2 x_{2i} + \cdots + d_k x_{ki}.$$

4. For each observation i, calculate the residuals:

$$r_{1i} = x_{1i} - \hat{x}_{1i}$$

5. Substitute in the equation above.

Observe that we could do the same for all the other b's. For example, for b_2 exchange the subscripts 1 and 2 in the formula and in the steps, and then do everything exactly the same.

By the way, remember the equation of b_1 in the univariate case? It was

$$b_1 = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)y_i}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}.$$

In the univariate regression case, the residuals are actually a regression of x_1 onto a constant (in step 1 we would write $x_1 = d_1$.) You can verify yourself that, in this case, $\hat{x}_{1i} = \bar{x}_1$. Therefore, $r_{1i} = x_{1i} - \bar{x}_1$, and thus the two equations for b_1 are identical.

3.1 Interpretation of the partialling-out formula

The following two statements are equivalent:

1. ____

   ____

2. ____

   ____

What are the residuals $r_{1i} = x_{1i} - \hat{x}_{1i}$? Take the actual value of x_1 for person i. Then take the predicted value of x_1 given the person's values of $x_{2i}, \ldots, x_{ki}$. The residuals are the difference. In other words: they are the part of x_{1i} which wasn't predicted by the controls.

So, what is the meaning of b_1?

- In the univariate case, the coefficient of x_1 measures how the expected y varies when x_1 varies.

  For example, how much much change in the final grade we can expect from an increase in class attendance of one unit. Notice that this is not the effect on grades if we force a student to attend one further class. No, it is simply the differences in grade that we expect to observe among, for example students that chose to go to 11 classes when compared to those that chose to go to 10 classes. That difference in grade may be due to the difference in class attendance, but it may also be due to natural differences between those students.

- In the multivariate case, we the coefficient of x_1 measures how the expected y varies when the part of x_1 which is not predicted by the controls varies. In other words: the coefficient of x_1 measures how the expected y varies when x_1 varies, but we fix the value of $x_2, \ldots, x_k$.

  In this case, we can analyze things in the way we did in the beginning of the course: fix, for example, the students grade in pre-requisites, SATs, number of other courses they attend, and time spent in social media. Then b_1 is how much we expect the grade to differ between students that attended 11 classes and students that attended

10 classes, but had all the characteristics above in common. That is a lot closer to the causal effect, because the students are more comparable. Are they entirely comparable? This depends on whether the stuff for which we are controlling can predict all the other things for which we are not controlling.

Well, the partialling-out formula allows us to think about this problem in a slightly different way. What is r_1 in our example? It is everything about class attendance which was not predicted by grade in pre-requisites, SATs, number of other courses they attend, and time spent in social media. One can think carefully about what this may be. One thing that is clearly there is the actual experience of class itself. Is natural ability in econometrics there as well? This is debatable. The hope is that after controlling for enough variables the only remaining thing inside the residual is the experience of coming to class. Then we regress the grades onto the residuals, which means that we are now regressing the grades onto the experience of coming to class. This would give us the expected difference in grades when we compare the experience of coming to 11 classes versus 10 classes. This is exactly what we want, this is the causal effect! The issue though is the premise that we cleaned up everything else with our controls. Did we?

We are now a little closer to saying that b_1 is the causal effect of x_1 on y.