Notes 5: More on regression and residuals

ECO 231W - Undergraduate Econometrics

Prof. Carolina Caetano

1 Regression Method

Let's review the method to calculate the regression line:

1. Find the point of averages $(\bar{x}_1, \bar{y})$.
2. Go one SD_{x_1} horizontally, and one $r_{y.x_1} \cdot SD_y$ vertically. This is the second point.
3. Trace the line that connects both points.

The regression line crosses two points: _____

Let's get the formula of the regression line by substituting these two points in a generic line equation:

The inclination of the line is given by b. Let's understand it better:

2 The meaning of the regression line

Let's write down the regression line, by substituting a and b calculated above:

Predictor: _____

If somebody has $x_{1i} = 5$, then the regression line predicts that their outcome will be It

doesn't mean that y_i is exactly that, but if the relation between x_1 and y is fairly linear, it should be a good prediction.

Notice that predictors are not necessarily good at predicting what would happen if we intervened and changed someone's x_{1i}. There is a difference between guessing what is the most likely final grade (y_i) of somebody that went to 5 classes ($x_{1i} = 5$), and guessing what will be the final grade of someone that we forced to attend 5 classes. In the first case, people selected to go to as many classes as they wanted. In the second case we intervened. The second case is about the causal effect.

In essence, suppose that the regression prediction is that someone that went to 5 classes will have a final grade of 35. Does that mean that if I get a random student in the course and force him to come to 5 classes, we can expect him to get a final grade of 35?

No. The reason is because of choice.

Students that chose to come to 5 classes on their own are probably not great students. The predicted grade of 35 is not only because students that attended 5 classes got enough knowledge in class to get 35 as final grade. It is also because students who attended 5 classes probably didn't study very much, didn't apply themselves in the homeworks, didn't get TA help, and so on. The regression prediction is taking all of this into account together.

A causal question is different. What would happen if I force a student to go to 5 classes? If the student was already planning to go to 5 classes, probably nothing will change. However, if the student was planning to go to 10 classes, he is probably a better student, one that would study more, participate in class, work harder on the homeworks. The fact that I forced him to go to 5 won't change who he is. Thus he is probably going to get a better grade than 35.

In other words, 5 classes don't cause a grade of 35. To find the causal effect of 5 classes on the final grade, we have to control for the confounders. We have to fix how much the student studies, participates, how well they do the homeworks. The next class is about that.

3 Regression Residuals

Regression Residual: is the difference between the actual value of y and the predicted value in the regression.

The regression line prediction (the expected y for x_{1i}) is

$$\hat{y}_i = \left(\bar{y} - \frac{rSD_y}{SD_{x_1}}\bar{x}_1\right) + \frac{rSD_y}{SD_{x_1}}x_{1i}$$

The residual (error of prediction) is: _____

3.1 Plotting the residuals

I said above that the residuals would cancel each other out. This is actually a fact, they do! If we average the residuals, the average is zero, every time. Checking that this is true is an excellent exercise to check your skills using summation signs. (If you would like to try, begin with the average residuals $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$ and substitute the equation for $\hat{y}_i$.)

It is always a good idea to plot the residuals. The plot should look like this:

The pattern above is a sign that the regression method is a good description of the data. This pattern is often a sign of **homoskedastic errors**. We will study more about this later in the course.

Other patterns often seen in the data are:

What do these plots have in common? _____

_____

The patterns above are often a sign of **heteroskedastic errors**. In this case, the regression method is not the best way to describe the data, but it is not a bad way either. We will also discuss more about this later in the course.

A yet different pattern is

What do these plots have in common? _____

_____

The patterns above are often a sign that the relationship between x_1 and y is not very linear. In this case, the regression line should probably not be used to describe the data.

The method of looking at residuals seems a bit silly now. It is an indirect way of judging the relationship between y and x_1. Why look at residuals when one can look directly at scatter plots and graphs of averages instead? Indeed at this point it makes no sense. However, plotting residuals becomes very useful indeed once we start studying multivariate regression. Hang in there.

3.2 Root Mean Square (RMS) error

How wrong is the regression line in predicting the outcomes? For one observation, the answer to this question is the residual error. We would like to get a general sense of the error, how wrong is it in average? That is not a good way to look at it, because there are positive and negative residuals. Hence, residuals could be very large, but in average they would cancel each other out. To solve this, we square the residuals, so that they are all positive, and then we average them. The problem is that the result would be the average squared residual. In order to return them to the right unit, we get the square root:

You can compute the RMS error in the sample. It is quite easy, but there is also an alternative trick:

4 The method of least squares

The regression method is often called Ordinary Least Squares (OLS). In fact, I like to use the term OLS, so you will see it frequently. The reason for this name is that the regression line is in fact the line that minimizes the RMS errors.

To say this again using our notation: suppose there is a generic line $y = a + bx_1$. The question is what would be the values of a and b such that this line is the one with the smallest possible RMS errors? In English: what is the line such that the prediction error is as small as possible?

The mathematical problem to solve is:

It turns out that the answer to that question is

So, among all the linear predictors, the regression line is the one with the smallest RMS errors. In practical terms, it means that it is the most precise among the linear predictors, the one with the smallest errors overall.

NOTE: To come up with the answer to the problem in (1) on your own is a pretty challenging exercise in dealing with the summation notation. You are not required to know this in this course. However, if you feel like proving that you are awesome, and at the same time practice your knowledge on calculus and summation notation, here are some things to remember:

1. Minimizing the square root of something is the same as minimizing what's inside, so you can get rid of the square root in equation (1).

2. To minimize a convex function (and $\frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_{1i})^2$ is a convex function of a and b) you must take the first order conditions. In other words, differentiate it with respect to a and set it equal to zero, then differentiate it with respect to b and set it equal to zero. Now you have a system of equations with 2 unknowns (a and b) and 2 equations (the first order conditions). Solve it.

3. The derivative of a sum is the sum of the derivatives.