Notes 7: Goodness of fit, and the linear model

ECO 231W - Undergraduate Econometrics

Prof. Carolina Caetano

1 Quick Review

Last class we were talking about the multivariate regression method. The point was to find the line

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

which best predicts the average y for a given value of $(x_1, x_2, \ldots, x_k)$. We spoke about the OLS method, and we even discussed the useful formula (analogous for b_2, ..., b_k).

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. You should also remember two other points:

1. ____

2. ____

2 Goodness of Fit: R^2

The R^2 is a measure of how well the regression line fits the data. There are several equivalent formulas for R^2, but two of them provide most of the intuition. The first is based on the definition of R^2.

The R^2 of a regression is ____

____

Now to the formula:

This measure has a lot of intuition in it. When is the correlation highest? When the two variables are the same. In this case, when y_i and $\hat{y}_i$ are the same.

When does this happen? ____

____

Additionally, remember that the correlation is a number between -1 and 1. Since the R^2 is the correlation squared, it is a number between 0 and 1. 1 when the data is on a line, all the way down to zero, when the data is not at all in a line.

Here are examples of R^2 as it relates to data plots in the univariate regression case:

(go to next page)

The fact that the data is not all on a line does not mean that the regression is bad. We still stand by the principles developed in the previous classes. We said it before, and will continue expanding on this later: regression is a good method when the relation between the variables is linear. It doesn't need to be perfectly linear, like in plot 1, it is enough that the part that is not predicted (the residuals) be fairly random. For this to be the case, it is necessary that residuals for a given value of $x_1, \ldots, x_k$ be just as likely to be positive as they are to be negative (as in plot 2). It is not necessary that they all be zero.

The R^2 is thus not a measure of whether the regression method should be used or not to describe relationships between variables. Two regression equations can have entirely different R^2 and be just as good (or bad). In fact, when studying causal relations, a regression with $R^2 = .2$ can be more useful than one with $R^2 = .95$. We will go back to this point in the future.

The following formula is less intuitive, but it is very used, and thus you should be familiar with it:

Notice: ____

____

3 Review of concepts (not part of the course)

These are a few things you should know from the pre-requisite courses. I am presenting them here as a review and so that you have a reference. For the variables Y, W and X, and the constant a,

1. $\mathbb{E}[Y + W] = \mathbb{E}[Y] + \mathbb{E}[W]$
2. $\mathbb{E}[a] = a$
3. $\mathbb{E}[aY] = a\mathbb{E}[Y]$

More useful for us, remember the conditional expectation rules

1. $\mathbb{E}[Y + W|X] = \mathbb{E}[Y|X] + \mathbb{E}[W|X]$
2. $\mathbb{E}[a|X] = a$
3. $\mathbb{E}[aY|X] = a\mathbb{E}[Y|X]$
4. $\mathbb{E}[f(X)|X] = f(X)$
5. $\mathbb{E}[f(X)Y|X] = f(X)\mathbb{E}[Y|X]$

These seem abstract, however, they are quite intuitive. The way to always know what to do is not to read them the right way. The right way to read $\mathbb{E}[Y|X]$ is to say: "expectation of Y conditional on X." This is too abstract. It helps if instead, you think: "what is the expected Y if I fix X?" or "what is the expected Y if I know X?"

Even better, think inside of an example. What is the expected final grade (Y) if I fix the number of classes attended (X)? In other words, what is the expected final grade among students that went to X classes?

Using examples makes the rules quite obvious. For example, suppose that $X =$ classes attended, then what is $\mathbb{E}[\log(X)|X]$? The best way to read this is: "what is the log of classes attended when I *know* the classes attended?" Duh, it is the log of classes attended! Alternatively, you can read it as "what is the log of classes attended when I fix the classes attended?" Well, if I fixed the classes to a given value, then the expected log of classes will be itself!

4 The Linear Model

We studied the regression method. Now we begin to study how it can help us understand causal relations between variables. The ability of the regression method to help us understand causal relations depends on the true nature of the causal relation. Before we get ahead of ourselves, consider the following equation:

where y, x_1, x_2, $\ldots$, x_k are as before.

- What is this equation saying about reality? ______________________________

  ____

  ____

A model restricts reality, in the sense that it says that reality is one way as opposed to another.

Now, suppose that we say that the world satisfies

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u,$$

and $\mathbb{E}[u|x_1, x_2, \ldots, x_k] = 0$. You should understand this to mean that we are saying that the world is such that y, x_1, $\ldots$, x_k, and u satisfy the equation and the expectation condition.

Let's understand what this model says about the world. From the rules of the conditional expectation:

This means that the expected outcome for a given value of the covariates is a linear equation of the covariates. In other words, that the best way to predict the outcome for a given value of the regressors is a linear equation.

Wait, what? What was the regression line again? It was the best linear prediction of the outcome for a given value of the covariates. What good was that for us if the true prediction was not linear? Not good at all. Who cares if we know what is the best line to describe something that should not be described with a line? Nobody.

However, the two conditions above told us that the prediction is a line. It's a line! And what is the best method to find that line? OLS!!!

So, summarizing, we just magically (not quite...) found the restrictions in reality that allow us to understand causal relations using OLS regression. But, what do they mean? Are they too restrictive? Are we asking too much? Is the world like that? Hard to answer at this point. What you should know is that yes, we are asking quite a lot. Intuitively, we are asking two things:

1. ____

   ____

   ____

2. ____

   ____

   ____