Notes Template

Shom Mazumder

Last updated June 16, 2018

(These notes are designed to accompany a social science statistics class. They emphasize important ideas and tries to connect them with verbal explanation and worked examples. However, they are not meant to be comprehensive, and they may contain my own errors, which I will fix as I find them. I rely on multiple sources for explanation and examples[1]. Thanks to the students of API-201Z (2017) for their feedback and Matt Blackwell for inspiration.)

WHERE ARE WE? WHERE ARE WE GOING?

In probability we studied how to predict outcomes based on a data generating model; in inference we studied how to infer something about the data generating model based on the outcomes. We use regression is in the same spirit of inference, but instead of focusing on one parameter at a time, it is mainly concerned with the *relationship* between two or more variables. Because policymakers and social scientists are keenly interested in the relationship between complex phenomena, regression is an attractive tool to analyze data. Regression, in this sense, is the workhorse tool to quantify the relationship of a combination of variables. Here we focus on the most important type of regression – linear regression.

CONTENTS

[1] DeGroot and Schervish (2012), Imai (2017), Diez, Barr, and Cetinkaya-Rundel (2015), Moore, McCabe, and Craig (2002), and notes by Matt Blackwell (2016)

CHECK YOUR UNDERSTANDING

- How does OLS choose its fitted line?
- What are some properties of OLS that make it a good model?
- How would you interpret a regression coefficient? Its standard error? Its p-value?
- How is regression and hypothesis connected? Confidence intervals?
- What are the predicted values from a regression?
- What are regression residuals and why are they useful?

SETUP

Regression is a tool that has multiple interpretations, and so there are multiple ways to introduce it. The most typical way is to start by saying, *suppose* a sequence of random variables Y is generated by the equation

$$Y_i \sim \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ...\beta_k X_{ik} + \epsilon_i$$

where the subscript i stands for observations which range $i = 1, 2, ..., n$, and there are k different types of X variables. X_{ik} is notation for the ith observation in variable X_k. $\beta_0 ... \beta_k$ are unknown constants, that is to say unlike X or Y they are not random. Finally, ϵ_i is also a random variable, and for now we will posit that they are distributed Normal with a mean of 0,

$$\epsilon_i \sim \text{Normal}(0, \sigma^2)$$

Notice there is a subscript i on epsilon, which means that the value of this random variable changes from individual to individual. σ^2 is another unknown constant.

We make these assumptions because we think they are a good guess about the data. And what is the data? It is always the realizations of Y, $X_1, ...X_k$, with n observations of each version. In contrast, we don't observe β_0 or ϵ_i. We need to use the data to estimate those.

The part that readers find most unsatisfying about this setup so far is probably the assumption that the Y is in fact a function of pairs of β and X_k summed together. We call this functional form a *linear combination*, thus the word "linear" in linear regression. The world is probably not linear, so isn't this too simplistic an assumption?

There are a couple of responses to this critique. First it is worth emphasizing that the ϵ_i term, which students starting to learn statistics often under-appreciate does a lot of work. ϵ_i is a random variable with some unknown variance. So even if the rigid function $\beta_0 + \beta_1 X_1 + ...\beta_k X_k$ is quite different from Y, the remaining ϵ_i can potentially be a value that "fills in the gap". The follow-up question is how large should σ^2 be for this to work out, and whether the mean 0 parameter is a good model. But, it's worth remembering that ϵ_i is in someways the crux of the regression.

Another response is that the linear combination is the *best linear approximation* to the true, more complex data generating process. All models are wrong, some are useful — if we get a good approximation of how Y was generated, that would be still be quite useful, especially if we knew it was the best approximation.

What do we mean by "best"? The answer to this completely depends on the measure of accuracy you choose to use. Intuitively, we want a measure that captures how well the estimates of Y come close to the actual Y. In linear regression, we always use one particular metric of accuracy,

$$\text{Sum of squared residuals} = \sum_{i=1}^{n}(Y_i - \text{Predicted Y}_i)^2$$

The subsequent sections will better explain what this formula means. But quickly, the intuition for why there is a square is so that negative and positive differences don't cancel out, and why we use squares instead of absolute values is because squares are mathematically easier to minimize.

SIMPLE REGRESSION ANATOMY

The fundamentals of OLS can be illustrated in the simple case where there is only one explanatory variable. We would like to estimate β_0 and β_1 in the equation

$$Y_i = \beta_0 + \beta_1 X + \epsilon_i$$

Using the data Y and X.

Example 1 (The Reversal of Fortune)**.** In a 2002 article, economists Acemoglu, Johnson, and Robinson argue that countries that were more wealthy and urbanized in the 1500s saw their fortunes reverse in the subsequent centuries. Countries such as Rwanda and Tanzania were high-density areas in the 1500s but in the 20th century had low GDP per capita. The authors argue that this is because European colonialism settled more in areas that were less developed in the 1500s, but then went on to become strong economies. A simple bivariate relationship motivates their argument.