

Topic 3 Introductory Note: Modelling Non-Continuous Dependent Variables

These notes are provided as a supplement to the lecture slides in BSM946 and are not expected to be used as your sole understanding of this topic. The main arguments are presented in the lecture, with extra notes here and the key reading to ensure your understanding of the topic is complete.

This week is starting to move away from the idea of having a continuous dependent variable. We have never explicitly discussed that the dependent variable is continuous, but everything you will have seen so far will most likely have been continuous. What do we mean by this? We just mean a variable that is not defined to take only a limited number of observations. It can therefore take any value between its minimum and maximum.

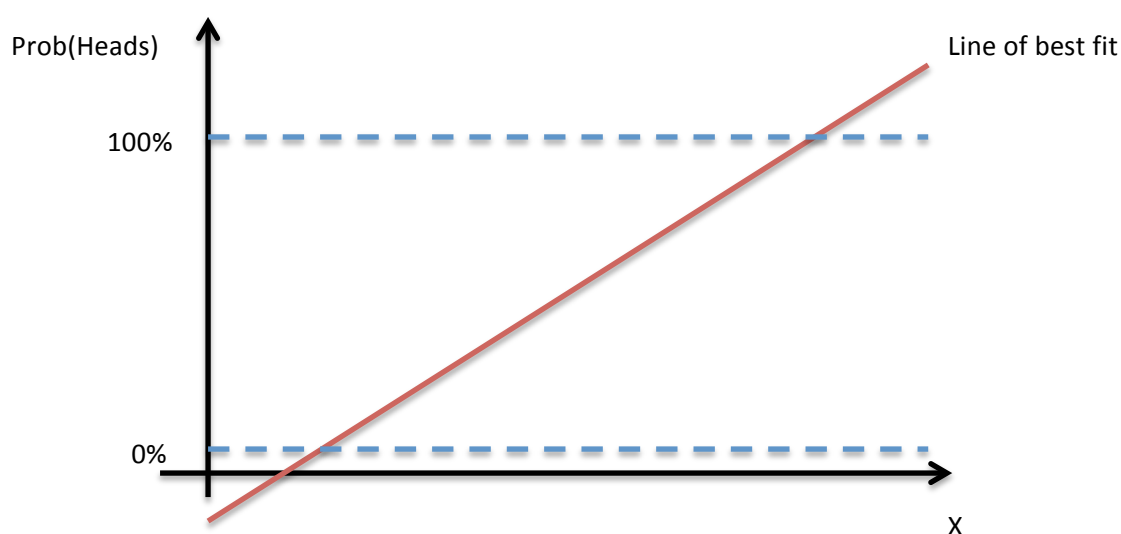
Let's think about an example or two. If we consider the time it takes a computer to complete a task. You might think you *can* count it, but time is often rounded up to convenient intervals, like seconds or milliseconds. Time is actually a continuum: it could take 1.3 seconds or it could take 1.33333333333333... seconds. We can imagine measuring the time to complete the task in ever more precise detail and thus the value is not limited in any real way except our measurement.

There are many more examples like this. A person's weight will also be continuous: Someone could weigh 180 pounds, they could weigh 180.10 pounds or they could weigh 180.1110 pounds. The number of possibilities for weight are limitless. What about income? You might think that income is 'countable' (because it's in pounds sterling) but who is to say someone can't have an income of a million dollars a year? Two billion? Fifty nine trillion? And so on... here it can take an infinite amount of values upwards. What about age? Surely that is not continuous? So, you're 23 years-old. Are you sure? How about 23 years, 19 days and a millisecond or two? Like time, age can take on an infinite number of possibilities and so it's a continuous variable.

While a lot of data is continuous there will be many times we want to model data that does not fit this mould. Discrete data is the alternative where data has to take certain values, often integers. Suppose we flip a coin six times and count the number of heads. The number of heads could be any integer value between 0 and six. However, it could not be any number between 0 and six. We could not, for example, get 2.5 heads. Therefore, the number of heads must be a discrete variable.

There are certain types of discrete data we are more concerned about this week, dichotomous variables, where an outcome has to be yes or no, one or zero, heads or tails. In other words the outcome has to be discrete and only one of two values. If we are using this type of data for a dependent variable it can cause certain issues in our regressions.

Consider the diagram below where we are considering the relationship between Y and X. If we are looking at the outcome of obtaining a heads from flipping a coin (1) rather than a tails (0), given an independent variable X, then we are trying to put a straight line in through this relationship. It often makes sense when thinking about if something happens or not to consider probabilities. Thus we may then model the probability of getting a heads (1).



What is the problem here? Well....our straight line does not stop at a probability of 100%, it goes beyond this. It also extends lower than 0%. Thus if we did fit a normal straight line regression in a situation like this we are likely to get results that do not make sense as they lie outside our range of possible values. We may obtain a probability that heads is equal to one that is greater than 100 percent.... Or less than zero....

Thus with binary choice (zero or one) data such as this, a standard OLS regression may cause issues. In the lecture we do discuss a few more issues that arise with OLS in this situation but this is the easiest to get our head around.

More formally, with OLS we are trying to measure the probability that the coin flip gives a head, given our X variable as we can see below:

$$Prob(\text{Coin toss lands on heads} | X) = \beta_0 + \beta_1 X$$

This is not bounded by 0 and 1 and thus we introduce two different models in the lecture that are a transformation of the OLS equation. The idea with these two models is to transform this model by a given function (F), so it is bounded between zero and one.

So instead we are looking at:

$$Prob(\text{Coin toss lands on heads} | X) = F(\beta_0 + \beta_1 X)$$

Where if the term $\beta_0 + \beta_1 X$ tends to negative infinity then F will give an answer of zero, and if $\beta_0 + \beta_1 X$ tends to infinity then F will give the answer of one. Thus our regression can never give an outcome of less than zero or greater than one. This is the basic idea behind the probit and logit models we introduce in the lecture.

Is this the only type of dependent variable where OLS will not work well? No!

We may have other forms of discrete data as our dependent variable that could cause issues. Think for example about the number of children someone has? This is discrete but it is also confined to be non-negative. While you can have zero children, you cannot have a negative amount of children. This is an example of what we would call count data and again OLS will not work well as it is not bound at zero and thus could predict a negative number of children for an individual.

So... before we start thinking about running a regression we need to consider what form our dependent variable comes in and whether OLS is the most appropriate model for that type of data. For binary dependent variables we should start thinking about probits or logits!

That is it for these introductory notes, please see the lecture slides and the extra simulation notes for more details on this topic!

David Morris