Pset 2_ 5 and 6

Matt Easton

3/7/2022

Problem Set 2

Question 5

Our response schedule for the Gibson study (from pg 324 of Freedman) is:

$$R = a + Mb + Ec + errorterm$$

Where:

R = Repression

M = Mass Tolerance

E = Elite Tolerance

a = the constant parameter

b = Mass Tolerance coefficient (effect/impact)

c = Elite Tolerance coefficient (effect/impact)

error term = the margin of error within our response schedule

To interpret this response schedule, we can see that when we plug in values for M (Mass Tolerance) and E (Elite Tolerance), we get an estimated value for R (Repression). The value for R is the sum of our constant a, the observed M multiplied by the effect of M (or b), the observed E multiplied by the effect of E (or c), and the margin of error (or error term).

For example, if M goes up by one unit with E held fixed, then R goes up by b units. Similarly, if E goes up one unit with M held fixed, then R goes up by c units.

There are two main assumptions embedded into our response schedule: 1) R is determined from the response schedule (meaning the errors are iid, the mean is 0 and the st. dev. is σ^2) 2) M and E are chosen at random by nature, independent of the error terms (in other words, there is no confounding).

These assumptions address (1) causal and (2) statistical issues with our response schedule. (pg 92 & 94, Freedman).

Both of these assumptions raise theoretical, conceptual, and empirical difficulties. Perhaps most obvious is concerns regarding confounding variables and endogeneity. Although we assume there are no confounding variables, empirically this may not hold. For example, it is easy to think of several variables that may actually be confounding our response schedule, such as access to media or state party (i.e. red states versus blue states). This is one of the most poignant conceptual difficulties with OLS regression analysis–there is an infinite number of potential confounders when dealing with observational data. Theoretically, we can try to address the largest ones in our model, but there is always the chance that we do not identify all of them–or even simply the largest ones.

Question 6

a)

Generally speaking, the second statement, “Regression analysis assumes causation but can be used to estimate the size of a causal effect—if the assumptions of the regression models are correct.” is closer to the truth.

This statement is closer to the truth because regressions rely on a set of assumptions in order to draw any meaningful conclusions (this is particularly true for observational data, which we will get to in part b). Regressions simply tell us about the relationship of the IV’s with the DV, and it is up to us to utilize assumptions to draw causal effects. If these assumptions do *not* hold–for example, if there are confounding variables–then the regression is not demonstrating causality (which is why the first statement doesn’t hold true).

Notably, if the regression is run on a properly conducted randomized experiment, then we can assume that the regression analysis demonstrates causality. In this scenario, the first statement would be just as true as the second one. However, in most cases we can be confident that the second statement is closer to reality, given the constraints of most data (particularly observational data).

b)

Yes, our answer changes depending on if we are using observational or experimental data.

As described briefly in part (a), experimental data has the advantage of true randomization and can therefore be used with a regression analysis to identify a causal effect. Therefore, when considering experimental data our answer in (a) should change to “Regression analysis can demonstrate causation”.

When considering observational data however, our answer in (a) remains the same. This is because there will always be the risk of endogeneity with observational data, which means that our regression will never be able to directly demonstrate causation.

c)

Our answer in (a) remains the same even if we replace “regression analysis” with “Analysis under the Neyman potential outcomes model”. We know this because the same issues presented in (a), such as endogeneity and confounding variables, would still be present in the Neyman model. The only way to circumvent this is if we can verify that we are using experimental data in the model (as discussed in part (b)).