Matching and causal inference
Demography 215

Carl Mason

March 9, 2008

1 Introduction

In a (pseudo)-profound philosophical sense, one can never *observe* causation. Scientists observe covariates and outcomes and then infer causality by invoking whatever prejudice suits their fancy. When Galileo dropped objects of dissimilar mass from the tower of Pisa and observed them to fall at nearly identical rates, he concluded that some mystical force called "gravity" *caused* this to happen.

As social scientists, we know better. We know that **correlation is not causation**. If Galileo et. al. did that experiment a million times in a million different places, we would still say with a smirk: "So what?" you haven't proved a thing. The Catholic Church took a similar approach on the matter of the heliocentric solar system.

But I digress. This week we will explore a statistical procedure for getting a little closer to what most people would accept as "causation". Specifically, we will take the kind of data that we get in the social sciences and try to convert it to something that looks and feels like Galileo's experiment, or it's modern equivalent: a clinical trial.

What distinguishes clinical trial like experiments from the typical social science "experiment" is that clinical trials have *treatment* and *control* groups. By randomly (or at least carefully) assigning one group of subjects to receive the "treatment" and other to receive the null treatment or possibly a placebo, the *effect* of the *treatment* can be measured as the difference between what happened to the treatment group and what happened to the control group. If the treatment involved say a mysterious incantation that might cause hair to grow on bald heads then the effect can be measured as:

$$\text{effect} = \frac{1}{n_{\text{treatment}}} \sum_{\text{treatment}} \text{new hair} - \frac{1}{n_{\text{control}}} \sum_{\text{control}} \text{new hair} \qquad (1)$$

or simply the mean experience of the treatment group - mean experience of the control group. What could be more straight forward?

Setting aside epistemological nuance, the clinical trial type experiment works because the subjects in the treatment group are in every relevant way assumed to be just like those in the control group *except* for the fact that they got the treatment. Thus any observed differences between the two groups has to be due to the treatment.

This sort of experiment happens in the social sciences just about never. That's because this sort of experiment is very expensive and nothing social scientists study is important enough to justify such expense. What we get are typically data where assignment to the treatment group is the result of self selection or else the action of a pernicious god.

1.1 What is a social scientist to do?

The point of this week's exercise is to explore some procedures that artificially transform social sciency data into something just like clinical trial like data only much cheaper and not really quite as good. In other words, we will figure out how to artificially construct a treatment and a control group–*after* the "experiment" is run and after the data have been collected. Of course This cannot always be done, but it might be possible in situations where:

1. You can specify a variable that indicates some sort of treatment.
2. The "treatment" in question happens at some point in time
3. The other covariates are in effect *before* the "treatment" occurs
4. You have access to a pool of observations where the variables of interest did not take effect

Variables like race do not work in this framework.

2 The problem

Consider a dataset like the artificial data that we created to play with proportional hazard models. Suppose we add a classic *non proportional* effect –

let's say that after age 50, all observations for which witnessed_lunar_eclipse == 1 are made to experience significantly lower mortality. If we wished to find this causal (or treatment) effect in the resulting data, we might naively separate our data into those observations which have `wle` and those which do not. Then we might assert that the difference in mean age at death of these two groups would give us the true effect *caused* by `witnessed_lunar_eclipse`.

In this very simple example, the above procedure would work pretty well– but that is because `witnessed_lunar_eclipse` is truly randomly assigned *independent* of all the other covariates. In a more realistic example, those who have `witnessed_lunar_eclipse` will also have higher values of perspicacity, `curiosity` and possibly `education` and `income` as well as lower values of `lethargy` and so on. In this more realistic example, the treatment group would consist of people who for complex socio-economic reasons *decided* to see an eclipse. And of course these complex socio-economic variables also have mortality effects.

Simply running a regression in this slightly more realistic case would lead you astray. The true effect of having seen an eclipse will be uh eclipsed by the presence or absence in the regression of all those other covariates.

3 The solution ...sort of

"Matching" techniques are used divide a dataset into a treatment and control group based on the covariates available (which are of course all the covariates that matter). In the simplest case, one can find one or more exact matches that is for every observation that has in the above example seen a lunar eclipse, one might identify an otherwise absolutely identical individual who has not–and then discard all the rest of the data. In this happy story, the treatment and control groups are perfectly balanced and a comparison of the mean age at death of each group should give us a reasonably good measure of the effect of `witnessed_lunar_eclipse` on mortality.

But that is way too easy. In most cases it will be necessary to do something much more ad hoc. Fortunately R has tools for this.

4 Matching, MatchIT and friends

R has a couple of tools for separating observations into artificial treatment and control groups. The most commonly use matching technique is called "distance" or "propensity score" matching. It works by assigning a weight to

each potential control group observation that is based on a logistic regression where the dependent variable is inclusion in the treatment group.

This week, we will create a dataset that obscures the treatment effect of one variable in a typical sort of way, and then we will use `MatchIt` to create a control group that will allow us to recover the true effect.

Next week Melissa Sills will show us how to do a more sophisticated version of matching that uses a genetic algorithm.