A PRIMER ON CAUSALITY*

Marc F. Bellemare†

Introduction

This is the second of two handouts written so as to help students understand the use of quantitative methods in the social sciences. This handout is dedicated to discussing (some) of the ways in which one can identify causal relationships in the social sciences. In keeping with the notation introduced in the handout on linear regression, let D be our variable of interest; y be an outcome of interest; and the vector $x = (x_1, \dots, x_K)$ represent other factors – or control variables – for which we have data. For the purposes of this discussion, let D measure a given policy, y measure welfare, and the vector x measure the various control variables the researcher has seen fit to include. See my "Primer on Linear Regression" for a more basic handout.

Mechanics

Recall that the regression of y on $(D, x_1, \dots, x_K)$ is written as

$$y_i = \alpha + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \gamma D_i + \epsilon_i, \quad (1)$$

where *i* denotes a unit of observation. In the example of wages and education, the unit of observation would be an individual, but units of observations can be individuals, households, plots, firms, villages, communities, countries, etc. Just as the research question should drive the choice of what to measure for y, D, and x, the research question also drives the choice of the relevant unit of observation.

The problem is that unless the researcher runs an experiment in which she randomly assigns the level of D to each unit of observation *i*, the relationship from D to y will not be causal. That is, γ will not truly capture the impact of D on y, as it will be "contaminated" by the presence of unobservable factors. Some of those factors can be included in $x = (x_1, \dots, x_K)$, of course, but it is in general impossible to fully control for every relevant factor. This is especially true when unobservable factors (e.g., preferences, risk aversion, technical ability, soil quality, etc.) play an important role in determining D and y. So even if we get an estimate of γ that is statistically significant, we cannot necessarily assume that the relationship between the variable of interest and the outcome variable is causal. In other words, correlation does not imply causation.

For example, suppose D is an individual's consumption of orange juice and y is (some) indicator of health. We have often discussed in lecture how a simple regression of y to D would provide us with a biased estimate of γ because orange juice consumption is nonrandom and not exogenous to health. That is, there are factors other than orange juice consumption which determine health. Some are observable (e.g., how much someone exercises; whether they smoke; their diet; etc.), but several are unobservable (e.g., their willingness to pay for orange juice; their subjective valuation of health; their level of risk aversion; their

---

* Version 3.0, August 2012.

† Assistant Professor, Sanford School of Public Policy, Duke University, Durham, NC 27708-0312, United States, marc.bellemare@duke.edu. This handout was prepared for the students in my PPS603 – Microeconomics of International Development Policy seminar.

genes; etc.) Thus, it really isn't sufficient to run a kitchen-sink regression (i.e., a regression in which everything observable is thrown in as a control) to properly identify the causal impact of D on y.

Identification

So how *do* we identify causality? The best way to do so is to run a randomized controlled trial (RCT). In this case, the idea would be to get a random sample of individuals of size N and to assign half of the sample (i.e., $N/2$) to a control group and half to a treatment group. The latter group would be told to consume, say, one glass of orange juice every morning, and the other half would be told not to do so. Then, after a suitable period of time, we would compare the mean of y between groups. The null hypothesis would of course be that the mean health of the treatment group is equal to the mean health of the control group. A rejection of the null in favor of finding that the mean health of the treatment group is higher than the mean health of the control group would then be evidence in favor of the hypothesis that orange juice is good for one's health. More than that – it would be evidence in favor that orange juice consumption *causes* good health.

The problem is that it is not always possible to run an RCT, and even the simple example described above would be subject to important problems. For example, the individuals in the treatment group may not comply with the experimenter's instructions, especially if they don't like orange juice (i.e., they may not consume one glass of orange juice every morning). More generally, they may simply forget to consume orange juice every morning. Similarly, individuals in the control group may not comply with the experimenter's instructions, especially if they really like orange juice (i.e., they may not abstain from consuming one glass of orange juice every morning). More generally, the individuals in the control group may end up inadvertently consuming orange juice when they are not supposed to. These reasons – and others – would contaminate one's estimate of γ in equation 1 and would invalidate the test of equality of means described above. So what is one to do?

Instrumental Variables Estimation

When one only has observational (i.e., nonexperimental) data at one's disposal, one of the best way to identify causality is to find an instrumental variable (IV) for the endogenous variable. In the example above, the endogenous variable is D, which is said to be *endogenous to* y.

What is an IV? It is a variable z that is (i) correlated with D; but (ii) uncorrelated with ϵ and which is used to make D exogenous to y. How does an IV exogenize an endogenous variable? By virtue of being correlated with the endogenous variable, yet uncorrelated with the error term, which is the definition of an instrument.

In other words, an instrument z, because it is uncorrelated with ϵ, "cleans out" D of its correlation with ϵ while keeping the relevant variation, i.e., the variation that allows studying the causal impact of D on y.

I realize that some of this might sound tautological, so let's look at an example. Angrist (1990) studies the impact of education (D) on wages (y). The problem is that education is endogenous to wage, if anything because people acquire education in expectation of the wage they think this will get them. In

other words, even if we find a positive coefficient for education in a regression of wage on explanatory variables, this is merely a correlation, and it does not necessarily indicate that education causally affects wages.

To instrument for this, Angrist had to find a variable that would be correlated with how much education someone would get, but uncorrelated with wage (except, of course, through how much education they acquire). The instrument he settled upon was an individual's Vietnam draft lottery number, since this correlates with whether one goes to war and is then subject to the GI Bill, but since those numbers are randomly generated, they are uncorrelated with one's wage.

How does IV estimation work, mechanically speaking? Recall that our equation of interest is

$$y_i = \alpha + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \gamma D_i + \epsilon_i. \quad (1)$$

The way IV estimation proceeds is to first regress the endogenous variable D on the instrument z as well as on the control variables in $x = (x_1, \ldots, x_K)$, such that

$$D_i = \delta + \pi_1 x_{1i} + \cdots + \pi_K x_{Ki} + \theta z_i + \epsilon_i. \quad (2)$$

Once equation 2 is estimated, it is possible to predict the variable D, whose prediction we label $\widehat{D}$ (the circumflex accent – or "hat" – denotes a predicted variable in econometrics) and to then estimate equation 1 as follows

$$y_i = \alpha + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \gamma \widehat{D}_i + \epsilon_i. \quad (1')$$

Note what has been done here: we have replaced the endogenous variable with an exogenized (i.e., "cleaned out" of its relationship with the error term) version of the same variable. The way it has been exogenized has been by regressing it on the IV, which is exogenous to the outcome of interest, and to obtain its predicted value, which we then use instead of the original endogenous variable.

The first requirement of an instrument – i.e., that it be correlated with D – is easily testable: we only need to check that the coefficient θ in equation 2 is significantly different from zero. The second requirement of an instrument – i.e., that it be uncorrelated with the outcome of interest y – cannot be tested for. Rather, one must make the case that it is truly exogenous to the outcome of interest. This is easier said than done in most cases, and some people have devoted entire careers to finding good IVs.

References

Angrist, Joshua D. (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from the Social Security Administrative Records," *American Economic Review* 80(3): 313-336.