Causal Inference with Observational Data

8. Tricks of the Trade II

Marc F. Bellemare

May 2018

Population Instead of Sample

What do you do when you are dealing with the population itself instead of dealing with a sample that is representative of a population?

Your first reaction might be to think "Well, I have the entire population, so I don't need to compute standard errors anymore, and everything I estimate is statistically significant."

Population Instead of Sample

Though that kind of reasoning is intuitively appealing would you be willing to submit for publication a paper where you make that claim?

Suppose, for example, that you have data on food-borne illness and farmers markets for all 50 states plus the District of Columbia for 2004-2013 (Bellemare and Nguyen, 2018). Would you really submit an article for publication in which you tell the editor and reviewers that you don't need to compute standard errors and run t-tests because you have the entire population of states?

Population Instead of Sample

In fact, if you look at published papers using data on all 50 states, those papers still report standard errors in tables of regression results. Why?

On the one hand, I understand analytically why having access to a whole population might obviate the need to compute standard errors. On the other hand, there are a few reasons why you might still want to treat your population as a sample.

Population Instead of Sample

One reason why you might want to treat your population as a sample is to test whether some estimated relationship is meaningful.

In other words, you might want to check that the relationship between your dependent variable and some regressor is statistically significant as a means of testing whether there really is a relationship between the two in your population, or whether the estimated relationship is indistinguishable from zero and the result of chance.

Population Instead of Sample

Another reason why you might want to treat your population as a sample and calculate standard deviations around the means of the various variables you are interested in is simply because those standard deviations are means in themselves—they are the average departure from the mean of each variable, or how far from the average you can expect each observation to be.

Population Instead of Sample

Lastly, a more compelling reason why you might want to treat your population as a sample is because you might be interested in prediction.

For example, a policy maker might ask you to predict the effect on your dependent variable of changing an explanatory variable by a certain amount. Without treating the population as a sample, you would be making a very sharp prediction: In essence, you'd give the policy maker a single number, without any uncertainty around it.

In practice, you would likely want to qualify that number with a range of credible values. And what better to do that than a confidence interval?

Comparing Distributions

There are methods to compare whole distributions. Let's cover a few of those methods.

There are two possibilities here. Suppose you have a single variable y broken into two groups. Suppose you have information on the duration of the hungry season experienced by rural households (y), and you also know who grows some crops under contract ($D = 1$) and who does not ($D = 0$) (Bellemare and Novak, 2017).

Comparing Distributions

You might be interested in two things:

1. Is it the case that $F(y|D=0) = F(y|D=1)$? That is, is the distribution of the duration of the hungry season statistically the same between those who participate and those who do not?
2. Is it the case that $F(Y)$ statistically follows a normal, logistic, Weibull, etc. distribution?

The former case is a two-sample test; the latter is a test with reference distribution.

Comparing Distributions

As with everything else, a first option is to apply the intra-ocular trauma test (so-named because it requires that something in a graph hits you "right between the eyes").

What if you would rather not rely on a subjective criterion like intra-ocular trauma? There is a test for that, the Kolmogorov-Smirnov (K-S). The K-S test is a nonparametric test that allows conducting both of the tests delineated above, and which can easily be conducted in Stata with the `ksmirnov` command.

Lag Identification

For a long time, it was common in economics to deal with endogeneity issues by lagging endogenous variables. For example, in the regression

$$y_{it} = \alpha + \beta x_{it} + \epsilon_{it} \tag{1}$$

where x is endogenous to y, people would try to solve the endogeneity-of-x issue by estimating the following equation instead:

$$y_{it} = \alpha + \beta x_{it-1} + \epsilon_{it}. \tag{2}$$

Lag Identification

The reasoning behind this type of “identification” is that there is no way that y in period t can cause x in time period $t-1$.

That is true, but that reasoning also illustrates a misunderstanding of what statistical endogeneity is about. Recall that statistical endogeneity has three sources: (i) reverse causality, (ii) unobserved heterogeneity, and (iii) measurement error.

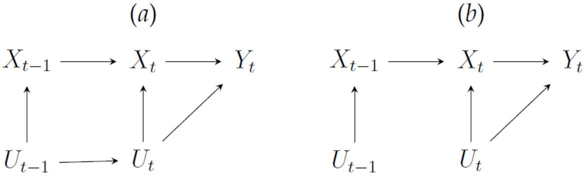
So replacing x_{it} with x_{it-1} will no doubt take care of reverse causality issues, but is unlikely to deal with unobserved heterogeneity or measurement error problems.

Lag Identification

In Bellemare et al. (2017), my coauthors and I look at the practice of lag identification and show that not only does it feel cheap, it also does not yield causal identification on the cheap.

To see this, consider the following DAG.

Lag Identification



Notes: This is a representation of the causal relationship from X to Y that is implied when using a lagged value of X to overcome the identification problem in figure 1. In (*a*), U_t depends on its previous value U_{t-1}. In (*b*) the two are independent.

Figure: (Source: Bellemare, Masaki, and Pepinsky, *Journal of Politics*, 2017.)

Lag Identification

What we show (i) graphically, (ii) analytically, and (iii) via Monte Carlo simulations is that regressing on x_{it-1} can not only introduce more bias than there would be if you were merely ignoring the problem and regressing on x_{it} instead, it also leads to inference problems in that it makes it more likely that you will reject the null that $\beta = 0$ when, in fact, it should not be rejected (i.e., it makes type II errors more likely).

Lag Identification

As such, it is not a solution to endogeneity problems.

Unfortunately, the practice remains widespread in political science, and even in some fields of economics: For the year 2014, my coauthors and I found 11 such cases of lag identification in *AER*, *Econometrica*, *JPE*, *QJE*, *REStud*, and *REStat*.

We also found 54 such cases in comparable journals in political science.

Interpreting Coefficients

For all the advanced technical knowledge we impart students in standard econometrics classes, we often don't do a very good job of teaching them how to interpret what they are estimating. Here are some particular scenarios that require specific interpretations.

Suppose you are interested in studying the effect of education on wage in the following modified Mincer equation

$$y = \alpha + \beta e + \gamma x + \delta e \times x + \epsilon, \tag{3}$$

where y denotes a person's wage, e denotes that person's education, x denotes their experience, and ϵ is an error term with mean zero.

Interpreting Coefficients

What is the effect of education on wage for the average individual in the data here? Too many would be quick to say that that effect is measured by the coefficient β when in fact $\frac{\partial y}{\partial e} = \beta + \delta x$ because the interaction term is included in the equation.

Interpreting Coefficients

Another application relates to how you interpret variables and their square.

For example, when using individual-level data, it is not uncommon to include a person's age and the square of their age as regressors in order to account for potential nonlinearities (here, U-shaped or inverse U-shaped relationships) between their age and the dependent variable.

When studying the accumulation of assets, for example, there usually is an inverse U-shape relationship between age and asset accumulation: People in their late teens and early 20s usually have few assets; they accumulate assets throughout their work life, buying real estate, saving for retirement, etc. Once they retire, they sell off their assets in order to consume.

Interpreting Coefficients

So suppose you want to study the effect of age a person's assets while controlling for a number of variables, you would estimate

$$y = \alpha + \beta A + \gamma A^2 + \delta x + \epsilon, \tag{4}$$

where A denotes age and x is a vector of controls, both affecting assets y. Here, the marginal effect of age is $\frac{\partial y}{\partial A} = \beta + 2A$.

Interpreting Coefficients

The age thing is a bit of a no-brainer, and few people make the mistake of only looking at age while ignoring its square.

Recently, however, I came across a paper looking at the inverse farm size–productivity relationship where the authors regressed productivity y (measured in kilograms of output per hectare) on the size of a plot of land A and the square of that plot's size A^2, so that they estimated.

Interpreting Coefficients

In order to test whether there was an inverse relationship between farm size and productivity, they simply looked at whether the estimated β was significantly different from zero and negative.

Obviously, the proper test was to test the null that $H_0 : \frac{\partial y}{\partial A} = \beta + 2A = 0$ versus the alternative hypothesis!

If you are interested in reading more about testing for U-shaped relationships, see Lind and Mehlum (2009).

Interpreting Coefficients

It is common to have

$$\ln y = \alpha + \beta x + \gamma D + \epsilon, \tag{5}$$

where D is a binary treatment variable, y is the dependent variable, x is a vector of control variables, and ϵ is an error term whose mean is equal to zero.

To take a classic example, y could be an individual's wage, D a variable equal to one if they have a college degree and equal to zero otherwise, and x their age, gender, etc. The equation above is called "semi-logarithmic" because we take the logarithm of only one side of the equation.

Interpreting Coefficients

(A log-log equation would regress a logarithm on the left-hand side on a logarithm on the right-hand side, in which case the estimated coefficient is directly interpretable as an elasticity, i.e., a percentage change in for a 1% increase in the variable of interest. It is unfortunately not possible to take the log of a binary treatment, because the log of zero is undefined.)

Perhaps because of the foregoing, a common mistake in interpreting β in the equation above is to treat it as a percentage. That is, to claim that β tells us by how much y changes in percentage terms when an observation goes from untreated to treated, i.e., when goes from $D = 0$ to $D = 1$.

Interpreting Coefficients

In what is perhaps the shortest AER paper ever, however, Kennedy (1981), correcting an earlier mistake in an earlier paper by Halvorsen and Palmquist (1980), derives a formula that allows deriving the effect $\widehat{g}$ of the treatment in percentage terms, which is such that

$$\widehat{g} = \exp\left\{\widehat{\beta} - \frac{1}{2}V(\widehat{\beta})\right\} - 1, \tag{6}$$

wherein $\widehat{g}$ is, in Kennedy's words "the percentage impact of the dummy variable on the variable being explained."

Log of Zero

Speaking of semi-logarithmic or log-log equations, it is not uncommon for a variable whose log we take to have a certain number of observations be equal to zero. Unfortunately, taking the log of zero yields an undefined value, which means that an observation for which you have $\ln(0)$ is dropped from estimation, which can introduce serious bias in your estimates.

In an old paper, MaCurdy and Pencavel (1986) suggested adding a small quantity to each observation before taking the log. That is, they suggested taking the transformation $\ln(x + 0.001)$ or $\ln(x + 1)$ instead of just $\ln(x)$, since the former allows keeping the zero-valued observations and eliminating the bias that would come from dropping those observations.

Log of Zero

As it turns out, this is no longer acceptable, and what people do nowadays is to take an inverse hyperbolic sine (IHS) transformation instead. The advantages of the IHS are twofold:

1. The IHS transformation is log-like, and in Bellemare and Wichman (2018), we derive exact elasticities for it.
2. The IHS allows not only keeping zero-valued observations, it also allowed keeping observations whose value is negative.

Log of Zero

In Bellemare et al. (2013), this allowed us to take an IHS transformation of each household's marketable surplus of each good. Recall that a marketable surplus M can be such that $M \gtreqless 0$ depending on whether the household is a net seller, autarkic, or a net buyer.

The IHS transformation is such that

$$IHS(x_i) = \ln\left(x_i + \sqrt{x_i^2 + 1}\right), \tag{7}$$

For more on the IHS, see Burbidge et al. (1988), McKinnon and Magee (1990), and Pence (2006). To see it in practice, see Bellemare et al. (2013), although note Ravallion's (2017) caveat.

"Determinants of ..." Papers

There was a time, when computing power was relatively weak and good data were rare, where you could get into good journals merely by writing a paper looking at the determinants of some outcome variable.

Nowadays, it is extremely difficult—if not impossible—to publish that kind of paper in decent journals, and for a good reason.

(One exception is if you're first to ask an important question.)

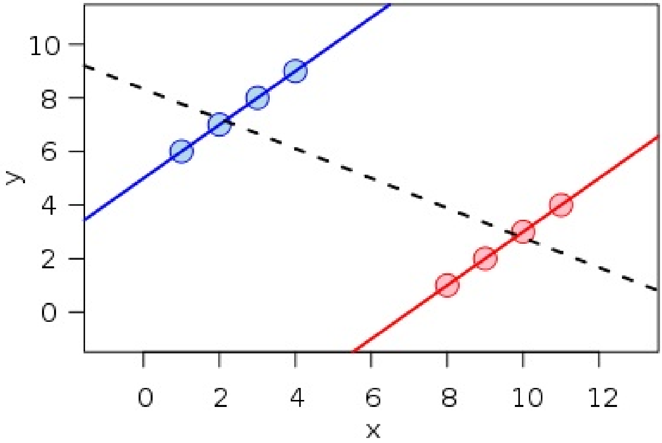
“Determinants of ...” Papers

That reason is Simpson’s paradox:

> *a paradox in probability and statistics, in which a trend appears in different groups of data but disappears or reverses when these groups are combined. . . . This result is often encountered in social-science and medical-science statistics, and is particularly confounding when frequency data is unduly given causal interpretations. The paradoxical elements disappear when causal relations are brought into consideration.*

"Determinants of ..." Papers

The quote above says it all, but an image is worth a thousand words, too:



“Determinants of ...” Papers

“Determinants of ...” papers are problematic, because they usually consist of linearly projecting some outcome on a number of variables, estimating the resulting regression, and then making up stories about those variables that are significant.

Though this counted as doing applied econometrics 20 years ago, good social scientists now see this kind of paper with a healthy dose of skepticism.

If you absolutely must write that type of paper, make sure you estimate several specifications of the same regression—from most to least parsimonious—and be honest about the fact that you’re only estimating partial correlations.

Thinking About Selection Systematically

Reading Smith and Sweetman (2016) introduced me to the Roy model, which is useful to think systematically about selection.

For instance, if you are interested in estimating the causal effect of a treatment in which people self-select, it helps to think about that selection as follows.

Thinking About Selection Systematically

The general problem is that we are interested in estimating the average treatment effect $E(y_1 - y_0)$, where y_1 denotes the outcome variable when the treatment is taken up and y_0 denotes the outcome variable when the treatment is not taken up.

The identification problem stems from the fact that for any given observation i, we cannot observe y_1 and y_0 simultaneously.

Thinking About Selection Systematically

The Roy model simply posits that a unit i (e.g., an individual, a household, etc.) will take up the treatment iff

$$y_{1i} - y_{0i} \geq c_i, \tag{8}$$

where c_i denotes the cost of taking up the treatment for unit i. The three relevant quantities y_{1i}, y_{0i}, and c_i can be used to think systematically about selection.

Thinking About Selection Systematically

Smith and Sweetman explain:

1. Holding y_{1i} and c_i fixed, units are less likely to take up the treatment as y_{0i} increases.
2. Holding c_i fixed, units are increasingly likely to take up the treatment as $y_{1i} - y_{0i}$ increases.
3. Holding y_{1i} and y_{0i} fixed, units are increasingly likely to take up the treatment as c_i decreases.

Thinking About Selection Systematically

A lot of this might seem obvious, because it is.

Still, Smith and Sweetman's discussion allows thinking systematically about selection instead of providing a haphazard discussion thereof, as is so often seen in papers where selection is an issue.

Moreover, the Roy model may help think about what kind of control variables are likely to help control for selection.

Testing for Mechanisms

The Credibility Revolution has stressed out the importance for policy and for social science of estimating causal relationships.

Now that we know how to estimate such relationships, what policy makers and social scientists want to know is the mechanism(s) whereby a causal effect operates.

Testing for Mechanisms

So suppose you are estimating the relationship

$$y = \alpha_0 + \beta_0 x + \gamma_0 D + \epsilon_0, \tag{9}$$

and assume you are interested in the causal effect of treatment D on the outcome y. Assume further that γ_0 is causally identified (say, because you have a selection-on-observables design, or because D is assigned at random).

Testing for Mechanisms

What a lot of people have been doing up until recently has been to estimate the following version of the equation of interest

$$y = \alpha_1 + \beta_1 x + \phi_1 M + \gamma_1 D + \epsilon_1, \tag{10}$$

where M is a mechanism whereby D is thought to cause y. The usual test here has been to see if γ drops out of significance once M is included; if so, then M was thought to be a mechanism for D.

Testing for Mechanisms

As it turns out, that old method is wrong. In a recent article, Acharya et al. (2016) show that the previous equation can lead to biased estimates, and they develop a method for whether M is a mechanism through which D causes y which only relies on determining which control variables are pre-treatment, and which control variables are post-treatment.

The real strength of Acharya et al.'s contribution is that it sometimes allows determining whether M is the *only* mechanism through which D causes y.

Testing for Mechanisms

The method is pretty simple:

1. Split controls into pre- and post-treatment controls, i.e., x_0 and x_1, respectively, where $x = (x_0, x_1)$.
2. Estimate $y = \alpha_1 + \beta_1 x + \phi_1 M + \gamma_1 D + \epsilon_1$.
3. Compute $\widetilde{y} = y - \widehat{\phi}_1 M$.
4. Estimate $\widetilde{y} = \alpha_2 + \beta_2 x_0 + \gamma_2 D + \epsilon_2$ and bootstrap the entire procedure.

Testing for Mechanisms

The coefficient of interest is γ_2. Here, if you fail to reject the null that $\gamma_2 = 0$, you have effectively shown that M is (statistically) the only mechanism whereby D causes y.

Note that you should read the paper for yourself—there is a lot more to it than what I describe (especially in terms of assumptions), but this provides the gist of the method.

Dealing with Imperfect IVs

When working with IVs, we often wish we had an IV whose exogeneity is unquestionable.

Unfortunately, in the real world where we live, this is rarely the case with observational data.

Dealing with Imperfect IVs

So how can you give your imperfect (read: only plausibly exogenous) IVs a bit more credibility?

Let's start with Conley et al. (2012). As always, we are interested in the coefficient on treatment D in the regression

$$y = \beta_0 D + \epsilon_0, \tag{11}$$

but we happen to have an IV z.

Dealing with Imperfect IVs

With an IV that is perfectly exogenous we have in theory that the coefficient γ in

$$y = \beta_1 D + \gamma_1 z + \epsilon_1 \tag{12}$$

is equal to zero—if the IV meets the exclusion restriction, then the IV only affects the outcome through the treatment. Here, β and γ are not jointly identified because D is endogenous.

Dealing with Imperfect IVs

With only a plausibly (i.e., sort of, kind of) exogenous IV, the problem is that γ will not be zero—though we hope that it will be small enough, since the smaller it is, the better.

So how do we go about this problem?

Dealing with Imperfect IVs

Conley et al. propose three solutions, all requiring you to incorporate extra information in the problem:

- We can specify only a range of possible values for γ,
- We can impose a distribution on γ, or
- We can adopt a full Bayesian approach (i.e., have a prior for both β and γ).

Dealing with Imperfect IVs

Then, it is possible to either obtain a point estimate or confidence interval, depending on the method chosen, for β, the estimand of interest.

If the point estimate is different from zero, or if the confidence interval excludes zero, then this is a sign that the 2SLS estimate is robust to a small departure from the strict exogeneity assumption—one wherein the IV is only plausibly but not strictly exogenous.

Dealing with Imperfect IVs

This is not a cure for a bad IV, and no amount of using this method will turn a bad IV into a good one.

Moreover, for all its benefits, this method can involve a certain amount of arbitrary decisions when it comes to incorporating prior information.

The Empirical Content of Heteroskedasticity

Suppose you are estimating

$$y_{it} = \alpha + \beta x_{it} + \epsilon_{it}. \tag{13}$$

Under idea circumstances, $V(\epsilon_i | x_i) = \sigma^2$, i.e., we have constant error variance, or homoskedasticity. In most cases, however, it will be the case that $V(\epsilon_i | x_i) = \sigma_i^2$.

The Empirical Content of Heteroskedasticity

Now, when it comes to inference about β, we want homoskedasticity. To get it, we can use the Huber-sandwich-White correction for standard errors.

But it happens that heteroskedasticity has empirical content. For instance, in Bellemare et al. (2018), y is the logarithm of household income, and the treatment variable is participation in agricultural value chains via participation in contract farming.

The Empirical Content of Heteroskedasticity

So when we estimate

$$\sigma_i^2 = \delta + \theta x_{it} + \xi_{it}, \tag{14}$$

the dependent variable is now household-specific income variability—a variable that is of direct interest if you are interested in looking at second-order welfare effects.

The bottom line is this: Heteroskedasticity sometimes has useful empirical content, and thinking about it in those terms can lead to interesting research projects.

Using Bits and Pieces of Likelihood to Study Behavior

I did mention early in this class that there is an unspoken order in which we tackle problems in applied econometrics: First, we worry about internal validity.

Second, we worry about standard errors. Third, we worry about external validity. (The latter two might be interchangeable, depending on who you talk to.)

Finally, we may worry about getting the DGP right for the dependent variable.

Using Bits and Pieces of Likelihood to Study Behavior

Sometimes, it is possible to tackle things a bit backward by combining bits and pieces of likelihood to study some behavior.

That's what my coauthor and I did in my very first published article. In that article, we were interested in studying the marketing behavior (i.e., sales or purchases of a given commodity; in our application, livestock) of agricultural households in East Africa.

Using Bits and Pieces of Likelihood to Study Behavior

The problem we were studying is this: We wanted to study the determinants (*autre temps, autres moeurs*) of household marketing behavior, which is best studied by looking at a household's net sales, which are equal to

$$\text{Net Sales} = \text{Sales} - \text{Purchases} \tag{15}$$

for a given commodity. Obviously, net sales *NS* can span the entire real line—agricultural households can be net buyers, autarkic, or net buyers.

Using Bits and Pieces of Likelihood to Study Behavior

One way to study the problem, then is to regress NS on some covariates. But the way I saw it, the decision making process involved in whether to participate on the market as a net buyer, autarkic, or net seller was very different than the decision making process involved in how much to sell or how much to buy conditional on having decided to be a net seller or a net buyer.

(That selection hypothesis, by the way, is testable.)

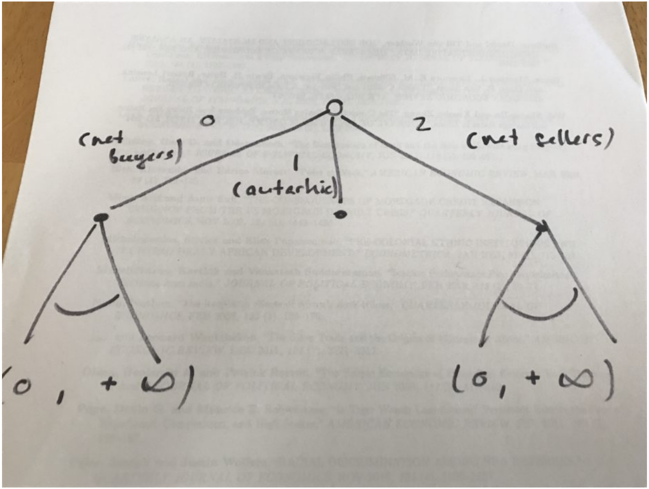
Using Bits and Pieces of Likelihood to Study Behavior

So I broke down the problem as follows:

1. In the first stage, a household decides whether it'll be a net seller, autarkic, or a net buyer. This is an ordered decision, with $NS < 0$, $NS = 0$, and $NS > 0$ mapping respectively into $y_1 = 0, 1, 2$.
2. In the second stage, conditional on y_1 being equal to 0 or 2, we respectively have sales $y_2 > 0$ and purchases $y_3 > 0$.

Note that there is selection taking place out of autarky ($y_1 = 1$) and into being a net seller or a net buyer, respectively. We called the end result an ordered tobit, but it should more accurately be seen as an ordered Heckit.

Using Bits and Pieces of Likelihood to Study Behavior



Using Bits and Pieces of Likelihood to Study Behavior

The likelihood function is a bit messy:

$$
\begin{aligned}
&\ell(\alpha', \beta', \sigma') \qquad (16)\\
= &\sum_{i=1}^{N} \left\{ \begin{array}{c}
I(y_{1i} = 0) \times \\
\left[\begin{array}{c}
\ln \Phi \left(\frac{\alpha_1 - x_{1i}\beta_1 + (y_{2i} - x_{2i}\beta_2)\rho_{12}/\sigma_2}{\sqrt{1-\rho_{12}^2}} \right) \\
-\frac{1}{2} \left(\frac{y_{2i} - x_{2i}\beta_2}{\sigma_2} \right) - \ln \left(\sqrt{2\pi}\sigma_2 \right)
\end{array} \right] \\
+ I(y_{1i} = 1) [\ln[\Phi(\alpha_2 - x_{1i}\beta_1) - \Phi(\alpha_1 - x_{1i}\beta_1)] \\
I(y_{1i} = 2) \times \\
+ \left[\begin{array}{c}
\ln \Phi \left(\frac{x_{1i}\beta_1 - \alpha_2 + (y_{3i} - x_{3i}\beta_3)\rho_{13}/\sigma_3}{\sqrt{1-\rho_{13}^2}} \right) \\
-\frac{1}{2} \left(\frac{y_{3i} - x_{3i}\beta_3}{\sigma_3} \right) - \ln \left(\sqrt{2\pi}\sigma_3 \right)
\end{array} \right]
\end{array} \right\}
\end{aligned}
$$

Using Bits and Pieces of Likelihood to Study Behavior

Obviously, as with any Heckit, you need to have an exclusion restriction to identify selection, which makes the use of the above procedure difficult. And you need to adjust the standard errors for the inclusion of the inverted Mills ratios in each second-stage equation. But if you are interested in studying decision making, combining likelihood functions can be an interesting way to do it.

In a recent article, Burke et al. (2015) add an extra layer of selection to the above model: Before asking whether a household is a net buyer or a net seller of a commodity, they add a stage modeling whether a household is a producer of that commodity.

One IV for Two Endogenous Variables

Earlier on in this deck of slides, I talked about Acharya et al.'s (2016) contribution, which allows testing whether a variable M is a mechanism for the treatment variable D.

One of the limitations of Acharya et al. (2016) is that it pretty much has to be the case that you have selection on observables in order to use their method—it is unclear whether their method applies to any other setup.

Dippel et al. (2017) have a similar contribution, but in the context of IV.

One IV for Two Endogenous Variables

Here's the idea: When a mediator variable M is a presumed mechanism whereby treatment variable D causes outcome y, it is possible to use a single instrument Z to estimate:

1. The total effect of D on y.
2. The indirect effect of D on y. This is the effect of D on y through M.
3. The direct effect of D on y. This is the effect of D on y net of the indirect effect.

One IV for Two Endogenous Variables

It has almost surely been drilled into your mind that every endogenous variable needs its own instrumental variable (IV).

How can Dippel et al. use the same IV for two endogenous variable?

Quite simply, it is possible to use the same IV for two endogenous variable when one of those endogenous variables is on the path (in a DAG sense) between the treatment and outcome variables.

One IV for Two Endogenous Variables

The three estimands above (total, indirect, direct effects) rely on three easy to perform 2SLS estimates, but Dippel et al.'s method relies on a crucial assumption, viz. that the unobserved confounders are "separable" between (i) those that affect the treatment and mediator variables, and (ii) those that affect the mediator and the outcome variables.

The nice thing here is that Dippel et al. provide the reader with a simple statistical test of that hypothesis, which only relies on the three estimands listed above.

When to Weight

When should you use sampling (or probability) weights? A good read on the topic is Solon et al.'s (2015) review in the *JHR*.

Suppose you oversample a specific group in order to get more precise estimates for that group. For instance, suppose you are interested in the opinion of LGBTQ students.

If you randomly sample individuals from a given population of students, you may not have enough LGBTQ respondents in your sample, and so whatever descriptive statistics you come up with for that sub-group might be too noisy.

When to Weight

Thus, you may wish to over-sample LGBTQ respondents in order to improve precision.

What I mean by this is that you would randomly sample respondents from each group–LGBTQ and non-LGBTQ–until you have the right number.

So if you target a sample size of n=100 and you'd like 50% respondents from each group, you split the population in two groups (assuming that's easy to do; in the case of LGBTQ students, it might not be easy to do) and sample from each until each group has 50 observations.

When to Weight

Here, sampling weights are easy to compute: population proportion divided by sample proportion.

So if your sample has 50% LGBTQ respondents and 50% non-LGBTQ respondents but the population has 10% LGBTQ respondents and 90% LGBTQ respondents, the weight on an LGBTQ observation is equal to $0.10/0.50 = 0.2$ and the weight on a non-LGBTQ observation is equal to $0.90/0.50 = 1.8$.

In a sample of $n = 100$, this means that the sample mean of the sampling weight is equal to $\frac{(0.2 \cdot 50 + 1.8 \cdot 50)}{100} = 1$. The mean of your sampling weight variable should be equal to one.

When to Weight

So when should you use sampling weights? Solon et al. divide empirical work in two rough categories, viz. descriptive statistics and causal inference.

For descriptive statistics, when you have a sample that is non-random because some groups were oversampled for precision as in my LGBTQ example, if you want to compute descriptive statistics for the entire population, you need to use sampling weights.

When to Weight

For causal inference, Solon et al. list three reasons you'd want to use sampling weights in your estimations:

- *Precision*: Weights can be used to correct for heteroskedasticity. Most students learn about this in their first econometrics class—this is the weighted least squares (WLS) estimator—but they soon forget about it once they learn about the White (1980) correction for heteroskedasticity. A recent article by Romano and Wolf purports to resurrect WLS. Here, Solon et al. suggest comparing OLS, WLS, and OLS with robust (either White or clustered) standard errors and discussing the differences in precision when conducting applied work.

When to Weight

- *Consistency*: If you have endogenous sampling (that is, if units of observation are selected on the basis of your outcome of interest; in Bellemare (2012), for instance, I selected units of observation on the basis of their choice of land-rental contract, which was my outcome of interest) you need to weight in order to get consistent estimates. There is a slight caveat in the Solon et al. article in cases where your model is correctly specified, but... when does that actually happen?

When to Weight

- *Identifying Average Partial Effects*: This is for cases where you're interested in a particular average of heterogeneous treatment effects. Since I have little to no experience doing this, I won't be discussing it beyond encouraging you to read that part if that's what you're interested in.

As always, there are exceptions to those rules, and Solon et al. encourage you to always "do both," show results with and without weights even in cases where they are undoubtedly necessary.

When (Not) to Cluster

This discussion is based off of Abadie et al. (2017), who start with two misconceptions about clustering:

1. Clustering matters only if the residuals and the regressors are both correlated within clusters, and
2. If clustering makes a difference in your standard errors, you should cluster.

When (Not) to Cluster

On 1, Abadie et al. show that even when the within-cluster correlation of the residuals and the within-cluster correlation of the regressors are both close to zero, clustering will matter.

What is important is the product of the within-cluster correlation of the residuals and the within-cluster correlation of the regressors.

If that correlation is nonzero, clustering matters. What this means is that cluster will make a difference—not that it is necessary.

When (Not) to Cluster

On 2, Abadie et al. show that in order to determine whether you should cluster, it's not sufficient to compare standard errors with and without clustering and see whether clustering makes a difference.

Rather, some additional information needs to be used, such as whether there are clusters in the population that have been left out of the sample due to sampling reasons.

When (Not) to Cluster

Abadie et al. recast clustering as a design problem. In some cases, it is a sampling design issue. In others, it is an experimental design issue.

Clustering is a sampling issue if sampling follows a two-stage strategy where clusters (e.g., villages) are first sampled at random and then observations within clusters (e.g., households) are then sampled at random.

When (Not) to Cluster

In this case, there are some (possibly many) clusters in the population which aren't included in the sample. Here, clustering is justified on the basis of the fact that some clusters in the population aren't included in the sample.

Clustering is an experimental design issue if the assignment to treatment is correlated within clusters, with the most obvious case being block randomization, when all the households (units) in a village (cluster) are either assigned to treatment or not.

When (Not) to Cluster

So when is clustering not necessary? When there is no clustering in the sampling (i.e., when you randomly select units from the whole population, without first randomly selecting clusters from which you will randomly select units) and there is no clustering in the assignment of treatment, or when there is no heterogeneity in the treatment effect and there is no clustering in the assignment of treatment.

Or if the sampling process is not clustered and the treatment assignment is not clustered, you should not cluster standard errors even if clustering changes your standard errors.

When (Not) to Cluster

Clustering will yield approximately correct standard errors in the following three possible cases.

First, when there is no heterogeneity in the treatment effect.
Second, when only few clusters are observed from the population.
And third, when there is only one unit sampled per cluster.

When (Not) to Cluster

The article also revisits the question of whether clustering is really necessary with fixed effects.

One comment I hear frequently from students (and even from some colleagues) is that with fixed effects, you shouldn't cluster standard errors at the level of the fixed effects. So for example, with state fixed effects, you shouldn't have to cluster standard errors at the state level.

Abadie et al. show that this is mistaken. Specifically, heterogeneity of the treatment effect (and really, when is a treatment effect not heterogeneous?) makes clustering necessary.