14.771 Development Economics: Microeconomic Issues and Policy Models
Fall 2008

# 14:771: Recitation Handout #4
Randomized Trials and IV, Probit

## A Quick Review of Probit/Binary LDV MLE

This is by no means comprehensive, but I wanted to be sure that everyone has at least some reference for probit for the upcoming problem set. I'll do the review in general terms so this can serve as a reference for all (binary) limited dependent variable (LDV) MLE models. We usually motivate these models by assuming that there is some latent dependent variable, $y_i^*$ which we cannot see:

$$y_i^* = \beta_0 + x_i'\beta - \varepsilon_i$$

The idea is that this latent dependent variable directly related to what we do see - which is just a binary indicator of something

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

Some examples are $y_i$ is employment status and $y_i^*$ is net utility from working, $y_i^*$ is how confused you are in lecture and $y_i$ is whether you show up for office hours, and so on - the idea is that once your latent index passes some threshold, you take a binary action.

Now, what we must do to make progress is assume that we know the distribution of $\varepsilon_i$. Specifically, assume that the CDF is given by the "link function" $\Lambda(.)$. So for probit, we assume that $\varepsilon_i \sim N(0, \sigma^2)$ and we use the CDF/link $\Phi(.)$. Now, let's try and find the pmf of $y_i$ :

$$\begin{aligned}
\Pr(y_i = 1) &= \Pr(y_i^* > 0) \\
&= \Pr(\beta_0 + x_i'\beta - \varepsilon_i > 0) \\
&= \Pr(\beta_0 + x_i'\beta > \varepsilon_i) \\
&= \Lambda(\beta_0 + x_i'\beta)
\end{aligned}$$

so for the probit case, note that in order to use the standard normal CDF, we have to divide $y_i^*$ by $\sigma$ (this is why you will sometimes here that probit coefficients are identified "up to scale" - what we estimate is always normalized by the variance of the errors). Back to the equation:

$$\begin{aligned}
\Pr(y_i = 1 \mid x_i) &= \Pr(\beta_0 + x_i'\beta > \varepsilon_i) \\
&= \Pr\left(\frac{\beta_0 + x_i'\beta}{\sigma} > \frac{\varepsilon_i}{\sigma}\right) \\
&= \Phi\left(\frac{\beta_0 + x_i'\beta}{\sigma}\right)
\end{aligned}$$

because $\frac{\varepsilon_i}{\sigma} \sim N(0,1)$. Similarly $\Pr(y_i = 0 \mid x_i) = 1 - \Phi\left(\frac{\beta_0 + x_i'\beta}{\sigma}\right)$. So now that we know this, we can write the pmf of $Y_i$ :

$$f(y_i \mid x_i, \theta) = \left\{ \begin{array}{cc} \Phi\left(\frac{\beta_0 + x_i'\beta}{\sigma}\right) & \text{if } y_i = 1 \\ 1 - \Phi\left(\frac{\beta_0 + x_i'\beta}{\sigma}\right) & \text{if } y_i = 0 \\ 0 & 0 \end{array} \right.$$

and now that we know this, we can set up a maximum likelihood estimator:

$$L\left(\frac{\beta_0}{\sigma}, \frac{\beta}{\sigma} \mid x\right) = \prod_{i=1}^{n} \left[ y_i \Phi\left(\frac{\beta_0 + x_i'\beta}{\sigma}\right) + (1 - y_i)\left(1 - \Phi\left(\frac{\beta_0 + x_i'\beta}{\sigma}\right)\right)\right]$$

If you think of what the FOC for the logged likelihood will look like, you should be able to see why we cannot identify $\sigma$ and $\beta_0, \beta$ separately. From here on out, things are standard MLE - you take logs, take the FOC, and find the solutions.

# The Fundamental Problem of Identification

All sciences (pure and social) that attempt to identify causal channels face the same fundamental problem: what is the *counterfactual*? What would have happened to person X if he had not been subject to the following treatment. Although this can normally be solved more easily in the context of laboratory experiments, an assumption is still made although rarely acknowledged. For example, let's say you are trying to find the effect of turning on a lightswitch in a dark room. You can run the same experiment many times and conclude that it generates light for the room. However, you still need to assume that it is not something else that I am doing at the same time as turning the lightswitch that is causing the light to appear (for those of you who are more philosophically inclined, I suggest reading Holland (1986) for a more detailed review of this argument).

Let's write down things a little bit more formally. Treatment status will be denoted $T_i = 1$ if treated, $T_i = 0$ if not. Then each person has a potential outcome:

$$\begin{aligned} Y_i(1) &= \text{i}^{\text{s}} \text{ outcome if } T_i = 1 \\ Y_i(0) &= \text{i}^{\text{s}} \text{ outcome if } T_i = 0 \end{aligned}$$

*So - what is the counterfactual for those individuals with $T_i = 1$? What is the counterfactual for those individuals with $T_i = 0$?*

Let us assume that we have a program that is affecting part of the population (we will call those treated) but not another share (called control). We are interested in knowing what is the effect of this particular program on a particular outcome variable that we will denote by Y. The average treatment effect (denoted by ATE) is then equal to:

$$ATE = E(Y_i(1) - Y_i(0))$$

In words, the average of the difference between what would be the outcome variable for the same individual if he was submitted to treatment and if he was. Sometimes, we might instead prefer to measure the average treatment on the treated (ATT) because what matters to us is what happens for those who are in the program rather than to the population in general. This is defined as:

$$\begin{aligned} ATT &= E(Y_i(1) - Y_i(0)\,|T_i = 1) \\ &= E(Y_i(1)\,|T_i = 1) - E(Y_i(0)\,|T_i = 1) \end{aligned}$$

However, this is never observable. *Why?* Take most lab experiences made in the context of physical sciences. They might subject two pieces of the same material to two different treatments and compare the outcome and claim that to be the differential effect of the treatment. However, it is not exactly the same material that is subject to the two treatments. Thus, we then need to assume that both materials would, at least on average, react in the same way to the treatment and the absence of such treatment for us to make such a claim, which is fairly easy to claim for most physical sciences.

In the context of social sciences, what we normally observe is something completely different. We normally let individuals self-select in the treatment group and thus what we observe is:

$$D = E(Y_i(1)\,|T_i = 1) - E(Y_i(0)\,|T_i = 0)$$

Adding and subtracting $E(Y_i(0)\,|T_i = 1)$, we can decompose D into two parts:

$$\begin{aligned} D &= E(Y_i(1)\,|T_i = 1) - E(Y_i(0)\,|T_i = 1) + E(Y_i(0)\,|T_i = 1) - E(Y_i(0)\,|T_i = 0) \\ &= ATT + selection\ bias \end{aligned}$$

The first element is what we are interested in while the second is a bias. All the methods we will review in the next couple of weeks are intended to disentangle both elements to isolate the "real" effect of the program. It is a good research habit to always try to guess what is the expected sign of the selection bias before using a method to control for it. Let us review a couple of classical cases and identify the expected sign of the bias:

1) Y is earnings, T is college attendance
2) Y is earnings, T is training program
3) Y is some health measure, T is a medical procedure

# Solutions to the Identification Problem

Many different solutions have been devised to solve the problem we have identified above. For this recitation, the ones related to randomized study settings will be highlighted.

## Pure randomized study

A fairly simple solution to the identification problem we presented above is if the treatment is purely randomized. If this is done perfectly, then by construction, the person who is treated is, on average, identical to that who is not treated. Thus, we have:

$$E(Y_i\left(0\right)|T_i = 1) = E\left(Y_i\left(0\right)|T_i = 0\right)$$

and there is no selection bias.

To obtain standard errors for our program effect D, a simple regression model can be estimated using OLS. What is the model we would need to estimate in this case?

$$Y_i = \alpha + \beta * 1(T_i = 1) + \varepsilon$$

This equation will provide the sample equivalent of the idealized equation for the ATT (see proof below). Also, if we randomized conditional on a list of covariates such as gender, geographical areas, age, etc, we could include those in this regression. We will find that this improves power - or decreases our standard errors. We could also verify whether the randomization was performed adequately by introducing such covariates and verifying that our estimate of $\hat{\beta}$ does not change significantly.

### Mathematical derivation

$$\begin{aligned}
\hat{\beta}_{OLS} &= (X'X)^{-1}(X'Y) \\
&= \begin{pmatrix} N & N_T \\ N_T & N_T \end{pmatrix}^{-1} \begin{pmatrix} \sum Y_i \\ \sum\limits_{i\in T} Y_i \end{pmatrix} \\
&= \frac{1}{NN_T - N_T^2} \begin{pmatrix} N_T & -N_T \\ -N_T & N \end{pmatrix} \begin{pmatrix} \sum Y_i \\ \sum\limits_{i\in T} Y_i \end{pmatrix} \\
&= \frac{1}{N_T(N-N_T)} \begin{pmatrix} N_T \sum Y_i - N_T \sum\limits_{i\in T} Y_i \\ N \sum\limits_{i\in T} Y_i - N_T \sum Y_i \end{pmatrix} \\
&= \frac{1}{N_T N_C} \begin{pmatrix} N_T \sum\limits_{i\in C} Y_i \\ (N_T + N_C) \sum\limits_{i\in T} Y_i - N_T \left( \sum\limits_{i\in T} Y_i + \sum\limits_{i\in C} Y_i \right) \end{pmatrix} \\
&= \begin{pmatrix} \dfrac{1}{N_C} \sum\limits_{i\in C} Y_i \\ \dfrac{1}{N_T N_C} \left( N_C \sum\limits_{i\in T} Y_i - N_T \sum\limits_{i\in C} Y_i \right) \end{pmatrix} \\
&= \begin{pmatrix} E\left(Y_i \widehat{|T_i} = 0\right) \\ E\left(Y_i \widehat{(1)}|T_i = 1\right) - E\left(Y_i \widehat{(0)}|T_i = 0\right) \end{pmatrix}
\end{aligned}$$

**How can we interpret it?**

*In this context is there a difference between the ATT and the ATE?*

The estimate we get might also provide little information for policy purpose. We are forcing the treated into the treatment. The effect on them might be extremely different than the one that would be observed on the people who would actually choose the program. Let's say we force individuals to take irons pills. We'll get the effect of giving iron pills on the entire sample. However, if the policy was actually implemented, it is highly probable that the individuals who have iron deficiency would be more likely to participate. Since the beneficial effect of iron would only be visible on anemic individuals, our estimate would be much lower than the actual benefit that would be given to the population that would participate in the actual program once implemented.

**Major problems with this approach**

- *Attrition*: People who have less benefits might be less likely to stay in touch with the program and this would bias us in the direction of a smaller effect (e.g. NIT experiments)
- *Supervision*: In this set-up, we need to ensure that all treated individuals participate. This can be very costly and time-consuming.
- *Mix-up in treatment and control groups*: Sometimes, maintaining the allocation to control and treatment to be random is almost impossible. Example: (Krueger 2000) evaluation of the Tennessee Star small class size

  experiment: children were moved to small classes (due to parental pressures, bad behavior, etc..).

# Using Randomized assignment as an instrument

For the reasons we have identified above, it is often very difficult to maintain the division between the treatment and control group such that we can simply compare the means of the two groups to obtain the estimate we would like to measure. Thus, a more common approach is to use the randomized assignment as an instrument. We will here review some basic principles of instrumental variables before presenting the particular elements of this strategy.

Here, we'll stick with binary instruments, because the interpretation (and math) is a lot cleaner. Then an instrument is something that we think influences the probability that someone gets treated. Specifically:

$$\begin{aligned} Z_i &= 1 \text{ means someone was exposed to the instrument} \\ Z_i &= 0 \text{ means someone was not exposed to the instrument} \end{aligned}$$

*Do you understand the difference between $Z$ and $T$? Which one is randomly assigned in most randomized trials?*

### ITT vs. TOT

If we randomize assignment but not treatment, we can then produce two different measures. The first one is called Intention to Treat (ITT). The formula is given by:

$$ITT = E\left(Y_i|Z_i = 1\right) - E\left(Y_i|Z_i = 0\right)$$

We therefore only look at the average among all those who were originally allocated to the treatment group and compare that to all those who were not offered access to the program. This is can be very telling if the actual number of participants is very high in the group randomly offered access and very low among those who were not offered access. This is also called the "reduced form".

The Treatment on the Treated (TOT) scales this measure up by how much the random assignment actually influenced program take-up:

$$TOT = \frac{E\left(Y_i|Z_i = 1\right) - E\left(Y_i|Z_i = 0\right)}{E\left(T_i|Z_i = 1\right) - E\left(T_i|Z_i = 0\right)}$$

### Wald Estimate and Binary Instruments

When we have an instrument that is binary, the instrumental variable formula is very simple and is called the "Wald estimate". Let's define our instrument as Z, the outcome of interest is Y and the endogenous covariate is X. The Wald estimate is given by:

$$\beta_W = \frac{E\left(Y_i|Z_i = 1\right) - E\left(Y_i|Z_i = 0\right)}{E\left(T_i|Z_i = 1\right) - E\left(T_i|Z_i = 0\right)}$$

The equivalence with this formula and the usual IV formula can be see below.

### Mathematical derivation

$$
\begin{aligned}
\beta_{IV} &= \left(Z'T\right)^{-1}\left(Z'Y\right) \\
&= \begin{pmatrix} N & \sum T_i \\ N_1 & \sum\limits_{i\in Z=1} T_i \end{pmatrix}^{-1} \begin{pmatrix} \sum Y_i \\ \sum\limits_{i\in Z=1} Y_i \end{pmatrix} \\
&= \frac{1}{N\sum\limits_{i\in Z=1} T_i - N_1\sum T_i} \begin{pmatrix} \sum\limits_{i\in Z=1} T_i & -\sum T_i \\ -N_1 & N \end{pmatrix} \begin{pmatrix} \sum Y_i \\ \sum\limits_{i\in Z=1} Y_i \end{pmatrix} \\
&= \frac{1}{N\sum\limits_{i\in Z=1} T_i - N_1\sum T_i} \begin{pmatrix} \sum\limits_{i\in Z=1} T_i \sum Y_i - \sum T_i \sum\limits_{i\in Z=1} Y_i \\ N\sum\limits_{i\in Z=1} Y_i - N_1\sum Y_i \end{pmatrix} \\
&= \frac{1}{(N_0+N_1)\sum\limits_{i\in Z=1} T_i - N_1\left(\sum\limits_{i\in Z=0} T_i + \sum\limits_{i\in Z=1} T_i\right)} \\
&\quad \begin{pmatrix} \sum\limits_{i\in Z=1} T_i \left(\sum\limits_{i\in Z=0} Y_i + \sum\limits_{i\in Z=1} Y_i\right) - \left(\sum\limits_{i\in Z=0} T_i + \sum\limits_{i\in Z=1} T_i\right)\sum\limits_{i\in Z=1} Y_i \\ (N_0+N_1)\sum\limits_{i\in Z=1} Y_i - N_1\left(\sum\limits_{i\in Z=0} Y_i + \sum\limits_{i\in Z=1} Y_i\right) \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{N_0\sum\limits_{i\in Z=1} T_i - N_1\sum\limits_{i\in Z=0} T_i} \begin{pmatrix} \sum\limits_{i\in Z=1} T_i \sum\limits_{i\in Z=0} Y_i - \sum\limits_{i\in Z=0} T_i \sum\limits_{i\in Z=1} Y_i \\ N_0\sum\limits_{i\in Z=1} Y_i - N_1\sum\limits_{i\in Z=0} Y_i \end{pmatrix} \\
&= \frac{1}{N_0N_1\left(E\left(T_i\widehat{|Z_i}=1\right) - E\left(T_i\widehat{|Z_i}=0\right)\right)} \times \\
&\quad \begin{pmatrix} N_0N_1\left(E\left(T_i\widehat{|Z_i}=1\right)E\left(Y_i\widehat{|Z_i}=0\right) - E\left(T_i\widehat{|Z_i}=0\right)E\left(Y_i\widehat{|Z_i}=1\right)\right) \\ N_0N_1\left(E\left(Y_i\widehat{|Z_i}=1\right) - E\left(Y_i\widehat{|Z_i}=0\right)\right) \end{pmatrix} \\
&= \begin{pmatrix} \frac{E(T_i\widehat{|Z_i}=1)E(Y_i\widehat{|Z_i}=0) - E(T_i\widehat{|Z_i}=0)E(Y_i\widehat{|Z_i}=1)}{E(T_i\widehat{|Z_i}=1) - E(T_i\widehat{|Z_i}=0)} \\ \frac{E(Y_i\widehat{|Z_i}=1) - E(Y_i\widehat{|Z_i}=0)}{E(T_i\widehat{|Z_i}=1) - E(T_i\widehat{|Z_i}=0)} \end{pmatrix}
\end{aligned}
$$

### Conditions for validity

Even if the program admissibility is randomized, the estimator we have just presented might not be valid. The two generals conditions for an IV instrument to be valid is that

$$
\begin{aligned}
E(Z'T) &\neq 0 \\
E\left(Z'\varepsilon\right) &= 0
\end{aligned}
$$

The first one indicates that we have a first stage. The instrument is somehow correlated with the endogenous variable in our model. You can look at the R-square and the F-test of that first regression. It must be high enough.
The second is the exclusion restriction: the only way that our instrument is influencing the outcome is through its effect on X. This is more difficult to test. If you have more instruments than endogenous variables, you can use overid tests and/or Hausman tests. However, these tend to have low power (accept the null that the instrument is valid too often). Also, they would reject the validity of good instruments if there are heterogenous effects and the instruments are triggering a response by different sub-groups.

### Interpretation

The interpretation of an IV variable was first explored by Imbens and Angrist (1994). They suggested that the IV estimate is the Local Average Treatment Effect (LATE), mathematically defined as:

$$E\left(Y_i\left(1\right) - Y_i\left(0\right) \left| T_i\left(1\right) \neq T_i\left(0\right)\right.\right)$$

Thus, the treatment effect is that for the individuals who are enticed to change their treatment category because of the instrument. For example, in the case of compulsory schooling laws as an instrument for schooling in an earnings regression, the returns to schooling that we measure is that for those who completed more years of schooling ONLY because the law compelled them to do so. The proof of this equivalence will be shown below. Sometimes, this is not the effect that we are interested in for policy purposes. Is it as much of a problem in the context of random intention to treat?

In order to identify LATE as a causal channel, we need to make 2 additional assumptions: independence and monotonicity. You'll see this in Esther's notes. They are

$$\begin{aligned}\text{Joint Independence} &: \quad \left(Y_i\left(1\right), Y_i\left(0\right), T_i\left(1\right), T_i\left(0\right)\right) \perp Z_i \\ \text{Monotonicity} &: \quad T_i\left(1\right) \geq T_i\left(0\right) \ \forall i\end{aligned}$$

Just to make this clear, note that $Y_i\left(1\right) = Y_i \mid T_i = 1$ while $T_i\left(1\right) = T_i \mid Z_i = 1$. And the same holds for $Y_i\left(0\right)$ and $T_i\left(0\right)$. Note that we still need to require that we have a first stage: $E\left[Z'T\right] \neq 0$.

The first condition states that potential outcomes and treatments $(Y_{ii}, Y_{0i}, X_{1i}, X_{0i})$ are jointly independent of Z$_i$. This means that Z$_i$ is randomly assigned (or as good as so). It cannot be that individuals with higher potential gains from the program are more likely to be treated. This is very close to the restriction exclusion we presented above.

Monotonicity is a little bit more precise. It says that either the instrument makes all individuals more likely to participate or less likely to participate. A good example of such a violation is Angrist and Evans (1998). They use sex composition of the first two children to instrument for the birth of a third child (to estimate the effect of fertility on earnings). For the monotonicity assumption to be satisfied, we would need that all parents are more likely to have a third child if they have a non-balanced gender composition of children (or all less likely). It is very possible that parents have different preferences for sex composition of their children and thus that this condition is violated in this case. Can you think of ways this condition is violated in the case of a randomized assignment?

Be sure you understand that joint independence bundes *both* random assignment and the exclusion restriction. *Can you think of examples where we have random assignment, but the exclusion restriction fails?*

### Mathematical derivation

Recall the treatment is also binary: T is either equal to 1 or to 0. Then, $Y_i = Y_i(0) + (Y_i(1) - Y_i(0)) T_i(1)$ if $Z_i = 1$ and $Y_i = Y_i(0) + (Y_i(1) - Y_i(0)) T_i(0)$ if $Z_i = 0$. Also assume wlog here that: $T_i(1) \geq T_i(0)$, ie that random assignment makes you more likely to join the program.

$$
\begin{aligned}
\beta_W &= \frac{E(Y_i|Z_i=1) - E(Y_i|Z_i=0)}{E(X_i|Z_i=1) - E(X_i|Z_i=0)} \\
&= \frac{E(Y_i(0) + (Y_i(1) - Y_i(0)) T_i(1) | Z_i = 1) - E(Y_i(0) + (Y_i(1) - Y_i(0)) T_i(0) | Z_i = 0)}{E(T_i|Z_i=1) - E(T_i|Z_i=0)}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{E(Y_i(0) + (Y_i(1) - Y_i(0)) T_i(1)) - E(Y_i(0) + (Y_i(1) - Y_i(0)) T_i(0))}{E(T_i|Z_i=1) - E(T_i|Z_i=0)}, \text{ by independence} \\
&= \frac{E((Y_i(1) - Y_i(0))(T_i(1) - T_i(0)))}{E(T_i|Z_i=1) - E(T_i|Z_i=0)} \\
&= \frac{\begin{array}{c} 1 * E(Y_i(1) - Y_i(0) | T_i(1) - T_i(0) = 1) * \Pr(T_i(1) - T_i(0) = 1) + \\ 0 * E(Y_i(1) - Y_i(0) | T_i(1) - T_i(0) = 0) * \Pr(T_i(1) - T_i(0) = 0) + \\ -1 * E(Y_i(1) - Y_i(0) | T_i(1) - T_i(0) = -1) * \Pr(T_i(1) - T_i(0) = -1) \end{array}}{E(T_i|Z_i=1) - E(T_i|Z_i=0)} \\
&= \frac{\begin{array}{c} E(Y_i(1) - Y_i(0) | T_i(1) - T_i(0) = 1) * \Pr(T_i(1) - T_i(0) = 1) - \\ E(Y_i(1) - Y_i(0) | T_i(1) - T_i(0) = -1) * \Pr(T_i(1) - T_i(0) = -1) \end{array}}{E(T_i|Z_i=1) - E(T_i|Z_i=0)} \\
&= \frac{E(Y_i(1) - Y_i(0) | T_i(1) - T_i(0) = 1) * \Pr(T_i(1) - T_i(0) = 1)}{E(T_i|Z_i=1) - E(T_i|Z_i=0)}, \text{ by monotonicity} \\
&= \frac{E(Y_i(1) - Y_i(0) | T_i(1) - T_i(0) = 1) * E(T_i(1) - T_i(0))}{E(T_i|Z_i=1) - E(T_i|Z_i=0)} \\
&= \frac{E(Y_i(1) - Y_i(0) | T_i(1) - T_i(0) = 1) * [E(T_i|Z_i=1) - E(T_i|Z_i=0)]}{E(T_i|Z_i=1) - E(T_i|Z_i=0)}, \text{ by independence} \\
&= E(Y_i(1) - Y_i(0) | T_i(1) - T_i(0) = 1)
\end{aligned}
$$

**LATE VS ATT** If treatment is only offered to the treatment group and is not offered to the control group (and there is no contamination in the groups), we can then prove that LATE=ATT. This is because $T_i(0) = 0$ $\forall i$. Thus using the last line of our derivation, we have that:

$$\begin{aligned}\beta_W &= E\left(Y_i(1) - Y_i(0)\,|T_i(1) - T_i(0) = 1\right)\\ &= E\left(Y_i(1) - Y_i(0)\,|T_i(1) = 1\right)\\ &= ATT\end{aligned}$$

**Is random assignment always a good instrument?**

Although random assignment is often thought of as the perfect instrument, we have to be careful. Some random assignment might not at all satisfy the conditions we have highlighted above.

Angrist (1990) uses the Vietnam era draft lottery to measure the effect of serving in the army on later earnings. Is this a good instrument? An individual who gets a high lottery number may make him more likely to be drafted but it may also make him more likely to go to college in order to avoid the draft. Which of the conditions highlighted above does this violates?

If program participation has externalities, then being part of a group that receives the treatment might have outcome effects that are separate from receiving the treatment. This particular type of problem will be the focus of Kremer and Miguel (2004), the required reading for next week.

## Problems of Randomized Experiments

- Financial costs: experiments are very costly and difficult to implement properly. Very often this cost makes them either poorly managed or too small.
- Ethical problems: we cannot always implement the type of experiment we are mostly interested in if this has a very serious effect on either the treated or the control group. NGOs and governments are often reluctant to deprive the controls from treatment which they consider valuable. Often, these entities are convinced by a staggered option whereby the program is introduced progressively and the late adopters are serving as a control group for those who get treated first.
- Non-response bias: people may move during the experiment and this may be related to the treatment/outcome (for example, see Hausman and Wise (1979))
- Limited duration: experiments are in general temporary. The effect that we capture may be very different than what we would observe once the program becomes permanent.
- Geographical/Demographic specificity: It is not clear that the results obtain in one setting apply elsewhere.
- Hawthorne and John Henry effects: People may behave differently because they are observed (both control and treatment).

- General equilibrium effects: small scale experiments do not generate equilibrium effects that might become important once the policy is implemented in the entire economy.

- Low power: due to cost, sample size might be too small to reject the null of no effect. Similarly, correlation within groups can require us to increase the size of our sample.

### Problems of IV

- IV can be very biased if we have even a small violation of the exclusion restriction

- LATE may be of little value for policy purposes

- Specification search and publication bias: papers with T-stats above 2 are more likely to be published. IV have larger standard errors than OLS, therefore they also need larger point estimates to be significant. Reported IV will therefore have a natural tendency to be "too high". Ashenfelter, Harmon and Oosterbeek (1999) explain why IV returns to education tend to be higher than OLS using this argument.

## Selecting Sample size

Let's assume we now know we would like to run a randomized evaluation. How do you do it? One of the most difficult question is how do you pick the size of your treatment and control group(s)? This is a short introduction to the theory of sample size selection.

### Basic Theory

A randomized trial tries to establish whether a null hypothesis can be rejected. Define the following terms:

- $\alpha$ : Probability of rejecting $\mathrm{H}_0$ when it is true (significance level)-Set to 5%

- $1-\alpha$ : Confidence level

- $\beta$ : Probability of not rejecting $\mathrm{H}_0$ when it is false

- $1-\beta$ : Power-Set by the experimentor

  The power is thus the probability of correctly rejecting our null hypothesis. By increasing the sample size, we can increase the power of our experiment and thus reduce the probability of obtaining inconclusive results. However, increasing the sample size is costly and so we usually set a fairly conservative power of 80%.

Notice also that there is a trade-off between $\alpha$ and $\beta$. The higher the confidence level of the test I want to perform, the lower the power.

## Simple Formula

Consider testing the hypothesis that the program effect is larger than $\theta_0$ using the conventional normal approximation test where one rejects the null if

$$\bar{X} < \theta_0 + \frac{z_\alpha \sigma}{\sqrt{n}}$$

The power of a test that rejects to detect a given effect size ($\theta_0$) when the real effect size is $\theta$ is given by:

$$\begin{aligned}
\beta(b) &= \Pr\left(H_0 \; not\ rejected | \; \theta\right) \\
\beta(b) &= \Pr\left(\bar{X} < \theta_0 + \frac{z_\alpha \sigma}{\sqrt{n}} \middle| \theta\right) \\
\beta(b) &= \Pr\left(\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} < \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} + z_\alpha \middle| \theta\right) \\
\beta(b) &= \Phi\left(\frac{\theta_0 - \theta}{\sigma/\sqrt{n}} + z_\alpha\right)
\end{aligned}$$

What's the power of a test when the null is true? What happens as $\theta_0$ becomes larger than $\theta$?

We can then collect data from a pre-survey and try to figure out the values of $\sigma$ and the desired effect size $\theta_0 - \theta$ and compute the sample size required for a power of 80%. Sometimes, the sample size is fixed by feasibility constraint (there aren't enough people in the area where you want to study) and then this computation allows you to know how big your program effect must be to achieve a power of 80%.

## Clustered design

Often, randomized trials are not done at the individual level but at the group level. Why is that?
How does that change the size of the sample required? Does it require more or less individuals? Does the number of individuals matter or the number of groups is the major element?

Formally, the formula to compute the sample size is given by:

$$k = \frac{(z_\alpha + z_\beta)^2 * 2\sigma^2 * (1 + (m-1)\rho)}{m(\theta_0 - \theta)}$$

where k is the number of clusters and m the size of each cluster. What happens when m increases? Does k increases or decreases? By less or more?
The key parameter to be estimated here is the intra-cluster correlation parameter $\rho$. This is usually difficult to compute and bounds are often given rather than real estimates.

## Simulations

Practitioners often find these formulas to be complicated and cumbersome. Although Stata has a command samplesi, it is often more suited for simple design. Many simply simulate using a data generating process and the input parameters defined above and repeat the simulation 1000 times for a given sample size. They tend compute the power as the number of times they rejected the null when it was false. Repeating this with various sample sizes will give you the minimum sample size necessary to achieve a given power.