Econ 1123: Section 8
Elena Llaudet

November 3, 2010

Experimental Data vs. Observational Data

Let's assume that you want to estimate the effect of a treatment on an outcome, for example, imagine that you want to estimate the effect of a drug on people's health.

Assuming that you had the following observational data on a particular population:

- health status: a continuous variable from 0 to 5 determined by a doctor
- drug: a dummy indicating whether the individual consumed the drug in the last 5 years
- age
- exercise: a dummy indicating whether the individual exercises periodically, and
- diet: a dummy indicating whether the individual has a healthy diet

How would you attempt to estimate the effect of the drug on health?

We could run the following regression:

Health = β₀ + β₁ drug + β₂ age + β₃ exercise + β₄ diet

What would then be the estimate of the treatment effect? Treatment Effect Estimate = β₁

β₁ = E [ health | drug = 1 ] - E [ health | drug = 0 ]

What would be the major threat to the internal validity of your treatment effect estimate? Omitted Variable Bias. For example, it could very well be that only individuals with very poor health to begin with were inclined to take the drug.
How would you design an experiment to measure the effect of the drug on people’s health? We would randomly select subjects from a population of interest, then, we would randomly assign them either to a treatment group, which receives the drug, or to a control group, which does not receive the drug.

What would then be the estimate of the treatment effect? Treatment Effect Estimate (TEE) = average health of those individuals in the treatment group - average health of those individuals in the control group

\[ TEE = E[\text{health | drug} = 1] - E[\text{health | drug} = 0]. \]

Which regression would you run with the experimental data if you wanted to estimate the treatment effect with OLS?

\[ \text{Health} = \beta_0 + \beta_1 \text{drug} \]

\[ TEE = \hat{\beta}_1 = E[\text{health | drug} = 1] - E[\text{health | drug} = 0] \]

This is called the differences estimator.

What is the major advantage of your experimental data estimate vs. your observational data estimate? Generally speaking, our experimental data estimates are a lot less likely to suffer from omitted variable bias because the treatment was assigned at random. If all worked well, receiving the drug should not be correlated with any of the background characteristics of the individuals that would be determinants of their health.

Analysis of Experimental Data with Controls

If you had also collected data on other characteristics of the populations (such as age, exercise, and diet), would there be any reason for you to include those in the regression? Yes.

There are usually three reasons for us to include additional regressors in the regression (two of them would apply in this case):

1. Improve the Precision/Efficiency of the estimates. Adding the additional regressors makes our estimator more efficient. That is, it makes the standard errors smaller.

The reason is that including the additional determinants of \( Y \) reduces the variance of the error term.

2. Try to control for failed randomization. Randomization might have failed, in which case there is potential for omitted variable bias again. If the probability of being assigned to the treatment group depends only on the observable variables \( W \), then controlling for \( W \) will adjust for this nonrandom assignment.

However, if the probability of being assigned to the treatment group depends on unobserved variables as well, then the adjustment made by including the \( W \)'s would be incomplete.

3. Adjust for conditional randomization. (Not applicable in this case)
Randomization

If randomization worked, what would you expect \( \hat{\pi}_1 \) to be in the following regressions?

1. \( \text{age} = \pi_0 + \pi_1 \text{ drug} \) ? We would expect \( \hat{\pi}_1 = 0. \)
2. \( \text{exercise} = \pi_0 + \pi_1 \text{ drug} \) ? We would expect \( \hat{\pi}_1 = 0. \)
3. \( \text{diet} = \pi_0 + \pi_1 \text{ drug} \) ? We would expect \( \hat{\pi}_1 = 0. \)

If randomization worked, and we ran:

\[
\text{drug} = \delta_0 + \delta_1 \text{age} + \delta_2 \text{exercise} + \delta_3 \text{diet}
\]

Do you think that we would reject the joint hypothesis: \( H_0: \delta_1 = \delta_2 = \delta_3 = 0 \) (using an F-test)? No.

How can we test for whether randomization worked based on our observed characteristics?

We could do one of two things:

- (1) Run as many regressions as observed characteristics of the type:
  \[
  W_1 = \pi_0 + \pi_1 X, \\
  W_2 = \pi_0 + \pi_1 X, \\
  \ldots
  \]
  where \( W \) are the observed characteristics and \( X \) is the dummy indicating treatment.
  If randomization worked, all of the \( \hat{\pi}_1 \) should be not statistically significant.

- (2) Run one regression of the following type:
  \[
  X = \delta_0 + \delta_1 W_1 + \delta_2 W_2 + \ldots
  \]
  If randomization worked, we would fail to reject the joint hypothesis that all of the \( \delta \) coefficients are zero.

Can we always know whether randomization failed? No.

If randomization worked, the two groups - treatment and control - should be composed of individuals with very similar characteristics. In other words, the probability of being assigned to the treatment should not be correlated with any background characteristics of the individuals.

Because we only have data on a subset of those background characteristics (in this case, age, exercise, and diet), we can only test whether randomization failed to produced two similar groups based on these three variables. There still exists the possibility that randomization failed to produced two similar groups based on other characteristics that we do not observe (which would bias our results, if those variables are determinants of \( Y \)).
Potential Threats to Internal Validity

What are some of the potential threats to the internal validity of analyses of experimental data?

- Randomization fails. Potential omitted variable bias.
- Some individuals do not follow protocol. We have compliance issues.

For example, in our case, this would happen if some of the individuals assigned to take the drug, decide not to take it, and if some of the individuals assigned to not take the drug, take it. Potential omitted variable bias. One solution is to use the randomization assignment as the instrument in an IV analysis.

Potential Threats to External Validity

What are some of the potential threats to the external validity of analyses of experimental data?

- Non-representative sample of the population of interest
- Non-representative program or policy
- General equilibrium effects. An internally valid small experiment might correctly measure a causal effect, holding constant the market or policy environment, but general equilibrium effect mean that these other factors are not, in fact, held constant when the program is implemented broadly.

- Some individuals leave our study. We have a problem of attrition. This will bias our results if the reason for dropping out of the study is related to the treatment itself. For example, in our case, this would happen if some of our individuals get so sick after taking the drug that they need to abandon the experiment.
- Experimental effects or Hawthorne effects. This happens when the mere fact of being part of a study (or knowing to be receiving the treatment or not to be receiving the treatment) changes the subjects’ behavior. For example, in our case, if those in the control group know that they are not taking the drug, they might be inclined to feel sicker. Double-blind protocol might mitigate this effect. In our case, if neither the doctors nor the patients know who is taking the treatment, we would mitigate this threat.

- Treatment vs. eligibility effects. Participation in an actual (nonexperimental) program is usually voluntary. An experimental study that measures the effect of the program on randomly selected members of the population will not, in general, provide an unbiased estimator of the program effect when the recipients of the actual implemented program are permitted to decide whether to participate.
The Differences-in-Differences Estimator

If in our example we had data not only on the health of the individuals after the drug was taken but we also had data on the health of the individuals before the drug was taken, we could also estimate the effect of the treatment using the diffs-in-diffs estimator:

\[(\text{health}_{\text{after}}) - (\text{health}_{\text{before}}) = \beta_0 + \beta_1 \text{drug}\]

\[\Delta \text{health} = \beta_0 + \beta_1 \text{drug}\]

where \(\beta_1\) is the diffs-in-diffs estimator.

\[\beta_{\text{diffs-in-diffs}} = \frac{(\text{health}^{\text{drug}=1, \text{after}} - \text{health}^{\text{drug}=1, \text{before}})}{(\text{health}^{\text{drug}=0, \text{after}} - \text{health}^{\text{drug}=0, \text{before}})}\]

Two potential advantages of the diffs-in-diffs estimator:

1. Efficiency/Precision: The diffs-in-diffs estimator can be more efficient than the differences estimator.
2. Eliminate the pretreatment differences in \(Y\): If treatment is correlated with the initial level of \(Y\), then the differences estimator is biased but the diffs-in-diffs is not.

Note that as before, we could also use the differences-in-differences estimator with additional regressors to either (a) improve efficiency, (b) check for randomization, or (c) adjust for conditional randomization.