

## Estimating Causal Effects Using Randomized Experiments

### **Randomized Controlled Experiments**

Suppose we wanted to estimate the causal effect of a variable  $X$  on an outcome variable  $y$ . If you could choose the source of variation in  $X$  your data how would you do it?

The answer is that there are good reasons why you would want to allocate  $X$  at random i.e. to give treatment to all sample members with equal probability<sup>1</sup>. This is what is known as a randomised controlled experiment. Evidence from randomised controlled experiments are sometimes referred to as the ‘gold standard’ - the reason for this is that random assignment avoids the problems of identifying causal effects (many of which have been discussed earlier in the course). By design your  $X$  will be independent of any other influences on  $y$ , whether observed or unobserved.

How would you estimate the causal effect of  $X$  on  $y$  with experimental data? To keep things simple, assume that  $X$  is binary (i.e. can only take the values 0 or 1) though nothing is much more complicated if  $X$  can take on more values or even if it is continuous. We will refer to those sample members for whom  $X=1$  as being in the ‘treatment’ group and those for whom  $X=0$  as being in the ‘control’ group. The causal effect is sometimes referred to as the treatment effect in line with this terminology.

### **Estimating Causal Effects in Randomized Controlled Experiments**

If we are just interested in the causal effect of  $X$  on the mean of  $y$  then we would like to have a good estimate of:

$$E(y_i | X_i = 1) - E(y_i | X_i = 0) \quad (1)$$

Of course,  $X$  might not just have an effect on the mean of  $y$ , it might affect the whole distribution. In some situations this might be something we are interested in, in others a possibility that we need to have in the back of our mind.

How can we estimate the causal effect in (1). One estimator would be to simply take the sample equivalent of the population moments and to estimate the causal effect as the difference in the sample means for those in the treatment and control groups.

Denote these means by  $\mu_1$  and  $\mu_0$ . One way of writing this is:

$$\mu_1 - \mu_0 = \frac{\sum X_i y_i}{\sum X_i} - \frac{\sum (1 - X_i) y_i}{\sum (1 - X_i)} \quad (2)$$

There is no real problem in computing the estimate of the causal effect in this way, but it is useful to derive it via a regression. Suppose we run the regression:

$$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (3)$$

---

<sup>1</sup> Actually you may want to do (or cannot avoid doing) conditional randomisation in which treatment is randomised conditional on some observable variables but is different for people with different values of those variables.

The OLS estimate of  $\beta_0$  will be the mean of the value of  $y$  for those in the control group and the estimate of  $\beta_1$  will be the difference in the mean of the value of  $y$  between the control and treatment group i.e. the causal effect of  $X$  on the mean of  $y$ . This is what Stock and Watson call the ‘differences’ estimator (p376).

The regression method for deriving the causal effect has the advantage that it naturally generalizes to the case where  $X$  is not binary as one can simply run a regression of  $y$  on  $X$  whatever the form of  $X$  (though there is an issue about the functional form of the relationship between  $y$  and  $X$  e.g. is it linear?).

And the regression output will also give you an estimate of the standard error on the coefficient on  $X$  which can be used to test hypotheses about the causal effect. Let us reflect for a minute on the way in which that standard error is computed.

Unless instructed otherwise your computer package is going to compute the standard errors of the coefficients based on the assumption that the error  $\varepsilon$  in (3) is homoskedastic. What this means is that:

$$Var(y_i | X_i = 1) = Var(y_i | X_i = 0) = Var(\varepsilon_i) \quad (4)$$

The equality on the left-hand side is the assumption that the treatment has no effect on the variance of  $y$  even though it may have an effect on the mean. That is a strong assumption to make and typically there will be no very good reason to believe it. If there is an effect of the treatment on the variance of  $y$  then this will imply that  $\varepsilon$  is heteroskedastic. You should know from basic econometrics that this does not affect the consistency of the OLS estimator as long as  $\varepsilon$  is mean-independent of  $X$  which it will be because of the randomisation assumption. But it does mean that the estimate of the standard errors will be inconsistent so we would like to have a consistent estimate.

There are a number of ways to derive consistent estimates of the standard errors. I will describe 3.

1. Get an estimate of  $Var(y_i | X_i = 1)$  from the data. This will just be the sample variance for the treatment group probably with a degrees of freedom correction. Do the same for the control group and then the variance of the difference in the group means will be:

$$Var(\mu_1 - \mu_0) = Var(\mu_1) + Var(\mu_0) \quad (5)$$

as the two samples are independent of each other. If you have done a course in statistics then this is the way in which one tests for the equality of means in two samples.

2. A regression way of doing this would be to estimate two separate regressions, one for the treatment group and one for the control group. The intercepts in these regressions will be the means for the treatment and control groups and the regression output will also give standard errors for these estimates – they will actually just be the sample variances as referred to above. Then do the hypothesis test as above.

Both of these methods exploit the fact that  $X$  is binary and we are only interested in how the variance of  $\varepsilon$  varies with  $X$ . They cannot be very easily generalised to other

applications. But there is a regression-based method that is of very wide application – what was described to you last term as White standard errors. They are often referred to by different names – Huber standard errors, robust standard errors, heteroskedastic-consistent standard errors – but the same thing is always meant.

The robust standard errors are computed in the following way. The OLS estimator can be written as:

$$\hat{\beta} = (X'X)^{-1} X'y = \beta + (X'X)^{-1} X'u \quad (6)$$

Hence we have that the asymptotic variance of  $\hat{\beta}$  can be written as:

$$\begin{aligned} p \lim \left[ (\hat{\beta} - \beta)(\hat{\beta} - \beta)' \right] &= p \lim (X'X)^{-1} X'uu'X (X'X)^{-1} \\ &= p \lim (X'X)^{-1} p \lim (X'uu'X) p \lim (X'X)^{-1} \end{aligned} \quad (7)$$

The estimate of this asymptotic variance then replace u with the residuals from the equation, sometimes with a small sample correction. If X is one-dimensional one can write the robust estimate of the variance of  $\hat{\beta}$  as:

$$\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{1}{n} \cdot \frac{\frac{1}{n-2} \sum (X_i - \bar{X})^2 \hat{u}_i^2}{\left[ \frac{1}{n-1} \sum (X_i - \bar{X})^2 \right]^2} \quad (8)$$

If the errors are homoskedastic i.e. their variance is independent of X this will reduce to the usual formula. Stock and Watson have a good discussion of these issues.

This method is used very widely so a package like STATA has an option in its regress command (and most other commands) to produce these robust standard errors – simply type

```
. regress y x, robust
```

What you should gather from this discussion is that econometrics would be very easy if all data was from randomised controlled experiments – one can get consistent estimates of causal effects simply by a comparison of means – no need for any matrix algebra or even multiple regression, or even collection of any variables other than treatment status and the outcome variable. However there are good reasons why data on additional variables can be useful.

### Additional Regressors

The regressions I have suggested so far do not include any variables other than X. Note that there is no assumption here that these other omitted variables do not matter for the determination of y just that the omission of variables that do influence y from our regression does not prevent us from obtaining consistent estimates of the causal effect of X on y. The reason is the randomisation that ensures that X is independent of all other variables so that a problem of omitted variable bias does not arise.

Does this mean there is no point in collecting and/or using information on other variables that affect y? The answer is that while other data is not necessary to obtain

consistent estimates of the causal effect, there are reasons why it may be useful. The discussion here follows Stock and Watson (p384). I will use their notation as well.

So, continue to denote the treatment variable by X but use the vector W to denote other variables that potentially influence y. Note that we are only going to be interested in the causal effect of X on y so we do not have to worry about whether any observed correlation between y and W is causal or spurious.

### *Improved Efficiency*

An important advantage of including other regressors is that it will typically reduce the standard error of the estimated treatment effect i.e. there is a gain in precision and the estimate will be more efficient. If we think of the error in (1) as homoskedastic (for simplicity) then the variance of the OLS estimate of the treatment effect is going to be, using the formula  $\sigma^2 (Z'Z)^{-1}$ , where  $\sigma$  is the standard error of the regression and Z is the vector of all the regressors. Now:

$$(Z'Z) = \begin{pmatrix} X'X & X'W \\ W'X & W'W \end{pmatrix} \quad (9)$$

but, because the treatment is randomized: we have that:

$$p \lim (Z'Z) = \begin{pmatrix} \Sigma_{xx} & 0 \\ 0 & \Sigma_{ww} \end{pmatrix} \quad (10)$$

so that:

$$p \lim (Z'Z)^{-1} = \begin{pmatrix} \Sigma_{xx}^{-1} & 0 \\ 0 & \Sigma_{ww}^{-1} \end{pmatrix} \quad (11)$$

so that:

$$p \lim Var \hat{\beta} = \sigma^2 \Sigma_{xx}^{-1} \quad (12)$$

which will fall as  $\sigma$  falls.

### *Check for randomization*

In many real world examples of randomised experiments there are serious questions about how well the randomisation was implemented in practice. One way of allaying these fears and to give your results added credibility is to check that the treatment variable is uncorrelated with the W variables (though there may be other important omitted variables remaining). There are a number of ways in which one might do this.

- compare the estimate of the treatment effect when including and excluding the W variables. If randomization has been carried out well then these will be consistent estimates of the same parameter- if there is a problem then one estimate will be suffering from omitted variables bias.
- Compare the values of X in treatment and control groups
- Run a regression of X on the W variables and test the hypothesis that the coefficients are zero (could do probit model)

Sometimes a failure of randomisation is just bad luck – the omitted variable bias is then random, sometimes leading to an over-estimate of the true causal effect and sometimes an under-estimate. In this situation the gain from including W is just the gain in precision discussed previously, one removes this as a potential source of variation in one's estimates.

### *Adjust for Conditional Randomization*

Sometimes the probability of treatment can be different for different sub-groups. So far we have assumed that:

$$\Pr(X_i = 1|W_i) = \Pr(X_i = 1) \quad (13)$$

i.e. the probability of treatment is independent of W. Conditional randomisation refers to the case where the probability of treatment depends on W but is independent of any other variable.

There are several reasons why this might be the case.

First, necessity - it might only be feasible to do conditional randomisation when, in an ideal world, we would like to do unconditional randomisation. For example, in Project STAR pupils within each school were randomly assigned to one of the treatment or control groups so that the probability of treatment was the same for everyone within the same school. But, because each school only contains a small number of classes, there is a relationship between the number of classes in the school and the probability of treatment. And the size of school is correlated with all sorts of other variables that might affect outcomes. In this case treatment is only random conditional on the W variables.

Secondly, one might choose to do conditional randomisation. This because one might be able to get more precise estimates in this way. The reason is similar to the reason we sometimes use stratified rather than random sampling, although I don't want to go into this in detail here.

How can we estimate the treatment effect in this case? The first point is that as X is no longer independent of W the 'differences' estimator will suffer from omitted variables bias. But, if we control for W, X will be independent of any other omitted variables so we will get a true estimate of the causal effect. It is important here to get the functional form for the way in which W affects y right here – a misspecification will lead to an inconsistency in the estimate of the causal effect. Note that this is not the case for unconditional randomisation.

Matching (that you have considered earlier in the course) is best thought of as a method for dealing with these issues – getting good estimates of causal effects when one is unsure about the functional form of the relationship between y and W.

### **Heterogeneity in Treatment Effects**

So far I have assumed that the treatment effect is the same for everyone. But there is no very good reason to believe this – indeed, it is quite likely that some groups have larger treatment effects than others. If this is the case, then we might want to know about it, because it is of some intrinsic interest in itself and because if one wants to generalise from the sample under question to a population in which the distribution of W is quite possibly different then one would want to re-weight the treatment effects.

Let us start with the case in which we do not include any additional regressors in our estimation of the treatment effect but there is heterogeneity in the treatment effect. Modify (3) to:

$$y_i = \beta_0 + \beta_{1i}X_i + \varepsilon_i \quad (14)$$

where the notation indicates that the treatment effect can vary from individual to individual (check you understand why we do not do a similar change for the intercept). Because  $X_i$  is randomly assigned it will be independent of  $\beta_{1i}$  something we will use in a minute.

In an ideal world we would like to estimate the treatment effect for each individual in the sample. But this is not possible – we have N observations and (N+1) parameters (the N treatment effects and the intercept). One way in which one might think of doing this is to estimate a separate regression for each individual – but this is only going to be possible if we have two observations on each individual, one in a treatment group and one in a control group<sup>2</sup>.

So we are going to have to be content with providing summaries of the treatment effect. This raises the question of whose treatment effect we are interested in – when it is the same for everyone this is not a meaningful question but when there is heterogeneity it is. One obvious thing to be interested in is the average treatment effect (ATE) - the mean of the treatment effects across individuals in the sample i.e.:

$$ATE = E(\beta_{1i}) = \bar{\beta}_1 \quad (15)$$

How can we estimate the ATE? It turns out that if we simply run a regression of y on X then the coefficient on X will be a consistent estimate of the ATE (see Stock and Watson, p409). We know that:

$$p \lim \hat{\beta}_1 = \frac{Cov(X_i, y_i)}{Var(X_i)} = \frac{Cov(X_i, \beta_0 + \beta_{1i}X_i + \varepsilon_i)}{Var(X_i)} = \frac{Cov(X_i, \beta_{1i}X_i)}{Var(X_i)} = E(\beta_{1i}) \quad (16)$$

Another way to see this is to re-write (14) as:

$$\begin{aligned} y_i &= \beta_0 + \bar{\beta}_1 X_i + (\beta_{1i} - \bar{\beta}_1) X_i + \varepsilon_i \\ &= \beta_0 + \bar{\beta}_1 X_i + u_i \end{aligned} \quad (17)$$

Note that the composite error  $u_i$  is mean-independent of  $X_i$  so that the OLS estimate of the coefficient on X will be the ATE<sup>3</sup>. Also notice that the error in (17) can be thought of as being the original error plus a term which is the difference of the individual's treatment effect from the sample average. Note that this error will inevitably be heteroskedastic – indeed, one interpretation of the model with heteroskedasticity in which the variance of the error depends on X is that it is really a model of heterogeneous treatment effects in which the coefficient on X is the average treatment effect.

The model we have just discussed sometimes appears in econometrics textbooks under the name of the random coefficients model.

<sup>2</sup> One might think that multiple observations on the same individual over time when they might sometimes be in a treatment group and sometimes a control group would help here. Such data can be very useful but one has to make an assumption that limits the degree of heterogeneity in the treatment effect e.g. it may vary across individuals but is constant over time for a given individual.

<sup>3</sup> However X appears in the composite error so the distribution of u will not be independent of X – it will be heteroskedastic.

It was argued above that one can improve one's estimates of the treatment effects by including additional regressors,  $W$ . But once we recognise that treatment effects are heterogeneous we need to also recognize the possibility that treatment effects vary systematically with  $W$ . So let's modify our model to allow for this possibility.

I will introduce a notation that is sometimes called the full outcomes notation. Suppose that for individual  $i$  the outcome if they are in the control group can be written as:

$$y_{0i} = \gamma_0' W_i + u_{0i} \quad (18)$$

where  $\gamma_0' W_i = E(y_i | W_i, X_i = 0)$ . Similarly write the outcome for individual  $i$  if they are in the treatment group as:

$$y_{1i} = \gamma_1' W_i + u_{1i} \quad (19)$$

where  $\gamma_1' W_i = E(y_i | W_i, X_i = 1)$ .

Taking the difference between (19) and (18) we have that the treatment effect for individual  $i$  can be written as:

$$\beta_{1i} = (\gamma_1 - \gamma_0)' W_i + (u_{1i} - u_{0i}) \quad (20)$$

Note that this has both an observable and unobservable component.

How can we estimate the treatment effects in this case. I will describe two approaches.

1. Estimate separate regressions for treatment and control groups and then compare the coefficients.
2. Combine the treatment and control groups into a single regression. Note that we can write:

$$y_i = X_i y_{1i} + (1 - X_i) y_{0i} \quad (21)$$

so that, combining (18) and (19) we have that:

$$\begin{aligned} y_i &= X_i [\gamma_1' W_i + u_{1i}] + (1 - X_i) [\gamma_0' W_i + u_{0i}] \\ &= \gamma_0' W_i + (\gamma_1 - \gamma_0)' X_i W_i + u_{0i} + X_i (u_{1i} - u_{0i}) \end{aligned} \quad (22)$$

i.e. a regression that includes  $W$  and the interactions of  $W$  with  $X$ . The coefficients on these interaction variables can be interpreted as the observable part of the treatment effect. Also note that if there is any unobservable component then the error will be heteroskedastic so one might want to compute standard errors in a way that allows for this possibility.

### **Problems with Experiments** (Stock and Watson, pp377-382)

Social experiments involving random assignment of treatment seem very attractive. But there are problems, mostly of a practical nature.

### *Expense*

Randomized Controlled trials are often very expensive – project STAR cost \$12m - whereas non-experimental data is often available at little or no additional cost e.g. because the government collects the statistics for some other purpose.

### *Ethical Issues*

One of the reasons social experiments have been rare is that people often have ethical issues related to some people receiving treatment and others not. Some of these objections seem rather odd to me – we accept randomised trials for the evaluation of the effectiveness of drugs even though whether you are in the treatment or control group might literally be a matter of life or death. And resistance to randomised experiments does seem to be falling over time as they are becoming much more common.

But the combination of cost and ethical issues means that experiments are rare and often small in scale. We simply would not have enough work to keep us busy if we only allowed ourselves to work with experimental data. And the small scale means that the estimates we often get from experimental data are not as precise as we would like so that the advances in knowledge are not that great.

But there also often problems with the experiments that are implemented. Stock and Watson usefully divide these into threats to internal validity and threats to external validity.

### *Threats to Internal Validity*

#### Failure to Follow Experiment

Sometimes randomisation fails because the researcher needs the help of people on the ground to implement the experiment. These people may not be as persuaded as to the virtues of experimentation as the researcher and may be under pressure from those who are in the control group to put them in the treatment group (pushy parents who want their child to be in the small class). In addition it may be very difficult to actually persuade some people in the treatment group to follow the treatment. Also experiments that last a long time may suffer from attrition that, if it is not random, will cause bias in the results.

#### Experimental Effects (the Hawthorne Effect)

The excitement of being in an experiment – being ‘watched’ - might bring forth greater effort that would not normally be present.

### *Threats to External Validity*

These problems relate to whether the conclusions drawn from experiments can be generalised outside the experimental situation. There are a number of reasons why this might be problematic.

#### Nonrepresentative Sample

Experimental data can only provide estimates of average treatment effects for those in the experiment. One cannot necessarily generalise the conclusions to other populations. Obviously one can if the treatment effect is the same for everyone and, even if it is not, then one can estimate the treatment effect conditional on  $W$  and then re-weight for the new population. To see this suppose that in our experiment the

distribution of  $W$  is given by  $f_e(W)$  and the distribution in the population is given by  $f_p(W)$ . Furthermore suppose the average treatment effect conditional on  $W$  is given by  $\beta(W)$ . Then the overall average treatment effect in the experimental sample will not be the same as for the population, but one can get an estimate if one estimates the treatment effect conditional on  $W$ . But as one slices the data finer and finer ones precision is going to fall.

### Nonrepresentative Programme

All social programmes are different and there is inevitably something of a leap of faith in assuming that a particular programme will have the same impact in other places at other times when the context might be very different. For example a programme to help the unemployed find work might have very different effects in a boom and recession, in a depressed or booming area.

### General Equilibrium Effects

If some intervention is found to be successful in an experiment then the intention is often to expand its scale e.g. to go nationwide with it. But the scale may have consequences for the effects. One prominent reasons for this is that there may be general equilibrium effects or the programme may have effects on the control group as well as the treatment group.

For example consider a programme that offers incentives to the unemployed to get back into work. If we see that those in the treatment group do get back into work faster than those in the control group then we might be tempted to conclude that the programme ‘works’ and offer these incentives to all the unemployed. But perhaps the number of jobs available is fixed and who gets them is determined by who fights the hardest. In this case the model will be given by:

$$y = \beta_0 + \beta_1 X - \beta_1 \bar{X} + \varepsilon \quad (23)$$

One will find a treatment effect but the gain of the treatment effect is at the expense of the control group. If the programme is expanded to cover anyone one will find no change in the job-finding rate.

### Treatment vs. Eligibility Effects

Participation in many social programmes is often voluntary. Often we give people an opportunity and do not force them to do it. Sometimes we try to force them e.g. the unemployed may be threatened with the loss of their welfare benefits if they do not take a job training programme – they may turn up but it is harder to make them pay attention.

Whether this is a problem or not may depend on what it is you want to estimate. We may want to estimate the impact of an opportunity to do something – in this case an estimate based on ‘intention to treat’ will be good enough. But if we really want the impact of the actual programme then we will need to worry about the selection implied by the fact that not everyone in the treatment group actually takes the treatment.

The bottom line is that, while well-implemented randomised controlled trials do represent the ‘gold standard’ of research, many practical implementations have their problems and they are going to be rare for the foreseeable future. On many important questions we lack evidence from social experiments and will continue to do so for a long time, perhaps ever. So, like it or not, we are going to have to continue to work with non-experimental data.