

1 Treatment Effects

Last Updated: April 22, 2008

1.1 Introduction

We are now going to focus on the estimation of “treatment effects.” This is a literature that originated in the medical literature, but has recently been appropriated by economists for the task of program evaluation. The basic idea is that we want to identify how participation in some treatment will affect an outcome. These treatments can be anything: using a new drug, being vaccinated for worms, going to college, participating in a job training program, being tested for illegal drug use by an employer, etc. When we can randomly assign the treatment to individuals, life is good as the treatment effects are easily identified. However, in actuality we usually work with observational data which force us to rely on more ingenious and less reliable means of identification.

1.2 Counterfactual Notation and the Selection Problem

We introduce counterfactual notation. Let y_1 and y_0 denote an outcome of interest both with and without the treatment. For example, y_1 could be wages with a college degree and y_0 could be wages without a college degree for a given individual. The problem is that we never observe both y_1 and y_0 ; we only observe one or the other. Going back to our previous example, we never observe the wage that a college graduate would have earned had they not gone to college. In other words, we never observe the counterfactual.

This literature typically assumes that we observe an *i.i.d.* sample from some underlying

population. This implies that a treatment that is applied to one individual does not affect the outcomes of another. This rules out general equilibrium effects, peer effects and externalities. Recent work by Miguel and Kremer (2004) has done an excellent job of identifying treatment effects in the presence of externalities.

We let w denote the treatment indicator. If $w = 1$, then the individual has been treated. If $w = 0$, then he has not been treated. The triple (y_0, y_1, w) denotes a random vector from the population.

Our measure of the treatment effect for a given individual is $y_1 - y_0$. Note that this is a random variable and so, we will have a distribution of it across the population. In other words, the effects of the treatment will be different depending on the individual. Because we never observe the counterfactual outcome, we will have to come with some summary measure of the treatment effect for the population. One of the most popular is $E[y_1 - y_0]$ or the **average treatment effect (ATE)**. Another commonly estimated treatment is $E[y_1 - y_0|w = 1]$ or the **average treatment effect on the treated (ATE1)**. There are circumstances where the two are the same, however, in general, they will be different.

The fundamental problem with estimating either ATE or ATE1 is missing data. We either observe y_1 or y_0 , but never both. Specifically, we observe

$$y = (1 - w)y_0 + wy_1 = y_0 + w(y_1 - y_0).$$

Because we only observe half of the data, we must provide assumptions that will enable the identification of ATE or ATE1.

The simplest way to identify a treatment effect is to randomly assign the treatment. In this

scenario, w will be independent of (y_1, y_0) and, thus, we will have that

$$E[y_1 - y_0 | w = 1] = E[y_1 - y_0]$$

so that ATE and ATE1 are identical. Estimation is straight-forward as

$$E[y_1 | w = 1] - E[y_0 | w = 0] = E[y_1 - y_0] = ATE = ATE1.$$

The quantities $E[y_1 | w = 1]$ and $E[y_0 | w = 0]$ can be estimated using the observed sample so that the missing data are not a problem. Note that independence is actually stronger than needed. Mean independence will suffice.

Randomization is very useful because without it individuals will select into the treatment. This makes the identification problem very difficult as people will select into the treatment based on their beliefs about $y_1 - y_0$. So, if the outcome is wages and the treatment is getting a college education, then people who believe that $y_1 - y_0$ is large will go to college and those that do not will not. This will make it look as if the returns to obtaining a college education are larger than they really are.

While randomization does have many desirable properties, there are limitations as well. First, randomization is not always feasible. Indeed, one of the most common criticisms of randomization is not always ethical. Second, even if randomization is possible, it is often difficult to extrapolate the results of a study on a sample from one population to a completely different population. This is the problem of **external validity**.

The objects ATE and ATE1 can be related in the following way. Write $y_g = \mu_g + v_g$ and

$E[y_g] = \mu_g$ for $g = 1, 2$. Then, we will have that

$$y_1 - y_0 = (\mu_1 - \mu_0) + (v_1 - v_0) = ATE + (v_1 - v_0).$$

This then gives us that

$$ATE1 = ATE + E[v_1 - v_0|w = 1]$$

So, we will have that ATE and ATE1 will differ by the quantity $E[v_1 - v_0|w = 1]$ which is the expected person-specific gains to participation.

1.3 Ignorability of Treatment

We now introduce a set of control variables which are contained in the vector x . Accordingly, the population is now described by the vector (y_1, y_0, w, x) and we observe (y, w, x) . The key identifying assumption is now going to be that conditional on x, w and (y_0, y_1) are independent. This is the classic assumption from the propensity score literature and is often called **ignorability of treatment**. This is also sometimes referred to as **selection on observables** since the assumption assumes that conditioning on enough variables solves the selection problem. Essentially, this assumption states that (y_1, y_0) and w are uncorrelated once we control for x . Please do not be confused by this. The strategy is simple: control for enough variables so that we preclude the possibility of an omitted variables bias.

Clearly, this assumption implies that ATE and ATE1 are the same *conditional on x* i.e.

$$E[y_1 - y_0|x, w = 1] = E[y_1 - y_0|x].$$

However, ATE and ATE1 need not be the same if we are no longer conditioning on x . To see this, we define $r(x) \equiv E[y_1 - y_0|x]$. Then, we will have that

$$ATE = E[r(x)]$$

and

$$ATE1 = E[r(x) | w = 1].$$

These objects are calculated simply by averaging $r(x)$ over the entire population for ATE and over the sub-population for whom $w = 1$ for ATE1.

1.4 Regression Based Methods

The ignorability of treatment assumption can be used to estimate ATE and ATE1 under quite general conditions. First, we note that the ignorability of treatment assumption implies that

$$\begin{aligned} E[y|x, w] &= E[y_0|x, w] + w[E[y_1|x, w] - E[y_0|x, w]] \\ &= E[y_0|x] + w[E[y_1|x] - E[y_0|x]]. \end{aligned}$$

Accordingly, we will have that

$$\underbrace{E[y|x, w = 1]}_{r_1(x)} - \underbrace{E[y|x, w = 0]}_{r_0(x)} = E[y_1|x] - E[y_0|x]$$

which is simply the ATE (or ATE1) conditional on x and, thus, these objects are non-parametrically identified (i.e. we can estimate, at least in principal, them without any assumptions on the dis-

tribution of (y_1, y_0, w, x) . Suppose that we have consistent estimators of $r_1(x)$ and $r_0(x)$ which we denote with $\widehat{r}_1(x)$ and $\widehat{r}_0(x)$. We can then estimate ATE and ATE1 via

$$\widehat{ATE} = N^{-1} \sum_{i=1}^n [\widehat{r}_1(x) - \widehat{r}_0(x)]$$

and

$$\widehat{ATE1} = \left(\sum_{i=1}^n w_i \right)^{-1} \sum_{i=1}^n w_i [\widehat{r}_1(x) - \widehat{r}_0(x)].$$

Note that the second estimator makes sense because

$$E [w_i [r_1(x) - r_0(x)]] = E [r_1(x) - r_0(x) | w_i = 1] P(w_i = 1).$$

Of course, estimating $r_1(x)$ and $r_0(x)$ may be difficult. The most general estimator would be a non-parametric estimator such as a kernel regression which we will discuss at greater length later on. However, this may be difficult to implement especially if x has high dimension. An alternative and simpler approach would be to loosely parameterize the functions with a number of polynomial and interaction terms. This would, in spirit, be a non-parametric regression.

In practice, even if the treatment was assigned randomly, it will be useful to calculate both ATE and ATE1. The reason is that if the randomization did, in fact, work, then the two treatment effects will be the same or, at least, similar. Hence, substantial differences between ATE and ATE1 suggest that there may have been some selection into the treatment.

If we are willing to impose stronger assumptions then we can estimate ATE and ATE1 with

a simple linear regression. We start out by writing

$$y_1 = \mu_1 + v_1$$

and

$$y_0 = \mu_0 + v_0$$

where $E[v_1] = E[v_0] = 0$. This then gives us that

$$y = \mu_0 + (\mu_1 - \mu_0)w + v_0 + w(v_1 - v_0).$$

In addition to assuming the ignorability of treatment, we also assume that $E[v_1|x] = E[v_0|x]$.

Essentially, this assumption tells us that, conditional on observables, there are no unobserved gains to treatment. These assumptions imply that

$$E[y|w, x] = \mu_0 + (\mu_1 - \mu_0)w + E[v_0|x] \equiv \mu_0 + \alpha w + g(x).$$

The coefficient on w is the ATE. We call $g(x)$ a **control function** and, in practice, its functional form will be unknown. A sensible solution to this problem would be, once again, to flexibly parameterize the function.

As it turns out, α identifies both ATE and ATE1 under these assumptions. To see this, note that ignorability of treatment implies that

$$E[y_1|w, x] = \mu_1 + E[v_1|x]$$

and

$$E[y_0|w, x] = \mu_0 + E[v_0|x].$$

We then obtain that

$$E[y_1|w, x] - E[y_0|w, x] = \mu_1 - \mu_0$$

which then implies by the Law of Iterated Expectations that

$$E[y_1|w] - E[y_0|w] = \mu_1 - \mu_0.$$

1.5 Propensity Score Methods

Propensity score methods are another means of estimating treatment effects. As before, we will still use the ignorability of treatment assumption. These methods involve modeling the propensity of getting treated conditional on x or

$$p(x) = P(w = 1|x).$$

We call this object the **propensity score**. Propensity score methods only require the ignorability of treatment and the assumption that the propensity score is neither zero nor unity. In contrast, the regression methods of the previous section involve imposing auxiliary assumptions on $E[v_1|x]$ and $E[v_0|x]$. Because of this, propensity score methods are often viewed as more robust than the regression-based methods from the previous section.

The propensity score has the desirable property that it balances the data in the sense that $x \perp w|p(x)$. This implies that conditional on the propensity score the distribution of the x 's is the

same in the treatment and control group. For this reason, the propensity score is a balancing score. Formally, we say that $b(x)$ is a **balancing score** if $x \perp w|b(x)$. For this reason, researchers using the propensity score provide evidence that their specification of it actually does balance the data. An ad hoc way to do this is to tests differences in means of the x 's across treatment and control for value of $p(x)$ within narrowly defined cells. The following proposition from Rosenbaum and Rubin (1983) tells us why the propensity score is a balancing score.

Theorem 1 $b(x)$ is a balancing score if and only if $b(x)$ is finer than $e(x) \equiv P(w = 1|x)$ in the sense that $e(x) = f(b(x))$ for some function $f(\cdot)$.

Proof. First, suppose that $e(x) = f(b(x))$. Then because $e(x) = P(w = 1|b(x), x)$ it will be sufficient to show that $e(x) = P(w = 1|b(x))$. To see this note that

$$\begin{aligned} P(w = 1|b(x)) &= E(P(w = 1|x)|b(x)) \\ &= E(e(x)|b(x)) \\ &= e(x) \end{aligned}$$

where the last equality follows because $e(x)$ is finer than $b(x)$. Next, to prove the other direction, let $b(x)$ be a balancing score but suppose that $b(x)$ is not finer than $e(x)$. This implies that there exists x_1 and x_2 such that $b(x_1) = b(x_2)$ but $e(x_1) \neq e(x_2)$. However, this means that $b(x)$ is not a balancing score which is a contradiction. ■

As it turns out, we can write ATE in terms of the propensity score using only the ignorability

of treatment assumption. To see this, first note that

$$\begin{aligned} [w - p(x)] y &= [w - p(x)] [(1 - w) y_0 + w y_1] \\ &= w y_1 - p(x) (1 - w) y_0 - p(x) w y_1. \end{aligned}$$

Next, we take expectations of this expression conditional on (w, x) and obtain

$$w m_1(x) - p(x) (1 - w) m_0(x) - p(x) w m_1(x)$$

where we have defined $m_j(x) = E[y_j|x]$ for $j = 0, 1$. Taking expectations conditional on x now, we obtain

$$p(x) m_1(x) - p(x) (1 - p(x)) m_0(x) - p(x)^2 m_1(x) = p(x) (1 - p(x)) [m_1(x) - m_0(x)].$$

Thus, we will have that

$$E \left[\frac{[w - p(x)] y}{p(x) (1 - p(x))} \right] = E [m_1(x) - m_0(x)] = ATE.$$

This result shows how we can use the propensity score to identify ATE. A similar identification result holds for ATE1.

Estimation of ATE using this method is straight-forward. First, researchers must estimate $p(x)$ via $\hat{p}(x)$. Non-parametric methods may be used, but generally a flexible probit or logit

model should suffice. In the next step, the researcher calculates

$$\widehat{ATE} = \frac{1}{N} \sum_{i=1}^n \frac{[w - \widehat{p}(x)] y}{\widehat{p}(x) (1 - \widehat{p}(x))}.$$

While this estimator is consistent, obtaining valid standard errors is difficult as it is a two-step estimator. Two possibilities for calculating valid standard errors are the δ -method and the bootstrap.

Simpler regression based methods that use the propensity score have also been proposed. These methods involve a simple regression of y_i onto $(1, w_i, \widehat{p}(x_i))$. Under additional assumptions, the coefficient on w_i recovers ATE. The basic idea of these methods is that the propensity score contains all of the information contained in x and, so the researcher only needs to include $\widehat{p}(x_i)$ in the regression. As stated by Wooldridge, the advantage of these regressions is that they are parsimonious. However, it is unclear that this procedure is superior to the regression methods discussed in the previous section where we include flexible parametrizations of the function $g(x)$. In addition, these “kitchen sink” regressions avoid the complication of having to worry about having a two-stage estimation procedure.

Another commonly used set of procedures for estimating treatment effects are matching methods. These methods hinge upon the observation that the ignorability of treatment assumption implies that, conditional on $p(x)$, w and (y_0, y_1) are independent which, in turn, gives us that

$$E[y|w = 1, p(x)] - E[y|w = 0, p(x)] = E[y_1|p(x)] - E[y_0|p(x)].$$

This implies that we can average the above object over $p(x)$ to recover ATE. Note that this

procedure would also work if we were conditioning on x . However, using the propensity score allows us to reduce the dimensionality of the problem. In practice, one difficulty with the implementation of this procedure is that it may be difficult to find both treated and untreated individuals for a given value of $p(x)$.

1.6 Local Average Treatment Effects

An alternative to using propensity score based methods to identify treatment effects is instrumental variables based methods. While these methods are useful in their own right, they have the added benefit that they allow us to sharpen our interpretation of the IV estimator.

We denote our IV by Z . Our IV is a variable that helps to determine participation. Accordingly, we will now write $w(Z)$ so that one's propensity to participate in the program will depend on the value of Z . To better understand this, suppose that $Z = z$, then $w(z)$ is the participation indicator when $Z = z$. We assume that Z is a random variable such that

$$(i) \text{ for all } z, (y_0, y_1, w(z)) \text{ is independent of } Z \tag{A1}$$

and

$$(ii) P(z) \equiv E[w|Z = z] \text{ is a non-trivial function of } z.$$

Part (ii) is testable. Part (i) is akin to an exclusion restriction and is not testable. It is important to note that randomization of Z is not sufficient for part (i).

To see this, consider the Josh Angrist paper on the impact of military service on earnings. His instrument stems from the fact that during the Vietnam War participation in the military was

to some degree random as it was determined by a lottery. Accordingly, Z should be independent of w . However, it may still be the case that Z is not independent of (y_0, y_1, w) . For example, in response to obtaining a high lottery number a person may have responded in such a way that would also affect their future earnings e.g. going to Canada or prolonging their education.

To fix ideas, we consider a simple parametric example from Heckman and Robb (1985). The participation decision is given by a latent index model

$$w^* = \gamma_0 + Z\gamma_1 + v$$

where

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

The response is given by

$$y = \beta_0 + w\beta_1 + \varepsilon$$

where $y = wy_1 + (1 - w)y_0$. We will then have that $y_1 = \beta_0 + \varepsilon$, and $y_0 = \beta_0 + \beta_1 + \varepsilon$. Thus, the ATE will be given by β_1 . Assumption A1 is satisfied as long as Z is independent of v and ε . These additional assumptions show us how instrumental variables can be used to identify ATE. It is important to note, however, that additional assumptions beyond A1 were required for identification. In particular, we assumed that the treatment effect was homogeneous in the population. Consequently, we do not have non-parametric identification in this example.

As it turns out, assumption A1 is not, by itself, sufficient for the identification of any treatment

effect. To see this, note that assumption A1 implies that

$$\begin{aligned}
E[y|Z = z] - E[y|Z = x] &= E[w(z)y_1 + (1 - w(z))y_0|Z = z] - E[w(x)y_1 + (1 - w(x))y_0|Z = x] \\
&= E[(w(z) - w(x))(y_1 - y_0)] \\
&= E[y_1 - y_0|w(z) - w(x) = 1]P(w(z) - w(x) = 1) \\
&\quad - E[y_1 - y_0|w(z) - w(x) = -1]P(w(z) - w(x) = -1).
\end{aligned}$$

This expression tells us that $E[y|Z = z] - E[y|Z = x]$ can be positive, zero or negative depending on how participation status responds to Z . In fact, it is possible to obtain a negative quantity despite having $y_1 - y_0 > 0$ for all individuals.

Consequently, identification requires additional assumptions. One assumption that is commonly used is that the treatment is constant for all individuals so that $y_1 - y_0 = \alpha$. In this scenario, we will have that

$$E[y|Z = z] - E[y|Z = x] = E[(w(z) - w(x))(y_1 - y_0)] = \alpha [P(z) - P(x)]$$

and the ATE is identified. Another assumption that is commonly invoked is that there exists a value of the instrument, x , such that the probability of being treated is zero i.e. $P(x) = 0$. In this case, we will have that $P(w(z) - w(x) = -1) = 0$ which implies that

$$E[y|Z = z] - E[y|Z = x] = E[y_1 - y_0|w(z) = 1]P(w(z) = 1).$$

This allows us to identify ATE1 for the subpopulation for whom $Z = z$. The ATE1 can then

be recovered by averaging over the entire population.

A third assumption used by Imbens and Angrist is to assume that the response of treatment status to the IV is monotonic in the sense that

$$\forall z, x \text{ either } w(z) \geq w(x) \text{ or } w(z) \leq w(x) \text{ for all } i. \tag{A2}$$

This assumption states that anyone who would participate given $Z = x$ must also participate given $Z = z$. Like part (i) of A1, this assumption is also fundamentally untestable. This assumption implies that $P(w(z) - w(x) = -1) = 0$ and, thus, implies that

$$\frac{E[y|Z = z] - E[y|Z = x]}{P(z) - P(x)} = E[y_1 - y_0 | w(z) - w(x) = 1].$$

This object is called the Local Average Treatment Effect or LATE. Note that the denominator on the left-hand side obtains because

$$P(w(z) - w(x) = 1) = E[w(z) - w(x)] = P(z) - P(x)$$

where the second equality followed because of the monotonicity assumption.

This object identifies ATE on the sub-population for whom $w(z) - w(x) = 1$ which are the people whose treatment status is altered by a move from $Z = x$ to $Z = z$. It is crucial to note that it says nothing about the treatment effect for the population whose treatment status is unaffected by the IV. This highlights one of the limitations of LATE which is that we will never know what part of the population is pivotal or, equivalently, has $w(z) - w(x) = 1$.

One of the interesting properties of LATE is that it helps us to sharpen our interpretation of the IV estimator. As it turns out, if the response to a treatment (or endogenous variable) is heterogeneous in the population then the IV estimator is a weighted average of LATE's. Moreover, if the instrumental variable is binary (i.e. takes on only two values) then the IV estimator is the LATE. Can you show this?