Heterogeneous Treatment Effects

Christopher Taber

University of Wisconsin

February 23, 2012

So far in this course we have focused on the case

$$Y_i = \alpha T_i + \varepsilon_i$$

Think about the case in which T_i is binary

Let

- Y_{1i} denote the value of Y_i for individual i when $T_i = 1$
- Y_{0i} denote the value of Y_i for individual i when $T_i = 0$

It is useful to define the treatment effect as

$$\pi_i = Y_{1i} - Y_{0i}$$

Note that in the case we have been thinking about so far

$$\begin{aligned}\pi_i &= \alpha + \varepsilon_i - \varepsilon_i \\ &= \alpha\end{aligned}$$

and thus we have imposed that it can not vary over the population

This seems pretty unreasonable for almost everything we have thought about in this class

A relatively recent literature has tried to study heterogeneous treatment effects in which these things vary across individuals

A clear problem is that even if we have estimated the full distribution what do we present in the paper?

We must focus on a feature of the distribution

The most common:

- Average Treatment Effect (ATE)

$$E(\pi_i)$$

- Treatment on the Treated (TT)

$$E(\pi_i \mid T_i = 1)$$

- Treatment on the Untreated (TUT)

$$E(\pi_i \mid T_i = 0)$$

(Heckman and Vytlacil discuss Policy Relevant Treatment effects, but I need more notation than I currently have to define those)

These each answer very different questions

In terms of identification they are related.

All we can directly identify from the data is :

$$E(Y_{1i} \mid T_i = 1), E(Y_{0i} \mid T_i = 0), Pr(T_i = 1)$$

and at this point, without anything else, that is all you can directly identify.

There are two key missing pieces:

$$E(Y_{1i} \mid T_i = 0), E(Y_{0i} \mid T_i = 1)$$

Knowledge of these would be sufficient to identify the parameters:

$$\begin{aligned} TT = E(\pi_i \mid T_i = 1) =& E(Y_{1i} \mid T_i = 1) - E(Y_{0i} \mid T_i = 1) \\ TUT = E(\pi_i \mid T_i = 0) =& E(Y_{1i} \mid T_i = 0) - E(Y_{0i} \mid T_i = 0) \end{aligned}$$

$$\begin{aligned} ATE = E(\pi_i) = & [E(Y_{1i} \mid T_i = 1) - E(Y_{0i} \mid T_i = 1)] \, Pr(T_i = 1) \\ & + [E(Y_{1i} \mid T_i = 0) - E(Y_{0i} \mid T_i = 0)] \, [1 - Pr(T_i = 1)] \end{aligned}$$

Now how do we estimate these?

Selection only on Observables

Lets start with the case in which we only have selection on observables

Assumption

For all x in the support of X_i and $t \in \{0, 1\}$,

$$
\begin{aligned}
E(Y_{1i} \mid X_i = x, T_i = t) =& E(Y_{1i} \mid X_i = x) \\
E(Y_{0i} \mid X_i = x, T_i = t) =& E(Y_{0i} \mid X_i = x)
\end{aligned}
$$

A “slightly” stronger version of this is random assignment of T_i conditional on X_i

This is often also called unconfoundedness

A very strong assumption

Interestingly this is still not enough if there are sets of observable covariates χ with positive measure for which $Pr(T_i = 1 \mid X_i \in \chi) = 1$ or $Pr(T_i = 0 \mid X_i \in \chi) = 0$ then clearly the full distribution of treatment effects is not identified.

For example suppose T_i is being pregnant and ment are never pregnant, we could never how to identify

$$E(\text{Income} \mid \text{Pregnant, Male})$$

This is perhaps not a very interesting counterfactual (actually relevant is probably a better word-it is kind of interesting)

But if you want to measure the average treatment effect you can't.

It wouldn't be a problem for the treatment on the treated

Thus we need the additional assumption

Assumption

For almost all x in the support of X_i,

$$0 < Pr(T_i = 1 \mid X_i = x) < 1$$

Theorem

Under assumptions 1 and 2 the ATE, TT, and TUT are identified

It is pretty clear to see why

Consider the treatment on the treated.

Note that $E(Y_{1i} \mid T_i = 1)$ is identified directly from the data so all we need to get is $E(Y_{0i} \mid T_i = 0)$.

Let $F(x \mid T_i)$ be the distribution of X_i conditional on T_i

$F(x \mid T_i)$ is identified directly from the data

Then under the first assumption above

$$E(Y_{0i} \mid T_i = 1) = \int E(Y_{0i} \mid X_i = x) dF(x \mid T_i = 1)$$

As long as assumption 2 holds, $E(Y_{0i} \mid X_i = x)$ is directly identified from the data so $E(Y_{0i} \mid T_i = 1)$ is identified

You can also get

$$E(Y_{1i} \mid T_i = 0) = \int E(Y_{1i} \mid X_i = x) dF(x \mid T_i = 0)$$

and use this to identify the ATE or the TUT

Estimation

There are a number of different ways to estimate this model

The most obvious is to just use OLS defining

$$\begin{aligned} Y_{0i} &= X_i'\beta_0 + u_{0i} \\ Y_{1i} &= X_i'\beta_1 + u_{1i} \end{aligned}$$

Then one could estimate

$$\widehat{ATE} = \frac{1}{N}\sum_{i=1}^{N} X_i'\left(\widehat{\beta}_1 - \widehat{\beta}_0\right)$$

or alternatively:

$$\widehat{ATE} = \frac{1}{N}\sum_{i=1}^{N} T_i\left[Y_{1i} - X_i'\widehat{\beta}_0\right] - (1 - T_i)\left[X_i'\widehat{\beta}_1 - Y_{0i}\right]$$

TT and TUT are analogous (although second method might be more natural)

Clearly, if you want to be more nonparametric you can either run nonparametric regression or allow a functional form that becomes more flexible with the sample size

Matching

Heckman and coauthors made a strong case for matching over regression

If say you are interested in TT, but the support of X_i conditional on $T_i = 1$ is very different than the unconditional support of X_i than the regression approach can be pretty screwed up

They made this argument in the context of JTPA where only low income people are eligible for treatment

The idea of matching with data with discrete support is relatively easy, lets focus on the TT case

Let N_1 be the the number of respondents with $T_i = 1$ and for simplicity label them $i = 1, .., N_1$

Similarly let N_0 be the number of respondents with $T_i = 0$ and label them $j = 1, ..., N_0$

1. For each i find a control j with exactly the same value of X_i. That is

$$J(i) = \{j \in \{1, .., N_0\} : X_i = X_j\}$$

   and $j(i)$ is a random element from this set

2. We can get a consistent estimate using

$$\widehat{TT} = \frac{1}{N_1} \sum_{i=1}^{N_1} \left(Y_{1i} - Y_{0j(i)} \right)$$

This is difficult to do in practice for two reasons:

1. If X_i is continuous we can't match exactly
2. If X_i is very high dimensional, even with discrete data we probably couldn't match directly

Propensity Score Matching

Propensity score matching is a way of getting around the second problem.

Rather than matching on the high dimensional X_i we can match on the lower dimensional

$$P(x) = Pr(T_i = 1 \mid X_i = x)$$

The reason why comes from Bayes Theorem

For any x,

$$
\begin{aligned}
&F(x \mid P(X_i) = \rho, T_i = 1) \\
&= Pr(X_i \leq x \mid P(X_i) = \rho, T_i = 1) \\
&= \frac{Pr(T_i = 1 \mid X_i \leq x, P(X_i) = \rho) Pr(X_i \leq x \mid P(X_i) = \rho)}{Pr(T_i = 1 \mid P(X_i) = \rho)} \\
&= \frac{\rho Pr(X_i \leq x \mid P(X_i) = \rho)}{\rho} \\
&= Pr(X_i \leq x \mid P(X_i) = \rho)
\end{aligned}
$$

and analogously for any *x*,

$$
\begin{aligned}
&F(x \mid P(X_i) = \rho, T_i = 0) \\
&= Pr(X_i \leq x \mid P(X_i) = \rho, T_i = 0) \\
&= \frac{Pr(T_i = 0 \mid X_i \leq x, P(X_i) = \rho) Pr(X_i \leq x \mid P(X_i) = \rho)}{Pr(T_i = 0 \mid P(X_i) = \rho)} \\
&= \frac{(1 - \rho) Pr(X_i \leq x \mid P(X_i) = \rho)}{1 - \rho} \\
&= Pr(X_i \leq x \mid P(X_i) = \rho) \\
&= F(x \mid P(X_i) = \rho, T_i = 1)
\end{aligned}
$$

Thus if we condition on the propensity score, the distribution of X_i is identical for the controls and the treatments.

But since the error term is uncorrelated with X_i,

$$\begin{aligned} & E(Y_{0i} \mid T_i = 1, P(X_i) = \rho) \\ & \quad = \int E(Y_{0i} \mid X_i = x) dF(x \mid T_i = 1, P(X_i) = \rho) \\ & \quad = \int E(Y_{0i} \mid X_i = x) dF(x \mid T_i = 0, P(X_i) = \rho) \\ & \quad = E(Y_{0i} \mid T_i = 0, P(X_i) = \rho) \end{aligned}$$

This means that we can match on the propensity score rather than the full set of $X's$.

This makes the problem much simpler, but

- You still need to estimate the propensity score which is a high dimensional non-parametric problem. People typically just use a logit
- Now you have to figure out how to match a control to treatment *i*.

There are essentially 3 ways to do that:

- Just take nearest neighbor (or perhaps caliper that you throw out observations without a close neighbor)
- Use all of the observations that are sufficiently close
- Estimate $E(Y_{0j} \mid T_j = 0, P(X_j) = P(X_i))$ say with local polynomial regression

Reweighting

Another approach is reweighting

Let $f_j(x)$ be the density of X_i conditional on $T_i = j$.

Note that using Bayes theorem

$$f_1(x) = \frac{P(x)f(x)}{Pr(T_i = 1)}$$
$$f_0(x) = \frac{(1 - P(x))\,f(x)}{Pr(T_i = 0)}$$

so

$$\begin{aligned} E(Y_{0i} \mid T_i = 1) &= \int E(Y_{0i} \mid X_i = x) f_1(x) dx \\ &= \int E(Y_{0i} \mid X_i = x) \frac{f_1(x)}{f_0(x)} f_0(x) dx \\ &= E\left(Y_{0i} \frac{P(X_i)}{1 - P(X_i)} \right) \frac{Pr(T_i = 0)}{Pr(T_i = 1)} \end{aligned}$$

Putting this together we can use the estimator

$$\begin{aligned} \frac{\sum_{i=1}^{N_1} Y_{1i}}{N_1} - \frac{\sum_{j=1}^{N_0} Y_{0j} \frac{P(X_j)}{1-P(X_j)}}{N_1} &= \frac{\sum_{i=1}^{N_1} Y_{1i}}{N_1} - \frac{\frac{1}{N_0} \sum_{j=1}^{N_0} Y_{0j} \frac{P(X_j)}{1-P(X_j)}}{\frac{N_1}{N_0}} \\ &\approx E(Y_{1i} \mid T_i = 1) - \frac{E(Y_{0i} \mid T_i = 1) \frac{Pr(T_i=1}{Pr(T_i=0}}{\frac{Pr(T_i=1)}{Pr(T_i=0)}} \\ &= TT \end{aligned}$$

Instrumental Variables

Define

$$
\begin{aligned}
Y_i &= \begin{cases} Y_{0i} & \text{if } T_i = 0 \\ Y_{1i} & \text{if } T_i = 1 \end{cases} \\
&= Y_{0i} + \pi_i T_i
\end{aligned}
$$

Assume that we have an instrument Z_i that is correlated with T_i but not with Y_{0i} or Y_{1i}.

Does IV estimate the ATE?

Lets abstract from other regressors

IV yields

$$\begin{aligned}
\text{plim}\widehat{\beta}_1 &= \frac{Cov(Z_i, Y_i)}{Cov(Z_i, T_i)} \\
&= \frac{Cov(Z_i, Y_{0i} + \pi_i T_i)}{Cov(Z_i, T_i)} \\
&= \frac{Cov(Z_i, Y_{0i})}{Cov(Z_i, T_i)} + \frac{Cov(Z_i, \pi_i T_i)}{Cov(Z_i, T_i)} \\
&= \frac{Cov(Z_i, \pi_i T_i)}{Cov(Z_i, T_i)}.
\end{aligned}$$

In the case in which treatment effects are constant so that $\pi_i = \pi_0$ for everyone

$$\begin{aligned} \text{plim}\widehat{\beta}_1 &= \frac{Cov(Z_i, \pi_0 T_i)}{Cov(Z_i, T_i)} \\ &= \pi_0 \end{aligned}$$

However, more generally IV does not converge to the Average treatment effect

Local Average Treatment Effects

Imbens and Angrist (1994) consider the case in which there are not constant treatment effects

The consider a simple version of the model in which Z_i takes on 2 values, call them 0 and 1 for simplicity and without loss of generality assume that
$Pr(T_i = 1 \mid Z_i = 1) > Pr(T_i = 1 \mid Z_i = 0)$

There are 4 different types of people those for whom $T_i = 1$ when:

1. $Z_i = 1, Z_i = 0$
2. never
3. $Z_i = 1$ only
4. $Z_i = 0$ only

Imbens and Angrist's monotonicity rules out 4 as a possibility

Let $\mu_1, \mu_2,$ and μ_3 represent the sample proportions of the three groups

and G_i an indicator of the group

Note that

$$\begin{aligned}\widehat{\beta}_1 \xrightarrow{p} & \frac{Cov(Z_i, \pi_i T_i)}{Cov(Z_i, T_i)} \\ & = \frac{E(\pi_i T_i Z_i) - E\left(\pi_i T_i\right) E\left(Z_i\right)}{E(T_i Z_i) - E\left(T_i\right) E\left(Z_i\right)}\end{aligned}$$

Let ρ denote the probability that $Z_i = 1$. Lets look at the pieces first the numerator

$$
\begin{aligned}
& E(\pi_i T_i Z_i) - E\left(\pi_i T_i\right) E\left(Z_i\right) \\
=& \rho E(\pi_i T_i \mid Z_i = 1) - E\left(\pi_i T_i\right) \rho \\
=& \rho E(\pi_i T_i \mid Z_i = 1) \\
& - \left[\rho E(\pi_i T_i \mid Z_i = 1) + (1-\rho) E(\pi_i T_i \mid Z_i = 0)\right] \rho \\
=& \rho(1-\rho)\left[E(\pi_i T_i \mid Z_i = 1) - E(\pi_i T_i \mid Z_i = 0)\right] \\
=& \rho(1-\rho)\left[E(\pi_i \mid G_i = 1)\mu_1 + E(\pi_i \mid G_i = 3)\mu_3 - E(\pi_i \mid G_i = 1)\mu_1\right] \\
=& \rho(1-\rho) E(\pi_i \mid G_i = 3)\mu_3
\end{aligned}
$$

Next consider the denominator

$$
\begin{aligned}
& E(T_i Z_i) - E\left(T_i\right) E\left(Z_i\right) \\
= & \rho E(T_i \mid Z_i = 1) - E\left(T_i\right) \rho \\
= & \rho E(T_i \mid Z_i = 1) \\
& - \left[\rho E(T_i \mid Z_i = 1) + (1 - \rho)\, E(T_i \mid Z_i = 0)\right] \rho \\
= & \rho(1 - \rho)\left[E(T_i \mid Z_i = 1) - E(T_i \mid Z_i = 0)\right] \\
= & \rho(1 - \rho)\left[\mu_1 + \mu_3 - \mu_1\right] \\
= & \rho(1 - \rho)\mu_3
\end{aligned}
$$

Thus

$$\begin{aligned}\widehat{\beta}_1 \xrightarrow{p} & \frac{\rho(1-\rho)E(\pi_i \mid G_i = 3)\mu_3}{\rho(1-\rho)\mu_3} \\ = & E(\pi_i \mid G_i = 3)\end{aligned}$$

They call this the local average treatment effect