Applied Econometrics

Lecture 11: Treatment Effects Part I

Måns Söderbom*

25 September 2009

*Department of Economics, University of Gothenburg. Email: mans.soderbom@economics.gu.se. Web: www.economics.gu.se/soderbom, www.soderbom.net.

1. Introduction

This and the next lecture focus on the estimation of treatment effects and the evaluation of programs. As you probably know this is an area of very active research in applied micro economics, and you often see papers using this methodology published in top journals. To some extent, the treatment evaluation literature uses a different language to what we are used to. In fact, much of what is done in treatment analysis can be related to standard regression analysis. But as we shall see, not everything.

References for this lecture are as follows:

Angrist and Pischke (2009), Chapters 3.2-3.3; 5 - read carefully.

Chapter 18.1-18.4 in Wooldridge (2002) "Cross Section and Panel Data" - read carefully.

Gilligan, Daniel O. and John Hoddinott (2007). "Is There Persistence in the Impact of Emergency Food Aid? Evidence on Consumption, Food Security and Assets in Rural Ethiopia," American Journal of Agricultural Economics, forthcoming - basis for empirical examples and computer exercise 4.

For a short, nontechnical yet brilliant introduction to treatment effects, see "Treatment Effects" by Joshua Angrist, forthcoming in the New Palgrave.

2. Concepts and quantities of interest

- A **treatment effect** is simply the causal effect 'treatment' (e.g. undergoing a training programme) on an outcome variable of interest (e.g. productivity at work).

- Typically the treatment variable is a binary (0-1) variable. Unless I say otherwise, this will be the assumption throughout the discussion.

- **The potential-outcomes framework**: For each individual there is a potential outcome with treatment, denoted y_1, and another potential outcome without treatment, denoted y_0. These can thus be thought of as outcomes in alternative states of the world, and the treatment (causal) effect is the difference between these two quantities: $y_1 - y_{0.}$

- Of course, it is impossible to measure treatment effects at the **individual** level, as we can never observe the full set of potential outcomes in alternative states of the world - basically, because we don't have access to parallel universes. Researchers therefore focus on various forms of **average treatment effects**.

- Following Wooldridge, I'm defining w as a binary treatment indicator (a dummy variable), where

$$\begin{aligned} w &= 1 \text{ if treatment,} \\ w &= 0 \text{ if no treatment.} \end{aligned}$$

The outcome variables, y_1 and y_0, as well as the difference $y_1 - y_0$, are random variables that potentially **vary across individuals** in the population. In seeking to estimate the effect of treatment on outcomes, it is therefore natural to focus on estimating the **average treatment effect**. We focus on two such measures:

1. The average treatment effect (ATE):

$$ATE = E\left(y_1 - y_0\right)$$

2. The average treatment effect on the treated (ATE_1):

$$ATE = E\left(y_1 - y_0 | w = 1\right)$$

- ATE is the expected effect of treatment for a randomly drawn individual from the population
- ATE_1 is the expected effect of treatment for a randomly drawn individual from those individuals in the population that have undergone treatment.

In some special cases, ATE and ATE_1 coincide.

- At this point it's worth pausing for a moment, and think about what the quantities just defined really mean. Notice in particular that what is being estimated here is the **overall impact** of a particular program on an outcome of interest (reduced form). This may be quite different from the impact of the treatment **keeping everything else constant**. Think of a structural relationship of the form $y = f(x, w)$, and suppose $x = x(w)$, i.e. x depends on w. The average treatment effect is an estimate of the total effect of w on y, i.e. both the direct effect and the indirect effect (the one operating through w).
- So how can we estimate these treatment effects? Recall that the treatment effect is the difference between two potential outcomes. We have data on actual outcomes. The **actual** outcome is equal to the **potential** outcome...
- ...with treatment for the treated individuals, i.e. we observe y_1 for individuals with $w = 1$; and
- ...without treatment for the untreated individuals, i.e. we observe y_0 for individuals with $w = 0$.

Thus we will not have data on both y_1 and y_0, which means we can't just compute sample averages of the difference $y_1 - y_0$.

- The problem is that we don't observe the **counterfactual** (the outcome that didn't happen).
- In other words, we do not observe outcomes

  - without treatment for the treated individuals (i.e. y_0 is unobserved whenever $w = 1$), or
  - outcomes with treatment for the untreated individuals (y_1 is unobserved whenever $w = 0$).

- In the data, the observed outcome y can be written

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

  This complicates the estimation of treatment effects. How can we estimate ATE or ATE_1, if this is all the data available?

3. Randomization: Experimental data

- **Randomization**: can be thought of as a process in which the outcome of a toss of a coin determines whether an individual get treatment ($w_i = 1$) or not ($w_i = 0$). If treatment is randomized across individuals, then estimation of the average treatment effect is simple, despite the unobservability problem just discussed.

- Suppose your sample consists of N observations, and your goal is to calculate $E(y_1)$ and $E(y_0)$. Your problem is that for each individual, either y_{1i} or y_{0i} is unobserved. Might it still be valid to calculate $E(y_1)$ by taking the average of the **observed** values of y_1, and vice versa for $E(y_0)$?

- Yes it would, since randomization ensures the potential outcomes (y_1, y_0) are statistically independent of treatment status.

- The reason is that independence implies $E(y_1|w=1) = E(y_1|w=0) = E(y_1)$, and so

$$\begin{aligned} ATE &= E(y_1 - y_0) \\ &= E(y_1) - E(y_0) \\ &= E(y|w=1) - E(y|w=0), \end{aligned}$$

where independence allows us to go from the second to the third line. It also follows that

$$\begin{aligned} ATE_1 &= E(y_1 - y_0|w=1), \\ &= E(y_1|w=1) - E(y_0|w=1), \\ &= E(y_1) - E(y_0), \\ &= E(y|w=1) - E(y|w=0), \end{aligned}$$

where independence allows us to go from the second to the third line, and from the third to the fourth line. Notice that, in this case,

$$ATE = ATE_1.$$

- Thus, a randomized experiment guarantees that the **difference-in-means** estimator is fine (unbiased and consistent). Notice that this estimator can be obtained by running the following simple OLS regression:

  $$y_i = \beta_0 + \beta_1 w_i + u_i,$$

  where the estimate of β_1 is the estimated ATE (and, by implication, ATE_1).

- You see how powerful the method of randomization is. Provided you get the design of your experiment right, all you need to do is to compare mean values across the two groups ($w = 0, w = 1$). If done right, a pure randomized experiment is in many ways the most convincing method of evaluation.

- It sounds easy, but, of course, life is never easy. Experiments have their own drawbacks:

  – They are rare in economics, and often **expensive** to implement. 'Social experiments' carried out in the U.S. typically had very large budgets, with large teams and complex implementation. However, quite a few randomized evaluations have recently been conducted in developing countries on fairly small budgets.

  – They may not be amenable to **extrapolation**. That is, there may be questionmarks as to the **external validity** of the results of a particular experiment. The main reasons are:

    ∗ it may be very hard to replicate all components of the program elsewhere

    ∗ the results may be specific to the sample (you might argue this is a general problem in empirical economics - that may well be true, but typically experiments are conducted in relatively small regions, which possibly exacerbates the problem);

    ∗ the results may be specific to the program (would a slightly different program have similar effects?).

  - There are lots of **practical problems** related to the implementation of experiments. Getting the design of the experiment right really is the big challenge, and as you can imagine much can go wrong in the field. Suppose you start to give free school meals randomly in 50% of the schools in a region where previously school meals were not free. One year later you plan to turn up and compare pupil performance in treated and nontreated schools. But how can you be sure parents whose kids are in nontreated schools have not reacted to your reform by changing schools? Or could treatment affect the decision as to when someone should leave school? The basic point is that you typically need time between initiating the treatment and measuring the outcome, and much can go wrong in the meantime. There may be ethical issues: why give some people treatment and not others? How justify not helping those that need it the most?

- For these reasons, most economic research still uses **non-experimental** (observational) data.

- When we have non-experimental data, we must assume that individuals at least partly determine whether they receive treatment. This may lead to problems with the simple difference-in-means estimator if the individual's decision to get treatment depends on the benefits of treatment. In such a case, we would say there is **self-selection** of treatment. Addressing this problem is largely what the literature on treatment effect estimation based on non-experimental data is about. Notice that this is precisely the problem solved - in principle - by randomization.

4. Non-experimental data

We now focus on the case where individuals potentially self-select into treatment. This breaks independence between (y_1, y_0) and w, and so the simple difference-in-means estimator discussed in the previous section does not estimate the average treatment effects consistently. Actually, we can interpret this problem as being posed by **omitted variables**.

- To see this, suppose your task is to evaluate the effect of a job training program on earnings. You have a random sample of workers, with data on earnings and whether the individuals have received training (the treatment).

- It would seem plausible that people self-select (or get self-selected by their boss) into training, depending on certain individual characteristics. It may be that people with a high level of education tend to select training more frequently than people with little education. In addition, we strongly suspect that (y_1, y_0) are positively correlated with education.

- Thus, (y_1, y_0) and w are no longer independent - they are both affected by a common factor, namely education - and any attempt to use the difference-in-means estimator will then result in bias of the estimated average treatment effect.

- To see how this links to omitted variables bias, recall that using the difference-in-means estimator is equivalent to estimating the regression

  $$y_i = \beta_0 + \beta_1 w_i + u_i.$$

  Since in the current example education is in the residual and assumed positively correlated with training, the OLS estimate of β_1 will be upward biased.

4.1. Selection on observables

In empirical economics, we worry about problems posed by omitted variables all the time. Of course, the simplest solution to this problem is to control for the role played by the omitted variables in estimation. Provided all variables that need to be controlled for **can** be controlled for, this solves the omitted variables problem completely, and we can estimate the treatment effect consistently. In the treatment literature this amounts to assuming **ignorability of treatment** (given x):

- **Ignorability of treatment:** *Conditional on x, w and (y_1, y_0) are independent.* [Wooldridge, Assumption ATE.1]. In words: If we are looking at individuals with the same characteristics x, then (y_1, y_0) and w are independent. Angrist and Pischke (2009) call this **conditional independence assumption** - CIA.

- **Conditional mean independence** [Wooldridge, Assumption ATE.1']:

$$\begin{aligned} E(y_0|x,w) &= E(y_0|x), \\ E(y_1|x,w) &= E(y_1|x) \end{aligned}$$

  In words: Comparing two individuals with the same x, the expected outcome under treatment is the same for treated individuals as for untreated individuals. This is often described as **selection on observables**.

- Clearly ignorability of treatment **implies** conditional mean independence (yes?).

- Notice that the ATE conditional on x, denoted $ATE(x)$, coincides with the ATE_1 conditional on x, denoted $ATE_1(x)$:

$$\begin{aligned} ATE_1(x) &= E(y_1|x,w=1) - E(y_0|x,w=1) \\ &= E(y_1|x) - E(y_0|x) \\ &= ATE(x). \end{aligned}$$

So you see how these average treatment effects are written simply as the difference between the expected value of y_1 and y_0, conditional on x.

- We haven't said anything at this point about a 'model' or the functional form relationship between x and the (conditional) expected values of y_1 and y_0. Remaining agnostic about functional form, we can write

$$\begin{aligned} E(y_1|x) &= r_1(x), \\ E(y_0|x) &= r_0(x), \end{aligned}$$

where $r_1(.)$ and $r_0(.)$ are functions. Assuming that r_1 and r_0 can be estimated, we can obtain a consistent estimator of ATE simply as follows:

$$A\hat{T}E = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{r}_1(x_i) - \hat{r}_0(x_i)\right],$$

where $\hat{r}_1$ and $\hat{r}_0$ are the estimates of r_1 and r_0.

- Notice that $A\hat{T}E$ is the estimated **unconditional** average treatment effect (because we average across individuals with different values of x in the sample).

- We said above that the conditional average treatment effect, $ATE(x)$, is equal to the conditional average treatment effect on the treated, $ATE_1(x)$. However, in general, the unconditional average treatment effect (ATE) is **not** equal to the unconditional average treatment effect on the treated (ATE_1). Subtle, isn't it? To estimate the latter we modify the above formula and simply average across the $j = 1, 2, ..., N_1$ treated individuals only

$$A\hat{T}E_1 = \frac{1}{N_1}\sum_{j=1}^{N_1}\left[\hat{r}_1(x_j) - \hat{r}_0(x_j)\right],$$

or, using Wooldridge's formulation,

$$A\hat{T}E_1 = \left(\sum_{i=1}^{N} w_i\right)^{-1} \sum_{i=1}^{N} w_i \left[\hat{r}_1\left(x_i\right) - \hat{r}_0\left(x_i\right)\right].$$

- Of course, in order to calculate any of these quantities, we need to be a little bit more specific about r_1 and r_0.

[Now turn to Section 1 in the appendix.]

- Suppose the observation with id=6 had not been included in the "data" just examined, so that there were no observations in the data for which $(w = 1, x = 1)$. What would be the implication of that? Think of a real example where something similar might happen.

- In the example above the treatment effects were calculated by comparing individuals for whom the values of x are identical. This is known as **exact matching** on observables.

- Typically in applied work, however, it is either impractical or impossible to divide up the data into (w, x)-specific cells (as in that example), because there are usually many x-variables and/or some or all of these may be continuous variables. Thus, there are typically no nontreated individuals in the data that have exactly the same x-values as a given treated individual. This makes it more difficult to estimate the counterfactuals.

- The two main ways of controlling for observable variables in practice are **estimation by regression** and **estimation by inexact matching**. We discuss these next. In terms of data availability, the premise of the discussion is initially that we have cross-section data. Later on, we discuss how longitudinal (panel) data offer some important advantages compared to cross-section data. Some authors, e.g. Richard Blundell, have argued that results based on panel data are more robust than results based on cross-section data.

4.1.1. Estimation by regression

- To see how regression techniques can be used in this context, consider the following two equations:

$$\begin{aligned} y_0 &= \mu_0 + v_0, \\ y_1 &= \mu_1 + v_1, \end{aligned}$$

where $E(v_0) = E(v_1) = 0$. These two equations describe the outcomes in the event of non-treatment and treatment, respectively, and can be re-written as a **switching regression**:

$$\begin{aligned} y &= w(\mu_1 + v_1) + (1 - w)(\mu_0 + v_0), \\ y &= \mu_0 + (\mu_1 - \mu_0) w + v_0 + w(v_1 - v_0). \qquad (4.1) \end{aligned}$$

- If v_0 and v_1 are independent of x, we have

$$ATE = (\mu_1 - \mu_0),$$

which is the regression coefficient on w in (4.1). In this case, we could estimate ATE simply by running a regression where y is the dependent variable and w is the only explanatory variable. As already discussed, this is how to proceed under randomization.

- Suppose now v_0 and v_1 are functions of x:

$$\begin{aligned} v_0 &= v_0(x), \\ v_1 &= v_1(x), \end{aligned}$$

and consider two assumptions:

1. $E(v_1|x) = E(v_0|x)$, [Wooldridge, Proposition 18.1]

2. $E(y_1|x,w) = E(y_1|x)$ and $E(y_0|x,w) = E(y_0|x)$.

The second of these is conditional mean independence, as we have already seen. The first assumption, $E(v_1|x) = E(v_0|x)$, implies that variation in x alters y_0 and y_1 in exactly the same way. For example, if our outcome of interest is earnings and x is experience, then an additional year of experience leads to the same change in earnings with treatment (perhaps training) as without.

- Under assumptions (1) and (2), we can show that

$$ATE = ATE_1,$$

and

$$E(y|w,x) = \mu_0 + \alpha w + g_0(x), \tag{4.2}$$

where $\alpha = (\mu_1 - \mu_0) = ATE$, and $g_0(x) = E(v_1|x) = E(v_0|x)$. A proof of this is provided in Section A1 in the appendix.

- The implication is that, once we have decided on a suitable functional form for $g_0(x)$, we can estimate the ATE by OLS. For example, if $g_0(x) = \eta_0 + x\beta_0$, the regression is

$$y = (\mu_0 + \eta_0) + \alpha w + x\beta_0 + \varepsilon, \tag{4.3}$$

where $x\beta_0$ can be interpreted as a control for self-selection into treatment.

- If we relax assumption (1), but continue to make assumption (2) and use linear specifications for g_0 and g_1:

$$\begin{aligned} E(v_1|x) &= g_1(x) = \eta_1 + x\beta_1, \\ E(v_0|x) &= g_0(x) = \eta_0 + x\beta_0, \end{aligned}$$

we can show that

$$E\left(y|w,x\right)=\mu_0+\alpha w+x\beta_0+w\left(x-\bar{x}\right)\delta,$$

where $\alpha=\left(\mu_1-\mu_0\right)$ and $\delta=\beta_1-\beta_0$. Again, I provide a proof in the appendix, Section A1. The latter equation can be estimated by means of OLS:

$$y=\mu_0+\alpha w+x\beta_0+w\left(x-\bar{x}\right)\delta+\varepsilon,$$

in which case

$$\begin{aligned} A\hat{T}E &= \hat{\alpha}, \\ A\hat{T}E_1 &= \hat{\alpha}+\left(\sum_{i=1}^{N}w_i\right)^{-1}\left(\sum_{i=1}^{N}w_i\left(x_i-\bar{x}\right)\hat{\delta}\right). \end{aligned}$$

- Notice that the regression in this case - where only assumption (2) is imposed - is more general than the one estimated when both assumptions (1) and (2) are imposed. The difference is that we now add interaction terms between w and $(x-\bar{x})$ on the right-hand side of the estimated equation. Notice also that, in general, $ATE\neq ATE_1$. So you see how equality between ATE and ATE_1 hinges on assumption (1).

4.1.2. Estimation by matching

- Estimation based on the **matching** involves matching treated and untreated individuals based on their observable characteristics x, and comparing how the outcome differs depending on treatment. As we have seen, **exact** matching involves comparing individuals for whom the values of x are **identical**. This estimator is rarely an option in practice. Why?

- With continuous variables in x, and/or many explanatory variables, we resort to inexact matching - instead of requiring the individuals across whom we compare outcomes to have identical values of x, we now require them to have **similar** values of x. In this section we discuss
  - how (inexact) matching works
  - how it is conceptually different from regression methods
  - what are the advantages and disadvantages, compared to other techniques

- When we have many x-variables, it will be difficult to match (even inexactly) on all of these simultaneously. Fortunately, there is a way around that, by matching on the **propensity score** instead. The reason matching on the propensity score is more attractive than matching on k different x-variables, is that the propensity score is a single (estimated) "variable" for each individual.

The propensity score Consider modelling the **likelihood of being treated** by means of a binary choice model (e.g. logit or probit):

$$\Pr\left(w_i = 1|x\right) = G\left(x\beta\right) \equiv p\left(x\right).$$

In the treatment literature, the function $p\left(x\right)$ is known as the **propensity score**. Wooldridge shows that ATE and ATE_1 can be written in terms of the propensity score (pp.615-617). Whilst interesting, the most useful property of the propensity score is probably in the context of estimating by matching, where the idea is to match individuals with similar propensity scores. A good discussion is provided in Angrist and Pischke, Chapter 3.3.2.

- **The Propensity Score Theorem**: *Suppose ignorability of treatment holds, so that conditional on x, w and (y_1, y_0) are independent. Then it must be that, conditional on the propensity score $p(x)$, w and (y_1, y_0) are independent.* See Angrist & Pischke pp. 80-81 for a proof (it's pretty trivial).

- This theorem says that you need only control for the probability of treatment itself.

- Matching by the propensity score can be thought of as follows. Suppose we choose a propensity score $p(x)$ at random, and suppose we select two individuals with the same propensity score, where the first individual receives treatment and the second does not. The expected difference in the observed outcomes for these two individuals is

$$\begin{aligned} & E(y|w = 1, p(x)) - E(y|w = 0, p(x)) \\ = \quad & E(y_1 - y_0|p(x)), \end{aligned}$$

  which is the ATE conditional on the propensity score, $ATE(x)$.

- Before we can do anything with the propensity scores, they need to be **estimated**. This is typically done by means of a logit or probit. After estimation (in Stata), the propensity scores can be obtained by typing *predict propscore, p*. In fact, we don't even have to do this - the Stata command pscore does this for us, as well as some basic analysis of its properties. We will have a look at this in the computer exercise.

- The basic idea behind the propensity score matching estimator is quite appealing. To estimate the counterfactual y_{0i} (i.e. the outcome that individual i, who was treated, would have recorded had s/he not been treated), use one or several observations in the (nontreated) control group that are similar to individual i, in terms of the propensity score.

- While this may sound relatively straightforward, there are a number of both conceptual and practical issues to keep in mind when implementing a propensity score matching estimator. We now turn to these.

Conditional mean independence.

- First of all, we know the assumption of conditional mean independence must hold (see above), otherwise we cannot interpret our estimates as average treatment effects. The practical implication of that is that you need a **complete set of variables** determining selection into treatment. That is, if your dataset does not contain the relevant variables determining selection, then your binary choice model (the first stage) will not generate useful propensity scores in this context, essentially because the propensity scores do not control fully for selection.

- Notice that, under pure randomization, no variable can explain treatment, and so in this case the pseudo-R-squared should be very close to zero.

- Of course, it's hard to know a priori what the right set of explanatory variables in the first stage are. Should draw on economic theory. The more you know about the process determining treatment, the more convincing is this particular identification strategy. Angrist & Pischke cite evidence suggesting that a logit model with a few polynomial terms in continuous covariates works well in practice, but note that some experimentation will be required in practice.

Common support.

- Now suppose that we have estimated the propensity scores by means of logit or probit. Remember that one of the cornerstones of matching estimators is that treated and nontreated individuals need to be comparable.[1]

- Suppose we find that there are a lot of treated observations with higher (lower) propensity scores than the maximum (minimum) propensity score in the control group. How do we match these treated observations? Because there are no observations in the control group that are similar to

[1] This is sometimes investigated formally by means of a **balancing test** (e.g. see the computer exercise, especially the results given by the pscore command). Essentially, for individuals with the same propensity scores, assignment to treatment should be random and not depend on the x characteristics. In other words, for a group of 'similar' individuals (in terms of their propensity scores) there should be no statistically significant difference in, say, the mean values of the x vector when we compare treated and nontreated observations. This can be tested by means of simple t-tests (see the pscore output, using the detail option).

these, matching will not be possible (extrapolation is not thought an option). Consequently all these treated observations that fall outside the **common support region** get dropped from the analysis.

- Figure 1 in Section 2 in the handout illustrates this, using the Ethiopian food aid data in Gilligan and Hoddinott. and focussing on the estimates reported in Table 3, column 1 in the paper (total consumption, food-for-work treatment). In this particular case, we only drop 31 out of 630 observations. Of course, in a different application the number of dropped treated observations may be much larger, which may lead to estimation (small-sample) problems.

- Also, notice that a conceptual issue arises here: we can never hope to estimate treatment effects on the treated outside the group of observations for which there is common support. Hence, the estimated effects should be interpreted as valid only for the sub-population of treated individuals for which there is support in the control group.

Finding the match and estimating the treatment effect If we are satisfied the propensity score is a good basis for matching nontreated and treated individuals, we are now ready to estimate the average treatment effect. The general formula for the matching ATE_1 estimator is

$$ATE_1^M = \frac{1}{N_T} \sum_{i \in \{w=1\}} \left(y_{1,i} - \sum_{j \in \{w=0\}} \phi(i,j) \, y_{0,j} \right),$$

where $\{w = 1\}$ is the set of treated individuals, $\{w = 0\}$ is the set of nontreated individuals (the control group), and $\phi(i,j)$ is a **weight**. Notice that $\sum_{j \in \{w=0\}} \phi(i,j) \, y_{0,j}$ is interpretable as the counterfactual for individual i, i.e. his or her outcome had s/he not been treated. This counterfactual is thus calculated as a weighted average of outcomes in the control group.

The issue now is how to calculate the weight. There are several possibilities.

- The simplest one is **nearest-neighbour matching**. This involves finding, for each treated individual in the data, the untreated observation with the most similar propensity score. That observation is then given a weight equal to one, and all other observations get zero weights. Once the data have been set up accordingly, one would then use the above general formula for the matching ATE_1.

- Another method - which is the one used by Gilligan and Hoddinott - is **kernel matching**. In this case

$$\phi(i,j) = \frac{K\left(p(x)_j - p(x)_i\right)}{\sum_{j=1}^{N_{C,i}} K\left(p(x)_j - p(x)_i\right)},$$

where K is a kernel function.

- A kernel function is an important tool in nonparametric and semiparametric analysis. K is a symmetric density function which has its maximum when its argument is zero, and decreases as the absolute argument of K increases. In other words, if $p(x)_j = p(x)_i$ in the formula above, then the value of K is relatively high, whereas if $p(x)_j$ is very different from $p(x)_i$ then K will be close to, or equal to, zero. You see how this gives most weight to observations in the control group for which the propensity scores are close to that of the treated individual i. If you want to learn more about kernel functions, I warmly recommend the book by Adonis Yatchew (2003), *Semiparametric Regression for the Applied Econometrician*, Cambridge University Press.

- To better understand how kernel matching works, now focus on the calculation of the counterfactual for the ith treated individual. By definition, the treatment effect for individual i is

$$y_{1,i} - \sum_{j \in \{w=0\}} \phi(i,j)\, y_{0,j},$$

where y_{ii} is observed in the data. The snag is that we need to compute the counterfactual of individual i, namely $y_{0,i}$. This is calculated as

$$\sum_{j \in \{w=0\}} \phi(i,j)\, y_{0,j}.$$

Section 3 in the handout provides details on how this works, using the Hoddinott-Gilligan food aid data from Ethiopia.

- Hence the kernel matching estimator of the average treatment effect for the treated is expressed as

$$\begin{aligned} ATE_1^M &= \frac{1}{N_T} \sum_{i \in \{w=1\}} \left(y_{1,i} - \Lambda \right), \\ \Lambda &= \sum_{j \in \{w=0\}} \frac{K\left(p\left(x\right)_j - p\left(x\right)_i\right)}{\sum_{j=1}^{N_{C,i}} K\left(p\left(x\right)_j - p\left(x\right)_i\right)} y_{0,j} \end{aligned}$$

  see also equation (5) in Gilligan and Hoddinott.

- Other matching methods include radius matching and stratification, but I leave it to you to follow up on the details for these methods, if you are interested.

- Somewhat disturbingly, different matching methods can give rather different results, especially if the sample size is small. For example, in Table 3, column 1, Gilligan and Hoddinott (2007) reports an estimate of ATE_1 equal to 0.215, which is significant at the 10% level. This estimate is based on Kernel weights. If we use a nearest neighbour approach, this estimate changes to 0.301 (still significant 10% level). There is, alas, little formal guidance as to which method to use.

Estimating standard errors using bootstrapping.

- Once we have calculated the average treatment effects of interest, we want estimates of the associated standard errors in order to do inference (e.g. we want to know if the estimated average treatment effect is significantly different from zero or not).

- In the second stage, however, the standard errors are typically computed assuming that the propensity score is not measured with sampling error (this is the case with Stata's psmatch2). Recall that the propensity score was estimated by means of a binary choice model, and so each recorded value of the propensity score is just an **estimate** of the score. That is, while the recorded propensity score for individual i may be 0.343, this will only be the true value with certainty if the parameter estimates in the first stage are in fact equal to the true values with certainty (in other words, that the standard errors in stage 1 are very close to zero). Of course, this will not generally be true, because there is sampling error in the first stage. The true propensity score for individual i may be higher or lower than 0.343. The standard errors in the second stage need to take into account the fact that the propensity scores are estimated with some degree of uncertainty. One popular and reasonably simple way of doing this is by means of **bootstrapping** (more on this in the computer exercise).

Very briefly, the bootstrapping process for the propensity score matching estimator is as follows:

1. Draw a new sample with replacement from the existing sample. Some individuals in the original sample will be included several times in this new sample, others once, others not at all.

2. Based on the new sample, estimate the binary choice model and the propensity scores. Notice that these estimates will be slightly different from those based on the original sample. Indeed, this reflects the effect of sampling error.

3. Estimate the average treatment effect in the second stage. Store this value.

Repeat this process, say, 100 times. You now have 100 'estimates' of the average treatment effect. Now calculate the standard deviation based on those numbers. That standard deviation is your bootstrapped standard error.

The bootstrapping approach is very general and very useful whenever it is hard to calculate standard errors using analytical formulae.

4.1.3. Regression or matching?

- The regression approach is easy to implement and interpret.
  - But regressions happily extrapolate between observations in the data, and so ignore the concept of common support. The idea that you need to compare the outcomes of two individuals with similar characteristics, except one was treated and the other wasn't, is not central to the regression approach. Suppose we write the regression as

    $$y_i = \gamma + \alpha w_i + \beta x_i + \varepsilon_i.$$

    You might say α is being estimated 'controlling for x', but it may be that most high values of x are associated with $w = 1$, and most low values of x are associated with $w = 0$. Suppose we want to calculate the (conditional) treatment effect $E\left(y_{1i} - y_{0i} | x_i \text{ is 'high'}\right)$. For treated observations, we observe y_{1i} in the data, but need the counterfactual y_{0i}. This counterfactual is thus the hypothetical value of the outcome variable under a) nontreatment; and b) a high value of x. The problem is that are very few observations in the control group with x high, and so the expected counterfactual $E\left(y_{0i} | x_i \text{ is 'high'}\right)$ is mostly based on combining observations on outcomes for which $\{w = 1, x \text{ high}\}$ and observations on outcomes for which $\{w = 0, x \text{ low}\}$. But whether this gives a good estimate of $E\left(y_{0i} | x_i \text{ is 'high'}\right)$ is uncertain, and hinges on the extrapolation not being misleading. And, as you know, relying on extrapolation not being misleading is always awkward.
  - Regressions also impose a functional form relationship between treatment and outcomes, because we need to write down the precise form of the specification in order to estimate the parameters by regression. But functional form assumptions are often arbitrary and can lead to misleading results.
- The matching estimator, in contrast to the regression approach, estimates treatment effects using only observations in the region of common support. There is thus no extrapolation. Furthermore,

there are no functional form assumptions in the second stage, which is attractive.

  - But we can never hope to estimate treatment effects on the treated outside the region of common support.
  - At least in small samples, it is often the case that estimated treatment effects change quite a lot when we change the matching method (e.g. Hoddinott & Gilligan, kernel matching vs. nearest neighbor matching).
  - Two-stage procedure means the standard errors in the second stage are unreliable. So more work is required - bootstrapping is often used.

- Moreover, as noted by Hahn (1998), cited in Angrist-Pischke (2009), the asymptotic standard errors associated with propensity score matching estimator will be **higher** than those associated with an estimator matching on any covariate that explains outcomes (regardless of it turns up in the propensity score or not). Angrist and Hahn (2004), also cited in Angrist-Pischke, note that Hahn's argument is less compelling in small samples.

4.2. Selection on unobservables

- We have concentrated on calculating average treatment effects when there is selection on observables. When there is selection on **unobservables**, however, the methods that we have reviewed will not yield consistent estimates of average treatment effects. The reason, essentially, is that the assumption of ignorability of treatment no longer holds: conditional on x, w and (y_1, y_0) are no longer independent, because there is some unobserved term that affects both selection into treatment and the potential outcomes.

- In the case where the relevant unobserved variable is time invariant, we may be able to use longitudinal (panel) data and remove the unobserved term by differencing. The most common estimator in this situation is known as the **Difference-in-Differences** estimator**.**

- Alternatively, we may be able to use **instrumental variables** to estimate the average treatment effect(s) consistently. In the context of the regression approach, this can be done using a standard 2SLS estimator, provided valid and informative instruments exist. We discuss this next time.

4.2.1. Difference-in-Differences

- Reference: Angrist-Pischke, Chapter 5.

- If we have data from two (or more) periods per individual, one possible way around the problem posed by selection on unobservables is to investigate if changes in the outcome variable over time are systematically related to treatment.

- Suppose treatment occurs between the first and the second time period for which we have data. For example, suppose our dataset contains information on individuals that have received training between time periods 1 and 2, and suppose we have data on their earnings in periods 1 and 2. Suppose the dataset also contains information on a **control group** of individuals, that are observed over the same time periods, but who did not receive any treatment.

- The difference-in-differences estimator, in its simplest form, is thus defined as the difference in the change in average earnings for the treatment group and the control group:

$$ATE_1 = \left(\bar{y}_a^T - \bar{y}_b^T\right) - \left(\bar{y}_a^U - \bar{y}_b^U\right),$$

where

$$\begin{aligned} \bar{y}_a^U &= \text{average earnings for nontreated after treatment} \\ \bar{y}_b^U &= \text{average earnings for nontreated before treatment.} \end{aligned}$$

- What about selection into treatment? Consider the following equations:

$$\begin{aligned} y_{it}^U &= \phi_i + \delta_t + \varepsilon_{it} \quad \text{(no treatment)} \\ y_{it}^T &= y_{it}^U + \alpha, \qquad\qquad \text{(treatment)} \end{aligned}$$

for $t = a, b$ (after and before), where ϕ_i is an individual-specific, possibly **unobserved**, time invariant fixed effect (thus a source of heterogeneity across individuals in observed outcomes - you could think of this as x_i), δ_t is a dummy variable equal to 1 in the time period after treatment and zero in the period before treatment ($\delta_a = 1, \delta_b = 0$), ε_{it} is a zero-mean residual, and α is the treatment effect (for everyone, as well as for the treated).

- Provided ε_{it} is **uncorrelated** with treatment status, it follows that

$$\begin{aligned} \left(\bar{y}_a^T - \bar{y}_b^T\right) &= \alpha + \delta_a, \\ \left(\bar{y}_a^U - \bar{y}_b^U\right) &= \delta_a, \end{aligned}$$

thus α is the treatment effect:

$$ATE_1 = \alpha.$$

Notice that even if ϕ_i is correlated with treatment, this would not lead to bias. And notice also that ϕ_i may be a continuous variable, or even a large set of time invariant variables (which would make exact matching infeasible), and yet we can identify the treatment effect simply by differencing the data. In effect, we are exploiting the time dimension of the data to define the counterfactual.

- But also notice that we require the source of 'selection bias' to be constant over time - otherwise the assumption of ignorability of treatment does not hold. If ε_{it} - which is time varying - is correlated with treatment, then the above DiD estimate of the treatment effect will be biased. In such a case we need to do more work.
- See section 4 in the appendix for some illustrations.
- To see this, return to the DiD estimator:

$$\begin{aligned} y_{it}^U &= \phi_i + \delta_t + \varepsilon_{it} \quad \text{(no treatment)}, \\ y_{it}^T &= y_{it}^U + \alpha, \qquad\qquad \text{(treatment)}, \end{aligned}$$

hence

$$\begin{aligned} y_{it} &= w_{it}\left(y_{it}^U + \alpha\right) + (1 - w_{it})\, y_{it}^U, \\ y_{it} &= \phi_i + \alpha w_{it} + \delta_t + \varepsilon_{it}, \end{aligned}$$

and so, in differences,

$$\Delta y_{it} = \alpha \Delta w_{it} + \Delta \delta_t + \Delta \varepsilon_{it},$$

which becomes

$$\Delta y_{it} = \alpha w_{it} + \delta_t + \Delta \varepsilon_{it},$$

if there are only two time periods (before and after), so that $\delta_a = 1$ and $\delta_b = 0$, and treatment happens after time b but before time a.

- As already noted, heterogeneity in the form of individual fixed effects will not bias our estimate of α. But non-zero correlation between $\Delta\varepsilon_{it}$ and treatment will bias the results.

- To counter this, we can **add observable variables** as controls to the specification. Of course, this set of control variables needs to fully control for selection into treatment, otherwise ignorability of treatment does not hold, in which case the estimates will be biased. We will revisit this issue in the computer exercise.

- Alternatively, we can combine DiD with matching. Recall the Kernel matching estimator in levels:

$$\begin{aligned} ATE_1^M &= \frac{1}{N_T}\sum_{i\in\{w=1\}}\left(y_{1,i}-\Lambda\right), \\ \Lambda &= \sum_{j\in\{w=0\}}\frac{K\left(p\left(x\right)_j-p\left(x\right)_i\right)}{\sum_{j=1}^{N_{C,i}}K\left(p\left(x\right)_j-p\left(x\right)_i\right)}y_{0,j} \end{aligned}$$

  We can take this one step further and write down the propensity score difference-in-differences estimator as follows:

$$ATE_1^{DIDM}=\frac{1}{N_T}\sum_{i\in\{w=1\}}\left(\left(y_{1it}-y_{1i,t-1}\right)-\tilde{\Lambda}\right),$$

  where

$$\tilde{\Lambda}=\sum_{j\in\{w=0\}}\frac{K\left(p\left(x\right)_j-p\left(x\right)_i\right)}{\sum_{j=1}^{N_{C,i}}K\left(p\left(x\right)_j-p\left(x\right)_i\right)}\left(y_{0it}-y_{0i,t-1}\right)$$

  This is what Gilligan and Hoddinott (2007) are using, and we will see how this estimator works in the computer exercise.

Appendix 1

A1. Proofs related to the regression approach

Consider

1. $E\left(v_1|x\right) = E\left(v_0|x\right)$,
2. $E\left(y_1|x,w\right) = E\left(y_1|x\right)$ and $E\left(y_0|x,w\right) = E\left(y_0|x\right)$.

a) Under Assumptions (1) and (2) (page 11), it follows that

- $ATE = ATE_1$,
- $E\left(y|w,x\right) = \mu_0 + \alpha w + g_0\left(x\right).$

Proof. The starting point is the switching regression

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

We know that $ATE = \left(\mu_1 - \mu_0\right)$. To prove that $ATE = ATE_1$, notice that

$$\begin{aligned} E\left(y_1|x,w\right) &= \mu_1 + E\left(v_1|x,w\right), \\ E\left(y_1|x\right) &= \mu_1 + E\left(v_1|x\right), \end{aligned}$$

and $E\left(y_1|x,w\right) = E\left(y_1|x\right)$ by assumption (2). Along similar lines,

$$\begin{aligned} E\left(y_0|x,w\right) &= \mu_0 + E\left(v_0|x,w\right), \\ E\left(y_0|x\right) &= \mu_0 + E\left(v_0|x\right), \end{aligned}$$

and $E\left(y_0|x,w\right)=E\left(y_0|x\right)$ by assumption (2). We therefore have

$$\begin{aligned} E\left(y_1|x,w\right)-E\left(y_0|x,w\right) &= \mu_1+E\left(v_1|x\right)-\mu_0-E\left(v_0|x\right) \\ E\left(y_1|x,w\right)-E\left(y_0|x,w\right) &= \mu_1-\mu_0, \end{aligned}$$

by assumption (1). The right-hand side of the last equation is independent of x, hence we can write

$$E\left(y_1|x,w\right)-E\left(y_0|x,w\right)=E\left(y_1|w\right)-E\left(y_0|w\right),$$

where the right-hand side is the ATE_1 by definition. It follows immediately that $ATE_1=\mu_1-\mu_0=ATE$. This concludes the first part of the proof.

To show that $E\left(y|w,x\right)=\mu_0+\alpha w+g_0\left(x\right)$, start from

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

and take expectations

$$\begin{aligned} E\left(y|w,x\right) &= \mu_0+\left(\mu_1-\mu_0\right)w+E\left(v_0|w,x\right)+w\left(E\left(v_1|w,x\right)-E\left(v_0|w,x\right)\right) \\ &= \mu_0+\left(\mu_1-\mu_0\right)w+E\left(v_0|x\right)+w\left(E\left(v_1|x\right)-E\left(v_0|x\right)\right) \\ &= \eta_0+\mu_0+\left(\mu_1-\mu_0\right)w+g_0\left(x\right), \end{aligned}$$

where $\eta_0+g_0\left(x\right)=E\left(v_1|x\right)=E\left(v_0|x\right)$. This concludes the second part of the proof. ■

b) Under Assumption (2) (page 11), it follows that

- $ATE \neq ATE_1$,
- $E\left(y|w,x\right) = \mu_0 + \alpha w + w\left[g_1\left(x\right) - g_0\left(x\right)\right].$

Proof. In the previous proof it was shown that

$$E\left(y_1|x,w\right) - E\left(y_0|x,w\right) = \mu_1 + E\left(v_1|x\right) - \mu_0 - E\left(v_0|x\right).$$

Now, however, $E\left(v_1|x\right) = E\left(v_0|x\right)$ is not imposed, and so the right-hand side is not independent of x. Therefore, we do not obtain $ATE = ATE_1$.

Further, use the result derived above that

$$E\left(y|w,x\right) = \mu_0 + \left(\mu_1 - \mu_0\right)w + E\left(v_0|x\right) + w\left(E\left(v_1|x\right) - E\left(v_0|x\right)\right),$$

and define

$$\begin{aligned} E\left(v_1|x\right) &= \eta_1 + g_1\left(x\right), \\ E\left(v_0|x\right) &= \eta_0 + g_0\left(x\right), \end{aligned}$$

so that

$$E\left(y|w,x\right) = \mu_0 + \left(\mu_1 - \mu_0\right)w + g_0\left(x\right) + w\left(g_1\left(x\right) + \eta_1 - g_0\left(x\right) - \eta_0\right).$$

Use a linear specification for g_0, g_1:

$$\begin{aligned} g_1\left(x\right) &= \eta_1 + x\beta_1, \\ g_0\left(x\right) &= \eta_0 + x\beta_0. \end{aligned}$$

Because the unconditional expectation of v_1 and v_0 is zero, we have

$$\begin{aligned} E\left(v_1\right) &= E_x E\left(v_1|x\right) = \eta_1 + \bar{x}\beta_1 = 0, \\ E\left(v_0\right) &= E_x E\left(v_0|x\right) = \eta_0 + \bar{x}\beta_0 = 0, \end{aligned}$$

thus

$$\begin{aligned} \eta_1 &= -\bar{x}\beta_1, \\ \eta_0 &= -\bar{x}\beta_0. \end{aligned}$$

Now it follows that

$$E\left(y|w,x\right) = \mu_0 + \left(\mu_1 - \mu_0\right) w + x\beta_0 + w\left(x\beta_1 - \bar{x}\beta_1 - x\beta_0 - \bar{x}\beta_0\right),$$

or

$$E\left(y|w,x\right) = \mu_0 + \left(\mu_1 - \mu_0\right) w + x\beta_0 + w\left(x - \bar{x}\right)\left(\beta_1 - \beta_0\right),$$

or, with a more tidy notation,

$$E\left(y|w,x\right) = \mu_0 + \alpha w + x\beta_0 + w\left(x - \bar{x}\right)\delta.$$

■

PhD Programme: Applied Econometrics
Department of Economics, University of Gothenburg
Appendix 2: Lecture 11
Måns Söderbom

Treatment Evaluation

1. Exact matching: A simple example

Suppose your dataset looks like this:

id	y	w	x
1	4	0	0
2	6	1	0
3	5	0	0
4	8	1	0
5	2	0	1
6	5	1	1
7	2	0	1

How would you estimate the ATE and the ATE$_1$? The formula for ATE is as follows

$$ATE = \frac{1}{N}\sum_{i=1}^{N}\left(r_1(x_i) - r_0(x_i)\right),$$

where

$$E(y_1 \mid x, w=1) = r_1(x)$$
$$E(y_0 \mid x, w=0) = r_0(x).$$

The estimated ATE is written

$$A\hat{T}E = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{r}_1(x_i) - \hat{r}_0(x_i)\right),$$

where the ^ indicate that the associated quantities are estimated. All we need to do, then, is

1) estimate the functions r_1 and r_0;
2) plug in the actual values of *x* into these estimated functions, for each individual *i* in the data, and obtain *N* different terms $\left(\hat{r}_1(x_i) - \hat{r}_0(x_i)\right)$;
3) calculate the sample average of the quantities computed in (2).

In this particular example, x can take only two values, 0 or 1. In this case there are only four **cells** in the data - i.e. there are only four different combinations of {x,w}. Hence we need to estimate only four quantities: $r_1(0)$, $r_0(0)$, $r_1(1)$ and $r_0(1)$. With these data:

$$\hat{r}_1(0) = (6+8)/2 = 7$$
$$\hat{r}_0(0) = (4+5)/2 = 4.5$$
$$\hat{r}_1(1) = 5/1 = 5$$
$$\hat{r}_0(1) = (2+2)/2 = 2$$

This is quite neat in the sense that none of these predications are obtained by extrapolation or interpolation in the data: **only** observations where {w,x} are exactly as conditioned in the expectation are used to estimate the latter. That is, to calculate $r_1(0)$, we **only** use observations for which {w=1, x=0}. The beauty of this is that we don't have to specify a functional form relationship between the expected value of y and {w,x}.

We can now add three columns to the data above, showing the estimated functions r_1 and r_0, given x, and the difference $(\hat{r}_1(x_i) - \hat{r}_0(x_i))$:

id	y	w	x	$\hat{r}_1(x_i)$	$\hat{r}_0(x_i)$	$\hat{r}_1(x_i) - \hat{r}_0(x_i)$
1	4	0	0	7	4.5	2.5
2	6	1	0	7	4.5	2.5
3	5	0	0	7	4.5	2.5
4	8	1	0	7	4.5	2.5
5	2	0	1	5	2	3
6	5	1	1	5	2	3
7	2	0	1	5	2	3

And now we can estimate the ATE simply by calculating the average of the numbers in the last column:

$$A\hat{T}E = 2.7143$$

To get an estimate of the average treatment effect for the treated, we use the following formula:

$$ATE_1 = \sum_{i=1}^{N} w_i (r_1(x_i) - r_0(x_i)) \left(\sum_{i=1}^{N} w_i \right)^{-1}$$

which essentially means discarding all non-treated observations when computing the average:

$$ATE_1 = (2.5 + 2.5 + 3)(3)^{-1} = 2.6667.$$

Finally, let's illustrate how this links to the regression approach. Because x takes only two values, there are only four categories - as defined by the values {w, x} - of observations in the data. Therefore, the following regression is completely unrestrictive in terms of the functional form relationship between {w, x} and the outcome variable y:

$$y_i = \beta_0 + \beta_1 w_i + \beta_2 x_i + \beta_3 \left(w_i \cdot x_i\right) + \varepsilon_i$$

Notice that

$$\begin{aligned} r_1(0) &= \beta_0 + \beta_1 \\ r_0(0) &= \beta_0 \\ r_1(1) &= \beta_0 + \beta_1 + \beta_2 + \beta_3 \\ r_0(1) &= \beta_0 + \beta_2 \end{aligned}$$

If I estimate this regression using the data above I obtain the following results:

```
      Source |       SS       df       MS              Number of obs =       7
-------------+------------------------------           F(  3,     3) =   10.09
       Model |  25.2142857     3   8.4047619           Prob > F      =  0.0447
    Residual |         2.5     3  .833333333           R-squared     =  0.9098
-------------+------------------------------           Adj R-squared =  0.8196
       Total |  27.7142857     6  4.61904762           Root MSE      =  .91287

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           w |        2.5   .9128709     2.74   0.071    -.4051627    5.405163
           x |       -2.5   .9128709    -2.74   0.071    -5.405163    .4051627
          wx |         .5   1.443376     0.35   0.752    -4.093466    5.093466
       _cons |        4.5   .6454972     6.97   0.006      2.44574     6.55426
------------------------------------------------------------------------------
```

(abstract from everything here except the point estimates). You can now confirm that this gives exactly the same estimates of ATE and ATE$_1$ as with the previous approach.

In cases where there are many x-variables, and/or the x-variable(s) can take many different values, it will be impractical to calculate the expected values of y for each possible combination of {w,x} in the data. In such cases we can use regression instead.

2. Propensity score matching: Food aid in Ethiopia

Table 1

> probit pwhh $PW_Xvars if psmpwsamp==1 & dlrconsae56~=.;

```
Probit regression                                 Number of obs   =        630
                                                  LR chi2(33)     =     176.20
                                                  Prob > chi2     =     0.0000
Log likelihood = -326.45198                       Pseudo R2       =     0.2125

------------------------------------------------------------------------------
        pwhh |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 dlrconsae45 |   .0901913   .0839291     1.07   0.283    -.0743067    .2546894
 dlrconsae34 |   .1136245   .0955152     1.19   0.234    -.0735818    .3008307
 dlrconsae23 |   .0954875   .0878546     1.09   0.277    -.0767044    .2676794
pwag_dw~1564 |  -.0058935   .0137676    -0.43   0.669    -.0328774    .0210905
pwag_dw_n014 |  -.0063456   .0088594    -0.72   0.474    -.0237096    .0110185
pwag_dw_n6~p |   .0212255    .023502     0.90   0.366    -.0248375    .0672885
 headed5_any |  -.0868071   .1794755    -0.48   0.629    -.4385725    .2649584
   lheadage5 |  -.5859348   .2240272    -2.62   0.009     -1.02502   -.1468497
    femhead5 |  -.2171898   .1447859    -1.50   0.134     -.500965    .0665854
    not_able |  -1.305092   .2411725    -5.41   0.000    -1.777781    -.832402
   lhhsize02 |   .1271798   .1346558     0.94   0.345    -.1367407    .3911002
   depratio5 |   .0023378    .048843     0.05   0.962    -.0933927    .0980683
   ownttlld5 |  -.1186871   .2271286    -0.52   0.601     -.563851    .3264767
 ownttlld5_2 |   .0731684    .059554     1.23   0.219    -.0435553    .1898922
   pwc_hhmet |   .2327307   .1587953     1.47   0.143    -.0785024    .5439637
     drt9395 |   .1139127   .1482278     0.77   0.442    -.1766085    .4044338
  death_9902 |   .0117345   .1401814     0.08   0.933    -.2630161     .286485
illmale_9902 |  -.3385186   .2272103    -1.49   0.136    -.7838426    .1068054
illfema~9902 |   .0528919   .2314581     0.23   0.819    -.4007576    .5065414
   born_here |  -.1909949   .1524115    -1.25   0.210    -.4897161    .1077262
fathermoth~c |   .3191964   .1322201     2.41   0.016     .0600497    .5783432
     n_iddir |   -.065811   .1195579    -0.55   0.582    -.3001403    .1685183
netsize_less |   .0268816   .1432336     0.19   0.851     -.253851    .3076142
netsize_more |  -.1792897   .1431511    -1.25   0.210    -.4598606    .1012812
total_netw~s |  -.0052962   .0052219    -1.01   0.310    -.0155309    .0049385
         pa1 |  -.2871679   .2938195    -0.98   0.328    -.8630436    .2887078
         pa3 |  -.6055412   .3740617    -1.62   0.105    -1.338689    .1276063
         pa6 |   .5688213   .3297202     1.73   0.084    -.0774183    1.215061
         pa8 |  -1.018245   .3376892    -3.02   0.003    -1.680104   -.3563865
         pa9 |   -.071929   .7028182    -0.10   0.918    -1.449427    1.305569
        pa13 |   .1331518   .3798018     0.35   0.726    -.6112462    .8775497
        pa15 |  -.6023727   .3094828    -1.95   0.052    -1.208948    .0042025
        pa16 |  -.7053323   .3901619    -1.81   0.071    -1.470036     .059371
       _cons |   2.670589   .9939073     2.69   0.007     .7225667    4.618612
------------------------------------------------------------------------------


predict pwhh_h if e(sample)==1, p;
(709 missing values generated)

. keep if e(sample)==1;
(709 observations deleted)

. sum pwhh_h if pwhh==0;

    Variable |       Obs        Mean    Std. Dev.       Min        Max
```

```
-------------+--------------------------------------------------------
      pwhh_h |       232    .4748997    .2175047   .0191887   .9705976

. local min=r(min);

. local max=r(max);

. scatter pwhh pwhh_h, xline(`min' `max');

. disp `min';
.01918869

. disp `max';
.97059757

/* NOTE: min,max determine the common support. Treated observations with
pscores outside this region are discarded from the analysis.

. /* treated on common support */
> count if pwhh==1 & pwhh_h<=`max' & pwhh_h>=`min';
  367

. /* nontreated on common support */
> count if pwhh==0 ;
  232

/* So a total of 599 observations on the common support. */
```

Figure 1: The region of common support

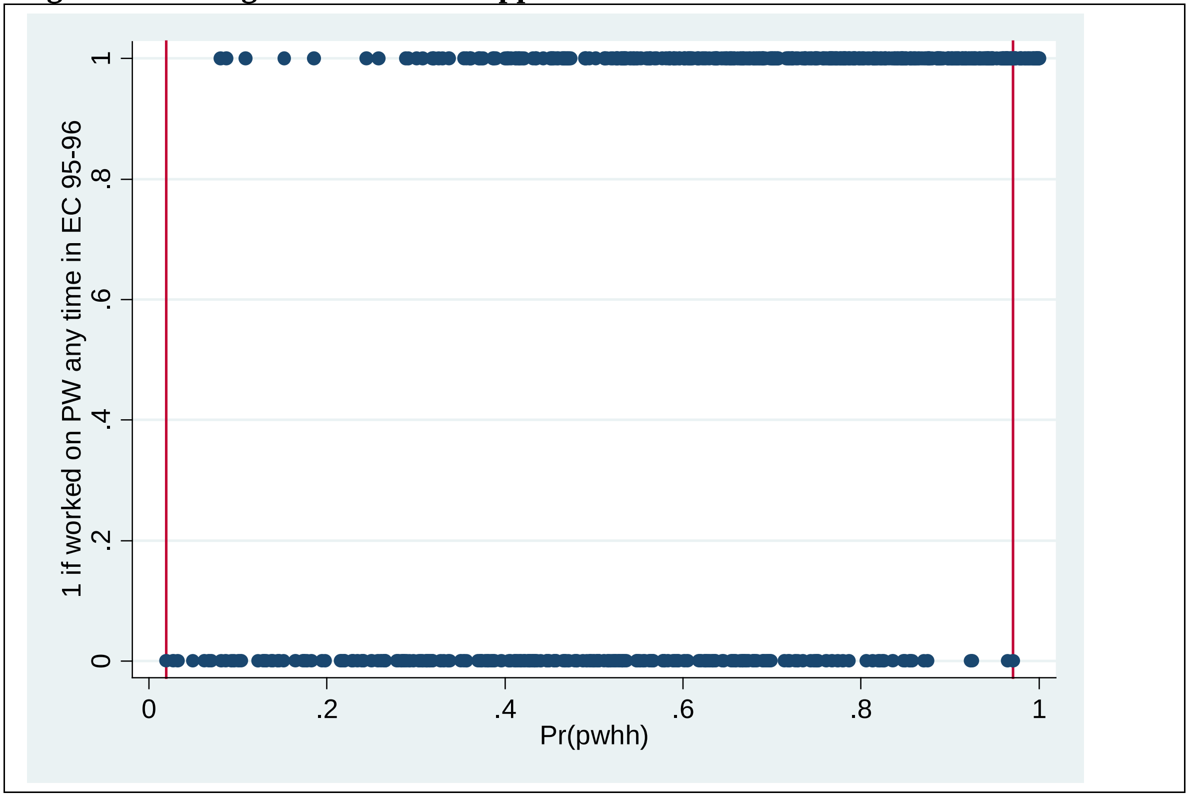
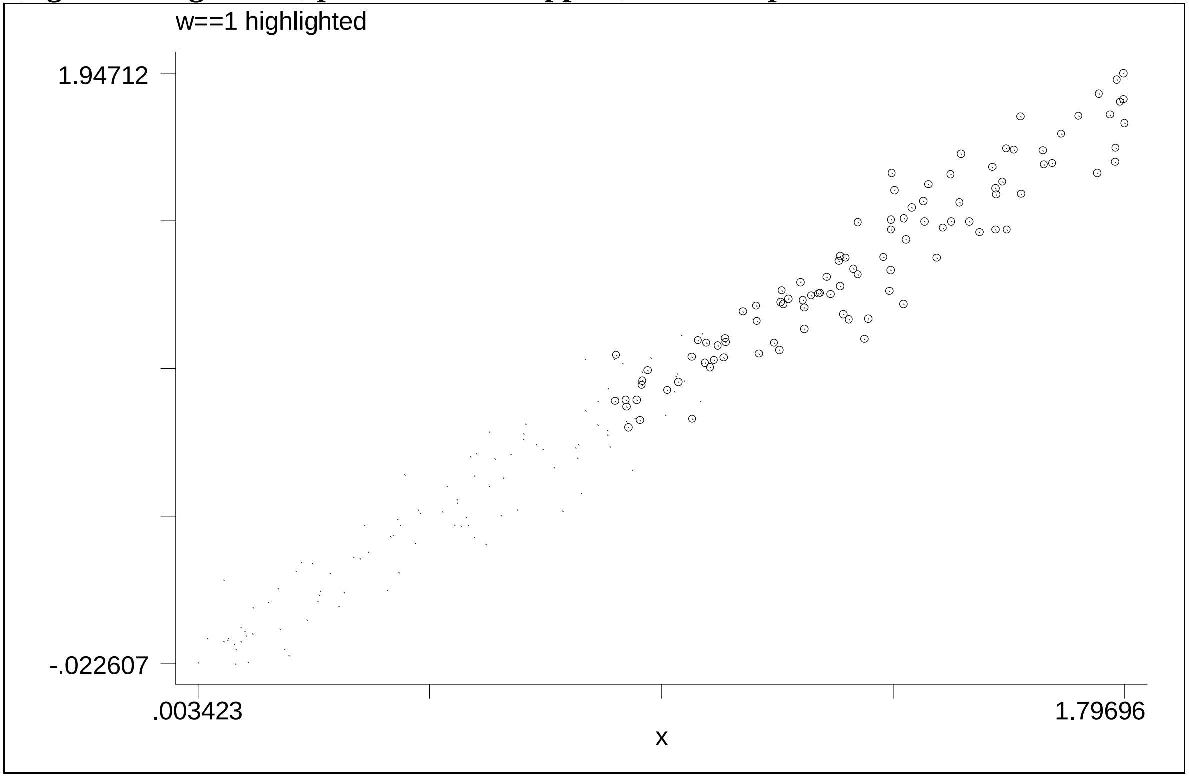


Figure 2: Regression, poor common support, and extrapolation

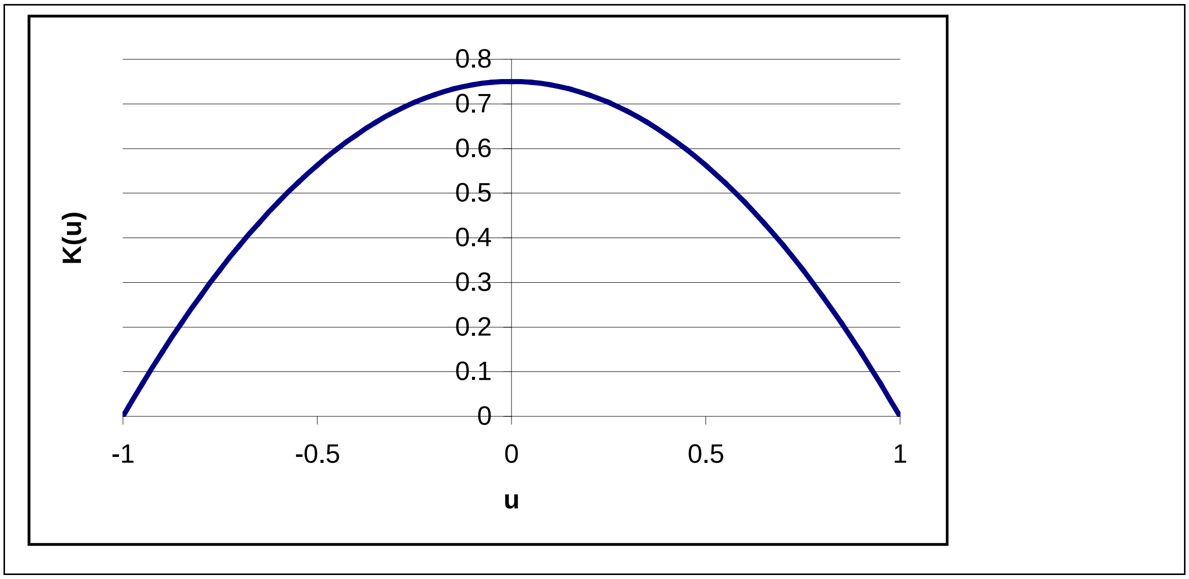


Note: Circled observations are treated, non-circled observations are nontreated. There are very few nontreated observations for which x is high.

3. Calculating the counterfactual for a treated observation using kernel matching: Illustration

The Stata command psmatch2 uses an Epanechnikov kernel. The Epanechnikov density function is equal to $0.75\left(1-u^2\right)$, where *u* takes values between -1 and 1 (for values of u outside this range, the density is zero). The density function looks as follows:

Figure 3: The Epanechnikov distribution



With kernel matching, recall that the weight for individual *i* is equal to

Notice that the argument of *K* is the difference between the propensity score of individual *i* (the treated individual) and the propensity score of individual *j*.

Now take a look at the data in Table (taken from Hoddinott and Gilligan, 2007), where I have computed the propensity score, and sorted the data from the lowest to the highest pscore value:

Table 2: Propensity scores and kernel weighting

pscore	pwhh	K	dlrconsae56	weight	weight x dlrconsae56 for matched obs only	Estimated counterfactual
0.0192	0	.	1.645944	.		
0.0271	0	0.1633	0.1656	0.0193	0.003196	
0.0323	0	0.2729	0.9741	0.0323	0.031463	
0.0496	0	0.555	0.4457	0.0656	0.029238	
0.0623	0	0.6833	0.6962	0.0808	0.056253	
0.0678	0	0.7181	0.5031	0.0849	0.042713	
0.0705	0	0.7305	2.5273	0.0864	0.218359	
0.0802	1	.	0.041	.		0.071846
0.0814	0	0.7497	-0.4217	0.0886	-0.03736	
0.0864	0	0.7419	-1.0075	0.0877	-0.08836	
0.0868	1	.	0.9315	.		
0.0927	0	0.7171	-0.176	0.0848	-0.01492	
0.0957	0	0.6999	-0.2276	0.0827	-0.01882	
0.1007	0	0.6625	0.2748	0.0783	0.021517	
0.1036	0	0.6354	-0.4609	0.0751	-0.03461	
0.1087	1	.	-1.7197	.		
0.1227	0	0.3735	-1.0766	0.0442	-0.04759	
0.1286	0	0.2608	1.1565	0.0308	0.03562	
0.1303	0	0.2266	-2.3975	0.0268	-0.06425	
0.132	0	0.1911	-2.2201	0.0226	-0.05017	
0.1379	0	0.0566	-1.0091	0.0067	-0.00676	
0.1393	0	0.0206	-1.5242	0.0024	-0.00366	
0.1451	0	.	0.9553	.		
0.1467	0	.	0.6368	.		
SUM				1.000	0.071846	

Suppose now we want to calculate the counterfactual of the first treated individual in the data, i.e. the shaded observation. I see that his value of dlrconsae56 (which in this context is his y_1) is equal to 0.0410.

- First, I calculate values of *K* for all observations in the **control group.** To be able to do so, I need to define the 'bandwidth'. I set this to 0.06, which is the default in psmatch2). These values are shown in the third (K) column. Notice that observations in the control group that have values of the propensity score close to 0.0802 get a relatively high value of K.
- I proceed by calculating the weights for the observations in the control group, using the formula

This gives me the values shown in the ‘weight’ column. Notice that they will sum to one.

- I then obtain the weighted average of dlrconsae56 for the individuals in the control group, using these weights. That is my estimate of the counterfactual for the treated individual here. That value turns out to be 0.0718.
- Thus, the treatment effect for this individual is 0.041-0.0718 = -0.0308.
- To get the average treatment effect for the treated, I proceed as above for each treated individual, and then calculate the average of the treatment effects. This gives me an estimate equal to 0.21496, which is the number reported by Hoddinott & Gilligan.
- If you were using a nearest neighbour approach instead of kernel matching, what would the counterfactual be?

4. Ignorability in the cross-section and in first differences

Figure 4: Evaluation with panel and cross-section data (ignorability holds)

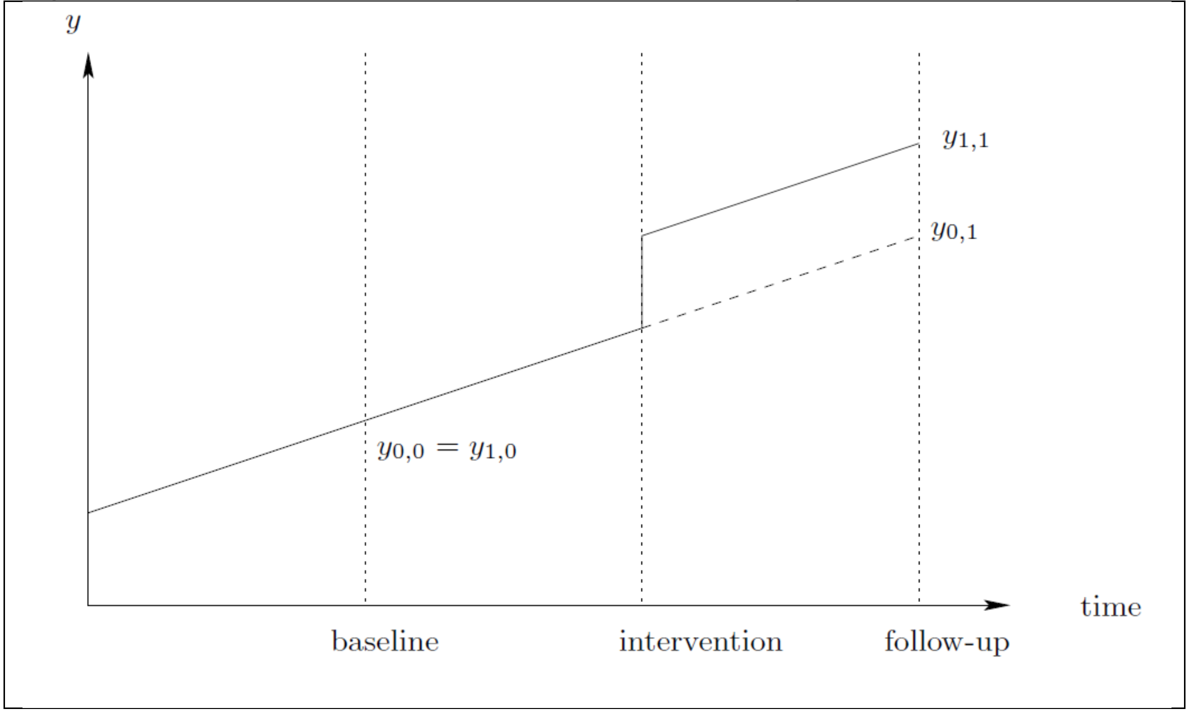
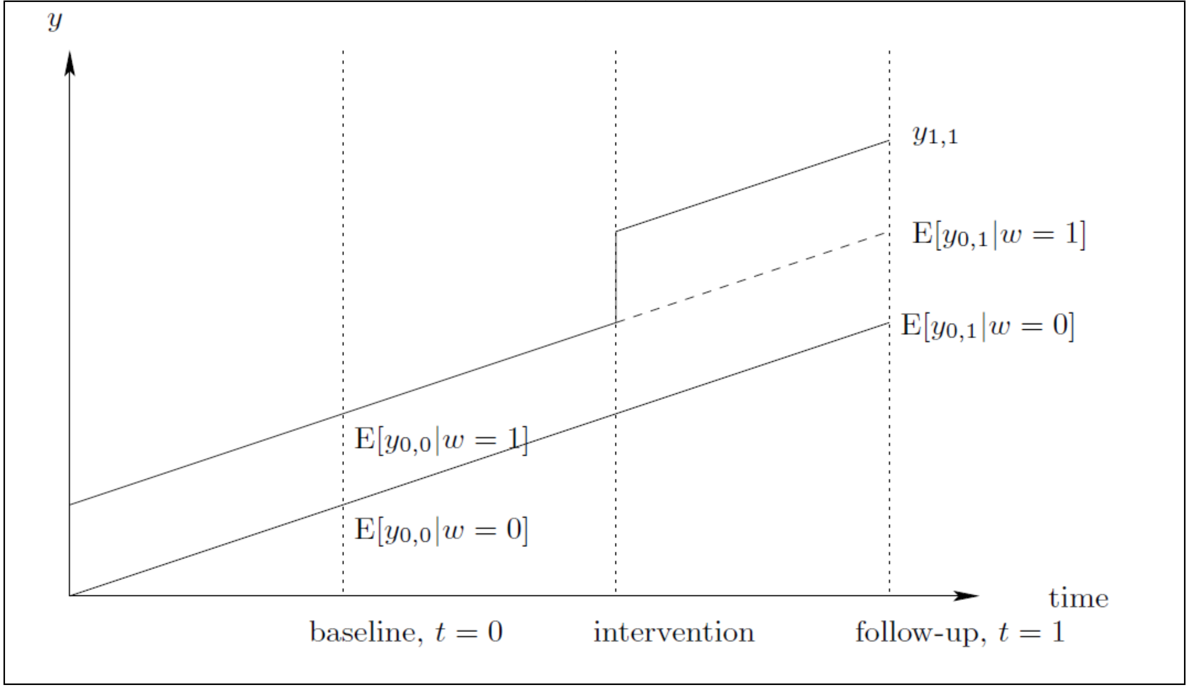


Figure 5: Evaluation with panel and cross-section data (ignorability fails in the cross-section but holds in first differences)


y
$y_{1,1}$
$\mathrm{E}[y_{0,1}|w = 1]$
$\mathrm{E}[y_{0,1}|w = 0]$
$\mathrm{E}[y_{0,0}|w = 1]$
$\mathrm{E}[y_{0,0}|w = 0]$
time
baseline, $t = 0$
intervention
follow-up, $t = 1$

Now illustrate the case where ignorability doesn't hold in first differences.