Econometrics II

Lecture 3: Regression and Causality

Måns Söderbom*

5 April 2011

*University of Gothenburg. mans.soderbom@economics.gu.se. www.economics.gu.se/soderbom. www.soderbom.net

1. Introduction

In lecture 1 we discussed how regression gives the best (MMSE) linear approximation of the CEF (regression justification III).

This, of course, doesn't necessarily imply that a regression coefficient can be given a **causal** interpretation. Because regression inherits its legitimacy from the CEF, it follows that whether causal interpretation of regression coefficients is appropriate depends on whether the CEF can be given a casual interpretation.

So what do we mean by causality? Angrist-Pischke (AP) think of causal relationships in terms of the potential outcomes framework, to describe what would happen to a given individual in a hypothetical comparison of alternative (e.g. hospitalization) scenarios. If we define "causal effect" as the differences in potential outcomes, it follows that the CEF if causal when it describes differences in **average** potential outcomes for a fixed reference population.

Once we've defined the CEF to be causal, the key question becomes if/how regression can be used to estimate the causal effects of interest. Lectures 3-7 will revolve around this particular question.

References for this lecture:

Angrist and Pischke (2009), Chapters 3.2-3.3.

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

I'll use data that have been analyzed in the following paper:

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

2. Regression and Causality

The Conditional Independence Assumption. Let's focus on the earnings-education relationship. Suppose our goal is to estimate the causal effect of schooling on earnings. Given our definition of causality, this amounts to asking what people would earn, on average, if we could either

- change their schooling in a perfectly controlled environment
- change their schooling randomly so that the those with different levels of schooling would otherwise be comparable.

As discussed in chapter 2, a randomized trial (experiment) ensures independence between potential outcomes and the causal variable of interest. In this case, the groups being compared - e.g. college graduates and non-graduates - are truly comparable (i.e. they don't differ systematically with respect to other characteristics determining earnings). But if all we have is non-experimental data, this may not be the case.

Let's return to the potential outcomes framework:

$$\text{Potential outcome} = \left\{ \begin{array}{c} Y_{1i} = \text{ outcome for } i \text{ if treated} \\ Y_{0i} = \text{outcome for } i \text{ if not treated} \end{array} \right\}.$$

Initially, think of treatment as binary: e.g. college schooling or not. Hence, Y_{1i} measures potential earnings for individual i if s/he has college education and Y_{0i} measures potential earnings for i if s/he does not have college education.

Hence, $Y_{1i} - Y_{0i}$ is the causal effect of college education on earnings for individual i.

The observed outcome Y_i can be written in terms of potential outcomes as

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})\, C_i$$

where C_i is a treatment dummy variable equal to 1 if individual i received treatment, and 0 otherwise.

We can't measure $Y_{1i} - Y_{0i}$ since we never observe both Y_{1i} and Y_{0i}.

Our goal is to measure the average of $Y_{1i} - Y_{0i}$ (the average treatment effect; ATE), perhaps for the sub-group of people who went to college (the average treatment effect on the treated; ATT).

We suspect we won't be able to learn about the causal effect of college education simply by comparing the average levels of earnings by education status because of selection bias:

$$\begin{aligned} E[Y_i|C_i = 1] - E[Y_i|C_i = 0] \quad = & \\ & E[Y_{1i}|C_i = 1] - E[Y_{0i}|C_i = 1] \qquad \text{(ATT)} \\ & + E[Y_{0i}|C_i = 1] - E[Y_{0i}|C_i = 0] \quad \text{(Selection bias).} \end{aligned}$$

We suspect that potential outcomes under non-college status are better for those that went to college than for those that did not; i.e. there is positive selection bias.

The **conditional independence assumption** (CIA): Conditional on *observed* characteristics X_i, the selection bias *disappears*. That is:

$$\{Y_{0i}, Y_{1i}\} \text{ independent of } C_i, \text{ conditional on } X_i.$$

In words: If we are looking at individuals with the same characteristics X, then $\{Y_{0i}, Y_{1i}\}$ and C_i are independent. It follows that, given CIA, conditional-on-X_i comparisons of average earnings across schooling levels have a causal interpretation:

$$E[Y_i|X_i, C_i = 1] - E[Y_i|X_i, C_i = 0] = E[Y_{1i} - Y_{0i}|X_i].$$

For obvious reasons, this quantity is interpretable as the average conditional treatment effect.

So far we've focused on binary treatment variables, i.e. variables that can take two values. Now generalize the framework so that the treatment variable can take more than 2 values. Focus now on years of schooling as the treatment variable, and define the potential outcome associated with schooling level

s as

$$Y_{si} \equiv f_i(S),$$

where we put an i-subscript on the $f(.)$ function to show that the potential earnings are individual specific. The CIA in this more general setup becomes

$$\{Y_{si}\} \text{ independent of } C_i, \text{ conditional on } X_i \text{ for all } s.$$

Conditional on X_i, the average causal effect of a one-year increase in schooling is

$$E(f_i(S) - f_i(S-1)|X_i), \tag{2.1}$$

for any value of s. Consequently, we will have separate causal effects for each value taken on by the conditioning variables X. To get the unconditional average causal effect of (say) high school graduation (which amounts to increasing S from 11 to 12), we take expectations using the distribution of X:

$$\begin{aligned} & E[E(f_i(S=12) - f_i(S=11)|X_i)] \\ = \; & E(f_i(S=12) - f_i(S=11)). \end{aligned} \tag{2.2}$$

How might we compute quantities like (2.1) or (2.2) in practice?

- One option would be to compare individuals for whom the values of X are identical. This is known as **exact matching** on observables.
- Whilst flexible, matching has problems of its own (can you think of any?). We will return to this below. A simpler approach is regression - but we need to think about how we can justify regression given the CIA.

Estimation by regression. Regression is an easy-to-use empirical strategy. There are essentially 2 ways of going from the CIA to regression:

1. We could assume that $f_i(S)$ is i) linear in S, and ii) the same for everyone except for an additive error term. However, this is quite a strong assumption.

2. If we assume that there is heterogeneity $f_i(S)$ across individuals and/or that f is nonlinear, regression can be thought of as a strategy for estimating a weighted average of the individual-specific difference $f_i(S) - f_i(S-1)$.

Focus on the first of these settings for simplicity. Our linear constant effects causal model is written as

$$f_i(S) = \alpha + \rho S + \eta_i.$$

Writing this in terms of observables (use Y_i instead of $f_i(S)$; and S_i instead of S), we get

$$Y_i = \alpha + \rho S_i + \eta_i.$$

Now, suppose schooling (S_i) is correlated with potential earnings outcomes Y_{si}. This would show up here as a correlation between S_i and the residual η_i (how?).

Further suppose that the CIA holds given a vector of observed covariates X_i. In fact, suppose the random part of potential earnings (η_i) can be written as a linear function of X_i plus an error term v_i:

$$\eta_i = X_i\gamma + v_i,$$

where γ is a vector of population coefficients satisfying

$$E(\eta_i|X_i) = X_i\gamma.$$

It follows from the CIA that

$$E(f_i(S)|X_i, S_i) = E(f_i(S)|X_i),$$

(conditional independence of potential outcomes), which in turn is written now as

$$\begin{aligned} E\left(f_i\left(S\right)|X_i\right) &= E\left(\alpha+\rho S+X_i\gamma+v_i\right) \\ &= \alpha+\rho S+X_i\gamma. \end{aligned}$$

Hence, the residual v_i in the linear causal model

$$Y_i=\alpha+\rho S_i+X_i\gamma+v_i \tag{2.3}$$

is uncorrelated with the regressors S_i, X_i. The regression coefficient ρ is the causal effect of interest.

Recall: The key assumption here is that the observable characteristics X_i are the only reason why η_i and S_i are correlated.

The Omitted Variables Bias Formula. The specification (2.3) thus contains a set of control variables X_i. We refer to this as being a "long" regression and we refer to a specification without the control variables as a "short" regression. The omitted variables bias (OVB) formula describes the relationship between the estimates across the two specifications

$$\begin{aligned} Y_i &= a_s+r_s\cdot S_i+u_i && \text{(Short)} \\ Y_i &= \alpha+\rho\cdot S_i+X_i\gamma+v_i. && \text{(Long)} \end{aligned}$$

In particular, if we leave out X_i (and thus estimate the short regression), we will get

$$r_s=\frac{Cov\left(Y_i,S_i\right)}{V\left(S_i\right)}=\rho+\gamma'\delta_{Xs} \tag{2.4}$$

where δ_{Xs} is the vector of coefficients from regressions of the elements of X_i on S_i; that is,

$$\begin{aligned} X_{1i} &= d_1 + \delta_{1Xs}S + e_{1i} \\ X_{2i} &= d_2 + \delta_{2Xs}S + e_{2i} \\ &(...) \\ X_{Ki} &= d_K + \delta_{KXs}S + e_{Ki}, \end{aligned}$$

$\delta_{Xs} = (\delta_{1Xs}, \delta_{2Xs}, ..., \delta_{KXs})$. Note: the vector of coefficients δ_{Xs} should not be given a causal interpretation. Equation (2.4) is the OVB.

Clearly the long and the short regression will give the same results if the omitted and included variables are uncorrelated.

- Now use the OVB framework to assess the likely consequences of omitting ability for schooling coefficients in earnings regressions.
- Apply this way of thinking when studying the regression results in the handout that I circulated in lecture 1.
- Have a close look at the results in Table 3.2.1 (p. 62) in AP, summarized here:

Table 3.2.1

Estimates of the returns to education for men in the NLSY

	(1)	(2)	(3)	(4)	(5)
Controls	None	Age dummies	Col. (2) + additional	Col. (3) +AFQT	Col. (4) + job dummies
	.132	.131	.114	.087	.066
	(.007)	(.007)	(.007)	(.009)	(.010)

Let's take stock.

- Assume the CEF is causal.

- Assume CIA holds for your empirical specification.
- Then there can be no omitted variables bias, and your regression thus has a causal interpretation.

So how do we know if CIA holds?

- Random assignment conditional on X (e.g. random assignment to training, conditional on covariates; Black et al. 2003, referenced in AP).
- Detailed institutional knowledge regarding the process that determines S_i.

Bad Control. Perhaps you get the impression that controlling for more covariates always increases the likelihood that regression estimates have a causal interpretation. Well, this is often true - but not always.

Some variables are **bad controls** and should not be included in a regression even when their inclusion might be expected to change the short regression coefficients (i.e. they are correlated with the explanatory variables of interest and with the outcome variable).

Bad controls are variables that are themselves **outcomes** of the treatment variable.

Earlier in the course, we talked about IQ test scores plausibly being determined by education - in which case IQ test scores would be a bad control.

Good controls, in contrast, are variables that we can think of as having been fixed at the time the treatment variable was determined.

Illustration: Suppose we are interested the effects of a college degree on earnings; and suppose there are two types of jobs - white collar and blue collar. Suppose we add controls for occupation in our earnings regression. Why might this not be a good idea?

Well it seems likely that the type of job someone gets depends on his or her education; presumably, you're more likely to get a white collar job if you have a college degree.

Now formalize the nature of the problem. Let W_i be a dummy variable = 1 for white collar workers and 0 for blue collar workers. Let Y_i denote earnings. Suppose these are both outcome variables driven

by education:

$$\begin{aligned} Y_i &= C_i Y_{i1} + (1 - C_i) Y_{i0} \\ W_i &= C_i W_{i1} + (1 - C_i) W_{i0}, \end{aligned}$$

where $C_i = 1$ for college graduates and is zero otherwise. Further, $\{Y_{0i}, Y_{1i}\}$ denote potential earnings outcomes, while $\{W_{0i}, W_{1i}\}$ denote potential white collar status (e.g. if $.W_{0i} = 0, W_{1i} = 1$, this means the individual needs college education to get a white collar job; whereas if $W_{0i} = 1, W_{1i} = 1$, the individual will get a white collar even if s/he doesn't have college education)

We assume C_i is randomly assigned so that it is independent of all potential outcomes. This implies we can estimate the causal effects of C_i on earnings and job type simply as the difference in means:

$$\begin{aligned} E(Y_i|C_i = 1) - E(Y_i|C_i = 0) &= E(Y_{1i} - Y_{0i}) \\ E(W_i|C_i = 1) - E(W_i|C_i = 0) &= E(W_{1i} - W_{0i}). \end{aligned}$$

Note the **absence** of a control for job type in the first of these expressions.

Bad control means that a comparison of earnings conditional on W_i does **not** have a causal interpretation.

To see why, consider the difference in means for individuals with and without college education, where everyone has a white collar job:

$$E(Y_i|W_i = 1, C_i = 1) - E(Y_i|W_i = 1, C_i = 0).$$

This can be written in terms of potential outcomes as

$$E(Y_{1i}|W_{1i} = 1, C_i = 1) - E(Y_{0i}|W_{0i} = 1, C_i = 0).$$

Random assignment of treatment implies this can be written as

$$E\left(Y_{1i}|W_{1i}=1\right)-E\left(Y_{0i}|W_{0i}=1\right)$$

$$\begin{aligned} = \quad & E\left(Y_{1i}-Y_{i0}|W_{1i}=1\right) \quad \text{(average causal effect)} \\ & +E\left(Y_{0i}|W_{1i}=1\right)-E\left(Y_{0i}|W_{0i}=1\right) \quad \text{(selection bias).} \end{aligned}$$

Now interpret the selection bias term.

- First term: expected potential non-college earnings, given that potential white collar status associated with college education is equal to 1.
- Second term: expected potential non-college earnings, given that potential white collar status associated with **non**-college education is equal to 1 - likely high; if, *despite* no college education, you get a white collar job, you are probably 'special' i.e. have a high Y_{0i}).

AP discusses a variant on this theme, which the refer to as 'proxy control'. Read if you are interested (personally, I don't think it adds new insights to what we've already learned about bad control).

3. Matching

This part of the lecture is based on Section 3.3.1 in AP but with a different emphasis: AP focus on establishing how regression and matching estimates are theoretically related; I focus on explaining matching.

Exact matching. Recall our favorite formula:

$$\begin{aligned} E\left[Y_i|D_i=1\right]-E\left[Y_i|D_i=0\right] \quad = & \\ & E\left[Y_{1i}|D_i=1\right]-E\left[Y_{0i}|D_i=1\right] \qquad \text{(ATT)} \\ & +E\left[Y_{0i}|D_i=1\right]-E\left[Y_{0i}|D_i=0\right] \quad \text{(Selection bias),} \end{aligned}$$

where D_i is a dummy variable for treatment.

The CIA in this context says that, conditional on X_i, the potential outcomes Y_{0i}, Y_{1i} are independent of actual treatment. In other words, selection bias disappears after conditioning on X_i.

We can write the average effect of treatment on the treated as

$$\begin{aligned}
\delta_{TOT} &= E\left[Y_{1i} - Y_{0i} | D_i = 1\right] \\
&= E\left\{E\left[Y_{1i} - Y_{0i} | X_i, D_i = 1\right] | D_i = 1\right\} \\
&= E\left\{E\left[Y_{1i} | X_i, D_i = 1\right] - E\left[Y_{0i} | X_i, D_i = 1\right] | D_i = 1\right\},
\end{aligned}$$

where the last term is the counterfactual.

Now observe that the CIA implies

$$E\left[Y_{0i} | X_i, D_i = 0\right] = E\left[Y_{0i} | X_i, D_i = 1\right],$$

enabling me to write

$$\delta_{TOT} = E\left\{E\left[Y_{1i} | X_i, D_i = 1\right] - E\left[Y_{0i} | X_i, D_i = 0\right] | D_i = 1\right\},$$

or

$$\delta_{TOT} = E\left[\delta_X | D_i = 1\right]$$

where

$$\delta_X \equiv E\left[Y_i | X_i, D_i = 1\right] - E\left[Y_i | X_i, D_i = 0\right],$$

is the (observable) difference in mean Y (e.g. earnings) at a given X_i.

How are we going to *estimate* δ_{TOT}?

Suppose X_i is discrete - then we can write

$$E\left[Y_{1i} - Y_{0i}|D_i = 1\right] = \sum_x \delta_X P\left(X_i = x|D_i = 1\right),$$

where $P\left(X_i = x|D_i = 1\right)$ is a probability density function. That is, we compute δ_X for each value of X_i, and then compute a weighted average of the different δ_X.

This approach is known as **exact matching**.

How would we compute the average treatment effect (unconditional of actual treatment)?

[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 $(D = 1, x = 1)$. What would be the implication of that? Think of a real example where something similar might happen.

As we've seen above, matching as a strategy to control for covariates is motivated by the CIA, amounting to covariate-specific treatment-control comparisons.

So, matching follows from the CIA.

We've also seen that causal interpretation of a regression coefficient is based on the CIA too.

So - matching and regression are both control strategies.

Does matching really differ from regression?

AP discuss the relationship between matching and regression on pp.74-77. This gets quite technical - consider this optional material. Personally, I find it somewhat useful to cast the matching estimator above in a regression framework:

$$Y_i = \phi_0 + \phi_1 D_i + X\phi_2 + (D_i X)\phi_3 + \text{residual},$$

where ϕ_0 and ϕ_1 are coefficients, ϕ_2, ϕ_3 are vectors, and X is a vector of dummy variables for each unique value of X. Note, for example, that this is exactly the type of regression shown on page 3 in the appendix

(and we've seen how to compute the average treatment effects from those estimates) . The specification above stands in contrast to what one would 'typically' adopt in regression analysis:

$$Y_i = \theta_0 + \theta_1 D_i + X\theta_2 + \text{residual},$$

where θ_0 and θ_1 are coefficients, and θ_2 is a vector. This specification is easier to interpret (just focus on the coefficient θ_1) and more restrictive (no $D_i X$ interaction terms are included).

Control for Covariates Using the Propensity Score. Consider modelling the **likelihood of being treated** by means of a binary choice model (e.g. logit or probit):

$$\Pr(D_i = 1|x) = F(X_i\beta) \equiv p(X_i).$$

In the treatment literature, the function $p(X_i)$ is known as the **propensity score**. A useful property of the propensity score emerges in the context of estimating by matching, where the idea is to match individuals with similar propensity scores.

- **The Propensity Score Theorem**: *Suppose the CIA holds, so that conditional on X, D and (y_1, y_0) are independent. Then it must be that, conditional on the propensity score $p(X_i)$, D and (y_1, y_0) are independent.* See AP, p.81 for a straightforward proof.
- This theorem says that you need only control for the probability of treatment itself.
- How might we adjust the matching approach outlined so as to enable us to match on the propensity score? What's the big advantage of matching on the propensity score compared to exact matching?
- 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.

- 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, keep in mind that you will need a **complete set of variables** determining selection into treatment for propensity score matching to work (this follows from CIA). 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.

- Of course, it's hard to know a priori what constitutes the right set of explanatory variables in the first stage. 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.

- 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.

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.

- 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 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.

- 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 ATT estimator is

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

where $\{D = 1\}$ is the set of treated individuals, $\{D = 0\}$ is the set of nontreated individuals (the control group), and $\phi(i,j)$ is a **weight**. Notice that $\sum_{j \in \{D=0\}} \phi(i,j)\, y_{0,j}$ 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 ATT.

- Another method 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 contribution to the ATT of treated individual i is

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

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 2 in the appendix provides details on how this works, using the Hoddinott-Gilligan food aid data from Ethiopia.

3.1. Regression or matching?

- The regression approach is easy to implement and interpret.
  - But there may be too much extrapolation. The idea underlying matching estimators is that you should compare the outcomes of two individuals with similar characteristics, except one was treated and the other wasn't. This idea is not really central to the regression approach.

Suppose we write the regression as

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

You might say α is being estimated 'controlling for x', but it may be that most high values of x are associated with $D = 1$, and most low values of x are associated with $D = 0$. Suppose we want to calculate the (conditional) treatment effect $E(Y_{1i} - Y_{0i}|x_i$ is 'high'). 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 $E(Y_{0i}|x_i$ is 'high') is mostly based on combining observations on outcomes for which $\{D = 1, x$ high$\}$ and observations on outcomes for which $\{D = 0, x$ low$\}$. But whether this gives a good estimate of $E(y_{0i}|x_i$ is 'high') is uncertain, and hinges on the extrapolation not being misleading.

- 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. 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 AP, p.84 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.

PhD Programme: Econometrics II
Department of Economics, University of Gothenburg
Appendix Lecture 3
Måns Söderbom

1. Exact matching: A simple example

Suppose your dataset looks like this:

id	y	D	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 average treatment effect on the treated (ATT) and the average treatment effect (ATE) here? Recall the formula for the ATT:

$$\delta_{TOT} = E\left[\delta_X | D_i = 1\right]$$

where

$$\delta_X \equiv E\left[Y_i | X_i, D_i = 1\right] - E\left[Y_i | X_i, D_i = 0\right]$$

Hence, all we need to do is to estimate these quantities.

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,D}. Define

$$E\left(y_1 | x\right) = r_1\left(x\right),$$

$$E\left(y_0 | x\right) = r_0\left(x\right),$$

Thus we need to estimate only four quantities: $r_1(0)$, $r_0(0)$, $r_1(1)$ and $r_0(1)$. With the present 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 {D,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 {D,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:

$$\hat{ATE} = 2.7143$$

To get an estimate of the average treatment effect for the treated, we simply discard all non-treated observations when computing the average:

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

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 {D, x} - of observations in the data. Therefore, the following regression is completely unrestrictive in terms of the functional form relationship between {D, x} and the outcome variable y:

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

Notice that

$$r_1(0) = \beta_0 + \beta_1$$
$$r_0(0) = \beta_0$$
$$r_1(1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$$

$$r_0(1) = \beta_0 + \beta_2$$

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 ATT 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 {D,x} in the data.

2. Propensity score matching: Example

The data in this illustration have been used in the following paper:

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

The outcome variable is consumption growth and the treatment variable is getting food aid. I have computed the propensity score, and sorted the data from the lowest to the highest pscore value. The full sample consists of 630 observations.

Table 2: Propensity scores and kernel weighting

Pscore	Treatment	K	Consumption growth	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

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

  This gives me the values shown in the 'weight' column. Notice that they will sum to one.
- I then obtain the weighted average of consumption growth 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?
- Note: These computations use and 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 is shown in Figure 1.

Figure 1: The Epanechnikov distribution