Econometric Methods for Ex Post Social Program Evaluation

Petra E. Todd[1]

[1]University of Pennsylvania

January, 2013

Chapter 1: The evaluation problem

Questions of interest in program evaluations

- Do program participants benefit from the program?
- Who chooses to participate in programs?
- What would be the program effects if extended to nonparticipants?
- Do people differ in how they benefit from the program?
- Do the benefits exceed the costs?
- What is the social return from the program?
- Would an alternative program design achieve greater impact at the same cost?

Goals

- to describe different estimators and their identifying assumptions
- to discuss the behavioral implications of these assumptions
- to illustrate how different kinds of estimators are related to one another
- to summarize the data requirements of different methods
- to provide examples of how the evaluation methods have been applied in the development, labor and health economics literatures.

Alternative approaches

- Randomization
- Regression estimators
- Matching
- Control function methods
- IV methods, MTE, LATE
- Regression-Discontinuity

The Evaluation Problem

- Let $D = 1$ for persons who receive the intervention and $D = 0$ for persons who do not receive it.
- Each person has associated a (Y_0, Y_1) pair that represents the outcomes that would be realized in the the untreated and treated states.
- At most one of the two potential outcomes is observed.
- The observed outcome is

$$Y = DY_1 + (1 - D)Y_0.$$

- The *treatment effect* is

$$\Delta = Y_1 - Y_0.$$

- Inferring gains from treatment therefore requires solving a missing data problem.

Parameters of interest

- Distinguish between
  - *direct effects*: effects of program on participants
  - *indirect effects*: effects of program on people who are not directly participating
- Example: job voucher program that gives employers a subsidy to hire workers may help program participants but may put nonparticipants at a disadvantage.
- Most of literature aims to estimate direct effects.

Parameters of interest

(a) the proportion of program participants that benefit from the program

$$\Pr(Y_1 > Y_0 | D = 1) = \Pr(\Delta > 0 | D = 1)$$

(b) the proportion of the total population benefitting from the program:

$$\Pr(\Delta > 0 | D = 1) \Pr(D = 1)$$

(c) quantiles of the impact distribution (such as the median), where q is the selected quantile

$$\inf_{\Delta} \{\Delta : F(\Delta | D = 1) > q\}$$

(d) the distribution of gains for individuals with some characteristics X_0

$$F(\Delta | D = 1, X = X_0),$$

Two key parameters

Much of the program evaluation literature develops methods for estimating two key parameters of interest:[1]

(e) the average gain from the program for persons with characteristics X

$$E(Y_1 - Y_0|X) = E(\Delta|X).$$

(f) the average gain from the program for program participants with characteristics X :

$$E(Y_1 - Y_0|D = 1, X) = E(\Delta|D = 1, X).$$

[1] See, *e.g.*, Rosenbaum and Rubin (1985), Heckman and Robb (1985), or Heckman, Lalonde and Smith (1999).

Distinction between ATE and TT

Suppose the outcomes in the treated and untreated states can be written as:

$$\begin{aligned} Y_1 &= \varphi_1(X) + U_1 \\ Y_0 &= \varphi_0(X) + U_0. \end{aligned}$$

The observed outcome $Y = DY_1 + (1-D)Y_0$ is:

$$Y = \varphi_0(X) + D(\varphi_1(X) - \varphi_0(X)) + \{U_0 + D(U_1 - U_0)\}.$$

Assume $E(U_0|X) = E(U_1|X) = 0$. The gain to an individual from participating in the program is:

$$\Delta = \varphi_1(X) - \varphi_0(X)) + (U_1 - U_0).$$

What is known when people enter program?

- Individuals may or may not know their values of U_1 and U_0 at the time of deciding whether to participate in a program.
- If people self-select into the program based on their anticipated gains, then we would expect that $E(U_0|X, D) \neq 0$ and $E(U_1|X, D) \neq 0$.
- If the gain from the program depends on U_1 and U_0 and people know future values of U_1 and U_0, or can forecast the values, then we would expect people to make use of this information when they decide whether to select into a program.

ATE and TT in terms of model

In the notation of the above model for outcomes:

$$\begin{aligned}\alpha_{ATE}(X) = E(\Delta|X) &= \varphi_1(X) - \varphi_0(X) + E(U_1|X) - E(U_0|X) \\ &= \varphi_1(X) - \varphi_0(X).\end{aligned}$$

The *average impact of treatment on the treated (TT)* is

$$\alpha_{TT}(X) = E(\Delta|X) = \varphi_1(X) - \varphi_0(X) + E(U_1 - U_0|X, D = 1).$$

As discussed in Heckman (2000), the average effect of treatment on the treated icombines the "structural parameters" (the parameters of the functions $\varphi_0(X)$ and $\varphi_1(X)$) with means of the unobservables.

Also, UT

For completeness, define the *average impact of treatment on the untreated (UT)* as

$$\alpha_{UT}(X) = E(\Delta|X) = \varphi_1(X) - \varphi_0(X) + E(U_1 - U_0|X, D = 0),$$

Parameter may be of interest if there are plans for expanding the program.

The relationship between TT, ATE and UT is:

$$\alpha_{ATE}(X) = \Pr(D = 1|X)\alpha_{TT}(X) + \Pr(D = 0|X)\alpha_{UT}(X).$$

Three types of assumptions

As discussed in Heckman, Lalonde and Smith (1999), there are three types of assumptions that can be made.
In order of increasing generality, they are:

- (A.1) conditional on X, the program effect is the same for everyone ($U_1 = U_0$)
- (A.2) conditional on X, the program effect varies across individuals but $U_1 - U_0$ does not help predict program participation
- (A.3) conditional on X, the program effect varies across individuals and $U_1 - U_0$ does predict who participates in the program.

ATE=TT under assumptions A.1 and A.2.
We will consider ways of estimating the $\alpha_{TT}(X)$ and $\alpha_{ATE}(X)$ parameters of interest under these three different sets of assumptions.

When does bias arise?

Consider the model

$$Y = \varphi_0(X) + D(\varphi_1(X) - \varphi_0(X)) + \{U_0 + D(U_1 - U_0)\}.$$

In terms of the two parameters of interest, the model can be written as:

$$Y = \varphi_0(X) + D\alpha_{ATE}(X) + \{U_0 + D(U_1 - U_0)\} \tag{1}$$

or

$$Y = \varphi_0(X) + D\alpha_{TT}(X) + \{U_0 + D[U_1 - U_0 - E(U_1 - U_0|X, D = 1)]\}.$$

When does bias arise?

Suppose the X are discrete.
We estimate an ordinary least squares regression:

$$Y = aX + b_x XD + v.$$

This model is known as the *common effect* model.
A special case of the model assumes that the coefficient on D is constant across X :

$$Y = aX + bD + v.$$

When does bias arise?

- Bias for the $\alpha_{ATE}(X)$ parameter arises if the mean of the error term does not have conditional mean zero, i.e.

$$E(U_0 + D(U_1 - U_0)|X, D)) \neq 0.$$

- Under assumption A.1 and A.2, potential bias arises only from $E(U_0|X, D) \neq 0$.
- Under A.3, there is also the potential for bias from $E(U_1 - U_0|D, X) \neq 0$.
- For estimating the $\alpha_{TT}(X)$ parameter, under A.1-A.3, bias arises if $E(U_0|X, D) \neq 0$.

Chapter 2: Randomization

Suppose select the comparison group using a randomization device (e.g. a lottery).

Main benefits

- Ensures that the treatment and control groups have the same distribution of observables and of unobservables.
- Ensures that the control group also satisfies program eligibility criteria

Potential problems in social experiments

- *Randomization bias* or so-called *Hawthorne effects*: randomization may change the way a program operates.
- *Contamination* or *cross-over effects*: occurs when some controls receive the treatment and/or some of the people assigned to treatment do not receive it.
- *Dropout*: when some of the treatment group drop out before completing the program.
- *Attrition*: Both controls and treatments may not respond to surveys and response patterns may differ by treatment status.
- *pioneer effects*: can occur if the program has not been in operation for long.(See Behrman and King, 2008, 2009). For example, program implementers could be less experienced or especially motivated.

Internal verses external validity

- If the experimental protocol was followed and the problems described earlier are not that significant, then the experiment is said to be *internally valid*.
- An experiment has *external validity* if the sample participating in the experiment is representative of the population of interest.
- If the sample in the experiment is not similar, for example, is younger, poorer or more likely to be female, then statistical adjustment can sometimes be used to extrapolate from the experimental results to the population of interest.

For a recent critical view on the value of randomized control trials in economic development studies, see Deaton (2009).

When to randomize?

At what stage should randomization be applied? There are two major approaches:

- Randomization after acceptance into the program
- Randomization of eligibility

Randomization after application

Let $R = 1$ if randomized into the program (treatment group).
Let $R = 0$ if randomized out (control group).
Let Y_0^* and Y_1^* denote the outcomes observed under the experiment.
Let $D^* = 1$ denote someone who applies to the program and is subject to the randomization. People with $D^* = 1$ are would-be participants, in the sense that they would participate in the program if offered to them.
No randomization bias and random assignment implies:

$$E(Y_1^*|X, D^* = 1, R = 1) = E(Y_1|X, D = 1)$$
$$E(Y_0^*|X, D^* = 1, R = 0) = E(Y_0|X, D = 1)$$

Randomization after application

Thus, the experiment gives the average effect of treatment for individuals who apply to the program.

$$TT(X) = E(Y_1 - Y_0 | X, D = 1).$$

The experiment also gives the marginal distributions of Y_1 and Y_0

$$F(Y_1 | X, D = 1)$$
$$F(Y_0 | X, D = 1)$$

It does not give the joint distribution
$F(Y_0, Y_1 | x, D = 1)$.

Randomization of eligibility

- Alternative approach is to randomize on eligibility.
- A subset of people may be told randomly that they are eligible for a program and then they can choose whether to participate or not.
- Let $e = 1$ denote that a person is eligible for a program and $e = 0$ if not eligible.
- People with $D = 1$ and $e = 0$ are people who would have liked to participate but they were randomly not eligible, so we only observe Y_0 for them.

Randomization of eligibility

We observe:

$$E(Y|X, e=1) = Pr(D=1|x, e=1)E(Y_1|X, e=1, D=1) + Pr(D=0|x, e=1)E(Y_0|X, e=1, D=0)$$
$$E(Y|X, e=0) = Pr(D=1|x, e=0)E(Y_0|X, e=0, D=1) + Pr(D=0|x, e=1)E(Y_0|X, e=0, D=0)$$

Because eligibility was randomized, we have

$$Pr(D=1|X, e=1) = Pr(D=1|X, e=0)$$
$$Pr(D=0|X, e=1) = Pr(D=0|X, e=0)$$
$$E(Y_0|X, D=1, e=1) = E(Y_0|X, D=1, e=0)$$
$$E(Y_1|X, D=1) = E(Y_1|X, D=1, e=0)$$

Randomization of eligibility

Thus, the difference in the previous two equations, $E(Y|X, e = 1) - E(Y|X, e = 0) = Pr(D = 1|X, e = 1)[E(Y_1|X, D = 1) - E(Y_0|X, D = 1)]$.
Therefore, we obtain

$$TT(X) = \frac{E(Y|X, e = 1) - E(Y|X, e = 0)}{Pr(D = 1|X, e = 1)}$$

.
The estimator replaces the means by their sample analogs.
When randomization is on eligibility, we can compare the means for those randomized-in and randomized-out, dividing by the proportion that selects into the program, given eligible.

The estimator can easily be modified to account for a fraction of the controls getting into the program despite not being eligible (e.g. contamination). In that case, we obtain

$$TT(X) = \frac{E(Y|X, e = 1) - E(Y|X, e = 0)}{Pr(D = 1|X, e = 1) - Pr(D = 1|X, e = 0)}$$

where contamination implies $Pr(D = 1|X, e = 0) > 0$. We do, however, require that $Pr(D = 1|X, e = 1) \neq Pr(D = 1|X, e = 0)$.

Experiments in the presence of drop-out

- Program dropout occurs when people assigned to the treatment group decide not to participate.
- If drop-out occurs early on, we can possible consider these persons to be untreated.
- Can treat program dropout in the same way as randomization of eligibility. The drop-outs were eligible ($e = 1$) for the program but decided not to participate ($D = 0$).

Intent-to-treat

- Alternatively, could define treatment as the "offer of treatment." All people offered the treatment are participants, regardless of whether they later attend the program.
- The relationship between the ITT program effect and the TT effect is

$$ITT(X) = TT(X)Pr(D = 1|e = 1, X) + 0Pr(D = 0|e = 1, X)$$

The second term implies that ITT penalizes a program for having a low participation rate by giving an impact of zero to a fraction of the group assigned to the program.

Drop-out

Program drop-out after partial participation is a more difficult problem. This happens when individuals attend the program for awhile and then drop-out before completing it.
In that case, we need to either decide at what point they become "treated" or else explicitly model the treatment outcome as a function of a *treatment dose* level.

Randomization Methods: Blocking

- If you are designing a randomized experiment, one option is to simply randomize. With a large sample, the distributions of the observables and unobservables within the $R = 1$ and $R = 0$ groups will be similar.
- Another option is to first divide the sample according the some observable X characteristics and then randomize within X subsets. This option is called *blocking*.
- The main advantage of blocking is that it ensures that the X distribution is the same even in sample samples, so it is a particularly useful method when the size of the sample being randomized is modest.
- Blocking eliminates the need to control for those X variables ex post, in a regression, and therefore can save on degrees of freedom and provide greater precision in estimating the treatment effect.

Place-based experiments

- Randomization can be done at an individual level or it may be preferable to do it at the level of a larger unit, such as a family or a school or a village.
- These are called *place-based* randomized experiments.
- One might choose a place-based design over an individual level design if you expect that there may be spillover effects from some individuals in the treatment to others, for examples, from some students to others within a school
- Also, sometimes it is much earlier to implement an intervention at the level of a higher unit, such as a school
- The main cost of doing a place-based randomized experiment instead of a individual level experiment is loss of power, because there are fewer individual units being randomized.

Randomized roll-out designs

- Sometimes, a program is being gradually implemented and one can use randomization to choose where it gets implemented first, even though eventually it may be implemented everywhere.
- The areas that are initially left out can temporarily serve as a control group.
- Under such *randomized roll-out* designs, it is important not to inform those units who are were initially left out that they will eventually be included in the program, because the expectation of receiving the program in the future could affect their behavior in the present.

Chapter 3: Simple regression estimators

- Nonexperimental estimators of program impacts use two types of data to impute missing Y_0 outcomes for program participants:
  - data on participants at a point in time prior to entering the program
  - data on nonparticipants.
- Three widely used methods for estimating $E(\Delta|X, D = 1)$, (TT)
  - (a) *before-after* estimator
  - (b) *cross-section* estimator
  - (c) *difference-in-difference* estimator
- Extensions to ATE parameter straightforward.

Notation

- Denote the outcome measures by Y_{1it} and Y_{0it}, where i denotes the individual and t the time period of observation,

$$\begin{aligned} Y_{1it} &= \varphi_1(X_{it}) + U_{1it} \\ Y_{0it} &= \varphi_0(X_{it}) + U_{0it}. \end{aligned} \tag{2}$$

- U_{1it} and U_{0it} distributed independently across persons and satisfy $E(U_{1it}|X_{it}) = 0$ and $E(U_{0it}|X_{it}) = 0$.
- X_{it} represents conditioning variables that may either be fixed or time-varying (such as gender or age), but whose distributions are assumed to be unaffected by whether an individual participates in the program.

- Write observed outcome at time t as

$$Y_{it} = \varphi_0(X_{it}) + D_{it}\alpha^*(X_{it}) + U_{0it}, \tag{3}$$

- D_{it} denotes being a program participant and $\alpha^*(X_{it}) = \varphi_1(X_{it}) - \varphi_0(X_{it}) + U_{1it} - U_{0it}$ is the treatment impact for an individual.
- Prior to the program intervention, we observe $Y_{0it} = \varphi_0(X_{it}) + U_{0it}$ for everyone.
- After the intervention we observe $Y_{1it} = \varphi_1(X_{it}) + U_{1it}$ for those who received it (for whom $D_{it} = 1$, for $t > t_0$, the time of the intervention) and $Y_{0it} = \varphi_0(X_{it}) + U_{0it}$ for those who did not receive it (for whom $D_{it} = 0$ in all time periods).

- This model is a random coefficient model, because the treatment impact can vary across persons.
- Assuming that $U_{0it} = U_{1it} = U_{it}$, yields the fixed coefficient or *common effect* model.
- The TT parameter is:

$$\alpha_{TT}(X_{it}) = E(\alpha^*(X_{it})|D_{it} = 1, D_{it'} = 0, X_{it}),$$

where the conditioning on $D_{it} = 1, D_{it'} = 0$ denotes that the person was not in the program at time t' but did participate by time t'.

Before-after estimators

- Uses pre-program data to impute missing Y_{0t} for program participants.
- Let t' and t denote two time periods, one before and one after the program intervention.
- The before-after estimator is the least squares solution obtained by

$$\begin{aligned} Y_{it} - Y_{it'} &= \varphi_0(X_{it}) - \varphi_0(X_{it'}) + \alpha^*_{TT}(X_{it}) + \varepsilon_{it} \\ \text{where } \varepsilon_{it} &= [U_{1it} - U_{0it} - E(U_{1it} - U_{0it} | D_{it} = 1, D_{it'} = 0, X_{it})] \\ &\quad + U_{0it} - U_{0it'} \end{aligned}$$

Before-after estimators

Consistency of the estimator for $\alpha_{TT}(X_{it})$ requires:

$$E(\varepsilon_{it}|D_{it}=1, D_{it'}=0, X_{it})=0.$$

The term in brackets has conditional mean zero by construction, so the key assumption required to justify application of this estimator is:

$$E(U_{0it}-U_{0it'}|D_{it}=1, D_{it'}=0, X_{it})=0.$$

Before-after estimators

- A special case where this assumption would be satisfied is if U_{0it} can be decomposed into a *fixed effect error structure*: $U_{0it} = f_i + v_{it}$, where f_i does not vary over time and v_{it} is a iid random error that satisfies $E(v_{it} - v_{it'} | D_i = 1, D_{it'} = 0, X_{it}) = 0$.
- This assumption allows selection into the program to be based on f_i (i.e. D_{it} is allowed to be correlated with f_i), so the estimation strategy admits to person-specific permanent unobservables that may affecting the program participation decision.

Before after estimators

- One drawback of a before-after estimation strategy is that identification breaks down in the presence of time-specific intercepts, making it impossible to separate effects of the program from other general time effects on outcomes.
- Such a common time effect may arise, e.g., from life-cycle wage growth over time or from time-varying shocks to the economy.
- Before-after estimates can also be sensitive to the choice of time periods used to construct the estimator.
- Minimal data requirements - two periods of cross-section data.

Ashenfelter's dip

- Many studies of employment and training programs in the U.S. and in other countries note that earnings and employment of training program participants dip down in the time period just prior to entering the program.
- The pattern can arise from serially correlated transitory downward shocks to earnings that may have been the impetus for the person applying to the training program.
- The dip pattern can also result from program eligibility criteria imposed that tend to select out the most disadvantaged persons for participation.
- A before-after estimation strategy that includes the preprogram "dip" period typically gives an upward biased estimate of the program effect

Cross-section Estimators

- Uses data on a comparison group of nonparticipants to impute counterfactual outcomes for program participants.
- Requires only post-program data on $D_{it} = 1$ and $D_{it} = 0$ persons.
- The least squares solution to

$$\begin{aligned} Y_{it} &= \varphi_0(X_{it}) + D_{it}\alpha_{TT}(X_{it}) + \varepsilon_{it}, \\ \text{where } \varepsilon_{it} &= U_{0it} + D_{it}[(U_{1it} - U_{0it}) - E(U_{0it} - U_{1it}|D_{it} = 1, X_{it})] \end{aligned}$$

estimated on $D_{it} = 1$ and $D_{it} = 0$ persons observed at time t.

Cross-section Estimators

- Consistency requires that $E(\varepsilon_{it}|D_{it}, X_{it}) = 0$..
- Same as assuming that $E(U_{0it}|D_{it}, X_{it}) = 0$.
- Rules out the possibility that people select into the program based on expectations about their U_{0it}, a strong assumption.

Difference-in-differences estimators

- Commonly used in evaluation work.
- Measures the impact of the program intervention by the difference in the before-after change in outcomes between participants and nonparticipants.
- Define an indicator I_i^D that equals 1 for participants (for whom $D_{it'} = 0$ and $D_{it} = 1$), and zero otherwise.
- The DID estimator is the least squares solution for $\alpha^*_{TT}(X_{it})$ in

$$\begin{aligned} Y_{it} - Y_{it'} &= \varphi_0(X_{it}) - \varphi_0(X_{it'}) + I_i^D \alpha_{TT}(X_{it}) + \varepsilon_{it} \\ \varepsilon_{it} &= D_{it}[U_{1it} - U_{0it} - E(U_{1it} - U_{0it} | D_{it} = 1, D_{it'} = 0, X_{it})] \\ &\quad + U_{0it} - U_{0it'}. \end{aligned}$$

- Identical to before-after regression, except uses both participant and nonparticipant data.

Difference-in-differences estimators

- Main advantage: allows for time-specific intercepts that are common across groups, included in $\varphi_0(X_{it})$. and separately identified from nonparticipant observations.
- The estimator is consistent if $E(\varepsilon_{it}|D_{it}, X_{it}) = 0$, which would be satisfied under a fixed effect error structure.
- Data requirements are either longitudinal or repeated cross-section data on both participants and nonparticipants.

Difference-in-differences estimators

Alternatively, DID can be implemented using a regression

$$\begin{aligned} Y_{it} &= \varphi_0(X_{it}) + I_i^D \gamma + D_{it}\alpha^*_{TT}(X_{it}) + \tilde{\varepsilon}_{it} \quad \text{for } t = t', .., t. \\ \tilde{\varepsilon}_{it} &= U_{0it} + D_{it}[U_{1it} - U_{0it} - E(U_{1it} - U_{0it} | D_{it} = 1, X_{it})] \end{aligned}$$

Difference-in-differences estimators

- Allow for unobservable determinants of program participation decisions and outcomes.
- But, fixed effect error structure only incorporates the potential influence of time-invariant unobservables.

Applications of DID estimators

- A study of the effect of school construction on education, and of education on wages, in Indonesia (Duflo, 2001)
- Evaluation of efficient use of inputs within households in Burkino Faso (Udry, 1996)
- Evaluation of impact of school meals on child nutrition in the Philippines (Jacoby, 2002)
- Impact of flip charts on student academic performance in Kenya (Glewwe et al. 2004)

Within estimator applications: Duflo (2001)

- Uses a DID estimator to evaluate the effects of a school construction program in Indonesia on education, and the effect of education (years of schooling) on wages.
- In 1973, the Indonesian government launched a major school construction program, the Sekolah Dasar INPRES program. From 1973-1974 and 1978-1979, more than 61,000 primary schools were constructed: an average of two schools per 1,000 children aged 5 to 14 in 1971.
- Enrollment rates among children aged 7 to 12 increased from 69 percent in 1973 to 83 percent by 1978.

Duflo(2001)

- Duflo exploited this policy change to estimate the impacts of this school construction program on education and earnings.
- Major estimation issue is that the placement of schools was not random. This is due to the fact that the construction of new schools was, in part, locally financed: more schools were built in more affluent communities.
- Exposure to the school construction program varied by region and year.
- Compares outcomes of older and younger individuals in regions where the school construction program was more and less active.

Duflo (2001): Results

The following table illustrates her identification strategy.

TABLE 3—MEANS OF EDUCATION AND LOG(WAGE) BY COHORT AND LEVEL OF PROGRAM CELLS

	Years of education			Log(wages)		
	Level of program in region of birth			Level of program in region of birth		
	High (1)	Low (2)	Difference (3)	High (4)	Low (5)	Difference (6)
Panel A: Experiment of Interest						
Aged 2 to 6 in 1974	8.49	9.76	−1.27	6.61	6.73	−0.12
	(0.043)	(0.037)	(0.057)	(0.0078)	(0.0064)	(0.010)
Aged 12 to 17 in 1974	8.02	9.40	−1.39	6.87	7.02	−0.15
	(0.053)	(0.042)	(0.067)	(0.0085)	(0.0069)	(0.011)
Difference	0.47	0.36	0.12	−0.26	−0.29	0.026
	(0.070)	(0.038)	(0.089)	(0.011)	(0.0096)	(0.015)
Panel B: Control Experiment						
Aged 12 to 17 in 1974	8.02	9.40	−1.39	6.87	7.02	−0.15
	(0.053)	(0.042)	(0.067)	(0.0085)	(0.0069)	(0.011)
Aged 18 to 24 in 1974	7.70	9.12	−1.42	6.92	7.08	−0.16
	(0.059)	(0.044)	(0.072)	(0.0097)	(0.0076)	(0.012)
Difference	0.32	0.28	0.034	0.056	0.063	0.0070
	(0.080)	(0.061)	(0.098)	(0.013)	(0.010)	(0.016)

Notes: The sample is made of the individuals who earn a wage. Standard errors are in parentheses.

Duflo (2001): Results

Results suggest that each new school constructed per 1,000 children was associated with:
- an increase of 0.12 to 0.19 in years of education, and
- a 1.5 to 2.7 percent increase in earnings for the first cohort fully exposed to the program.

This implies estimates of economic returns to education ranging from 6.8 to 10.6 percent. (Note: These estimates of economic returns to education were obtained using instrumental variables (IV) methods.)

Chapter 4: Within estimators

- Identify program impacts from differences in outcomes within some unit of observation, such as within a family, a school or a village.
- Let Y_{0ijt} and Y_{1ijt} denote the outcomes for individual i, from unit j, observed at time t,.
- For now, assume that $U_{1it} = U_{0it}$.
- Write the model for outcomes as:

$$Y_{ijt} = \varphi_0(X_{ijt}) + I_{ij}^{D}\gamma + D_{ijt}\alpha_{TT}(X_{ijt}) + \varepsilon_{ijt}$$

- Assume that the error term $\varepsilon_{ijt}(= U_{0it})$ can be decomposed as:

$$\varepsilon_{ijt} = \theta_j + v_{ijt}$$

where θ_j represents the effects of unobservables that vary across units but are constant for individuals within the same unit and v_{ijt} are *iid*.

Within estimators

- Taking differences between two individuals from the same unit observed in the same time period:

$$Y_{ijt} - Y_{i'jt} = \varphi_0(X_{ijt}) - \varphi_0(X_{i'jt}) + (I_{ij}^D - I_{i'j}^D)\gamma +$$

$$(\mathrm{D}_{ijt} - D_{i'jt})\alpha_{TT}(X_{ijt}) + (v_{ijt} - v_{i'jt}).$$

- Consistency for $\alpha_{TT}(X_{ijt})$ requires that

$$E(v_{ijt} - v_{i'jt} | X_{ijt}, X_{i'jt}, D_{ijt}, D_{i'jt}) = 0.$$

Within estimators

- Assumption implies that within a particular unit, which individual receives the treatment is random with respect to the error term v_{ijt}.
- If $U_{0it} \neq U_{1it}$, it has to be assumed that which individual receives treatment is random with respect to that individual's idiosyncratic gain from the program.
- Program may be nonrandomly targeted (e.g. at families or villages), but within units, which individuals participated must be unrelated to idiosyncractic program gain.
- Also, assumes no spillover effects from treating one individual on others within the same unit.
- Allows treatment to be selective across units ($E(\varepsilon_{ijt}|D_{ijt}, X_{ijt}) \neq 0$), because treatment selection can be based on the unobserved heterogeneity term θ_j.

Within estimators

- When the variation being exploited for identification of the treatment effect is variation within a family, village, or school at a single point in time, then requires a single cross-section of data.
- If all individuals in a unit receive treatment at the same time, then can take differences across different points in time, but will suffer same drawbacks as before-after approach.

Applications of Within estimators

- A study of the impact of a family planning and health counseling program on child outcomes in the Philippines (Rosenzweig and Wolpin, 1986)
- Evaluation of efficient use of inputs within households in Burkino Faso (Udry, 1996)
- Evaluation of impact of school meals on child nutrition in the Philippines (Jacoby, 2002)
- Impact of flip charts on student academic performance in Kenya (Glewwe et al. 2004)

Within estimator applications: Rosenzweig and Wolpin (1986)

- Rosenzweig and Wolpin (1986)–assesses the impact of a family planning and health counseling program on child outcomes in twenty villages (barrios) in the Phillipines.
- Discusses the statistical problems created when the placement of a program potentially depends on the outcome variable of interest.
- For example, family planning programs are often placed in areas where the need is considered to be the greatest. Not accounting for nonrandom placement would lead to the erroneous conclusion that family planning programs cause fertility.

Rosenzweig and Wolpin (1986): Data

- Information from surveys of 240 randomly selected households residing in these barrios on the age, height, and weight of every family member was collected in 1975 and 1979. Information was also obtained in the 1979 survey round on the dates of introduction of rural health clinics and family planning clinics financed by the national government for each of the barrios.

- To estimate the effects of the facilities on child health, Rosenzweig and Wolpin (1986) used a sample of 274 children (defined to be under age 18 as of 1979) in 85 house-holds for whom height and weight information exists in both years of the Laguna survey.

Empirical analysis adopts the following statistical model:

$$H_{ijt}^{a} = \rho_{ij}^{a}\beta + \mu_i + \mu_j + \varepsilon_{ijt},$$

where H_{ijt}^{a} is a child health measure (height, weight) for child i observed at age a, living in locality j at time t. ρ_{ij}^{a} represents the length of time that child was exposed to the program intervention.

μ_i is a time invariant, child-specific unobserved health endowment and μ_j is an unobserved locality level effect.

Compare results using different estimators: simple OLS regression (without controlling for either μ_i or μ_j), OLS regression controlling for community fixed effects (controlling for μ_j but not for μ_i), and first-differenced regressions (controlling for both μ_i and μ_j).

Rosenzweig and Wolpin (1986): Results

- The differences in estimated program exposure effects across specifications are striking. In the height regressions (panel A), both the cross-section and barrio fixed-effects estimate of health and family planning clinic effects are generally negative with standard errors that are at least as large as the point estimates.
- The child fixed-effect (longitudinal) estimates, however, indicate that exposure to health and family planning clinics increases height, with the family planning effect statistically significant and the health clinic effect marginally significant.
- The point estimates indicate that the height of a child for whom no health clinic existed would be 5 percent below that for a child always exposed to a clinic, while exposure to a family planning clinic increases height by 7 percent.
- The weight regressions tell a similar story (see Table 3).

Rosenzweig and Wolpin (1986): Results

Table 3 from Rosenzweig and Wolpin (1986)

Variable	*OLS* Cross Section[a]	Fixed Effect	
		Barrio	Child
A. Log of Standardized Height			
Rural Health Unit Exposure	−.00473	−.0205	.0511
	(0.53)	(0.40)	(1.21)
Family Planning Exposure	−.0131	−.00913	.0710
	(1.12)	(0.27)	(3.32)
R^2	.0339	.1695	.0660[b]
F	1.88	2.12	9.61
$d.f.$	268	249	272
B. Log of Standardized Weight			
Rural Health Unit Exposure	−.0313	−.162	.0992
	(1.35)	(1.20)	(1.52)
Family Planning Exposure	.0263	.0803	.121
	(0.87)	(0.90)	(2.76)
R^2	.0337	.1401	.0500[b]
F	1.87	1.69	7.16

[a] Equation also includes the age and educational attainment of each parent.
[b] From first-differenced equation.

Udry (1996)

- Used a within estimator to test for Pareto efficiency of allocations within households, an implication of cooperative bargaining models.
- Data: rural households in Burkina Faso (ICRISAT data).
- In those households, and in many other African households, agricultural production is carried out on several plots of land, with different plots being controlled by different members of the household.
- Pareto efficiency implies that farm inputs (seeds, fertilizer, etc.) be allocated efficiently, so the yield on a particular plot should not depend on which household member farms it.

Udry (1996) Estimation Strategy

$$Q_{htci} = X'_{hci}\beta + \gamma G_{htci} + \lambda_{htc} + \varepsilon_{htci}$$

where Q=plot yield,X=characterisics of the plot (land quality, size), h is household, G=gender of the individual who farms the plot (women=1,men=0), λ_{htc}=household year-crop fixed effect.

Finds that, on average, plots controlled by women have higher values of output per hectare than much smaller plots that are controlled by men. However, men and women grow different types of crops:

Udry (1996): Results

TABLE 1

MEAN YIELD, AREA, AND LABOR INPUTS PER PLOT BY GENDER OF CULTIVATOR
(N = 4,655)

	Crop Output per Hectare (1,000 FCFA)*	Area (Hectare)	Male Labor (Hours/ Hectare)	Female Labor (Hours/ Hectare)	Nonfamily Labor (Hours/ Hectare)	Child Labor (Hours/ Hectare)	Manure Weight (kg/ Hectare)
Men's plots	79.9	.740	593	248	106	104	2,993
	(186)	(1.19)	(1,065)	(501)	(407)	(325)	(11,155)
Women's plots	105.4	.100	128	859	46	53	764
	(286)	(.16)	(324)	(1,106)	(185)	(164)	(5,237)
T-statistic H_0: $\mu_m = \mu_w$	−3.27	29.03	22.16	−21.31	6.89	7.08	7.68

NOTE.—Standard deviations are in parentheses.
* In 1982, the exchange rate was approximately US$1 = FCFA 325.

Jacoby (2002)

- Studies whether public transfers targeted towards children *stick* to them, the *flypaper effect*, or whether their effect is diluted by intra-household reallocation of food at home away from the child and toward other household members.
- Studied the impact of a school feeding program in the Philippines using data on 3,189 children in 159 schools.
- DID estimation strategy compares inter-day (school-day vs. non-school day) calorie differentials across program participants and nonparticipants.

Jacoby (2002)

- Analysis assumes that the only reason that the calorie intake of program participants varies across school and non-school days, relative to nonparticipants, is because of the school feeding program.
- A potential threat to the validity of this DID strategy is that the feeding program might be targeted at poorer households, and poor children may spend more time working when not in school and therefore have higher calorie consumption, which would tend to bias the estimates in favor of finding a fly-paper effect.

Jacoby (2002)

$$C_{is}^{T} = \alpha_P D_s^P \times D_{is}^A + \alpha_A D_{is}^A + \delta_S + U_{is}$$

C_{is}^{T}=daily calorie data for child i in school s, D_{is}^{A}=indicator that calorie data for child i in school s is for a school day, D_s^P=indicator for whether school s offers a feeding program,δ_s=a school fixed effect,U_{is} unobserved child specific determinants of calorie intake.

Note: In some estimations, Jacoby (2002) replaced the term $D_s^P \times D_{is}^A$ with $D_s^P \times D_{is}^A \times C_{is}^P$, where C_{is}^P is calories from the program. Thus, $\alpha_P = 1$ implies program calories stick with the child.

Jacoby (2002): Results

Impact of School Feeding Programmes on Caloric Intake

Specification	Morning + Afternoon Snack			Morning Snack Only			Total Daily Calories		
	$\hat{\alpha}_P$	$\hat{\alpha}_A$	p-value	$\hat{\alpha}_P$	$\hat{\alpha}_A$	p-value	$\hat{\alpha}_P$	$\hat{\alpha}_A$	p-value
All Schools ($N = 3{,}189$)									
(1) OLS, $\boldsymbol{\beta} = \mathbf{0}$	0.985	88.3		0.892	84.2		1.059	79.9	
	(0.069)	(10.1)		(0.039)	(6.5)		(0.140)	(23.9)	
(2) OLS	0.983	84.7	0.000*	0.897	82.8	0.000*	1.104	61.1	0.000*
	(0.068)	(10.0)		(0.038)	(6.5)		(0.134)	(21.9)	
(3) 2SLS (IV1)†	1.019	83.1	0.662‡	0.912	82.4	0.757‡	1.358	49.5	0.443‡
	(0.200)	(12.8)		(0.161)	(8.1)		(0.452)	(28.7)	
(4) 2SLS (IV2)§	0.977	85.0	0.635‡	0.911	100.0	0.751‡	1.153	38.9	0.739‡
	(0.198)	(16.9)		(0.144)	(10.7)		(0.460)	(39.0)	
(5) 2SLS (IV2), Day of interview dummies	1.053	118.3	0.148‖	0.931	128.0	0.266‖	1.082	65.6	0.616‖
	(0.202)	(26.9)		(0.148)	(16.2)		(0.464)	(66.0)	
(6) 2SLS (IV2), Month of interview dummies	1.106	53.0	0.024**	1.004	82.8	0.146**	1.108	37.7	0.658**
	(0.198)	(18.4)		(0.145)	(11.2)		(0.458)	(41.0)	
Schools with $n_s \geq 20$ ($N = 2{,}489$)									
(7) Tobit††	0.961	104.8		0.920	130.2				
	(0.051)	(10.9)		(0.049)	(8.5)				
(8) Two-Stage Tobit (IV2)‡‡	0.855	115.8	0.659‡	0.891	160.7	0.923‡			
	(0.243)	(22.4)		(0.225)	(16.2)				
(9) Censored LAD§§	0.929	96.1		0.921	122.7				
	(0.068)	(14.8)		(0.063)	(16.4)				

Within estimator applications: Glewwe, Kremer, Moulin and Zitzewitz (2004)

- Questions reliability of DID estimation approach in an application evaluating the effectiveness of an educational intervention in Kenyan schools.
- Program provided flip-charts as teaching aids in certain subjects.
- Compares DID estimates to those obtained from a randomized social experiment.
    - DID estimator compares changes over time in test scores in flip-chart and non-flip-chart subjects within the schools that received the intervention.
    - The experiment randomly allocated the flip-charts to a subset of schools.
- The experimental results indicate that flip-charts had little effect on test scores, while the DID estimates are statistically significantly different from zero at conventional levels.

Kremer et. al. (2004): Results

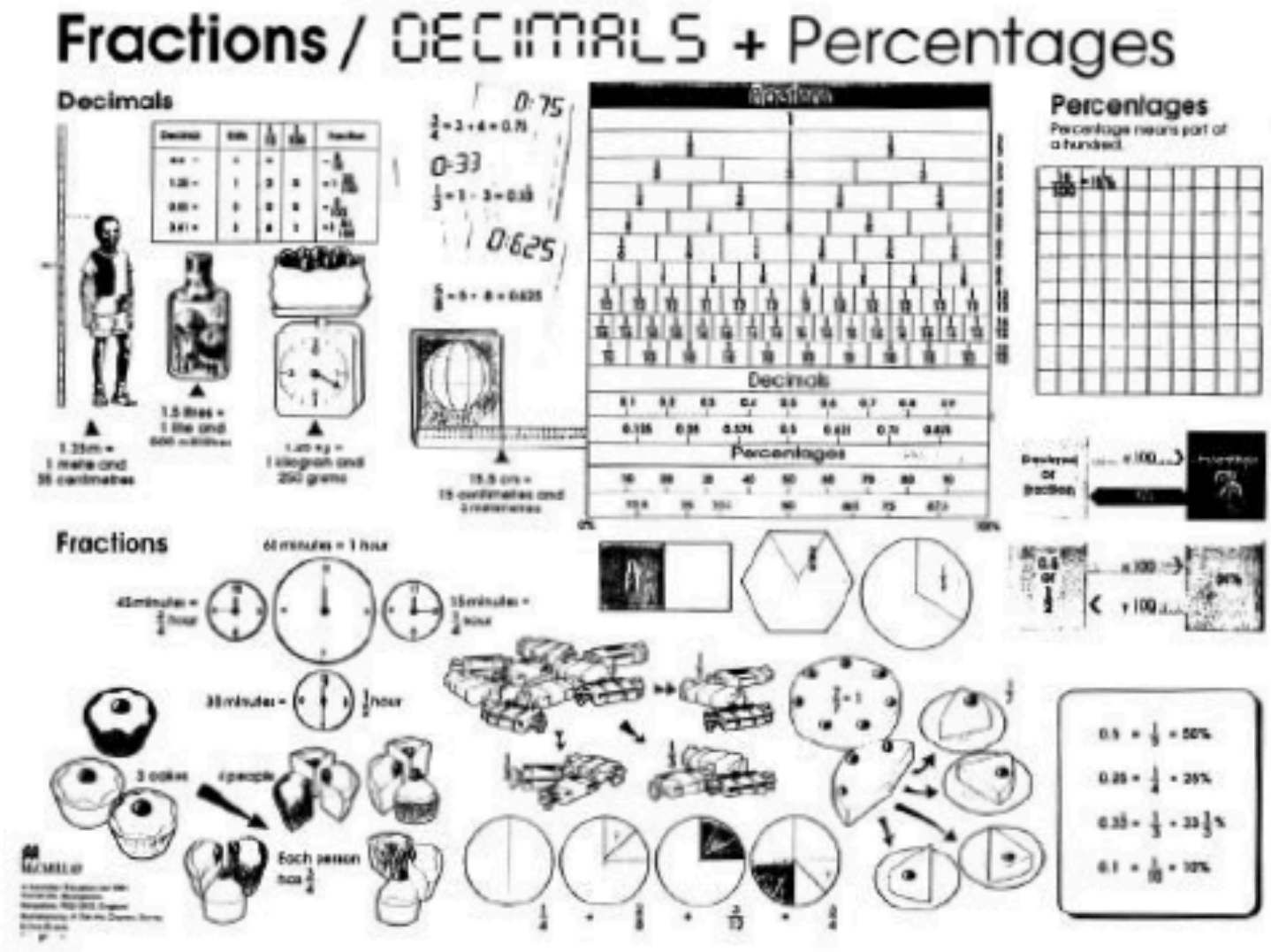


Fig. 1. Example of Mathematics Flip Chart (Macmillan Education Ltd. 1994)

Kremer et. al. (2004): Results

Results: The experimental results indicate no learning effect, but DID estimates are statistically significant. They conclude that the DID estimates are unreliable.

Retrospective estimates of effect of four flip charts in grades 6–8

Dependent variable: normalized 1998 test scores							
Specification		(1)	(2)	(3)	(4)	(5)	(6)
	Mean(S.D.)	Level estimates				Diffs-in-diffs	
Random effects							
School		Yes	Yes	Yes	Yes	Yes	Yes
School × subject	No	No	No	No	No	No	Yes
Schools		83	79	79	79	79	79
Pupils		5152	4998	4998	4998	4998	4998
Grades included		6–8	6–8	6–8	6–8	6–8	6–8
Subjects included		Sc, Mat, HS	Sc, Mat, HS	Sc, Mat, HS	Sc, Mat, HS	All	All
Flip chart variable							
Number of charts in school (divided by four)	1.1 (2.4)	0.192*** (0.080)	0.194*** (0.065)	0.205*** (0.064)	0.076* (0.041)	0.154*** (0.057)	0.157*** (0.056)
Charts × flip-chart subject (Science/ Agr., Math, HS–BE)						0.049** (0.021)	0.040* (0.024)
Other variables							

Note: Cols. (1) - (4) are level estimates. Cols. (5)-(6) are DID estimates. Col (2), (3), (5) and (6) control for school inputs (textbooks, teacher training); Col (1) does not. Col (4) controls for student scores on non-flip chart subjects.

Kremer et. al. (2004): Results

Prospective estimates of effect of flip charts—single subject multi-test regressions

Dependent variable: normalized test score

Subject	Past perf.	Flip-chart school		Obs.
	Controls	Coeff.	S.E.	
Flip-chart subjects				
Science/Agriculture	No	0.0005	0.0752	20,446
	Yes	−0.0007	0.0591	
Math	No	−0.0201	0.0600	20,441
	Yes	−0.0212	0.0486	
Health Science/Business Ed. (HS−BE)	No	−0.0295	0.0728	20,434
	Yes	−0.0276	0.0559	
Geography/History/Civics/Religious Ed. (GHC)	No	0.0018	0.0714	20,450
	Yes	−0.0012	0.0553	
Non-flip-chart subjects				
English	No	0.0038	0.0737	20,433
	Yes	−0.0100	0.0576	
KiSwahili	No	0.0110	0.0790	20,448
	Yes	0.0146	0.0737	
Arts/Crafts/Music (ACM)	No	−0.0679	0.0758	20,417
	Yes	−0.0723	0.0589	
Memo				
Math and Science; grades 6 and 7 in 1998 only	No	0.0508	0.0828	13,836
	Yes	0.0534	0.0655	

Regressions include school and school × year random effects and test fixed effects. Past performance controls are controls for the school-average performance on the July 1996 practice exam.

Kremer et. al. (2004): Results

Prospective estimates of effect of flip charts—single test multi-subject regressions

Dependent variable: normalized test score

Test	Grade	Past perf.	4 Flip chart subjects		3 Non-flip chart subjects		Obs.
		Controls	Coeff.	S.E.	Coeff.	S.E.	
All tests	6–8	No	−0.0117	0.0638	−0.0149	0.0649	143,069
		Yes	−0.0063	0.0484	−0.0144	0.0498	141,698
Jul-97	8	No	−0.0138	0.0716	−0.0388	0.0751	25,939
		Yes	−0.0347	0.0605	−0.0627	0.0644	25,827
Nov-97	8	No	−0.0474	0.0744	−0.0516	0.0758	25,418
		Yes	−0.0656	0.0601	−0.0700	0.0617	25,418
Jul-98	8	No	0.0135	0.0848	0.0102	0.0866	17,882
		Yes	0.0325	0.0718	0.0291	0.0739	17,791
Nov-98	8	No	−0.0018	0.0708	−0.0134	0.0722	25,396
		Yes	0.0145	0.0575	0.0043	0.0591	25,060
Oct-98	7	No	−0.0061	0.0910	−0.0029	0.0925	24,708
		Yes	0.0327	0.0669	0.0268	0.0690	24,288
Oct-98	6	No	0.0708	0.1005	0.0612	0.1019	23,726
		Yes	0.0579	0.0799	0.0485	0.0817	23,314

Regressions include school, school × subject, and pupil random effects subject and test fixed effects. Pupil random effects cannot be included when results are estimated for all tests due to computational constraints. For the single-test results, excluding pupil effects changes point estimates by no more than 0.0045 and standard errors by no more than 0.001.

Chapter 5: Matching estimators

- A widely-used method of evaluation that compares the outcomes of program participants with the outcomes of similar, matched nonparticipants.
- Methods were first used in economics to evaluate effects of job training programs, matching program participants to nonparticipants. (See, e.g., Heckman, Ichimura and Todd (1997, 1998), Dehejia and Wahba (1999), Smith and Todd (2005).)
- Other early applications were to evaluate economic development and anti-poverty programs.
- One of the main advantages of matching estimators is that they do not require specifying the functional form of the outcome equation and are therefore not susceptible to bias due to misspecification along that dimension.

- Traditional matching estimators, proposed in the statistics literature, pair each program participant with an observably similar nonparticipant and interpret the difference in their outcomes as the effect of the program intervention (see, e.g., Rosenbaum and Rubin, 1983).
- More recently developed methods pair program participants with more than one nonparticipant observation, using statistical methods to estimate the matched outcome.
- Propensity score matching - match on conditional probability of participating in the program (most popular approach).

Two main variants of matching estimators

- *cross-sectional matching*
  - allow for selection on unobservables only in a very limited sense.
  - applicable in contexts where the researcher is relatively certain that the major determinants of program participation are accounted for and that any remaining variation in who participates is due to random factors.
- *difference-in-difference matching*
  - identify treatment effects by comparing the change in outcomes for treated persons to the change in outcomes for matched, untreated persons.
  - allow program selection to be based on unobserved time-invariant characteristics of individuals.

Cross-sectional matching

Assume that the outcomes (Y_0, Y_1) are independent of participation status D conditional on a set of characteristics Z,

$$(Y_0, Y_1) \perp\!\!\!\perp D \,|Z \ . \tag{4}$$

In the terminology of Rosenbaum and Rubin (1983) treatment assignment is *strictly ignorable* given Z. Also assumed that

$$0 < \Pr(D = 1|Z) < 1. \tag{5}$$

Assumption required so that matches for $D = 0$ and $D = 1$ observations can be found.

Cross-sectional matching

If assumptions satisfied, then get mean program impacts by simply substituting the Y_0 distribution observed for the matched non-participant group for the missing Y_0 distribution for program participants, holding constant observables.

Heckman, Ichimura and Todd (1998) show that the above assumptions are overly strong if the parameter of interest is the mean impact of treatment on the treated (TT)- require only conditional mean independence on Y_0:

$$E(Y_0|Z, D = 1) = E(Y_0|Z, D = 0) = E(Y_0|Z). \tag{6}$$

When α_{TT} is the parameter of interest, only require

$$\Pr(D = 1|Z) < 1. \tag{7}$$

Cross-sectional matching

Under these assumptions, the mean impact of the program on program participants can be written as

$$\begin{aligned}\Delta &= E(Y_1 - Y_0|D=1) \\ &= E(Y_1|D=1) - E_{Z|D=1}\{E_Y(Y|D=1,Z)\} \\ &= E(Y_1|D=1) - E_{Z|D=1}\{E_Y(Y|D=0,Z)\},\end{aligned}$$

where the second term can be estimated from the mean outcomes of the matched (on Z) comparison group.

Note:
$E_{Z|D=1}\{E_Y(Y|D=0,Z)\} = \int_z \int_y yf(y|D=0,z)f(z|D=1)dydz)$.

Cross-sectional matching

- Conditional independence assumption implies that D does not predict values of Y_0 conditional on Z.
- Selection into the program cannot be based directly on anticipated values of Y_0, other than that which is forecastable given Z.
- No restriction is imposed on Y_1, so the method does allow individuals who expect high levels of Y_1 to select into the program.
- Accommodates selection on unobservables, but only in a very limited sense through Y_1.

Non-overlapping support

- With nonexperimental data, there may or may not exist a set of observed conditioning variables for which matching conditions hold.
- A finding of Heckman, Ichimura and Todd (1997) and HIST (1996,1998) in their application of matching methods to data from the JTPA experiment is that $0 < Pr(D = 1|Z) < 1$ was not satisfied, because no close match could be found for a fraction of the program participants.
- If there are regions where the support of Z does not overlap for the $D = 1$ and $D = 0$ groups, then matching is only justified when performed over the *region of common support*.
- The estimated average treatment effect must then be defined conditionally on the region of overlap.

Rosenbaum and Rubin (1983) provide a theorem that is useful in reducing the dimension of the conditioning problem. They show that for random var. Y and Z and a discrete random var. D:

$$E(D|Y, P(D=1|Z)) = E(E(D|Y,Z)|Y, \Pr(D=1|Z)),$$

so that

$$E(D|Y,Z) = E(D|Z) \Longrightarrow E(D|Y, \Pr(D=1|Z)) = E(D|\Pr(D=1|Z)).$$

Using the Rosenbaum and Rubin (1983) theorem, the matching procedure can be broken down into two stages:

- In the first stage, the propensity score $\Pr(D = 1|Z)$ is estimated, using a binary discrete choice model such as a logit or probit or a semiparametric estimation method (such as Ichimura's (1993) semiparametric least squares (SLS))
- In the second stage, individuals are matched on the basis of their first stage estimated probabilities of participation.

Alternative matching estimators

Let $P = P(D = 1|Z)$. A typical cross-sectional matching estimator takes the form:

$$\hat{\alpha}_M = \frac{1}{n_1} \sum_{i \in I_1 \cap S_P} [Y_{1i} - \hat{E}(Y_{0i}|D = 1, P_i)] \tag{8}$$

$$\hat{E}(Y_{0i}|D = 1, P_i) = \sum_{j \in I_0} W(i,j) Y_{0j},$$

- I_1 denotes the set of program participants, I_0 the set of non-participants
- S_P the region of common support.
- n_1 denotes the number of persons in the set $I_1 \cap S_P$.
- The match for each participant $i \in I_1 \cap S_P$ is constructed as a weighted average over the outcomes of non-participants, where the weights $W(i,j)$ depend on the distance between P_i and P_j.

Alternative matching estimators

- Define a neighborhood $C(P_i)$ for each i in the participant sample.
- Neighbors for i are non-participants $j \in I_0$ for whom $P_j \in C(P_i)$.
- Persons matched to i are those people in set A_i where $A_i = \{j \in I_0 \mid P_j \in C(P_i)\}$.
- Alternative matching estimators (discussed below) differ in how the neighborhood is defined and in how the weights $W(i,j)$ are constructed.

Nearest Neighbor matching

Traditional, pairwise matching sets

$$C(P_i) = \min_j \|P_i - P_j\|, \; j \in I_0.$$

- That is, the non-participant with the value of P_j that is closest to P_i is selected as the match and A_i is a singleton set. - The estimator can be implemented either matching with or without replacement. When matching is performed with replacement, the same comparison group observation can be used repeatedly as a match. A drawback of matching without replacement is that the final estimate will likely depend on the initial ordering of the treated observations for which the matches were selected.

Caliper matching (Cochran and Rubin, 1973)

- A variation of nearest neighbor matching that attempts to avoid bad matches (those for which P_j is far from P_i) by imposing a tolerance on the maximum distance $\|P_i - P_j\|$ allowed.
- A match for person i is selected only if $\|P_i - P_j\| < \varepsilon$, $j \in I_0$, where ε is a pre-specified tolerance.
- Neighborhood is $C(P_i) = \{P_j \mid \|P_i - P_j\| < \varepsilon\}$.
- Treated persons for whom no matches can be found excluded (way of imposing a common support condition).
- Difficult to know a priori what choice for the tolerance level is reasonable.

Stratification or interval matching

- Common support of P is partitioned into a set of intervals.
- Separate impact calculated by taking the mean difference in outcomes between the $D = 1$ and $D = 0$ observations within the interval.
- Weighted average of the interval impact estimates, using the fraction of the $D = 1$ population in each interval for the weights, provides an overall impact estimate.
- Dehejia and Wahba (1999) choose intervals selected such that the mean values of the estimated P_i's and P_j's are not statistically different from each other within intervals.

Kernel matching

Construct matches using a weighted average over multiple persons in the comparison group.

Consider a nonparametric *kernel matching estimator*, given by

$$\hat{\alpha}_{KM} = \frac{1}{n_1} \sum_{i \in I_1} \left\{ Y_{1i} - \frac{\sum_{j \in I_0} Y_{0j} G\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k \in I_0} G\left(\frac{P_k - P_i}{a_n}\right)} \right\}.$$

where $G(\cdot)$ is a kernel function and a_n is a bandwidth parameter. (See Heckman, Ichimura and Todd (1997, 1998) and Heckman, Ichimura, Smith and Todd (1998)

The weighting function, $W(i,j)$, is equal to $\frac{G\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k \in I_0} G\left(\frac{P_k - P_i}{a_n}\right)}$.

Kernel matching

- For a kernel function bounded between -1 and 1, the neighborhood is $C(P_i) = \{|\frac{P_i - P_j}{a_n}| \leq 1\}, j \in I_0$.
- Under standard conditions on the bandwidth and kernel function ($G(\cdot)$ integrates to one, has mean zero and that $a_n \to 0$ as $n \to \infty$ and $na_n \to \infty$.), $\frac{\sum_{j \in I_0} Y_{0j} G\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k \in I_0} G\left(\frac{P_k - P_i}{a_n}\right)}$ is a consistent estimator of $E(Y_0|D = 1, P_i)$.

Determining the overlapping support region

To determine the support region, Heckman, Ichimura and Todd (1997) use kernel density estimation methods:

$$\hat{S}_P = \{P : \hat{f}(P|D=1) > 0 \text{ and } \hat{f}(P|D=0) > c_q\},$$

where $\hat{f}(P|D=d)$, $d \in \{0,1\}$ are nonparametric density estimators given by

$$\hat{f}(P|D=d) = \sum_{k \in I_d} G\left(\frac{P_k - P}{a_n}\right),$$

and where a_n is a bandwidth parameter.

Determining the overlapping support region

To ensure that the densities are strictly greater than zero, it is required that the densities be strictly positive density (i.e. exceed zero by a certain amount), determined using a "trimming level" q. The set of eligible matches is:

$$\hat{S}_q = \{P \in \hat{S}_P : \hat{f}(P|D=1) > c_q \text{ and } \hat{f}(P|D=0) > c_q\},$$

where c_q is the density cut-off level that satisfies:

$$\sup_{c_q} \frac{1}{2J} \sum_{\{i \in I_1 \cap \hat{S}_P\}} \{1(\hat{f}(P|D=1) < c_q + 1(1(\hat{f}(P|D=0) < c_q\} \leq q.$$

J is the cardinality of the set of observed values of P that lie in $I_1 \cap \hat{S}_P$.

- Matches are constructed only for participants for which propensity scores lie in $\hat{S}_q$.

- The literature has developed some alternative, more efficient estimators. See, for example, Hahn (1998) and Hirano, Imbens and Ridder (2003).
- Heckman, Ichimura and Todd (1998) propose a regression-adjusted matching estimator that replaces Y_{0j} as the dependent variable with the residual from a regression of Y_{0j} on a vector of exogenous covariates.
- In principal, imposing exclusions restrictions can increase efficiency. In practice, there was not much gain from using the regression-adjusted matching estimator.

Difference-in-difference matching

- For a variety of reasons there may be systematic differences between participant and nonparticipant outcomes, even after conditioning on observables, which could lead to a violation of the matching assumptions.
- For example, could have program selectivity on unmeasured characteristics, or levels differences in outcomes across different labor markets in which the participants and nonparticipants reside.
- A difference-in-differences (DID) matching strategy, as defined in Heckman, Ichimura and Todd (1997) and Heckman, Ichimura, Smith and Todd (1998), better accomodates the potential for selection on unobservables by allowing for temporally invariant differences in outcomes between participants and nonparticipants.

Difference-in-difference matching

Estimator is analogous to the standard DID regression estimator defined, but reweights the observations according to the weighting functions implied by matching estimators.
Assumes that that

$$E(Y_{0t} - Y_{0t'}|P, D = 1) = E(Y_{0t} - Y_{0t'}|P, D = 0),$$

where t and t' are time periods after and before the program enrollment date.
Also requires the support condition, which must hold in periods t and t'.

Difference-in-difference matching

The local linear difference-in-difference estimator is:

$$\hat{\alpha}_{KDM} = \frac{1}{n_1} \sum_{i \in I_1 \cap S_P} \left\{ (Y_{1ti} - Y_{0t'i}) - \sum_{j \in I_0 \cap S_P} W(i,j)(Y_{0tj} - Y_{0t'j}) \right\},$$

with LLR weights. If repeated cross-section data

$$\hat{\alpha}_{KDM} = \frac{1}{n_{1t}} \sum_{i \in I_{1t} \cap S_P} \left\{ (Y_{1ti} - \sum_{j \in I_{0t} \cap S_P} W(i,j) Y_{0tj} \right\}$$

$$- \frac{1}{n'_{1t}} \times \sum_{i \in I_{1t'} \cap S_P} \left\{ (Y_{1t'i} - \sum_{j \in I_{0t'}} W(i,j) Y_{0t'j} \right\},$$

where $I_{1t}, I_{1t'}, I_{0t}, I_{0t'}$ denote the treatment and comparison group datasets in each time period.

Difference-in-difference matching

- Allows selectivity into the program to be based on anticipated gains from the program, but only in a limited way

- D can help predict the value of Y_1 given P., but D cannot predict changes Y_0 (i.e. $Y_{0t} - Y_{0t'}$) conditional on P.

Matching with choice-based sampled data

- Samples used in evaluating the impacts of programs are often choice-based, with program participants being oversampled relative to their frequency in the population.
- Weights are required to consistently estimate the probabilities of program participation, where the weights equal the ratio of the proportion of program participants in the population relative to the proportion in the sample.(see, e.g., Manski and Lerman (1977)).
- True population proportions usually are not obtainable from the sample and have to be derived from some other sources.

Matching with choice-based sampled data

- When weights are known, the Manski and Lerman (1977) procedure can be used to consistently estimate propensity scores.
- If weights not known, Heckman and Todd (1995) show that with a slight modification, matching methods can still be applied, because the odds ratio $(P/(1-P))$ estimated using a logistic model with incorrect weights (i.e. ignoring the fact that samples are choice-based) is a scalar multiple of the true odds ratio, which is itself a monotonic transformation of the propensity scores.
- Matching can proceed on the (misweighted) estimate of the odds ratio (or the log odds ratio).
- Failure to account for CBS will not affect nearest-neighbor point estimates, but will matter for kernel or local linear matching methods, because these methods take into account the absolute distance between the P observations.

When does bias arise in matching?

- Success of a matching estimator depends on the availability of observable data to construct the conditioning set Z, such that the matching assumptions are satisfied. - Suppose only a subset $Z_0 \subset Z$ of the variables required for matching is observed. The propensity score matching estimator based on Z_0 then converges to

$$\alpha'_M = E_{P(Z_0)|D=1}\left(E(Y_1|P(Z_0), D=1) - E(Y_0|P(Z_0), D=0)\right). \tag{8}$$

- The bias for the parameter of interest, $E(Y_1 - Y_0|D=1)$, is

$$bias_M = E(Y_0|D=1) - E_{P(Z_0)|D=1}\{E(Y_0|P(Z_0), D=0)\}.$$

Choosing the matching variables

- No statistical procedure for choosing the set. The set Z that satisfies the matching conditions is not necessarily the most inclusive one. Augmenting a set that satisfies the conditions for matching could lead to a violation of the conditions.
- Using too many conditioning variables could also exacerbate a common support problem.
- Heckman, Ichimura, Smith and Todd (1998), Heckman, Ichimura and Todd (1999) and Lechner (2001) show that which variables are included in the estimation of the propensity score can make a substantial difference to the estimator's performance.
- Biases tended to be more substantial when cruder sets of conditioning variables where used.
- The set Z can be chosen to maximize the percent of people correctly classified by treatment status under the model.

Other determinants of the performance of matching estimators

- Perform best when the treatment and control groups are located in the same geographic area, so that regional effects on outcomes are held constant.
- Important to use the same survey to gather data on the comparison group and treatment group, so variables measured in same way.
- Difference-in-difference matching methods are more reliable than cross-sectional matching methods when treatments and controls are mismatching geographically or in terms of the survey instrument.
- The success of matching depends strongly on the data capturing the key determinants of the program participation decision.

Balancing tests

- Rosenbaum and Rubin (1983) present a theorem that does not aid in choosing which variables to include in Z, but which can help in determining which interactions and higher order terms to include in the propensity score model for a given set of Z variables.
- The theorem states that

$$Z \perp\!\!\!\perp D | \Pr(D = 1|Z),$$

or equivalently

$$E(D|Z, \Pr(D = 1|Z)) = E(D|\Pr(D = 1|Z)).$$

- After conditioning on $\Pr(D = 1|Z)$, additional conditioning on Z should not provide new information about D. If there is still dependence on Z, this suggests misspecification in the model used to estimate $\Pr(D = 1|Z)$.
- Theorem holds for any Z, including sets Z that do not satisfy the conditional independence condition required to justify matching. As

Balancing tests

- Motivates a specification test for $\Pr(D = 1|Z)$, which tests whether or not there are differences in Z between the $D = 1$ and $D = 0$ groups after conditioning on $P(Z)$.
- Various testing approaches have been proposed in the literature.
- Eichler and Lechner (2001) use a test based on standardized differences in terms of means of each variable in Z, squares of each variable in Z and first-order interaction terms between each pair of variables in Z.
- Dehijia and Wahba (1999,2001) divide the observations into strata based on estimated propensity scores and do tests within strata. Common to use five strata.

Balancing tests

Another way of implementing the balancing test estimates a regression of each element of the set Z, Z_k on D interacted with a power series expansion in $P(Z)$:

$$\begin{aligned} Z_k &= \alpha + \beta_1 P(Z) + \beta_2 P(Z)^2 + \beta_3 P(Z)^3 + ... + \beta_j P(Z)^j + \\ &\quad \gamma_1 P(Z) D + \gamma_2 P(Z)^2 D + \gamma_3 P(Z)^3 D + ... + \gamma_j P(Z)^j D + \nu, \end{aligned}$$

and then tests whether the estimated γ coefficients are jointly insignificantly different from zero.

- When significant differences are found , higher order and interaction terms in those variables are added to the logistic model and the testing procedure is repeated, until such differences no longer emerge.

Assessing the variability of matching estimators

- Distribution theory for cross-sectional and DID kernel and local linear matching estimators is derived in Heckman, Ichimura and Todd (1998), although implementing the asymptotic formulae can be cumbersome.
- Bootstrapping can be used, but only if estimators use a fixed bandwidth.
- Imbens and Abadie (2004a) shows that standard bootstrap resampling methods are not valid for assessing the variability of nearest neighbor estimators and they present alternative standard error formulae.

Chapter 6: Modeling program participation

- Propensity score plays an important role in the implementation of matching estimators and in other estimators considered later.
- There is no statistical method for determining which observed variables belong in the propensity score model, but one can use economic theory as a guide.
- Use model similar to that developed in Heckman, Lalonde, and Smith (1999)

Modeling program participation

- Assume individuals have the option to take training in period k.
- Prior to k, we observe Y_{0j}, $j = 1, .., k$.
- After k, we observe two potential outcomes Y_{0t} and Y_{1t}
- To participate in training, individuals must apply and be accepted, so there may be several decision-makers governing participation.
- $D = 1$ if participates and $D = 0$ else.
- Participation decisions are based on maximization of future earnings and that future earnings are uncertain.

Modeling program participation

$D = 1$ if

$$E[\sum_{j=1}^{T=k} \frac{Y_{1,k+j}}{(1+r)^j} - C - \sum_{j=0}^{T=k} \frac{Y_{0,k+j}}{(1+r)^j} | I_k] \geq 0$$

- First term is the earnings stream if the person participates in the program
- C is the direct cost of training.
- The last term is the earnings stream if the person does not participate, which includes period k earnings.
- I_k is the information set at time k used to form expectations about future earnings.

Modeling program participation

Implications about who takes training and about the value of past earnings in modeling program participation:

1. Past earnings are irrelevant except for value in predicting future earnings
2. Persons with lower foregone earnings or lower costs are more likely to participate in programs
3. Older persons and persons with higher discount rates are less likely to participate
4. The decision to take training is correlated with future earnings only through the correlation with expected future earnings.

Modeling program participation

A special case of this model is when the treatment effect is constant, in which case

$$D = 1 \text{ if } E[\sum_{j=1}^{T=k} \frac{\alpha}{(1+r)^j} | I_k] \geq C + Y_{0k}$$

If period k earnings (Y_{0k}) are temporarily low (e.g. currently unemployed), then people are more likely to enroll in the program. This implication of the model is consistent with the Ashenfelter Dip pattern.

Modeling program participation

To get an empirically implementable model, as in Heckman, Lalonde and Smith (1999), assume that the expected future rewards are modeled as a function of some X variables that capture the information set used in forecasting future earnings:

$$H(X) = \sum_{j=1}^{T=k} \frac{Y_{1,k+j}}{(1+r)^j} - \sum_{j=1}^{T=k} \frac{Y_{0,k+j}}{(1+r)^j} | I_k]$$

and that costs of training, including of foregone earnings costs are unobserved:

$$V = C + Y_{0k}.$$

Modeling program participation

Then, $D = 1$ if $H(X) - V \geq 0$. If we further assume that V is stochastically independent of X and distributed either logistic or normal (with mean μ_v and variance σ_v^2), we get either a logistic or normal propensity score model:

$$Pr(D = 1|X) = \frac{e^{H(X)}}{1 + e^{H(X)}}$$

or

$$Pr(D = 1|X) = \Phi(\frac{H(X) - \mu_v}{\sigma_v})$$

Modeling program participation

1. To summarize, variables included in the propensity score should be those variables that an individual might use to forecast future outcomes, which determine the benefits from participating in a program.

2. Past outcomes are relevant to the extent that they are used to forecast future outcomes, with or without the treatment intervention.

Evidence on performance of matching estimators

- We can study the performance of nonexperimental estimators by comparing experimental and nonexperimental estimates.
- One strategy is to directly compare a randomized-out control group to a nonperimenal group.
- Because neither group received treatment, any impact estimator applied to those groups should give a value of zero.
- Can also study how performance varies with conditioning variables or data sources used.

Data quality

TABLE 1

Comparison groups used in different studies

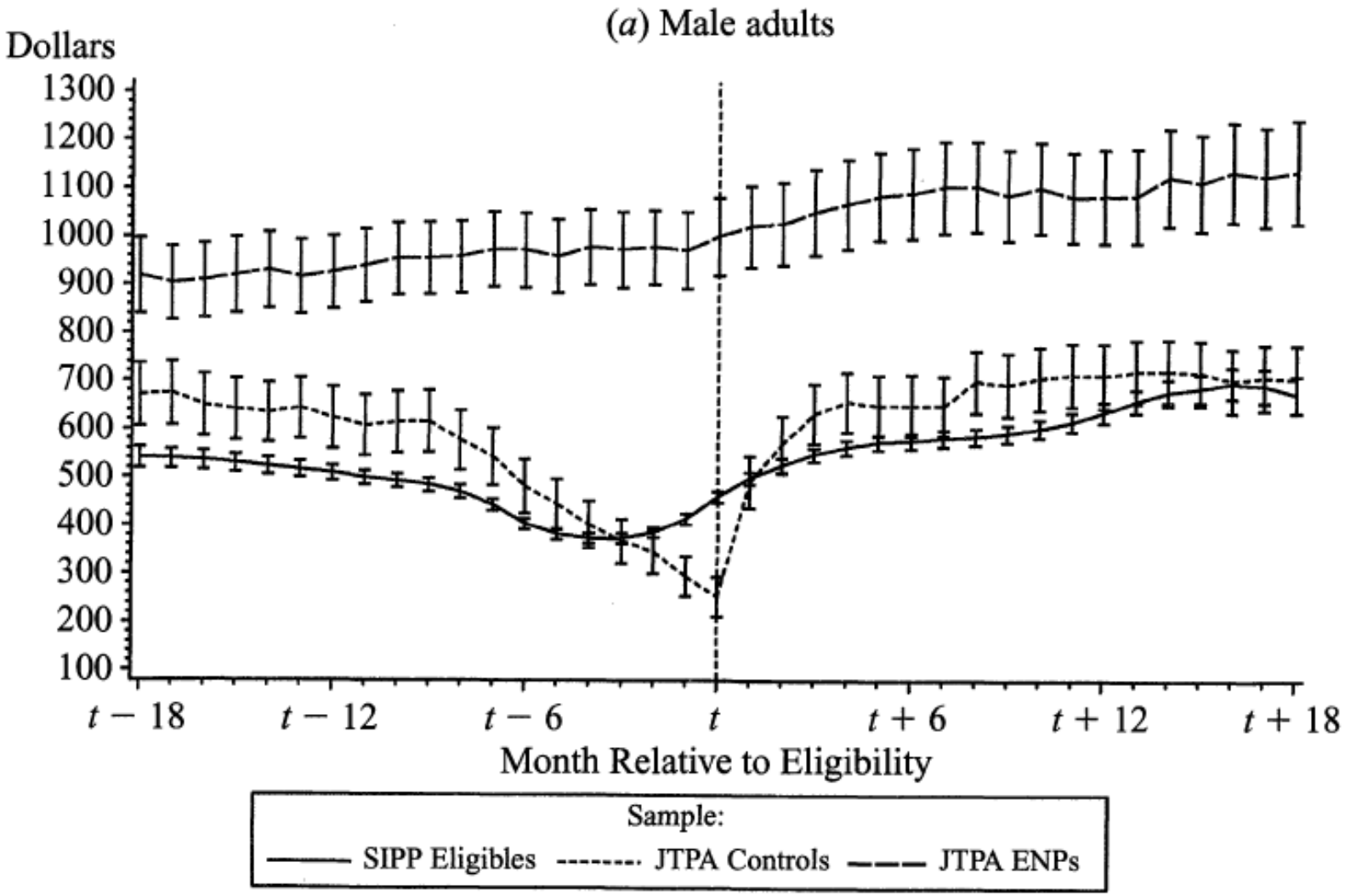
Study	Ashenfelter (1978)	Ashenfelter and Card (1985)	Dickinson–Johnson–West (1987)	Westat (1986) (Rupp and Bryant)
Programme, year, outcome variable	MDTA Classroom Trainees (First 3 months 1964), 1965–1969 annual social security record earnings	1976 CETA Trainees (1977, 1978 annual social security record earnings), CLMS data	1978 annual social security earnings. CETA trainees enrolled in 1976, CLMS data	2 cohorts of CETA Trainees 1977, 1978 annual social security record earnings, CLMS data
(1) Comparison group in the same labour market?	No	No	No	No
(2) Same questionnaire administered to comparison and treatment group	Yes	Yes	Yes	Yes
(3) Matching criteria (criteria for membership in comparison sample is also called "screening" criteria)	None specified	(a) 1975 earnings $\leq$ \$20K Household income $\leq$ \$30K (b) In labour force (March, 1976) Matched on age (persons $\geq$ 21 used)	(Matching based on a metric over vectors of variables) Matched on predictors of 1978 earnings including lagged earnings (1975–1970), worked in public sector, sex, and demographics. In labour force, March, 1976	Match on 1976 earnings, change in 1976 earnings (1975–1976, 1974–1975) change in earnings, demographics, 1975 labour force status, family income (for 1976–1977 cohort one year previous for 1975–1976 cohort). Either in the labour force, 1975 or at interview March 1976. Three matching groups based on income.
(4) Eligibility for programme known for comparison group members?	No	No	No	No
Variables used in analysis				
Age, race, sex	Yes (No age restriction)	Yes (Age $\geq$ 21 years old)	Yes (Age 21–65)	Yes (Age 14–60)
Education	No	Yes	Yes	Yes
Training history	No	No	No	No
Children	No	No	No	No
Employment histories	No	No	Yes (recent)	Yes (recent)
Hours worked	No	Yes	Yes	Yes
Unemployment histories	No	No	Yes (recent)	Yes (recent)
On welfare	No	No	Yes	No
Earning histories**	(Annual earnings) 5 years pre-programme 5 years post-programme	(Annual earnings) 2 years pre-programme 2 years post-programme	Same as Ashenfelter and Card (1985)	(Annual earnings) 4 years pre-enrollment earnings histories

** CLMS data matched Social Security Longitudinal Records to March CPS data for 1976 and 1977. The CPS data are for comparison group members. SSA data on longitudinal earnings are available for both groups. All of the personal and family information available in the CPS including short-term employment and labour-force participation histories are available but not necessarily used in the analysis. The CLMS studies all use the social security earnings data.

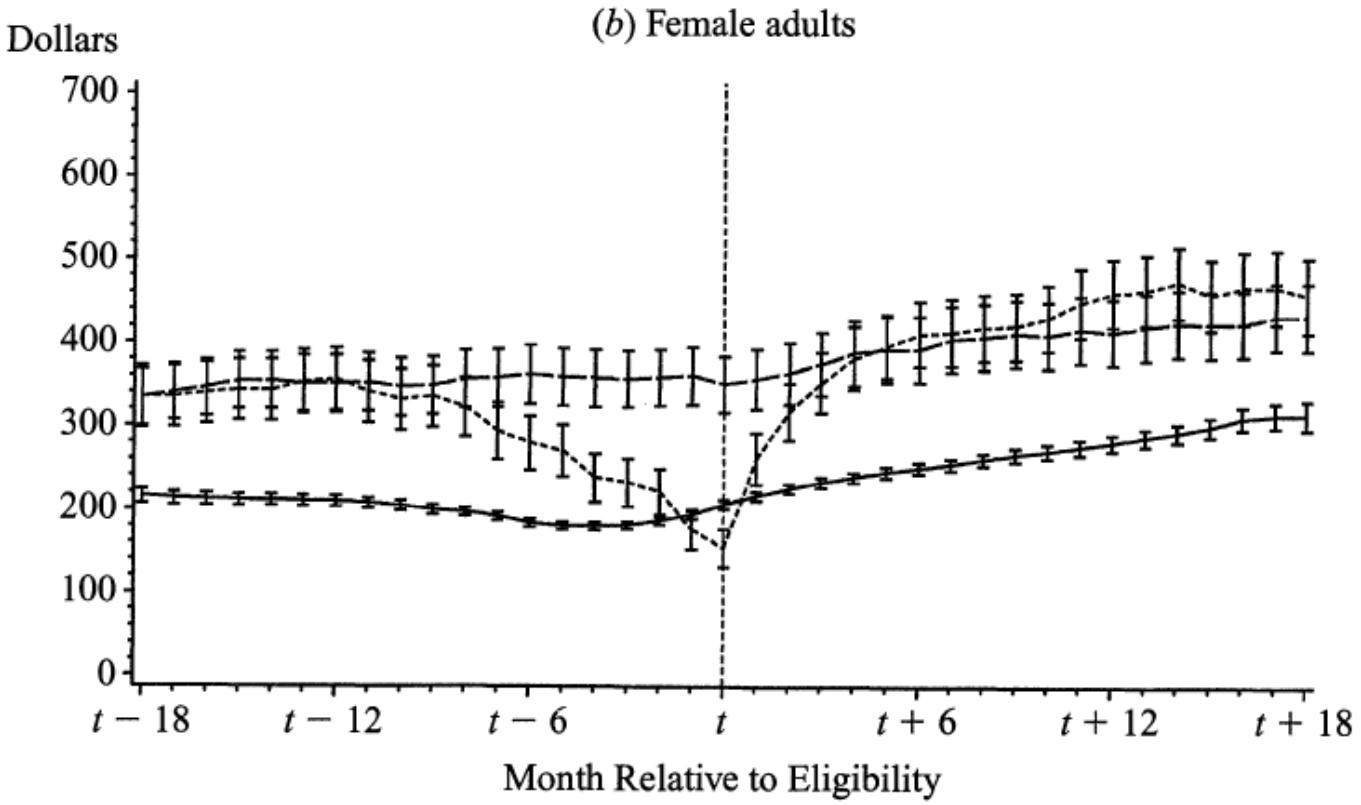
Data quality

NSW (supported work) data			
Study	LaLonde (1986)	Fraker and Maynard (1987) and LaLonde and Maynard (1986)	JTPA data
Programme, year, outcome variable	Annual earnings 1978 annual social security earnings and PSID earnings NSW (Supported Work) Data	1977, 1978, 1979 annual earnings for AFDC recipients and for youth NSW (Supported Work) Data	Quarterly and monthly earnings 1987–1989
(1) Comparison group in the same local labour market	No	No	Yes
(2) Same questionnaire administered to comparison and treatment group?	No	No	Yes
(3) Matching criteria (criteria for membership in comparison sample is also called "screening criteria")	PSID: (a) Men and Women who are household heads 1975–1979 CPS: Matches March 1976 CPS earnings with SSA earnings. Person with 1976 income $\leqq$ \$20K and household income $\leqq$ \$30K	Three Samples (i) Eligible in sample period: for youth: high school dropout–exclude in school youth. For AFDC: age of youngest child, receipt of welfare matching. (ii) Cell matching: based on predictors of 1979 SSA earnings of eligibles: (earnings prior to programme participation), demographics, education, family income, change in earnings. (iii) Stratified matches on imputed 1979 earnings: earnings estimated on eligible nonparticipant sample plus demographic criteria (race, sex). Same criteria for prediction as in (ii).	Persons screened to be eligible for JTPA; out of school youth, no disabled persons; Title IIA only
(4) Eligibility for programme known for comparison group members?	No	No	Yes
Variables used in analysis			
Age, race, sex	Yes: Women AFDC recipients 20–55, Males $\leqq$ 55	Same as LaLonde	Yes
Education	Yes	Same as LaLonde	Yes
Training history	No	Same as LaLonde	Yes
Children	Yes	Same as LaLonde	Yes
Employment histories	No	Same as LaLonde	Yes
Hours worked	No	Same as LaLonde	Yes
Unemployment histories	No	Same as LaLonde	Yes
Welfare receipt?	Yes	Same as LaLonde	Yes
Earnings histories	Two years post-programme Two years pre-programme	Same as LaLonde	(Five years of pre-programme earnings) monthly earnings

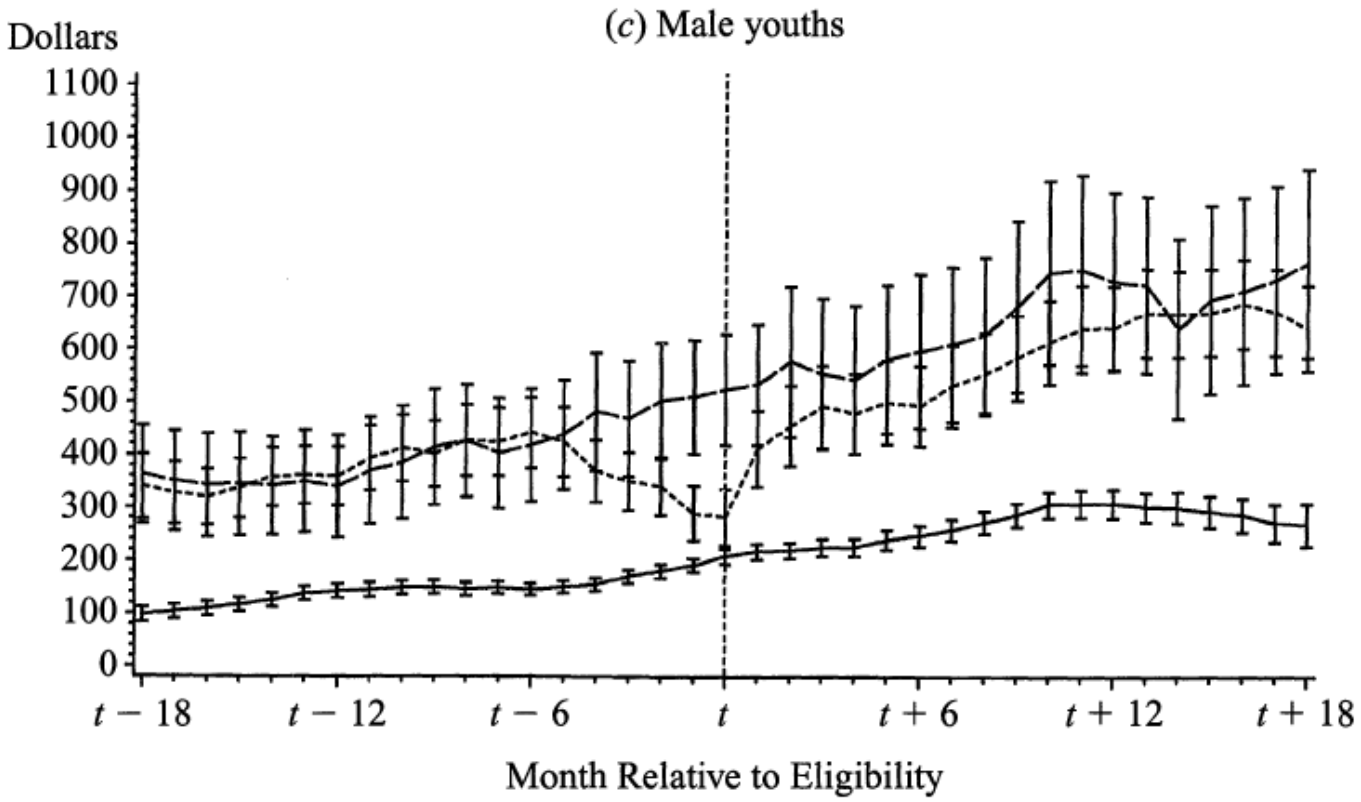
Dip pattern


(a) Male adults
Dollars
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
t − 18
t − 12
t − 6
t
t + 6
t + 12
t + 18
Month Relative to Eligibility
Sample:
SIPP Eligibles
JTPA Controls
JTPA ENPs

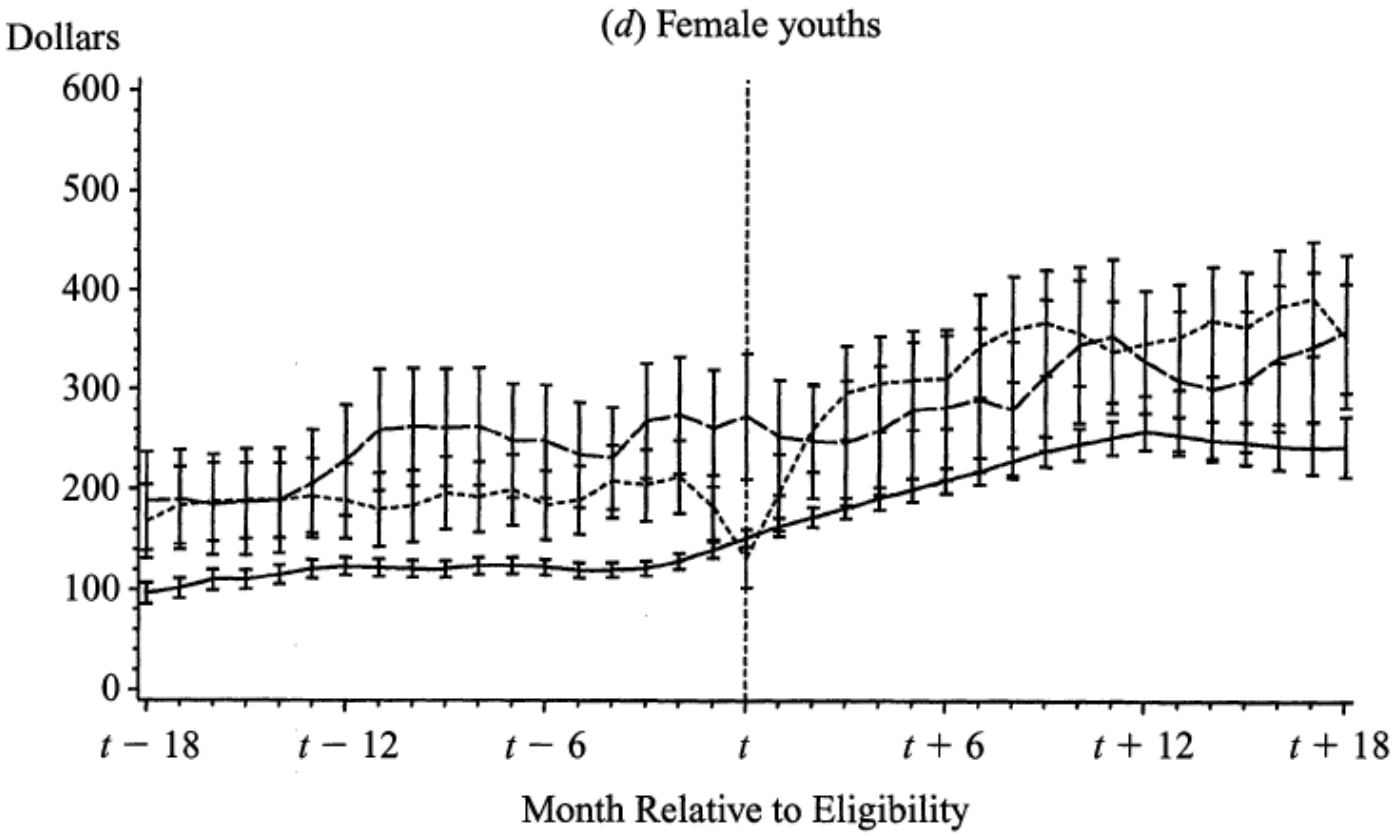
Dip pattern


(b) Female adults
Dollars
700
600
500
400
300
200
100
0
t − 18
t − 12
t − 6
t
t + 6
t + 12
t + 18
Month Relative to Eligibility

Dip pattern


(c) Male youths
Dollars
1100
1000
900
800
700
600
500
400
300
200
100
0
t − 18
t − 12
t − 6
t
t + 6
t + 12
t + 18
Month Relative to Eligibility

Dip pattern


(d) Female youths
Dollars
600
500
400
300
200
100
0
t − 18
t − 12
t − 6
t
t + 6
t + 12
t + 18
Month Relative to Eligibility

Prop Score Model

TABLE 3
COEFFICIENT ESTIMATES AND p-VALUES FROM WEIGHTED PARTICIPATION LOGIT †
BEST PREDICTOR MODEL FOR THE PROBABILITY OF PARTICIPATION ††
Experimental Control and Eligibile Nonparticipant (ENP) Samples
Dependent Variable: 1 for Experimental Control, 0 for Eligible Nonparticipant
Adult Males, 508 Controls and 388 ENPs

Variables	Coeff	Std Error	p-Value*
Intercept	-5.07	0.83	0.0000
Fort Wayne, IN	2.45	0.41	0.0000
Jersey City, NJ	0.66	0.43	0.1273
Providence, RI	2.19	0.44	0.0000
Black	0.49	0.33	0.1333
Hispanic	0.43	0.40	0.2837
Other race/ethnicity	0.61	0.55	0.2653
Age 30 to 39	-0.50	0.30	0.0926
Age 40 to 49	-0.60	0.38	0.1115
Age 50 to 54	-0.29	0.62	0.6361
Fewer than 10 years schooling	-0.83	0.40	0.0397
10-11 years schooling	0.66	0.34	0.0510
13-15 years schooling	0.90	0.35	0.0096
16 or more years schooling	-1.38	0.54	0.0101
Last married 1-12 months prior to RA/EL**	0.42	0.80	0.5995
Last married >12 months prior to RA/EL	-0.03	0.61	0.9648
Single, never married at RA/EL	0.71	0.36	0.0498
Child age less than 6 present in household	-0.16	0.38	0.6761
Unemployed -> Employed	1.52	0.42	0.0003
OLF -> Employed	0.79	0.76	0.3016
Employed -> Unemployed	2.46	0.46	0.0000
Unemployed -> Unemployed	2.67	0.58	0.0000
OLF -> Unemployed	3.27	0.60	0.0000
Employed -> OLF	2.55	0.61	0.0000
Unemployed -> OLF	2.30	0.87	0.0085
OLF -> OLF	-0.15	0.61	0.8002
One job in 18 months prior to RA/EL	0.41	0.39	0.2894
Two jobs in 18 months prior to RA/EL	0.57	0.50	0.2600
More than two jobs in 18 months prior to RA/EL	1.87	0.52	0.0003
Enrolled in vocational training at RA/EL	1.94	0.62	0.0019
Ever had vocational training?	-0.28	0.32	0.3815
Total number of household members	-0.25	0.10	0.0134
Earnings in the month of RA/EL	-0.00	0.00	0.0000

† Weights are used in the estimation procedure to account for choice-based sampled data. It is assumed that in a random sample Controls represent 3% and ENPs 97% of the eligible population.

†† The omitted training center is Corpus Christi, TX; the omitted race is white; the omitted age group is 22-29; the omitted schooling category is twelve years; the omitted marital status is currently married at RA/EL; the omitted labor force transition pattern is Employed -> Employed; the omitted number of job spells in the 18 months prior to RA/EL is zero.

* Reported p-values are for two-tailed tests of the null hypotheses that the true coefficient equals zero.

** RA/EL indicates the month of random assignment (RA) for the experimental controls and the date of eligibility (EL) for Eligible Nonparticipants (ENPs)

CS Matching

TABLE 5(a)

Estimated bias for alternative nonparametric matching methods Experimental controls and eligible nonparticipants (ENPs)†‡*

Quarter	Difference in means ($\hat{\beta}$)	Nearest neighbour without common support	Nearest neighbour with common support	Local linear P score matching	Regression-adjusted local linear matching	Difference-in-differences from local linear P score matching	Difference-in-differences from regression-adjusted local linear matching
			Adult males				
$t=1$	−418 (38)	221 (56)	123 (67)	33 (59)	39 (60)	97 (62)	104 (63)
$t=2$	−349 (47)	−166 (151)	77 (83)	37 (61)	39 (64)	77 (89)	77 (92)
$t=3$	−337 (55)	−58 (206)	53 (96)	29 (78)	21 (80)	90 (114)	74 (114)
$t=4$	−286 (57)	161 (178)	86 (96)	80 (77)	65 (82)	112 (90)	98 (91)
$t=5$	−305 (57)	167 (196)	87 (100)	64 (77)	50 (83)	19 (95)	−5 (99)
$t=6$	−328 (63)	45 (191)	34 (113)	37 (82)	17 (90)	4 (105)	−35 (111)
Ave. 1 to 6	−337 (47)	62 (127)	77 (80)	47 (60)	38 (64)	67 (71)	52 (74)
As a % of impact**	775%	142%	177%	108%	87%	153%	120%
As a % of adjusted impact	552%	102%	126%	77%	62%	109%	85%
			Adult females				
$t=1$	−26 (24)	115 (30)	67 (36)	45 (33)	55 (36)	65 (31)	74 (30)
$t=2$	29 (25)	113 (53)	47 (46)	48 (37)	55 (39)	53 (40)	60 (39)
$t=3$	38 (26)	124 (107)	63 (59)	26 (48)	31 (52)	10 (56)	14 (59)
$t=4$	55 (30)	106 (102)	58 (52)	36 (39)	35 (45)	12 (53)	7 (56)
$t=5$	62 (34)	92 (111)	47 (51)	48 (40)	48 (45)	29 (51)	23 (53)
$t=6$	40 (36)	79 (84)	−6 (54)	23 (40)	16 (42)	−5 (51)	−18 (51)
Ave. 1 to 6	33 (26)	105 (69)	46 (43)	38 (33)	40 (38)	27 (38)	27 (39)
As a % of impact**	113%	358%	157%	130%	137%	93%	91%
As a % of adjusted impact	94%	300%	131%	109%	114%	78%	76%

* The table reports the bias for alternative matching methods. The bias in the first column is $\hat{\beta}$. The estimator in the second column does not restrict matches to a common support region. The estimators in the third through seventh columns restrict matches to a common support region and the bias estimates correspond to $\hat{\bar{B}}_{S_P}$.

† The best predictor model given in the second footnote to Table 2, is used to estimate the probability of programme participation. The conditioning variables in the regression adjusted local linear models are site, race, age, education, previous training, work experience in months, the local unemployment rate, indicator variables for marital status and for the presence of a child aged less than 6 in the household, and indicators for the quarter of the year and the year.

‡ A 2% trimming rule is used to determine the region of overlapping support (see Appendix C). A fixed bandwidth equal to 0·06 and a biweight kernel, defined in Appendix A, are used for the nonparametric estimates.

** The impacts in the table are mean monthly impacts for the six post-programme quarters, estimated using the experimental treatment and control data for the four JTPA training sites in our study. See the experimental impacts and the adjusted impacts in Table 4.

CS Matching

TABLE 5(b)

Estimated bias for alternative nonparametric matching methods Experimental controls and eligible nonparticipants (ENPs)†‡*

Quarter	Difference in means ($\hat{\beta}$)	Nearest neighbour without common support	Nearest neighbour with common support	Local linear *P* score matching	Regression-adjusted local linear matching	Difference-in-differences from local linear *P* score matching	Difference-in-differences from regression-adjusted local linear matching
			Male youth				
$t=1$	−51 (58)	146 (92)	49 (75)	3 (64)	8 (61)	43 (72)	80 (77)
$t=2$	2 (60)	197 (92)	98 (82)	40 (64)	28 (55)	43 (60)	61 (60)
$t=3$	5 (73)	202 (105)	83 (119)	33 (81)	−8 (77)	92 (80)	70 (86)
$t=4$	17 (69)	246 (105)	98 (94)	44 (81)	4 (71)	9 (74)	−5 (77)
$t=5$	82 (73)	283 (118)	138 (89)	84 (93)	42 (76)	18 (88)	−11 (81)
$t=6$	65 (77)	258 (145)	129 (121)	28 (93)	−31 (92)	−23 (89)	−64 (84)
Ave. 1 to 6	20 (57)	222 (88)	99 (78)	39 (66)	7 (53)	30 (49)	22 (48)
As a % of impact**	34%	382%	170%	67%	12%	52%	38%
As a % of adjusted impact	56%	617%	275%	108%	19%	84%	61%
			Female youth				
$t=1$	6 (31)	67 (54)	−7 (60)	31 (42)	−8 (46)	−7 (38)	−14 (41)
$t=2$	54 (40)	85 (57)	23 (60)	79 (53)	27 (49)	60 (49)	27 (47)
$t=3$	89 (44)	142 (62)	97 (78)	121 (60)	49 (52)	135 (59)	83 (58)
$t=4$	42 (50)	89 (56)	24 (72)	37 (59)	−28 (59)	45 (57)	4 (59)
$t=5$	64 (41)	121 (57)	51 (63)	65 (54)	8 (54)	45 (61)	−7 (63)
$t=6$	31 (46)	107 (82)	34 (70)	34 (65)	1 (62)	31 (70)	6 (69)
Ave. 1 to 6	48 (36)	102 (49)	37 (56)	61 (45)	8 (42)	52 (39)	17 (39)
As a % of impact**	7059%	15000%	5441%	8971%	1176%	7574%	2426%
As a % of adjusted impact	195%	415%	150%	248%	33%	209%	67%

* The table reports the bias for alternative matching methods. The bias in the first column is $\hat{\beta}$. The estimator in the second column does not restrict matches to a common support region. The estimators in the third through seventh columns restrict matches to a common support region and the bias estimates correspond to $\hat{B}_{S_P}$.

† The best predictor model given in the second footnote to Table 2, is used to estimate the probability of programme participation. The conditioning variables in the regression adjusted local linear models are site, race, age, education, previous training, the local unemployment rate, indicator variables for marital status and the presence of a child aged less than 6 in the household, and indicators for the quarter of the year and the year.

‡ A 5% trimming rule is used to determine the region of overlapping support (see Appendix C), and a fixed bandwidth equal to 0·06 and a biweight kernel, defined in Appendix A, are used for nonparametric estimates.

** The impacts in the table are mean monthly impacts for the six post-programme quarters, estimated using the experimental treatment and control data for the four JTPA training sites in our study. See the experimental impacts and the adjusted impacts in Table 4.

DID matching

TABLE 6(a)

Bias from local linear regression matching estimator†
Under alternative predictor models for the probability of programme participation

Quarter	Regular††	Coarse I‡	Coarse II‡	Coarse III‡	SIPP§	Site mismatch§§	No-show§§§
				Adult males			
$t=1$	39 (60)	−390 (51)	−228 (67)	−84 (77)	249 (77)	−184 (110)	58 (38)
$t=2$	39 (64)	−312 (58)	−193 (61)	−39 (88)	123 (79)	−154 (120)	37 (39)
$t=3$	21 (80)	−286 (62)	−153 (57)	−36 (96)	76 (81)	−147 (127)	27 (42)
$t=4$	65 (82)	−231 (63)	−104 (66)	−9 (92)	13 (93)	−164 (132)	−6 (48)
$t=5$	50 (83)	−244 (73)	−146 (70)	20 (96)	* (*)	−211 (132)	1 (48)
$t=6$	17 (90)	−286 (84)	−172 (79)	−3 (111)	* (*)	−189 (112)	−21 (48)
Ave. 1 to 6	38 (64)	−291 (54)	−166 (56)	−25 (83)	115 (78)	−175 (108)	16 (37)
				Adult Females			
$t=1$	55 (36)	−69 (33)	−73 (29)	40 (30)	167 (35)	−84 (56)	26 (28)
$t=2$	55 (39)	−9 (33)	−15 (29)	63 (34)	122 (40)	−57 (69)	9 (36)
$t=3$	31 (52)	5 (34)	−6 (31)	42 (40)	98 (40)	−62 (70)	−13 (37)
$t=4$	35 (45)	5 (34)	−10 (34)	21 (47)	87 (43)	−42 (60)	2 (35)
$t=5$	48 (45)	14 (38)	−6 (37)	26 (48)	* (*)	−38 (63)	1 (31)
$t=6$	16 (42)	−10 (37)	−24 (37)	−2 (44)	* (*)	−35 (58)	−2 (34)
Ave. 1 to 6	40 (38)	11 (31)	−22 (29)	32 (35)	119 (39)	−53 (57)	4 (30)
				Male youth			
$t=1$	8 (61)	−41 (56)	−40 (53)	37 (61)	302 (120)	−29 (106)	104 (44)
$t=2$	28 (55)	10 (62)	9 (63)	45 (65)	275 (140)	12 (110)	36 (43)
$t=3$	−8 (77)	−29 (74)	−24 (76)	10 (83)	217 (153)	38 (136)	70 (48)
$t=4$	4 (71)	2 (69)	8 (70)	30 (81)	157 (176)	110 (162)	116 (45)
$t=5$	42 (76)	63 (72)	73 (71)	46 (99)	* (*)	132 (182)	95 (48)
$t=6$	−31 (92)	9 (76)	21 (75)	−68 (131)	* (*)	−63 (210)	108 (53)
Ave. 1 to 6	7 (53)	2 (52)	8 (52)	17 (70)	238 (144)	33 (128)	88 (38)
				Female youth			
$t=1$	−8 (46)	3 (34)	17 (32)	60 (45)	−11 (72)	74 (76)	55 (32)
$t=2$	27 (49)	46 (39)	54 (39)	81 (46)	−31 (79)	91 (77)	52 (32)
$t=3$	49 (52)	64 (42)	72 (41)	101 (51)	−37 (82)	84 (90)	74 (34)
$t=4$	−28 (59)	18 (48)	18 (47)	48 (57)	−55 (85)	−33 (119)	21 (36)
$t=5$	8 (54)	46 (43)	48 (41)	46 (56)	* (*)	21 (131)	37 (39)
$t=6$	1 (62)	37 (50)	40 (48)	38 (62)	* (*)	−3 (114)	57 (36)
Ave. 1 to 6	8 (42)	36 (36)	41 (35)	62 (42)	−34 (78)	39 (83)	49 (26)

† A 2% trimming rule is used for adult males and females to determine the overlapping support region (see Appendix C) and a 5% trimming rule is used for male and female youth. A fixed bandwidth of 0·06 and a biweight kernel, described in Appendix A, are used to compute the estimates for all four groups. Bootstrapped standard errors are shown in parentheses. They are based on 50 replications with 100% sampling.

* Data not available to compute for this quarter. Averages reported over available quarters.

†† The regular predictor model is the model for the probability of programme participation that maximizes the percent correctly classified. The regressors in the model for each demographic group are given in the footnote to Table 2.

‡ Coarse I predictor model includes indicator variables for site, race, age, education, marital status, and for the presence of children aged less than 6 in the household. Coarse II predictor model augments Coarse I with earnings from the year preceding enrollment into the programme. Coarse III predictor model augments Coarse I with indicators for labour force transition patterns.

§ SIPP predictor model includes indicators for age, race, education, marital status, children aged less than 6, labour force transition patterns and levels of earnings in the preceding year. The data used are SIPP JTPA eligibles matched with Experimental JTPA Controls.

DID matching

TABLE 6(b)

Bias from difference-in-differences local linear regression estimator†
Under alternative predictor models for the probability of programme participation

Quarter	Regular‡	Coarse I‡	Coarse II‡	Coarse III†	SIPP‡	Site mismatch‡
			Adult males			
$t=1$	104 (63)	167 (67)	31 (57)	67 (68)	−97 (38)	−135 (126)
$t=2$	77 (92)	143 (82)	−80 (62)	103 (107)	−230 (51)	−72 (130)
$t=3$	74 (114)	62 (95)	−158 (71)	105 (134)	−277 (52)	−9 (141)
$t=4$	98 (91)	33 (93)	−150 (82)	47 (109)	−338 (72)	19 (151)
$t=5$	−5 (99)	−73 (104)	−254 (86)	−29 (122)	* (*)	−136 (167)
$t=6$	−35 (111)	−143 (106)	−255 (96)	−36 (129)	* (*)	−82 (165)
Ave. 1 to 6	52 (74)	32 (78)	−144 (61)	43 (95)	−236 (45)	−69 (123)
			Adult females			
$t=1$	74 (30)	80 (24)	71 (23)	86 (25)	−43 (10)	38 (42)
$t=2$	60 (39)	69 (35)	54 (33)	85 (35)	−86 (25)	66 (56)
$t=3$	14 (59)	20 (39)	−1 (37)	25 (49)	−99 (27)	43 (66)
$t=4$	7 (56)	−16 (42)	−34 (40)	−15 (59)	−116 (36)	67 (62)
$t=5$	23 (53)	−9 (42)	−26 (42)	−5 (55)	* (*)	80 (68)
$t=6$	−18 (51)	−44 (42)	−52 (43)	−40 (53)	* (*)	48 (72)
Ave. 1 to 6	27 (39)	17 (31)	2 (30)	23 (39)	−86 (21)	57 (50)
			Male youth			
$t=1$	80 (77)	123 (56)	111 (56)	194 (85)	22 (53)	−92 (98)
$t=2$	61 (60)	102 (73)	81 (72)	58 (72)	12 (78)	1 (118)
$t=3$	70 (86)	−9 (88)	−23 (87)	38 (100)	−60 (102)	33 (152)
$t=4$	−5 (77)	−45 (81)	−54 (80)	−6 (85)	−85 (135)	32 (157)
$t=5$	−11 (81)	34 (85)	28 (82)	−3 (108)	* (*)	25 (188)
$t=6$	−64 (84)	18 (83)	19 (80)	−74 (126)	* (*)	−117 (211)
Ave. 1 to 6	22 (48)	37 (56)	27 (54)	34 (59)	−28 (81)	−20 (122)
			Female youth			
$t=1$	−14 (41)	59 (39)	62 (41)	14 (36)	−14 (31)	5 (66)
$t=2$	27 (47)	82 (42)	75 (42)	48 (47)	−67 (33)	29 (88)
$t=3$	83 (58)	116 (51)	106 (52)	91 (62)	−90 (46)	89 (111)
$t=4$	4 (59)	53 (48)	36 (48)	30 (56)	−96 (60)	−21 (139)
$t=5$	−7 (63)	−1 (43)	−8 (43)	−16 (60)	* (*)	44 (154)
$t=6$	6 (69)	2 (53)	5 (53)	−3 (71)	* (*)	2 (134)
Ave. 1 to 6	17 (39)	52 (35)	46 (35)	28 (39)	−67 (35)	25 (87)

† A 2% trimming rule is used to determine the overlapping support region for adult groups and a 5% trimming rule is used for the youth groups (see Appendix C). A fixed bandwidth equal to 0·06 and a biweight kernel, described in Appendix A, are used to compute the nonparametric estimates.
* Data not available to compute for these periods. Averages are reported over available quarters.
‡ The alternative predictor models for the probability of programme participation are described in the footnote to Table 6(a).

Using No-shows as comparison group

TABLE 7

Estimated bias for alternative nonparametric matching methods Experimental controls and no-shows†*

Quarter	Difference in means	Local linear P score matching	Regression-adjusted local linear matching	Difference in means	Local linear P score matching	Regression-adjusted local linear matching
	Adult males			Adult females		
$t=1$	64 (35)	66 (39)	58 (38)	17 (24)	28 (30)	26 (28)
$t=2$	32 (37)	45 (40)	37 (39)	7 (30)	11 (38)	9 (36)
$t=3$	26 (41)	36 (42)	27 (42)	−5 (30)	−9 (38)	−13 (37)
$t=4$	19 (46)	3 (49)	−6 (48)	12 (28)	5 (36)	2 (35)
$t=5$	22 (49)	10 (51)	1 (48)	14 (24)	4 (32)	1 (31)
$t=6$	7 (50)	−12 (52)	−21 (48)	11 (27)	1 (36)	−2 (34)
Ave. 1 to 6	29 (37)	25 (39)	16 (37)	9 (23)	7 (31)	4 (30)
As a % of impact**	66%	57%	37%	32%	23%	14%
As a % of adjusted impact	47%	41%	26%	27%	20%	11%
As a % of Control-ENP	8%	53%	42%	29%	18%	10%
	Male youth			Female youth		
$t=1$	92 (37)	116 (41)	104 (44)	12 (30)	53 (33)	55 (32)
$t=2$	33 (38)	49 (43)	36 (43)	17 (27)	52 (37)	52 (32)
$t=3$	56 (42)	80 (47)	70 (48)	39 (31)	80 (35)	74 (34)
$t=4$	111 (36)	133 (42)	116 (45)	−7 (32)	23 (39)	21 (36)
$t=5$	100 (39)	108 (46)	95 (48)	17 (34)	39 (46)	37 (39)
$t=6$	111 (43)	111 (47)	108 (53)	30 (31)	58 (39)	57 (36)
Ave. 1 to 6	84 (31)	99 (34)	88 (38)	18 (23)	51 (30)	49 (26)
As a % of impact**	144%	171%	152%	2639%	7474%	7273%
As a % of adjusted impact	233%	276%	245%	73%	207%	201%
As a % of Control-ENP	419%	255%	1258%	37%	83%	618%

* The table reports the bias for alternative matching methods. The bias in the first column is $\hat{\beta}$. The estimators in the second and third columns for each group restrict matches to the common support region and the bias estimates correspond to $\hat{B}_{S_p}$.

† The predictor model given in the second footnote to Table 2 is used to estimate the probability of programme participation. The conditioning variables in the regression adjusted local linear models are site, age, education, race, the local unemployment rate, indicator variables for marital status and for the presence of children aged less than 6 in the household, and indicators for the quarter of the year and the year.

‡ A 2% trimming rule is used to determine the region of overlapping support for adult groups and a 5% trimming rule is used for youth groups (see Appendix C). A fixed bandwidth equal to 0·06 and a biweight kernel, defined in Appendix A, are used for nonparametric estimates.

** The impacts in the table are mean monthly impacts for the six post-programme quarters, estimated using the experimental treatment and control data for the four JTPA training sites in our study. See the experimental impacts and the adjusted impacts in Table 4.

Matching applications in development

- Jalan and Ravaillon (2003a) use p0score matching techniques to evaluate effects of a workfare program in Argentina on wages. Jalan and Ravaillon (2003b) use p-score matching to study effects of piped water in rural India on child health outcomes.
- Handa and Mallucio (2006) study the performance of matching estimators by comparing matching-based estimates to estimates obtained from a randomized social experiment. Find that estimators perform well only for outcomes that are relatively easily measured, such as schooling attainment, less well for more complex such as expenditures. Imposing common support and choosing highly comparable comparison groups improvesperformance.

Matching applications in development

- used in evaluation of the urban *Oportunidades* CCT program in Mexico .
- Matches for treatment group households were drawn from two data sources: families who did not sign up for the program but who otherwise met the eligibility criteria, and families who met the eligibility criteria for the program but who were living in areas where the program was not yet available.
- The estimated propensity score model used to impute propensity scores in nonintervention areas. - DID matching estimators applied when possible.
- Find statistically significant program impacts on school enrollment, educational attainment, dropout rates, employment and earnings of youth, and on the numbers of hours spent doing homework. (Behrman, Garcia-Gallardo, Parker, and Todd, and Velez-Grajales (2012)

Matching applications in development

Galiani, Gertler, and Schargrodsky (2005) analyze effects of privatization of water services on child mortality in Argentina using DID matching.
- Godtland, Sadoulet, de Janvry, Murgai and Ortiz (2004) apply cross-sectional p-score matching estimators to evaluate effects of agricultural extension services in Peru.
- Gertler, Levine and Ames (2004) use CS matching to study of effects of parental death on child outcomes.
- Lavy (2004) study effects of a teacher incentive program in Israel on student performance.
- Angrist and Lavy (2001) study effects of teacher training on children's test scores in Israel
Chen and Ravaillon (2005), in a study of a poverty reduction project in China.

Chapter 7: Control function estimation

- Also known as *generalized residual methods*
- Proposed as a solution to the evaluation problem in Heckman and Robb (1986), but are related to Heckman (1976, 1979, 1980).
- Early applications are Willis and Rosen (1979), Heckman and Sedlacek (1985).
- Defined within the context of an econometric model for the outcome process.

Control function estimators

- Explicitly recognize that nonrandom selection into the program gives rise to an endogeneity problem and aim to obtain unbiased parameter estimates by explicitly modeling the source of the endogeneity.
- Allow selection into the program to be based on time varying unobservable variables, under some assumptions needed to secure identification of the treatment effect.

Write the model for outcomes as

$$Y = \varphi_0(X) + D\alpha_{TT}(X) + \tilde{\varepsilon},$$

where

$$\alpha_{TT}(X) = E(Y_1 - Y_0|X, D = 1) = \varphi_1(X) - \varphi_0(X) + E(U_1 - U_0|X, D = 1)$$

is the parameter of interest (TT(X)) and

$$\tilde{\varepsilon} = U_0 + D(U_1 - U_0 - E(U_1 - U_0|X, D = 1)).$$

- Decision to participate may be endogenous with respect to the outcomes, so expect that $E(U_0|X, D) \neq 0$.
- Heckman (1976,1979) showed that the endogeneity problem can be viewed as an error in model specification

Adding and subtracting $E(U_0|X,D) = DE(U_0|D=1,X) + (1-D)E(U|D=0,X)$, rewrite the outcome model as

$$\begin{aligned} Y &= \varphi_0(X) + D\alpha_{TT}(X) + E(U|D=0,X) + \\ &\quad D[E(U_0|D=1,X) - E(U_0|D=0,X)] + \varepsilon \\ &= \varphi_0(X) + D\alpha_{TT}(X) + K_0(X) + D[K_1(X) - K_0(X)] + \varepsilon \end{aligned} \tag{9}$$

where

$$\begin{aligned} K_0(X) &= E(U_0|D=0,X) \\ K_1(X) &= E(U_0|D=1,X) \\ \varepsilon &= D\{U_0 - E(U_0|D=1,X)\} + (1-D)\{U_0 - E(U_0|D=0,X)\} \\ + &\quad D\{U_1 - U_0 - E(U_1 - U_0|D=1,X)\} \end{aligned}$$

- By construction, $E(\varepsilon|X, D) = 0$.

- $K_1(X)$ and $K_0(X)$ are termed *control functions*.

- When these functions are known up to some finite number of parameters, they can be included in the model to control for the endogeneity and regression methods (either linear or nonlinear) applied to consistently estimate program.

- If no restrictions where placed on either $\alpha_{TT}(X)$, $K_1(X)$, or $K_0(X)$, then the treatment impact parameter ($\alpha_{TT}(X)$) could not be separately identified from the control functions.

- Different implementations of control function estimators impose different kinds of restrictions.

- Usually, functional form restrictions and/or exclusion restrictions (variables that determine the participation process (i.e. the choice of D) be excluded from the outcome equation).

Identification through index restrictions

- Heckman and Robb (1986) show how index restrictions can be used to secure identification of $\alpha_{TT}(X)$.

- Participation is assumed to depend on a set of characteristics Z through an index $h(Z\gamma)$ and on unobservables V:

$$D = 1 \text{ if } h(Z\gamma) + V > 0, = 0 \text{ if } h(Z\gamma) + V \leq 0$$

$h(Z\gamma) + V$ represents the net utility from participating in a program. (McFadden, 1981, and Manski and McFadden, 1981).

Under this model, the function $K_0(X) = E(U_0|D = 1, X)$ can be written as

$$E(U_0|D = 1, X) = E(U_0|h(Z\gamma) + V > 0, X) \quad (10)$$

$$= \frac{\int_{-h(Z\gamma)}^{\infty} \int_{-\infty}^{\infty} uf(u, v|X)dudv}{\int_{-h(Z\gamma)}^{\infty} \int_{-\infty}^{\infty} f(u, v|X)dudv}. \quad (11)$$

If $F(U_0, V|X)$ is assumed to be continuous with full support in R^2 and $F_V(\cdot)$ is invertible, then the index $Z\gamma$ can be written as a function of the conditional probability of participation.

Identification at infinity

- As $h(Z\gamma)$ approaches infinity, $E(U_0|D=1,Z)$ approaches 0 (recall that we assumed that $E(U_0|Z)=0$).
- For this reason, subgroups with a high probability of participating in the program (i.e. $h(Z\gamma)$ close to infinity) can be used to secure identification of model parameters.
- Essentially, there is no selection problem for groups who always participate. (Heckman, 1990).

$$\begin{aligned}
\Pr(D=1|Z) &= \Pr(V > -h(Z;\gamma)) \\
&= 1 - F_V(-h(Z;\gamma)). \\
\Longrightarrow\quad & h(Z;\gamma) = -F_v^{-1}(-\Pr(D=1|Z))
\end{aligned}$$

Heckman and Robb (1986) note that with the additional assumption that the joint distribution of the unobservables, U_0 and V, does not depend on X, except possibly though the index, $h(Z;\gamma)$:

$$f(U_0, V|X) = f(U_0, V|h(Z;\gamma)),$$

then $E(U_0|D = 1, X)$ can be written solely as a function of the probability of participating in the program, $\Pr(D = 1|Z)$:

$$\begin{aligned} E(U_0|D &= 1, X) = E(U_0|D = 1, P(Z)) = K_1(P(Z)) \\ E(U_0|D &= 0, X) = E(U_0|D = 0, P(Z)) = K_0(P(Z)). \end{aligned} \quad (12)$$

- A stronger assumption that would also imply index sufficiency is independence, $f(U_0, V|X) = f(U_0, V)$.

- *index sufficiency* greatly simplifies the problem of estimating the $K_d(X)$, $d \in \{0,1\}$ functions and also aids in the identification problem.

- Suppose $\varphi_0(X)$ and $h(Z\gamma)$ were both linear in the regressors. With one continuous variable included in Z but excluded from X , we can allow for overlap between X and Z and even for the case where X are fully contained in Z.(Cosslett, 1984).

- If the control functions are estimated nonparametrically, distinguishing the treatment effect from the control function requires the application of identification at infinity methods (more later..)

Heckman (1976,1979), assumed that U_0 and V are jointly normally distributed which implies a parametric form for $K_1(P(Z))$ and $K_0(P(Z))$:

$$
\begin{aligned}
E(U_0|D = 1, Z) = K_1(P(Z)) &= \frac{\sigma_{U_0 V}}{\sigma_{V^2}} \frac{\phi(-h(Z\gamma))}{1-\Phi(-h(Z\gamma))} \\
E(U_0|D = 0, Z) = K_0(P(Z)) &= \frac{\sigma_{U_0 V}}{\sigma_{V^2}} \frac{-\phi(-h(Z\gamma))}{\Phi(-h(Z\gamma))}.
\end{aligned}
$$

- Heckman, Ichimura, Smith and Todd (1996), invoke index sufficiency and nonparametrically estimate the $K(\cdot)$ functions are estimatedas a function of the probability of participating in the program., estimated by a probit model)

- This leads to a *partially linear* model and they use a variation of the Robinson (1988) estimator to estimate it.

- They test and do not reject the index sufficiency restriction.

Identification at infinity

- When the $K_1(P(Z))$ and $K_0(P(Z))$ are estimated nonparametrically, the intercept of the $K_1(P(Z)) - K_0(P(Z))$ cannot be separately identified from the treatment effect ($\alpha_{TT}(X)$).

- Under normality, functional form assumptions may be sufficient, assuming that the form of $\alpha_{TT}(X)$ is not co-linear with the $K_1(P(Z)) - K_0(P(Z))$ functions.

- Andrews and Schafgans (1998) develop an empirically implementable semiparametric version of Heckman's "identification at infinity" estimator.

Identification at infinity

- Approach is feasible when there is a subgroup in the data for which $\Pr(D = 1|Z) = 1$ for some set Z, meaning that individuals with that set of characteristics always select into the program.

- In the index model described above, this would be the group for which $h(Z\gamma)$ is close to infinity.

A Comparison of Control Function and Matching Methods

- Conventional matching estimators can in some cases be viewed as a restricted form of a control function estimator.
- Recall that traditional cross-sectional matching methods assume that selection is on observables, whereas control function methods explicitly allow selection into programs to be based on observables Z and unobservables V.

The assumption that justifies matching outcomes on the basis of Z is

$$E(Y_0|D=1,Z) = E(Y_0|D=0,Z).$$

If $X \subset Z$, then, in terms of previous model, this assumption implies that

$$E(U_0|D=1,Z) = E(U_0|D=0,Z).$$

This assumption is equivalent to assuming that the control functions are equal for both the $D = 0$ and $D = 1$ groups

$$K_1(P(Z)) - K_0(P(Z)) = 0, \tag{13}$$

in which case the model for outcomes can be written as

$$Y_0 = \varphi_0(X) + D\alpha^*(X) + K_0(P(Z)) + D\{U_1 - U_0 - E(U_1 - U_0 | D = 1, X)\}.$$

- This special case is *selection on observables*. (see Heckman and Robb, 1986; Heckman, Ichimura, Smith and Todd, 1995; and Barnow, Cain and Goldberger, 1980).

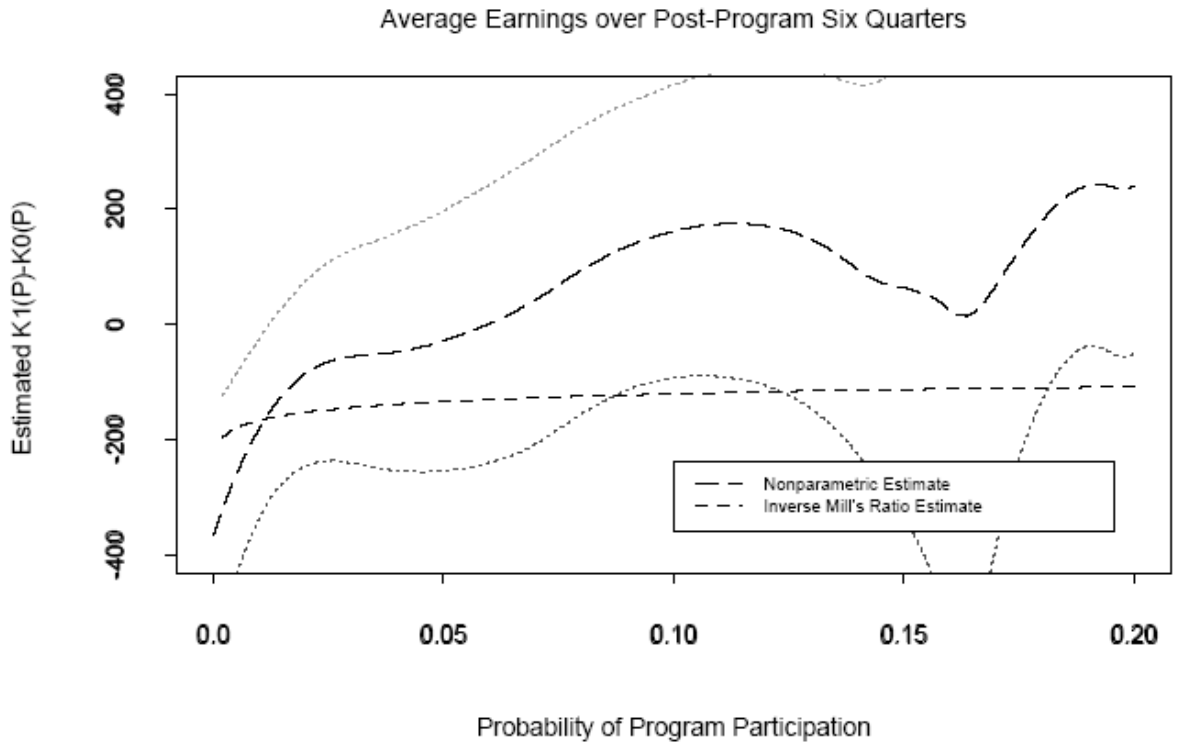
- When selection is of this form, many identification problems that arise in trying to separate the treatment impact $\alpha_{TT}(X)$ from the bias function $K_1(X)$ go away.

- Under the normal model, $K_0(P(Z)) = K_1(P(Z))$ will, in general, not be satisfied unless the errors have zero covariance, $\sigma_{U_0 V} = 0$.

Comparison of normal/nonpar Models (JTPA data)

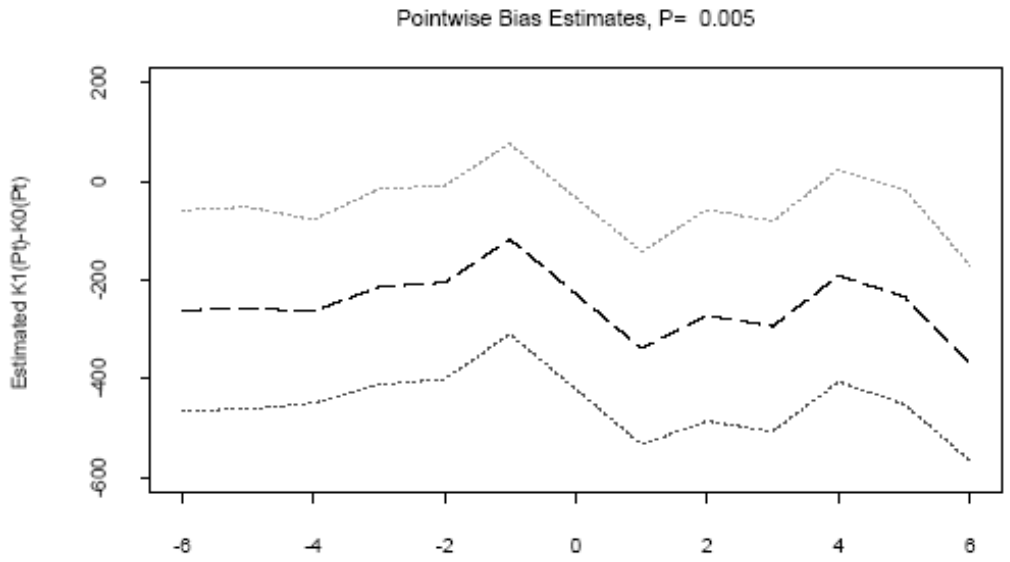
Pointwise Bias and Comparison with Normal Model

Figure 3: Local Linear Regression Estimates of Pointwise Bias (B(P))
Adult Males, Best Predictor P Model for The Probability of Program Participation



Stability of bias function over time

Pointwise bias over time, conditional on P



Chapter 8: Instrumental Variables and LATE estimation

- Instrumental variables methods provide another approach to estimating program effects in the presence of nonrandom self-selection
- Can accommodate selection on unobservables
- Will consider applications with discrete and with continuous instruments.

The Wald Estimator

Consider the treatment effect model:

$$Y = \varphi_0(X) + D\alpha^*_{TT}(X) + \tilde{\varepsilon},$$

where

$$\alpha^*_{TT}(X) = E(Y_1 - Y_0|X, D = 1) = \alpha(X) + E(U_1 - U_0|X, D = 1)$$

is the parameter of interest (TT) and

$$\tilde{\varepsilon} = U_0 + D(U_1 - U_0 - E(U_1 - U_0|X, D = 1)).$$

The Wald estimator

- Suppose that there is an exclusion restriction, a variable Z that affects the program participation decision but does not enter into the outcome equation.
- Also, assume that the conditioning variables X and the instrument Z are binary and that the instrument takes on the values Z_0 and Z_1.
- Assume that we condition on X by first partitioning the dataset by X and then use the instrument to estimate the program effect using the method of instrumental variables within X subsamples.
- The identifying assumption is that

$$E(U_0|X,Z) = E(U_0|X).$$

The Wald estimator

The so-called Wald estimator is:

$$\begin{aligned}\hat{\alpha}^*_{IV}(X) &= \frac{\hat{E}(Y|Z=Z_0,X)-\hat{E}(Y|Z=Z_1,X)}{\hat{E}(D|Z=Z_0,X)-\hat{E}(D|Z=Z_1,X)} \\ &= \frac{\hat{E}(Y|Z=Z_0,X)-\hat{E}(Y|Z=Z_1,X)}{\widehat{\Pr}(D=1|Z=Z_0,X)-\widehat{\Pr}(D=1|Z=Z_1,X)}.\end{aligned}$$

The denominator is the difference in the probability of participating in the program under the two different values of the instrument.

The Wald estimator

As noted in Heckman (1992), $\hat{\alpha}^*_{IV}(X)$ recovers the average impact of treatment on the treated (the TT parameter) only under one of two alternative assumptions on the error term (in addition to the assumption $E(U_0|X,Z) = E(U_0|X)$):

Case I: $U_1 = U_0$

or

Case II: $U_1 \neq U_0$ and
$E(U_1 - U_0|X,Z,D=1) = E(U_1 - U_0|X,D=1)$.

The Wald estimator

- In Case I, the average impact of treatment on the treated (TT) is assumed to be the same as the average treated effect (ATE).

- Under Case II, the ATE and TT parameters differ, but the instrument does not forecast the unobservable component of the gain from the program.

The Wald estimator

- Either of these assumptions would give $E(\tilde{\varepsilon}|X, Z = Z_1) = E(\tilde{\varepsilon}|X, Z = Z_0)$.
- Note that $E(D(U_1 - U_0 - E(U_1 - U_0|X, D = 1))|X, Z) = \Pr(D = 1|X)E(U_1 - U_0 - E(U_1 - U_0|X, D = 1))|X, Z, D = 1)$,so the required assumption is that $E(U_1 - U_0|X, Z, D = 1) = E(U_1 - U_0|X, D = 1)$.
- Heckman (1992) provides some examples where the assumption that the instrument does not help forecast the program gain can be problematic, some of which are described below.

Local Average Treatment Effects (LATE)

- If assumptions I or II are not satisfied, then the Wald estimator does not recover the TT nor the ATE parameters but still has a meaningful alternative interpretation as a *Local Average Treatment Effect (LATE)* (See Imbens and Angrist, 1994).
- However, LATE is the average treatment effect for a particular group of people - those induced by a change in the value of the instrument from Z_0 to Z_1 to participate in the program.
- The usefulness of LATE depends on whether this population is of interest.

Local Average Treatment Effects (LATE)

Some notation, following Imbens and Angrist (1994):

$$D_0 = \text{value of } D \text{ if } Z = Z_0$$
$$D_1 = \text{value of } D \text{ if } Z = Z_1$$

Recall that everyone has a value of Y_0 and a Y_1, though only one of these is realized. Similarly, everyone has a D_0 and a D_1, which represents a hypothetical participation status under different values of the instrument.

Local Average Treatment Effects (LATE)

The observed value of D is

$$D = 1(Z = Z_0)D_0 + 1(Z = Z_1)D_1 \ = D_0 + 1(Z = Z_1)(D_1 - D_0).$$

Putting this expression for D into $Y = Y_0 + D(Y_1 - Y_0)$ gives:

$$Y = Y_0 + D_0(Y_1 - Y_0) + 1(Z = Z_1)(D_1 - D_0)(Y_1 - Y_0).$$

LATE

We will assume that the instrument Z is independent of Y_0, Y_1, D_0 and D_1:

$$(Y_0, Y_1, D_0, D_1) \perp\!\!\!\perp Z.$$

It may seem odd to assume this for D_0 and D_1, because we are also assuming that Z affects D. However, Z having an effect on D does not mean that Z cannot be independent from D_0 and D_1. For example, in a randomized trial, random assignment of the offer of the program can be used as Z, but because this is random, it is not correlated with D_0 or D_1, which represents what a person would decide without the offer or with the offer.

LATE

- We also require that Z has no relationship with either Y_0 or Y_1.
- Even in the case where Z is generated by a randomized experiment, this assumption could be violated.

  - Example: Angrist et. al. (2002) analyze the effect of a Colombia private school voucher program, which randomly allocated vouchers for tuition at private school to a random fraction of eligible children.
  - Program stipulated that if a child repeats he/she is no longer eligible for the voucher, so private schools may have promoted children who randomly received vouchers, leading to a correlation between Z and Y_1.

LATE

We can divide the population into four types of people, depending on their D_0 and D_1 values:

1. *never-takers* - those for whom $D_0 = D_1 = 0$
2. *compliers* - those for whom $D_0 = 0, D_1 = 1$
3. *defiers* - those for whom $D_0 = 1, D_1 = 0$
4. *always-takers* - those for whom $D_0 = D_1 = 1$.

When the instrument is the randomized offer of a program, defiers are those who enter the program when it is not offered to them, but do not enter the program when it is offered. One could think of this behavior as being “irrational."

LATE

LATE assumes that everyone is affected by the instrument in the same way, essentially, that there are no defiers, which is called a *monotonicity assumption* (See Imbens and Angrist, 1994).

Without defiers, these different groups are identifiable in the data:

	Z_0	Z_1
$D = 0$	never taker or complier	never-taker
$D = 1$	always-taker	always-taker or complier

LATE

- In the data, some “always-takers" (those with $D = 1$ and $Z = 0$) are clearly recognized, while others (e.g. “compliers") are always mixed with “always-takers" or “never-takers."

- From the data, we can figure out the proportions of the data that are compliers, always-takers and never-takers.

That is, consider people for whom $Z = Z_0$. We observe what percentage of these people are always-takers. Because Z is assumed to be independent of D_0 and D_1, we have

$$P_a = Prob[D = 1|Z = 0].$$

LATE

By similar reasoning, get the proportion of never-takers (P_n)

$$P_n = Prob[D = 0|Z = 1].$$

Assuming no defiers, we get the proportion of compliers (P_c)

$$P_c = 1 - P_a - P_n.$$

The next step is to obtain the average treatment effect for compliers:

1. Estimate $E(Y|D=0, Z=Z_1)$. This is the mean of Y_0 for never-takers.
2. Estimate $E(Y|D=0, Z=Z_0)$. This mean is a weighted average of the mean Y_0 for never-takers and compliers, with weights equal to the the proportions of the types in the two populations (P_n and P_c).
3. From these two means, can infer the mean of Y_0 for compliers: $E(Y_0|D_0=0, D_1=1)$

4. Repeat the first three steps for always-takers and Y_1. Get the mean of Y_1 for always-takers and the mean of Y_1 for always-takers mixed with compliers. Use to get $E(Y_1|D_0 = 0, D_1 = 1)$.

5. Take the means for compliers to get:
$\alpha_{LATE} = E(Y_1 - Y_0|D_0 = 0, D_1 = 1)$

Alternatively, as shown by Imbens and Angrist (1994), an easier approach to getting $\alpha_{LATE}(X)$ is to use Z as an instrumental variable for D. If we condition on X by simply dividing the data by X cells, then this is the Wald estimator:

$$\alpha_{LATE}(X) = \frac{E(Y|Z = Z_1, X) - E(Y|Z = Z_0, X)}{E(D|Z = Z_1, X) - E(D|Z = Z_0, X)}$$

In the case of a randomized control trial with imperfect compliance, $\alpha_{LATE}(X)$ is equivalent to the intent-to-treat estimate divided by the difference between the probability of being treated for those assigned to the treatment group and those assigned to the control group. If most people are compliers, than you can use a bounds approach that uses the estimate of $\alpha_{LATE}(X)$ to obtain bounds for an estimate of ATE. (See Wooldridge (2009)).

Applications: example from labor economics

Angrist (1990) - evaluates the effect of serving in the Vietnam War on future earnings, uses the draft lottery number as an instrument for whether they participated. Never-takers = men who would not serve in the war under any circumstances, always-takers = men who would serve even if not assigned (e.g. career military). Compliers serve only if drafted.

Heckman (1997) notes that the draft lottery number is not necessary valid as an instrument for the $\alpha_{TT}(X)$ parameter. If firms took into account lottery numbers in making hiring decisions, then could induce correlation between the error term and the instrument. Even in that case, the IV estimate will have a valid interpretation as a LATE estimate.

Applications: example from development economics

Angrist et. al. (2002) study the impacts of a voucher program in Colombia (PACES), using both an intent-to-treat approach and an instrumental variables approach. The program gave more than 125,000 vouchers through lottery covering a little more than half the cost of attending a private secondary school. About 90% of the lottery winners used the voucher.

- use the win/loss status as an instrument for scholarship receipt. ITT estimates show that lottery winners were 10 percent more likely to complete the 8th grade and that they scored, on average, 0.2 standard deviations higher on standardized tests three years after the initial lottery. LATE estimates that are roughly 50 percent higher than the ITT estimates.

Chapter 9: Marginal Treatment Effects and Local IV

- Recent advances in the program evaluation literature have led to a better understanding of the relationship between the TT, ATE and LATE parameters and of new ways to estimate them.
- Heckman and Vytlacil (2005) develop a unifying theory of how the parameters relate to one another using a new concept, called a *marginal treatment effect* (MTE).

Model

Consider the treatment effect model of the previous sections, written in slightly more general form that does not assume additive separability:

$$\begin{aligned} Y &= DY_1 + (1-D)Y_0 \\ Y_1 &= \mu_1(X, U_1) \\ Y_0 &= \mu_0(X, U_0) \\ D &= 1 \text{ if } \mu_0(Z) - U_D \geq 0 \end{aligned}$$

It is assumed that $\mu_0(Z)$ is nondegenerate conditional on X, so that there is variation in who participates in the program holding X constant (i.e. that there is an exclusion restriction). The error terms are assumed to be independent of Z conditional on X.

MTE and IV

- Denote the propensity score as
$P(Z) = \Pr(D = 1|Z = z) = F_{U_D}(\mu_0(Z))$
- Assume that there is full support $(0 < \Pr(D = 1|Z) < 1)$

MTE and IV

Heckman and Vytlacil (2005) show that without loss of generality, one can assume U_D distributed uniformly. To see why, suppose that

$$D = 1 \text{ if } \varphi(Z) - v \geq 0$$

so that

$$\Pr(v < c) = F_V(c).$$

Because $F_V(\cdot)$ is a monotone transformation of the random variable v, we have

$$\Pr(F_V(v) < F_V(c)) = F_V(c).$$

Define $U_{D_i} = F_V(v)$ and note that $\Pr(U_{D_i} < t) = t$. Thus, U_{D_i} is uniformly distributed between 0 and 1.

When U_{D_i} is uniformly distributed,

$$E(D|Z) = \Pr(D = 1|Z) = F_{U_D}(\mu_0(Z_i)) = \mu_0(Z_i).$$

- Let Z and Z' be two values of the instrument such that $\Pr(D = 1|Z) < \Pr(D = 1|Z')$. The threshold crossing model of program participation implies that some individuals who would have chosen $D = 0$ with $Z = Z$ will instead choose $D = 1$ when $Z = Z,'$ but no individual with $D = 1$ when $Z = Z$ would choose $D = 0$ when $Z = Z'$.
- Vytlacil (2002) shows that the assumptions required to justify a threshold crossing model are the same as the monotonocity conditions typically assumed to justify application of LATE estimators, proposed in Imbens and Angrist (1994).

Parameters of interest in terms of $\alpha_{MTE}(X)$

Using this framework, we can define different parameters of interest. Let $\Delta = Y_1 - Y_0$.

(i) The average treatment effect (ATE):

$$\alpha_{ATE}(X) = E(\Delta|X = x)$$

(ii) The average effect of treatment on the treated, conditional on a value of $P(Z)$,:

$$\alpha_{TT}(X, P(Z), D = 1) = E(\Delta|X = x, P(z) = P(Z), D = 1)$$

Parameters of interest in terms of $\alpha_{MTE}(X)$

(iii) The *marginal treatment effect (MTE)* conditions on a value of the unobservable:

$$\alpha_{MTE} = E(\Delta | X = x, U_D = u)$$

(iv) The *local average treatment effect (LATE)* parameter

$$\alpha_{LATE}(X, P(Z), P(Z')) =$$

$$\frac{E(Y|P(Z) = P(Z), X) - E(Y|P(Z) = P(Z'), X)}{P(Z) - P(Z')}.$$

MTE is a new concept. If $U_D = P(Z)$, then the index $\mu_0(Z_i) - U_{D_i} = 0$ (by the above reasoning, $\mu_0(Z_i) = P(Z)$ when U_{D_i} is uniformly distributed).

How do we interpret $\alpha_{MTE}(X, U)$?

- People with the index equal to zero have unobservables that make them just indifferent between participating or not participating in the program.
- People with $U_{D_i} = 0$ have unobservables that make then most inclined to participate.
- People with $U_{D_i} = 1$ have unobservables that make them the least inclined to participate.

MTE and Local IV

All the parameters can be written in terms of MTE, first note that the following statements are equivalent, because conditioning on $P(Z)$ is the same as conditioning on Z:

$$\begin{aligned} Pr(Y_j \in A | X = x, Z = z, D = 1) &= Pr(Y_j \in A | Z = z, U_D \leq P(z)) \\ &= P(Y_j \in A | X = x, P(Z) = p(z), D = 1) \end{aligned}$$

Similarly,

$$\begin{aligned} Pr(Y_j \in A | X = x, Z = z, D = 0) &= Pr(Y_j \in A | Z = z, U_D \geq P(z)) \\ &= P(Y_j \in A | X = x, P(Z) = p(z), D = 0) \end{aligned}$$

MTE and Local IV

In terms of the model, the parameters are:

$$\alpha_{TT}(x, P(z)) = E(\Delta | X = x, U_D \leq P(Z))$$

$$\alpha_{LATE}(x, z, z') = \frac{E(Y|X = x, P(Z) = P(z)) - E(Y|X = x, P(Z) = P(z'))}{P(z) - P(z')}$$

MTE and Local IV

$$E(Y|X=x,P(Z)=P(z)) = Pr(z)E(Y_1|X=x,P(Z)=p(z),D=1)+$$
$$(1-P(z))E(Y_0|X=x,P(Z)=p(z),D=0)$$
$$= P(Z)\frac{\int_0^{P(Z)} E(Y_1|X=x,U_D=u)dU_D}{P(Z)} + (1-P(Z))\frac{\int_{P(Z)}^1 E(Y_0|X=x,U_D=u)dU_D}{1-P(Z)}$$

Thus, the numerator of LATE is equal to

$$\int_0^{P(z)} E(Y_1|X=x,U_D=u)dU_D + \int_{P(z)}^1 E(Y_0|X=x,U_D=u)dU_D$$
$$-\int_0^{P(z')} E(Y_1|X=x,U_D=u)dU_D - \int_{P(z')}^1 E(Y_0|X=x,U_D=u)dU_D$$

MTE and Local IV

Therefore, $\alpha_{LATE}(z, P(z), P(z'))$ equals

$$= \frac{\int_{P(z)}^{P(z')} E(Y_1|X = x, U_D = u)dU_D - \int_{P(z)}^{P(z')} E(Y_0|X = x, U_D = u)dU_D}{P(z) - P(z')}$$
$$= E(\Delta|X = x, P(z') \leq U_D \leq P(z)),$$

which is the average treatment effect for people with U_D within a given range.

These people would not participate if $Z = z'$ but do participate if $Z = z$. The change in the value of the instrument changes their participation status. This group is known as the *complier* group.

Heckman and Vytlacil (2005) show that all of the the parameters of interest can be written as an average of $\alpha_{MTE}(X, U_D)$ for values of U_D lying in different intervals.

$$\begin{aligned}
\alpha_{TT}(X) &= \frac{\int_0^{P(Z)} E(\Delta|X=x, U_D=u)dU_D}{P(Z)} \\
\alpha_{ATE}(X) &= \int_0^1 E(\Delta|X=x, U_D=u)dU_D \\
\alpha_{LATE}(X, P(Z), P(Z')) &= \frac{\int_{P(Z')}^{P(Z)} E(\Delta|X=x, U_D=u)dU_D}{P(Z)-P(Z')}
\end{aligned}$$

Knowledge of the MTE function therefore enables computation of all of the parameters of interest.

Estimation: MTE as a limiting form of LATE

- The α_{MTE} function depends on a value of an unobservable. Heckman and Vytlacil (2005) propose an estimation strategy that is implementable when the researcher has access to a continuous instrumental variable, Z, that enters into the participation equation but not the outcome equation.
- The MTE parameter can be seen as a limiting form of LATE.
- Heckman and Vytlacil define a *local instrumental variables* estimand as

$$
\begin{aligned}
\alpha_{LIV}(X, P(Z)) &= \frac{\partial E(Y|P(Z) = P(Z), X)}{\partial P(Z)} \\
&= \lim_{P(Z') -> P(Z)} \frac{E(Y|P(z) = P(Z), x = X) - E(Y|P(z) = P(Z'), x = X)}{P(Z) - P(Z')} \\
&= \alpha_{MTE}(X, U_D = P(Z)).
\end{aligned}
$$

Estimation proceeds in two steps.

1. First, estimate the program participation (propensity score) model to get $\hat{P}(Z)$.
2. Then, estimate $\frac{\partial E(Y|P(Z),X)}{\partial P(Z)}$ nonparametrically (which can be done by local linear regression of Y on $P(Z)$).

Evaluating this function (separately by data grouped by X) for different values of $P(Z)$ traces out the $\alpha_{MTE}(X, U_D)$ function.

The different estimands $\alpha_{TT}(X)$, $\alpha_{ATE}(X)$, $\alpha_{LATE}(X)$ can then be obtained by integrating under different regions of the $\alpha_{MTE}(X, U_D)$ function.

Applications

- For a recent application to estimating returns to education using U.S. data, see Carniero, Heckman and Vytlacil (2001).

- Doyle (2013, forthcoming in AER) - uses MTE to analyze effect of foster care placement on outcomes related to foster children (earnings, employment, teen motherhood, delinquency)

Doyle (2013)

Figure 3A: Delinquency MTE

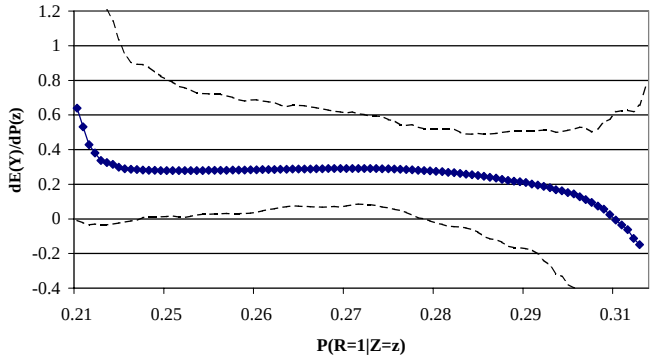


Figure 3B: Teen Motherhood MTE

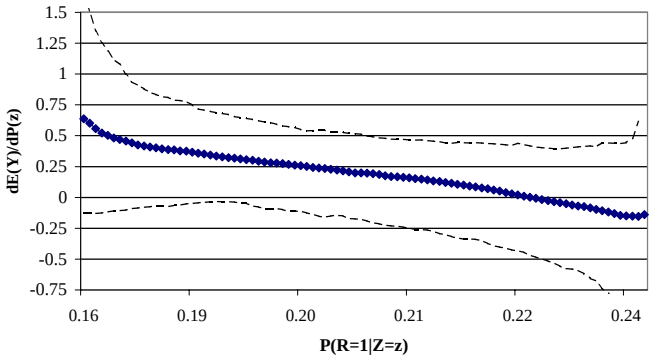


Figure 3C: Earnings MTE

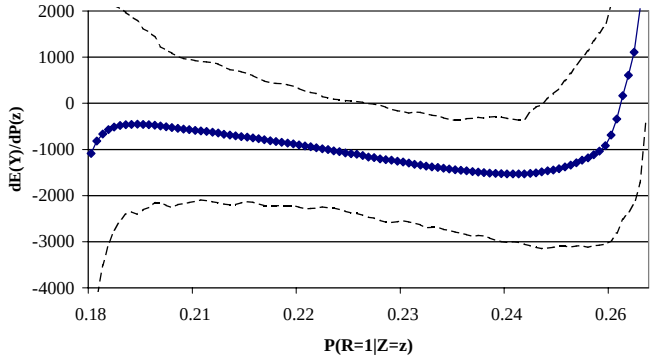
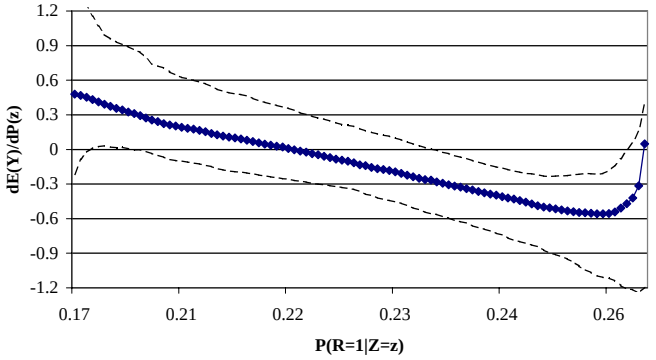


Figure 3D: Employment MTE



Figures report the results of a local quadratic estimator evaluated at each percentile of P(z).
5-95% confidence intervals reported, calculated using a bootstrap with 250 replications, clustered at the case manager level. Bandwidth=0.037.

Chapter 10: Regression Discontinuity Methods

- Goal is to evaluate causal impacts of an intervention
- Assignment to treatment is determined in part by the value of an observed covariate lying on either side of a fixed threshold (cut-off)
- Design first introduced by Thistlewaite and Campbell (1960) to evaluate the effect of National Merit awards on career aspirations of award recipients.
- Analyzed by Goldberger (1972) in the context of evaluating education interventions and Berk and Rauma (1983) in analzying effect of an unemployment benefit program on recidivism rates.

- Many studies implicitly rely on nonlinearities or discontinuities in the assignment rule (Black (1996) and Angrist and Krueger (1991)).
- Since 1990's, there has been a large number of studies in economics and other fields applying and extending RD methods (will discuss many examples later)
- New theoretical advances in interpretation and estimation

Potential Outcomes Framework

- Potential outcomes associated with treated and untreated states

$$Y_i(0), Y_i(1)$$

- Framework laid out in Fisher (1951), Roy (1951), Quandt (1972), Rubin (1978)
- Interest usually focuses on

$$Y_i(1) - Y_i(0)$$

- Let $W_i = 1$ if unit i exposed to treatment, else $W_i = 0$.
- Observed outcome

$$\begin{aligned} Y_i &= (1 - W_i)\, Y_i(0) + W_i\, Y_i(1) \\ &= Y_i(0) + W_i(\, Y_i(1) - Y_i(0)) \end{aligned}$$

Assignment to Treatment

- Let (X_i, Z_i) be a vector of covariates or pretreatment variables known not to be affected by treatment (e.g. pre-test score, age)
- Assignment to treatment is determined either completely or partly by the value of X_i being on either side of a fixed threshold.
- $Y_i(0)$ or $Y_i(1)$ may also be associated with X_i, but the dependence is assumed to be smooth
- Discontinuities in the conditional distribution of Y_i (or in its conditional expectation) are attributed to a causal effect of treatment.

Two Types of Designs

- *Sharp Regression Discontinuity (SRD) Design*

$$W_i = 1\{X_i \geq c\} \tag{14}$$

All individuals with covariates of c or greater are assigned to treatment.

- We can use the discontinuity in the conditional expectation of the outcome given the covariate to uncover an average causal effect of treatment

$$\tau_{SRD} = lim_{x \downarrow c} E(Y_i | X_i = x) - lim_{x \uparrow c} E(Y_i | X_i = x) \tag{15}$$

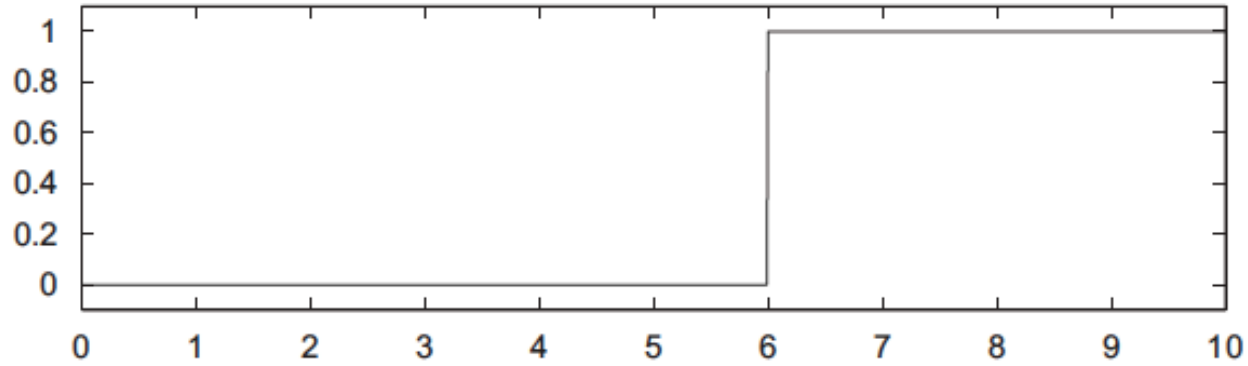


Fig. 1. Assignment probabilities (SRD).

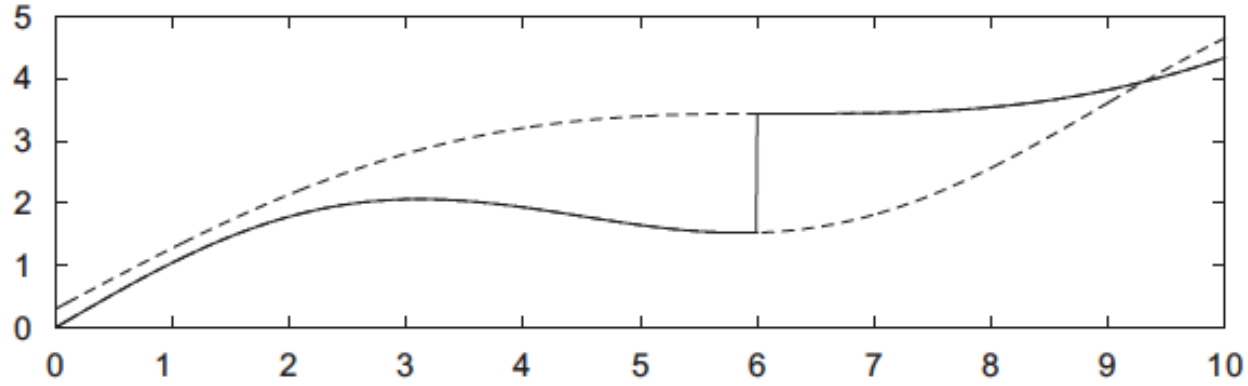


Fig. 2. Potential and observed outcome regression functions.

Two Types of Designs

- *Fuzzy Regression-Discontinuity (FRD) Design*
- Prob of receiving treatment need not change from 0 to 1 at the threshold, but there there is a discontinuous jump in the probability, so that

$$lim_{x\downarrow c} E(W_i|X_i = x) - lim_{x\uparrow c} E(W_i|X_i = x) \neq 0$$

$$or, equivalently,$$

$$lim_{x\downarrow c} Pr(W_i = 1|X_i = x) - lim_{x\uparrow c} Pr(W_i = 1|X_i = x) \neq 0$$

- Treatment effect can be obtained by the ratio

$$\tau_{FRD} = \frac{lim_{x\downarrow c} E(Y_i|X_i = x) - lim_{x\uparrow c} E(Y_i|X_i = x)}{lim_{x\downarrow c} E(W_i|X_i = x) - lim_{x\uparrow c} E(W_i|X_i = x)} \tag{16}$$

Why?

Assume constant treatment effect τ_{FRD}.

$$Y_i = Y_i(0) + W_i[Y_i(1) - Y_i(0)]$$
$$Y_i = Y_i(0) + W_i\tau_{FRD}$$
$$\lim_{x\downarrow c} E(Y_i|X_i = x) =$$
$$\lim_{x\downarrow c} E(Y_i(0)|X_i = x) + lim_{x\downarrow c} E(W_i|X_i = x)\tau_{FRD}$$
$$\lim_{x\uparrow c} E(Y_i|X_i = x) =$$
$$\lim_{x\downarrow c} E(Y_i(0)|X_i = x) + lim_{x\uparrow c} E(W_i|X_i = x)\tau_{FRD}$$

Take the difference, use the fact that $E(Y_i(0)|X_i = x)$ is continuous at $X_i = c$ and solve for τ_{FRD}.

$$\lim_{x\downarrow c} E(Y_i|X_i = x) - lim_{x\uparrow c} E(Y_i|X_i = x) =$$
$$\lim_{x\downarrow c} E(W_i|X_i = x) - lim_{x\uparrow c} E(W_i|X_i = x)\tau_{FRD}$$

$$\tau_{FRD} = \frac{lim_{x\downarrow c} E(Y_i|X_i = x) - lim_{x\uparrow c} E(Y_i|X_i = x)}{lim_{x\downarrow c} E(W_i|X_i = x) - lim_{x\uparrow c} E(W_i|X_i = x)}$$

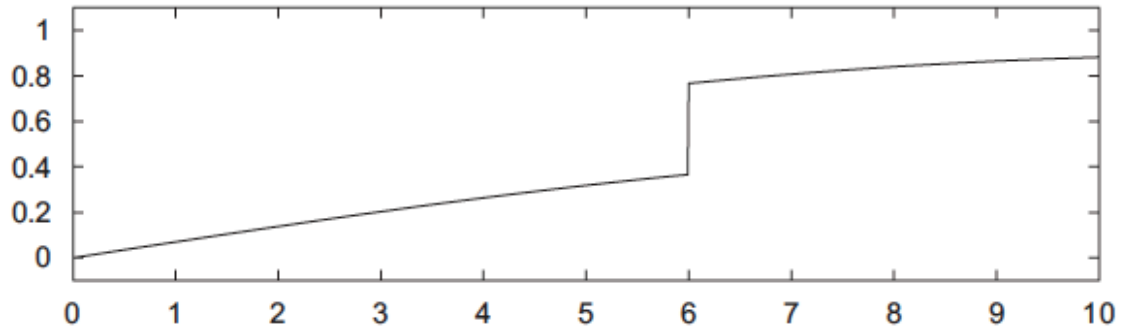


Fig. 3. Assignment probabilities (FRD).

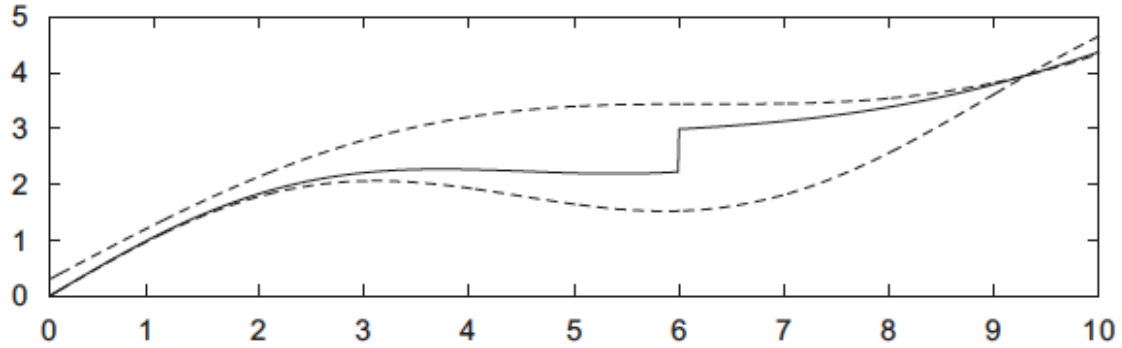


Fig. 4. Potential and observed outcome regression (FRD).

Interpretation of FRD when treatment response is heterogeneous

- Assume that treatment effect varies by unit τ_{FRD} random
- Let $W_i(x)$ be the potential treatment status given cut-off point x, for x in a neighborhood of c.
- $W_i(x) = 1$ if unit i would take treatment if cut-off equals x
- Assume monotonicity: $W_i(x)$ is nonincreasing in x at $x = c$

Define compliance status

- *Compliers:* have

$$lim_{x\downarrow X_i} W_i(X_i) = 0, lim_{x\uparrow X_i} W_i(X_i) = 1, \tag{17}$$

would get treatment if cut-off X_i or below, would not get treatment otherwise

- *Nevertakers:* do not get treatment either way

$$lim_{x\downarrow X_i} W_i(X_i) = 0, lim_{x\uparrow X_i} W_i(X_i) = 0 \tag{18}$$

- *Always takers:* get treatment either way

$$lim_{x\downarrow X_i} W_i(X_i) = 1, lim_{x\uparrow X_i} W_i(X_i) = 1 \tag{19}$$

For example, consider a program that assigns children with a pre-test score below a threshold to some remedial intervention (e.g. a summer reading program).

- *Compliers* are children who participate in the program only if their test score is below the threshold and not otherwise. They comply with their assigned treatment status.
- *Always-takers* are children who manage to receive the intervention regardless (e.g. parents request that they attend the program)
- *Never-takers* are children who do not attend the program even if assigned to it.

Interpretation of τ_{FRD}

In that case, τ_{FRD} gives the average treatment effect for compliers. (shown in Hahn, Todd and Van der Klaauw, 2001, building on insights of Angrist and Imbens, 1994, about LATE estimators).

Another example of FRD Design

- Van Der Klaauw (2002)
- Studies effect of financial aid on college admissions
- Association is ambiguous. More generous financial aid offers make students more likely to attend, but those students are also likely to have more generous offers from other places.
- x_i - numerical score assigned to college application based on the objective part of the application (SAT scores, grades)

$$G_i = 1 \text{ if } 0 \leq X_i < c_1$$
$$G_i = 2 \text{ if } c_1 \leq X_i < c_2$$
$$\vdots$$
$$G_i = L \text{ if } c_{L-1} \leq X_i$$

Comparison of RD Approach with a Matching Approach

- Matching assumes

$$Y(0), Y(1) \perp\!\!\!\perp W|X \tag{21}$$

- In that case, treatment effect can be obtained by comparing people with same x values who did and did not receive treatment

$$E(Y(1) - Y(0)|X = x) = E(Y|W = 1, X = c) - E(Y|W = 0, X = c)$$

- This approach would not exploit the jump in the probability of assignment at the discontinuity point
- It could not be implemented with a sharp design, where there is no overlap.
- Treated units with $x_i = c$ include both compliers and alwaystakers.
- Unconfoundedness is based on units being comparable if covariates are similar.

External and Internal Validity of RD Designs

- When treatment response is heterogeneous, RD approach provides estimates for subpopulation with x_i=c.
- If FRD and treatment effect heterogeneous, then effect is further restricted to the effect on compliers only (and compliers cannnot be identified in the data)
- The RD design has high internal validity (valid with the population studied), but potentially limited external validity (limited application to outside populations)

Estimation

- For sharp design, need estimators of two limits

$$\tau_{SRD} = lim_{x \downarrow c} E(Y_i | X_i = x) - lim_{x \uparrow c} E(Y_i | X_i = x) \tag{23}$$

- Could estimate each limit by kernel regression

$$\hat{\mu}_l(x) = \frac{\sum_{x_i < c} Y_i K(\frac{x_i - x}{h})}{\sum_{x_i < c} K(\frac{x_i - x}{h})} \tag{24}$$

$$\hat{\mu}_r(x) = \frac{\sum_{x_i \geq c} Y_i K(\frac{x_i - x}{h})}{\sum_{x_i \geq c} K(\frac{x_i - x}{h})} \tag{25}$$

With rectangular (uniform) kernel

- $K(u) = 1/2$ for $-1 \leq u \leq 1$, $= 0$ elsewhere

$$\tau_{SRD} = \frac{\Sigma_{i=1}^{n} Y_i 1(c \leq X_i \leq c+h)}{\Sigma_{i=1}^{n} 1(c \leq X_i \leq c+h)} - \frac{\Sigma_{i=1}^{n} Y_i 1(c-h \leq X_i < c)}{\Sigma_{i=1}^{n} 1(c-h \leq X_i < c)} \quad (26)$$

- Simple kernel regression suffers from boundary bias problem - slower rate of convergence at boundary points than in interior points.

Boundary bias

$$plim\hat{\mu}_r(c) = \frac{\int_c^{c+h} \mu(x)f(x)dx}{\int_c^{c+h} f(x)dx} = \mu_{r(c)} + lim_{x\downarrow c}\frac{\partial}{\partial x}\mu(x)\frac{h}{2} + O(h^2)$$

- bias is linear in the bandwidth, h.
- At interior points, bias is usually of order h^2. Convergence of bias to 0 is slower at boundary points.

Recommended alternative: Local linear regression

- Fan and Gijbels (1996) discuss local linear regression methods that have the same order of convergence at boundary points as in interior points.
- These methods fit a regression to observations within a distance h on either side of the discontinuity point.

$$min_{\alpha_r,\beta_r} \sum_{i=1,x_i\geq c}^{n} (Y_i - \alpha_r - \beta_r(X_i - c))^2 K(\frac{X_i - c}{h}) \qquad (27)$$

- $\hat{\alpha}_r$ provides an estimator of μ_r at the point x=c
- Obtain $\hat{\alpha}_l$ similarly and then obtain treatment effect as $\hat{\alpha}_r - \hat{\alpha}_l$.
- Local linear regression has same variance as kernel regression, but faster rate of convergence of bias at boundary points.(Fan and Gijbels, 1996).

Estimation under the FRD design

Again, we need to estimate the expected value of the outcome on both sides of the discontinuity point

$$min_{\alpha_r^y,\beta_r^y} \sum_{i=1,x_i\geq c}^{n} (Y_i - \alpha_r - \beta_r(X_i - c))^2 K(\frac{X_i - c}{h}) \tag{28}$$

$$min_{\alpha_l^y,\beta_l^y} \sum_{i=1,x_i< c}^{n} (Y_i - \alpha_l - \beta_l(X_i - c))^2 K(\frac{X_i - c}{h}) \tag{29}$$

Estimation under the FRD design cont...

- In addition, estimate the expected value of the treatment indicator on both sides of the discontinuity point

$$min_{\alpha_r^w,\beta_r^w} \sum_{i=1,x_i\geq c}^{n} (W_i - \alpha_r - \beta_r(X_i - c))^2 K(\frac{X_i - c}{h}) \tag{30}$$

$$min_{\alpha_l^w,\beta_l^w} \sum_{i=1,x_i< c}^{n} (W_i - \alpha_l - \beta_l(X_i - c))^2 K(\frac{X_i - c}{h}) \tag{31}$$

- The RD treatment effect estimate under the FRD is

$$\hat{\tau} = \frac{\hat{\alpha}_r^y - \hat{\alpha}_l^y}{\hat{\alpha}_r^w - \hat{\alpha}_l^w}$$

Smoothing parameter selection

- For a given bandwidth, h, let the regression function at x be

$$\begin{aligned}\hat{\mu}(x) &= \hat{\alpha}_l(x) \text{ if } x < c \\ &= \hat{\alpha}_r(x) \text{ if } x \geq c\end{aligned}$$

- Define the cross-validation criterion as

$$CV_y(h) = \frac{1}{N}\sum_{i=1}^{n}(Y_i - \hat{\mu}_{-i}(X_i))^2$$

- where $\hat{\mu}_{-i}(X_i)$ is the so-called *leave-one-out* estimator, that leaves out the ith datapoint in calculating the estimate at X_i.

Smoothing parameter selection

- Choose h to minimize $CV_y(h)$ over a grid of possible bandwidths.

$$\mathrm{h}_{CV}^{opt} = argmin_h CV_y(h)$$

- Typically, get a cross-validation "check function"
- It is also possible to choose the bandwidth locally, focussing only on data points within close distance to the cut-off point c.
- Can choose a separate bandwidth for estimating the regression function of W_i given X_i

Assessing the variance of the estimator

- Can obtain standard errors using bootstrap methods
- Bootstrap methods are useful when it is cumbersome to obtain asymptotic standard errors.

  (i) Generate B bootstrap subsamples from the original data (can use 100% sampling with replacement.)
  (ii) Estimate treatment effect within each bootstrap sample
  (iii) The estimate of the treatment effect is based on the original data. The empirical variation across bootstrap estimates provides an estimator of the variance.

$$var(\hat{\hat{\mu}}(\hat{x}_i)) = \frac{1}{B}\sum_{i=1}^{B}(\hat{\mu}_b(x_i) - \bar{\hat{\mu}}_b(x_i))^2$$

Should analysis condition on other covariates?

- There may be other covariates (Z) that are observed and that determine outcomes
- Presence of these covariates rarely changes the identification strategy. The distribution of outcomes is usually continuous in other covariates.
- Do not necessarily need to condition on other covariates.
- In practice, conditioning on Z may be helpful if we use observations on X that are not too close to c.

Graphical analysis

- Integral part of RD analysis
- RD -> treatment impact measured by a discontinuity in expected value of outcome at a particular point
- Inspect histogram estimate of avg value of the outcome around the threshold - is there evidence of a jump?
- Calculate averages that are not smoothed over the cut-off
- Also verify that there is a jump in the probability of treatment at the cut-off point
- It is also useful to inspect graphs for covariates and density of the "forcing" variable to assess credibility
- Plot average values of other covariates
- Plot the density of the forcing variable to look for evidence of manipulation (e.g. individuals know the threshold and can manipulate their value of x_i, for example, by retaking a test.)

RD Examples: Card, D., Dobkin, C.,(2008,AER)

Effect of health insurance coverage on health care utilization

- Medicare eligibility at age 65 leads to sharp changes in the health insurance coverage of the U.S. population and health care utilization increases after age 65.
- Paper compares health-related outcomes (such as different kinds of doctor visits and proceures) among people just before and just after 65, also examining results disaggregated according to group characteristics.
- It follows DiNardo and Lee (2004) and assumes the age profiles in equations (1), (2a) and (2b) are continuous polynomials with potential discontinuities in the derivatives at age 65.
- They also fit many of the models using local linear regression (as suggested by Hahn, Todd and van der Klaauw, 2001) and find results to be relatively robust.

RD Examples: Lalive (2007, J of Econometrics)

Examines whether extended benefits affect unemployment duration

- Analyzes effect of a targeted program that extends the max duration of unemployment benefits from 30 weeks to 209 weeks in Austria for individuals 50 and older living in certain geographic regions.
- There are sharp discontinuities in treatment assignment at age 50 and at the geographical border between eligible and ineligible regions.
- Uses social security data and data on unemployed.
- Two identification strategies: (i) compare individuals around the age cut-off, (ii) compare individuals across geographic borders
- Finds that job search is prolonged by 0.09 weeks per additional week of benefits for women and unemployment duration increase by 0.32 weeks per additional week of benefits for women.

RD Examples: Lalive (2007, J of Econometrics)

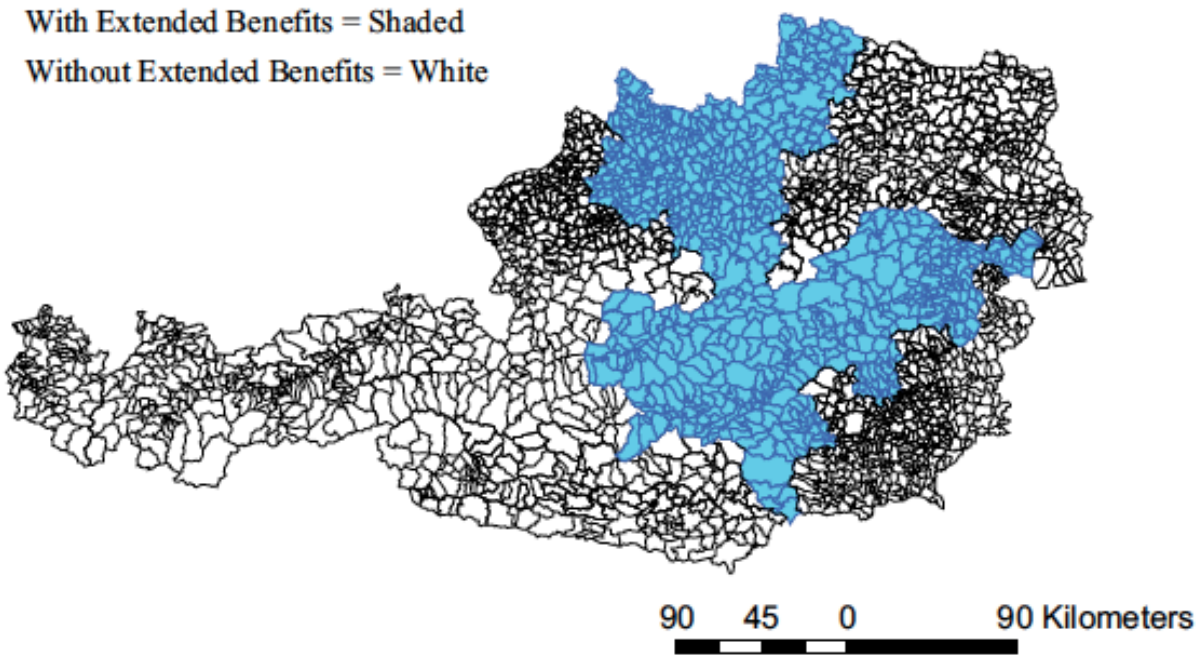


Fig. 1. Regional distribution of REBP.

RD Examples: Lalive (2007, J of Econometrics)

Table 1
Selected descriptive statistics (means)

	Column		
Living in Age bracket	(1) Treated region 50–53 years	(2) Treated region 46–49 years	(3) Control region 50–53 years
A. Men			
Age (years)	51.7	48.0	51.7
Distance to border (minutes)	28.2	27.2	−39.2
Married (share)	0.828	0.785	0.821
Construction (share)	0.481	0.492	0.600
Number of spells	4,759	4,975	8,537
B. Women			
Age (years)	51.5	48.1	51.9
Distance to border (minutes)	27.1	26.6	−37.1
Married (share)	0.780	0.696	0.721
Construction (share)	0.030	0.027	0.034
Number of spells	3,466	2,193	3,625

Source: Own calculations, based on ASSD.

RD Examples: Lalive (2007, J of Econometrics)

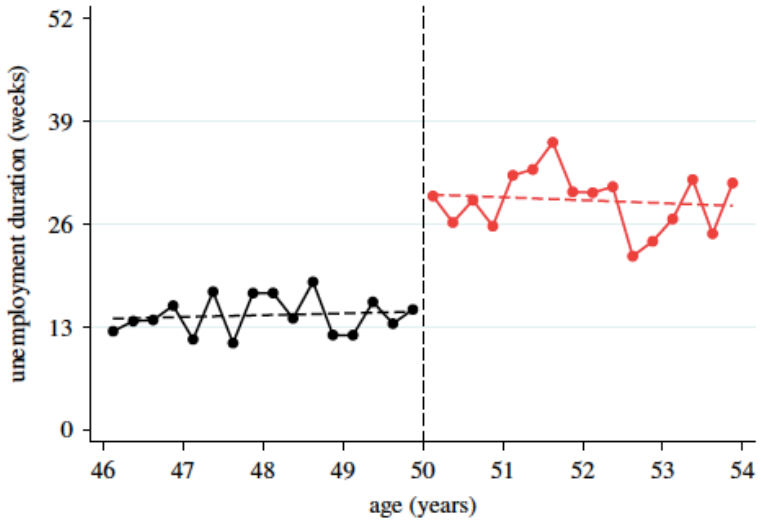


Fig. 2. The effect of REBP on unemployment duration for men: age threshold. Sample restricted to inflow into unemployment the period 8/1989 until 7/1991 (during REBP) and to individuals living in treated region.

RD Examples: Lalive (2007, J of Econometrics)

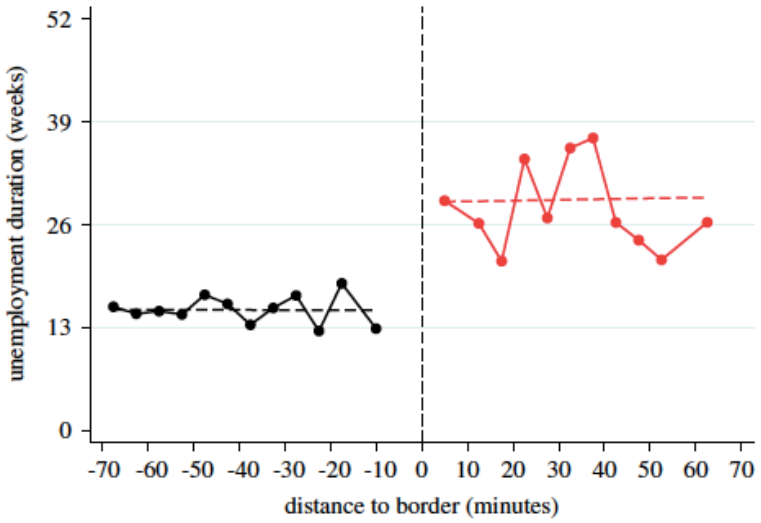


Fig. 3. The effect of REBP on unemployment duration for men: border threshold. Sample restricted to inflow into unemployment the period 8/1989 until 7/1991 (during REBP) and to individuals aged 50 years or older.

RD Examples: Lalive (2007, J of Econometrics)

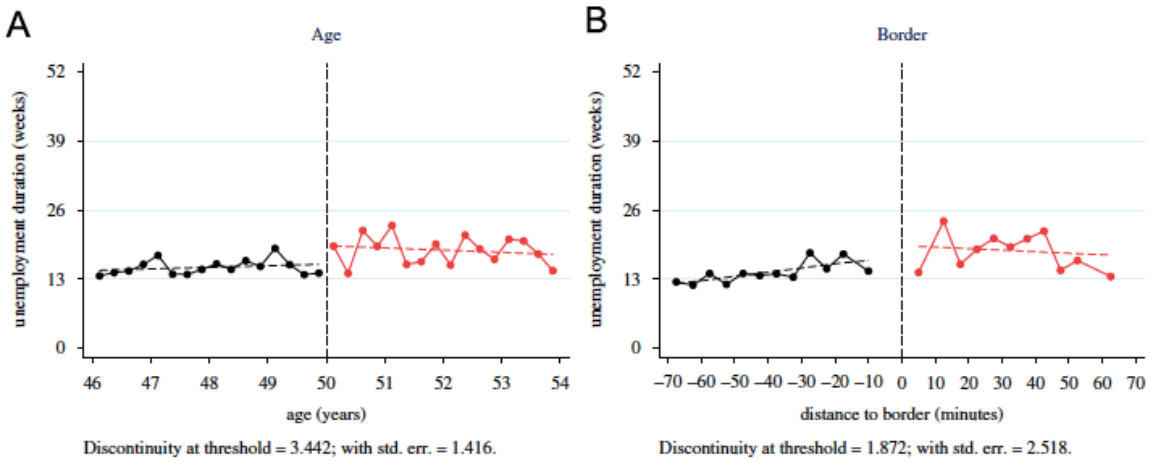


Fig. 4. The effects of age and distance before REBP: men. Sample restricted to inflow into unemployment in the period 1/1986 until 12/1987 (before REBP). Sample for age identification is restricted to treated region. Sample for border identification is restricted to individuals aged 50 years or older.

RD Examples: Jacob, B.A., Lefgren, L., (2004, Restat)

Effect of summer school and grade retention on student performance

- Analyzes the effectiveness of remedial education programs on test scores.
- In 1996, Chicago public schools instituted an accountability policy that tied summer school and grade retention to performance on standardized tests.
- Finds that summer school increased academic achievement in reading and mathematics and that these positive effects remain in the two years following the summer school program.
- Grade retention did not have negative consequences for third graders and increased short run performance.
- Retention had no impact on math performance of older students (sixth graders) and a negative impact on reading.

RD Examples: Jacob, B.A., Lefgren, L., (2004, Restat)

Effect of summer school and grade retention on student performance

- Uses administrative data from the Chicago Public School System
- 40% of third-graders and 30% of sixth graders failed to meet promotional standards.
- 3% of students who scored below the cut-off received waivers from summer school, so design was fuzzy.

FIGURE 1.—STUDENT PROGRESS UNDER THE CHICAGO ACCOUNTABILITY POLICY

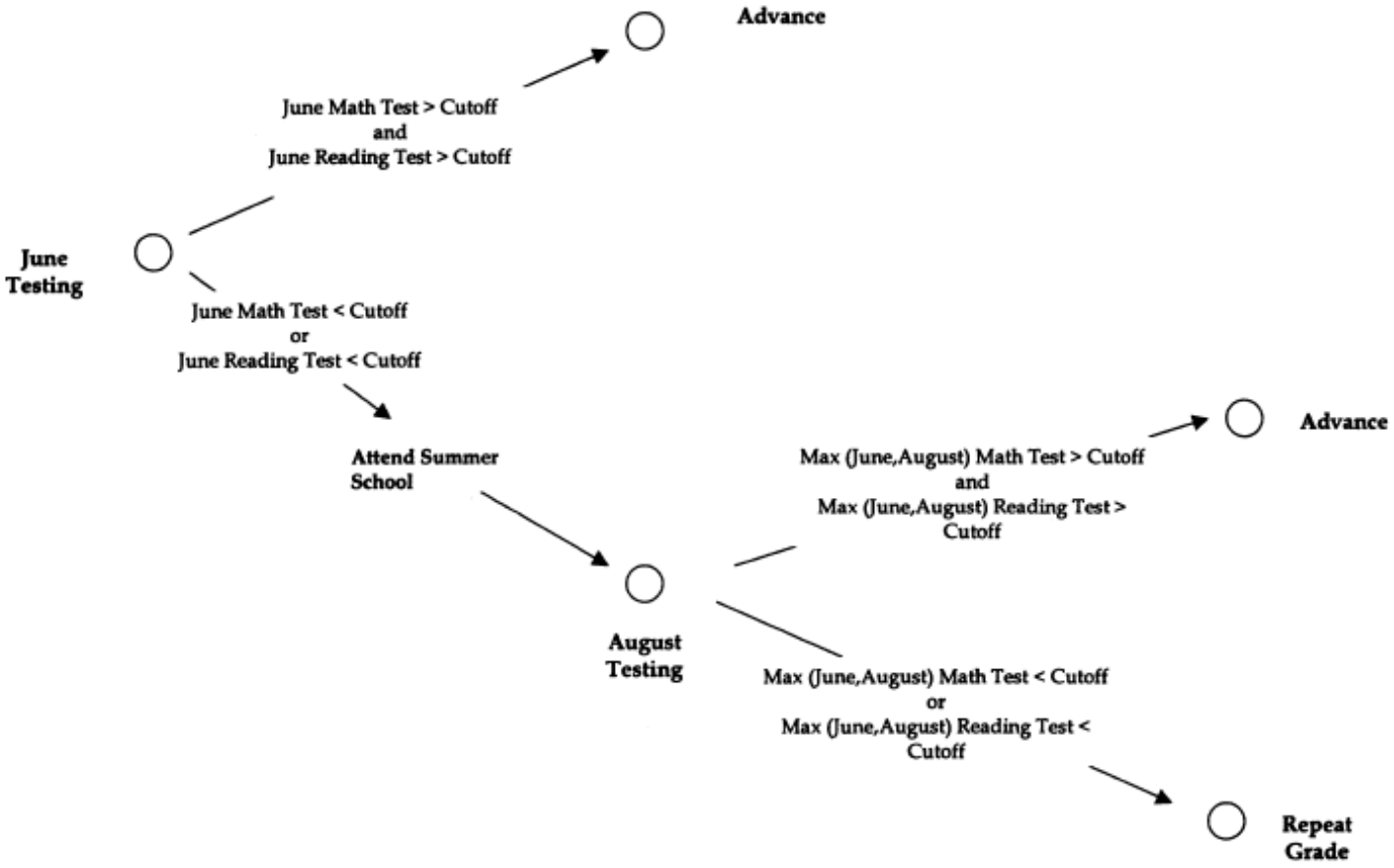
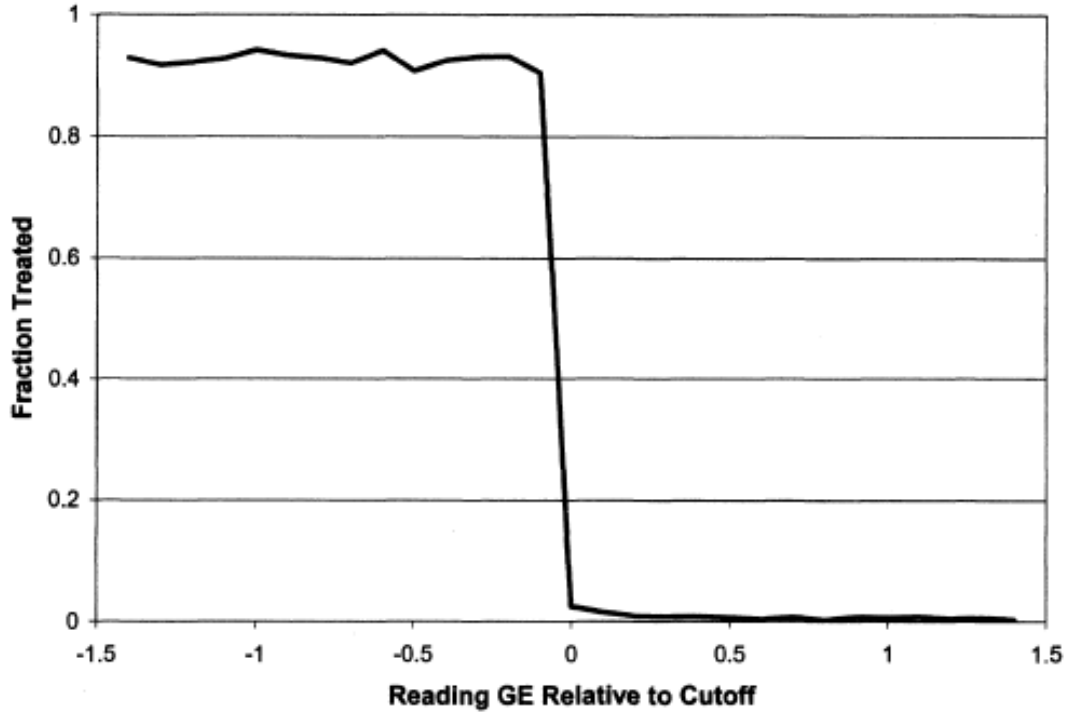


FIGURE 2.—THE RELATIONSHIP BETWEEN JUNE READING SCORES AND THE PROBABILITY OF ATTENDING SUMMER SCHOOL OR BEING RETAINED

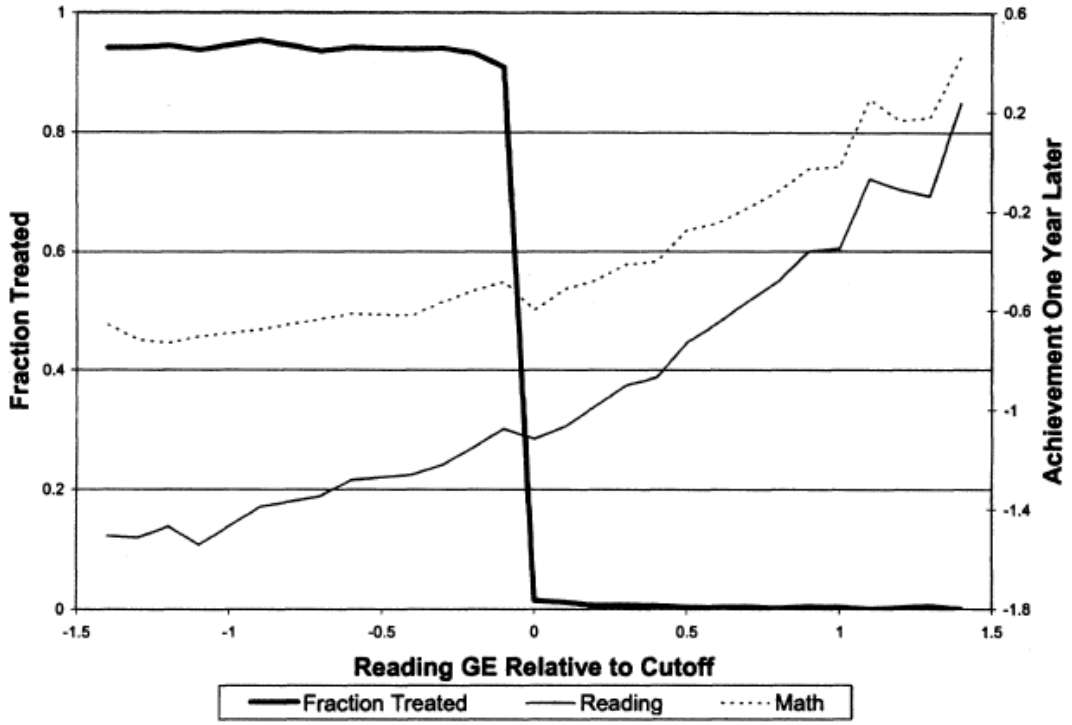


Sample of third- and sixth-grade students from 1997 to 1999 whose June math score exceeded the promotional cutoff but whose June reading score did not.

TABLE 3.—THE NET EFFECT OF SUMMER SCHOOL AND GRADE RETENTION ON STUDENT ACHIEVEMENT

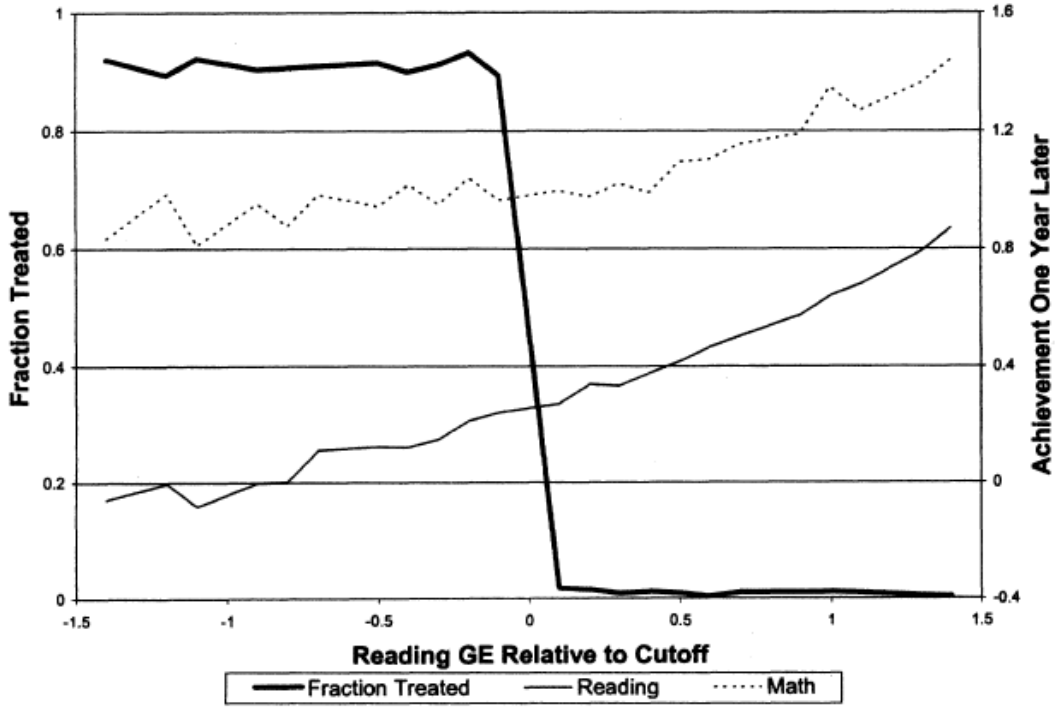
Dependent Variable	Specification		
	OLS (1)	IV (2)	IV (3)
Third grade:			
Reading:			
1 year (n = 13,687)	0.082	0.112	0.104
	(0.019)	(0.026)	(0.025)
2 years (n = 12,806)	0.032	0.064	0.062
	(0.020)	(0.027)	(0.026)
Math:			
1 year (n = 13,664)	0.155	0.132	0.136
	(0.019)	(0.026)	(0.024)
2 years (n = 12,802)	0.066	0.087	0.095
	(0.021)	(0.027)	(0.026)
Sixth grade:			
Reading:			
1 year (n = 7,920)	−0.013	0.012	0.024
	(0.022)	(0.029)	(0.027)
2 years (n = 7,262)	−0.027	−0.015	0.000
	(0.024)	(0.032)	(0.030)
Math:			
1 year (n = 7,904)	0.056	0.077	0.077
	(0.016)	(0.021)	(0.021)
2 years (n = 7,249)	0.007	0.018	0.019
	(0.019)	(0.025)	(0.023)
Additional performance and demographic covariates	No	No	Yes

FIGURE 4.—THE RELATIONSHIP BETWEEN READING AND MATH PERFORMANCE AND JUNE READING PERFORMANCE FOR THIRD-GRADE STUDENTS



Sample of third-grade students from 1997 to 1999 whose June math score exceeded the promotional cutoff but whose June reading score did not.

FIGURE 5.—THE RELATIONSHIP BETWEEN READING AND MATH PERFORMANCE AND JUNE READING PERFORMANCE FOR SIXTH-GRADE STUDENTS



Sample of third-grade students from 1997 to 1999 whose June reading score exceeded the promotional cutoff but whose June math score did not.

Examples: Hahn, J., Todd, P., Van Der Klaauw, W., (1999).

Evaluating the effect of an anti discrimination law

- Assesses the impact of an anti-discrimination law on minority hiring that mandates that firms with 15 or more employees make reports to the government about the ethnic/racial/gender composition of their work.
- Firms with at least 15 employees are covered by Title VII of the Civil Rights Act (1972 Amendment extended coverage from firms with 25 or more to firms with 15 or more employees).
- Uses a sharp RD design.
- Finds that law led to modest increase in minority hiring.

RD Examples

Effect of class size on scholastic achievement

Angrist, J.D., Lavy, V. (1999, QJE)

- Analyzes the effect of class size on student test scores using data from Israel and exploiting a discontinuity created by , which states that a class be added whenever average class size reaches 40 students.
- Finds that reducing class size induces a significant and substantial increase in test scores for fourth and fifth graders, although not for third graders.

RD Examples

Estimating the value parents place on school quality

Black, S., (1999,QJE)

- Uses house prices to infer the value parents place on school quality.
- Compares, within school districts, the prices of houses located on attendance district boundaries - houses that differ only by the elementary school the child attends.
- This comparison removes the variation in neighborhoods, taxes, and school spending.
- Finds that parents are willing to pay 2.5 percent more for a 5 percent increase in test scores.
- Possible that parents on either side are different, so that estimate is a lower bound on valuation.

RD Examples:
Chay, K., McEwan, P., Urquiola, M., (AER, 2005)

Effects of a school incentive program on test score performance

- Evaluates the effect of a school-incentive program in Chile (the Chile-900 program) in which resources were allocated based on cutoffs in schools' mean test scores.
- Shows how a regression discontinuity design that exploits the discrete nature of the selection rule can be used to evaluate the program.
- Finds that the P-900 program had significant but modest size effects on test score gains.

RD Examples: DiNardo, J., Lee, D.S., 2004, QJE)

Effect of unionization on labor market outcomes

- Using US establishment-level data on establishments that faced union organizing drives during 1984-1999, this paper uses a sharp RD design to estimate the impact of unionization on business survival, employment, output, productivity, and wages.
- Compares outcomes for employers where unions won the election by a close margin with those where the unions lost by a close margin (e.g. 49% compared to 51%).
- Impacts on all outcomes are small and impacts on wages are close to zero. Concludes that mandates for employers to bargain with unions had little effect.

RD Examples: Card, D., Mas, A., Rothstein, J., (2006, QJE)

Tests for discontinuities in the dynamics of neighborhood racial composition

- Theoretical models of social interactions (Schelling (1971)) predict tipping behavior in neighborhoods - e.g. once the minority share in a neighborhood exceeds a so-called tipping point,all the whites leave.
- This paper uses regression discontinuity methods and Census tract data from 1970 through 2000 to test for discontinuities in the dynamics of neighborhood racial composition.
- Finds evidence for tipping-like behavior in most cities, with a distribution of tipping points ranging from 5% to 20% minority share, but evidence for tipping points in on other outcomes, like house prices.

RD Examples: Lee (2007, J of Econometrics)

Effect of incumbancy advantage

- Uses data on US Congressional election returns from 1946 to 1998.
- Analyzes the effect of the incumbancy advantage at the level of the party at the district level, without regard to the identify of the nominee for the party.
- For example, analyzes the prob of winning the election in t+1 given that democrats won the election in t, coming districts where they won by a close margin to districts where they lost by a close margin.
- Paper recommends checking the density of observables to test for systematic selection around the cut-off point.
- Finds that democrats who just barely win the election are much more likely to run for office and succeed in the next election compared to democrats who barely lose, which implies a large incumbency advantage. (also see Moretti and Butler, 2004, QJE)

Recommended RD Practices

(Imbens and Limieux, 2007)

Sharp RD Designs

1. Graph the data by computing the average value of the outcome variable over a set of bins. The bin width should be large enough to have a sufficient amount of precision so that the plots looks smooth on either side of the cut-off value, but also small enough to be able to see the jump around the cut-off value.

2. Estimate the treatment effect by running linear regressions on both sides of the cut-off points using only data within a bin width h of the cut-off point. These are kernel regressions using a rectangular kernel.

-standard errors can be computed using standard least squares methods (using robust standard errors).

-optimal bandwidth can be chosen using cross-validation

3. Examine robustness of the results by
(a) looking at possible jumps in the value of other covariates around the cut-off point.
(b) using various values of the bandwidth, with and without controlling for other covariates in the regression.

4. The performance of the estimator can be improved by using nonparametric local linear regression and computing the standard errors either using a plug-in approach or by bootstrapping.

Fuzzy Regression Discontinuity Designs

1. Graph the average outcomes over a set of bins as in the case of SRD, but also graph the probability of treatment.

2. Estimate the treatment effect using TSLS applied only to data within h of the cut-off (above and below), which is numerically equivalent to computing the ratio in the estimate of the jump (at the cutoff point) in the outcome variable over the jump in the treatment variable.

3. Standard errors can be computed using robust TSLS estimates or using a plug-in estimator.

4. Robustness can be examined using similar approaches as in SRD.

5. The performance of the estimator can again be improved by using nonparametric local linear regression instead and computing the standard errors either using a plug-in approach or by bootstrapping.