Empirical Analysis III

The University of Chicago

April 11, 2019

Content: Week 2

- Today we focus on observational data, a single cross-section
- What can we learn?

- How to select controls and argue exogeneity
- How to perform matching
- Relation between matching and regression

Selection

Formal definition

- There is **selection** into the treatment state D if

$$\underbrace{Y_d|D=d}_{\text{observable}} \text{ is distributed differently from } \underbrace{Y_d|D=d'}_{\text{unobserved}} \text{ for } d' \neq d$$

- This is not the case under the random assignment assumption
- Expected to occur if agents choose D with knowledge of $\{Y_d\}_{d\in\mathcal{D}}$

Selection is common

- Particularly concerning if you are trained in neoclassical economics
- Agents choose a job training program ($D \in \{0,1\}$) to max utility
- Utility will incorporate expected future earnings (Y_0, Y_1)
- Agents who choose job training might do so because of low Y_0
- Data typically supports this story ("Ashenfelter's (1978) dip")
- Alternatively, might choose $D=0$ because of high Y_0

Selection Bias

- Consider the simple treatment/control mean contrast under selection
- This contrast would be the ATE under random assignment
- Decompose the contrast into a causal effect and selection bias:

$$\begin{aligned}
&\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0] \\
&\quad = \underbrace{(\mathbb{E}[Y_1|D=1] - \mathbb{E}[Y_0|D=1])}_{\text{ATT}} + \underbrace{(\mathbb{E}[Y_0|D=1] - \mathbb{E}[Y_0|D=0])}_{\text{slection bias}} \\
\text{or} \quad &= \underbrace{(\mathbb{E}[Y_1|D=0] - \mathbb{E}[Y_0|D=0])}_{\text{ATU}} + \underbrace{(\mathbb{E}[Y_1|D=1] - \mathbb{E}[Y_1|D=0])}_{\text{slection bias}}
\end{aligned}$$

- First term is the causal effect for those who were treated/untreated
- Under random assignment would have ATT = ATU = ATE
- Second term is how the treated would have been different anyway
  Under random assignment this would be 0
- The first expression is more natural if thinking of $D = 0$ as baseline

Selection on Observables

- A simple relaxation of random assignment is **selection on observables**
- Suppose that we observe (Y, D, X) where X are covariates
- The selection on observables assumption is that

$$\{Y_d\}_{d \in \mathcal{D}} \perp\!\!\!\perp D | X$$

- Says: Conditional on X, treatment is as-good-as randomly assigned
- Other terms: **unconfoundedness, ignorable treatment assignment**
- Underlies causal interpretations of linear regression
  We will look into this connection more later

Identification under Selection on Observables

Identification argument

- Conditional version of random assignment:

$$\begin{aligned} F_d(y|x) &\equiv \mathbb{P}[Y_d \leq y | X = x] \\ &= \mathbb{P}[Y_d \leq y | D = d, X = x] = \mathbb{P}[Y \leq y | D = d, X = x] \end{aligned}$$

- Second equality requires the **overlap condition**: $\mathbb{P}[D = d|X = x] > 0$
- By averaging over x, one can point identify the marginals

$$F_d(y) \equiv \mathbb{P}[Y_d \leq y] = \mathbb{E}(\mathbb{P}[Y_d \leq y|X]) = \mathbb{E}(\mathbb{P}[Y \leq y|D = d, X])$$

- Treatment effects are then constructed from $D = d_1 \rightarrow d_2$ contrasts

Thought experiment: An RCT given $X = x$

- Fix an $X = x$
- Find treated (d_1) and control agents (d_2) with $X = x$ (**"match"** on $X = x$)

Identification of Mean Contrasts

- Suppose $D \in \{0, 1\}$ is binary — by far the most common case
- Using essentially the same argument as on the previous page:

$$\text{ATE} \equiv \mathbb{E}\big[\mathbb{E}[Y_1|X] - \mathbb{E}[Y_0|X]\big] = \mathbb{E}\big[\mathbb{E}[Y|D=1,X] - \mathbb{E}[Y|D=0,X]\big]$$

- Similar expressions for the ATT and ATU have an important difference:

$$\text{ATT} = \mathbb{E}[Y|D=1] - \mathbb{E}\big[\mathbb{E}[Y|D=0,X]|D=1\big]$$
$$\text{ATU} = \mathbb{E}\big[\mathbb{E}[Y|D=1,X]|D=0\big] - \mathbb{E}[Y|D=0]$$

- Helps in estimation since only one conditional expectation (more later)
- Note that only **mean independence** is needed for these arguments:

$$\mathbb{E}[Y_d|D=0,X] = \mathbb{E}[Y_d|D=1,X]$$

- Difficult to think of arguments for mean (without full) independence

Conditional independence and choice of controls

When might conditional independence hold?

- You have detailed information about the assignment mechanism
- You have very rich data: personal traits, histories, etc.

Only condition on predetermined observables

- For selection on observables to be plausible, X should be **predetermined**
- Usually this really is a temporal issue (measured before vs. after D)
- Intuition is clear - we want to condition on selection into treatment

Examples

- Suppose we accidentally included Y as part of X
- Then clearly we aren't going to have $(Y_0, Y_1) \perp\!\!\!\perp D|X$. Thus:
- Don't include earnings after the program in X
- Don't include employment after the program in X
- Don't include marital status after the program in X

What happens if your controls are affected by treatment?

Neale and Johnson (1996, JPE)

- What is the role of premarket factors in Black-White wage gap?
- They regress adult earnings (Y) on race ($D = 1$ if white) and some covariates, with and without test scores T
- A key findings it that:
  $\mathbb{E}[Y|D = 1] - \mathbb{E}[Y|D = 0] >> \mathbb{E}[Y|D = 1, T = 1] - \mathbb{E}[Y|D = 0, T = 1]$

- Potential outcome model for earnings, test scores and race:

$$Y = DY_1 + (1 - D)Y_0$$
$$T = DT_1 + (1 - D)T_0$$

- For simplicity, abstacts from covariates and assume:

$$D \perp\!\!\!\perp (Y_1, Y_0, T_1, T_0)$$

What happens if your controls are affected by treatment?

- Unconditional regression:

$$\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0] = \mathbb{E}[Y_1 - Y_0]$$
$$\mathbb{E}[T|D=1] - \mathbb{E}[T|D=0] = \mathbb{E}[T_1 - T_0]$$

- Conditional on test scores:

$$\begin{aligned}
&\mathbb{E}[Y|D=1, T=1] - \mathbb{E}[Y|D=0, T=1] \\
=&\mathbb{E}[Y_1|D=1, T_1=1] - \mathbb{E}[Y_0|D=0, T_0=1] \\
\underbrace{=}_{\text{independence}}&\mathbb{E}[Y_1|T_1=1] - \mathbb{E}[Y_0|T_0=1] \\
=&\underbrace{\mathbb{E}[Y_1 - Y_0|T_1=1]}_{\text{causal effect conditional on test score}} + \underbrace{(\mathbb{E}[Y_0|T_1=1] - \mathbb{E}[Y_0|T_0=1])}_{\text{selection bias}}
\end{aligned}$$

- Selection bias would be negative if blacks need to be smarter to achieve the same test score, despite their disadvantages background

Does bias go down if you control for more?

- Suppose that $(Y_0, Y_1) \perp\!\!\!\perp D|X$
- Let X_2 be a subset of X
- Let X_1 be a subset of X_2
- So controlling on X is the most information, and X_1 is the least
- Suppose however that we only have X_2 (hence X_1) in our data

- Perhaps surprisingly, using X_2 can be more biased than using X_1
- That is, adding information (X_2 vs X_1) need not reduce bias
- If $X_2 = X$, then it does, but not more generally
- Point is not well-appreciated but should be concerning
- Means we really need to have the correct X
- Suggests one should be careful with automated model selection
- Insight seems to come from Heckman and Navarro-Lozano (2004)

Propensity Score

Definition

- Binary treatment case, $D \in \{0, 1\}$
- $p(x) \equiv \mathbb{P}[D = 1|X = x]$ is called the **propesity score**
- Let $P \equiv p(X)$ be the random variable $\mathbb{P}[D = 1|X]$

Rosenbaum and Rubin (1983) sufficiency argument

- Result: Selection on observables implies $(Y_0, Y_1) \perp\!\!\!\perp D|p$
- Can you prove it (see problem set)?
- The result implies that we can condition on $p(X)$ instead of X
- Still need overlap, but now with $p(X)$ (scalar) instead of X (vector)

Propensity Score Weighting

- Using p, we can write the ATE as a weighted average of Y
- The following is derived in the supplemental notes:

$$\text{ATE}(x) = \mathbb{E}\left[\frac{Y(D - p(x))}{p(x)(1 - p(x))} | X = x\right]$$

- Average over X to point identify

$$\text{ATE} = \mathbb{E}\left[\frac{Y(D - p(X))}{p(X)(1 - p(X))}\right]$$

- similar expression can be derived for the ATT and ATU
- can you derive these expressions (see problem set)?

Why Have Different Identification Arguments?

Summarizing identification under selection on observables

- Three different constructive identification result for the ATE
- Match on X, match on P, weight using p
- Each one shows that ATE is point identified
- And they are all derived under the same assumptions

So why have three?

- Identification arguments directly inform the construction of estimators
- Different arguments suggest different estimators
- In general, these different estimators may have different properties: Efficiency, rates of convergence, finite sample performance...
- See e.g. Imbens (2010, 2015)

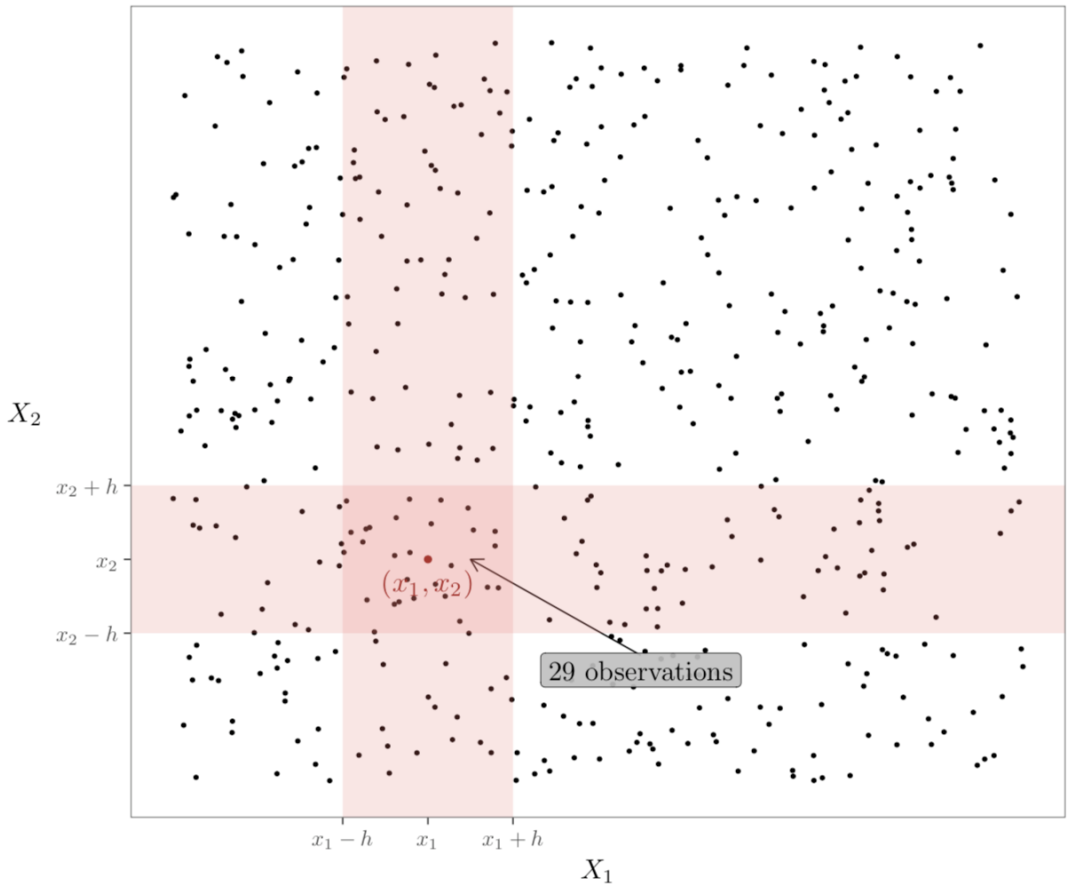
Curse of Dimensionality

- Matching requires the estimmation of $\mathbb{E}[Y|D=0,X]$
- Doing this non-parametrically is typically not feasible: **Curse of Dimensionality**
- e.g. K=30 binary covariates $\implies$ min $N = 2^{30+1} = 2147483648$

Overcoming the curse

- Regression analysis
- Match on Propensity score
- Propesity score weighting
- Inexact matching: metric on X (Mahalanobis)

Curse of Dimensionality



- Observations used to estimate $\mathbb{E}[Y|X = x_1, X = x_2]$ with bandwidth h
- Effective number of observations drops by an order of magnitude

Matching vs linear regression

- Consider the following saturated-in-X_i regression model:

$$Y_i = \sum_{k=1}^{K} d_{ik}\alpha_k + \beta D_i + U_i$$

$$d_{ik} = \mathbb{1}[X_i = x_k]$$

- α_k is the coefficient for $X_k = x_k$
- β is the regression estimand for treatment
- Note: no interactions between X_i and D_i

Regression as pseudo-matching

- Then it can be shown that:

$$\beta = \sum_{k=1}^{K} \{\mathbb{E}[Y_i|D_i = 1, X_i = x_k] - \mathbb{E}[Y_i|D_i = 0, X_i = x_k]\} w_k$$

$$w_k = \frac{\mathbb{P}[D_i = 1|X_i = x_k](1 - \mathbb{P}[D_i = 1|X_i = x_k])\mathbb{P}[X_i = x_k]}{\sum_{k=1}^{K} \mathbb{P}[D_i = 1|X_i = x_k](1 - \mathbb{P}[D_i = 1|X_i = x_k])\mathbb{P}[X_i = x_k]}$$

- weights depend on the conditional variance of treatment status

Linear regression and matching

- Differ in weighting of $\mathbb{E}[Y_i|D_i = 1, X] - \mathbb{E}[Y_i|D_i = 0, X]$
- Same if treatment effects do not vary with X
- Biased with selection on unobservable factors of non-treatment level

Matching and ATT

- Suppose X takes on values $x_1, ..., x_K$
- Then the matching estimator $\hat{\beta}^{ATT}$ can be written as:

$$\sum_{k=1}^{K} \{\mathbb{E}[Y_i|D_i = 1, X_i = x_k] - \mathbb{E}[Y_i|D_i = 0, X_i = x_k]\}\mathbb{P}[X_i = x_k|D_i = 1]$$

- Where $\mathbb{P}[X_i = x_k|D_i = 1]$ is the probability mass function for X_i given $D_i = 1$
- That is, partition the treated and control sample in K cells by X
  - Calculate the mean outcome difference in each cell
  - Take a weighted average of the mean differences
  - Using the fraction of treated observations in each cell as the weights

Matching and ATE

- Suppose X takes on values $x_1, ..., x_K$
- Then the matching estimator $\hat{\beta}^{ATE}$ can be written as:

$$\sum_{k=1}^{K} \{\mathbb{E}[Y_i|D_i = 1, X_i = x_k] - \mathbb{E}[Y_i|D_i = 0, X_i = x_k]\}\mathbb{P}[X_i = x_k]$$

- Where $\mathbb{P}[X_i = x_k]$ is the probability mass function for X_i in the whole population
- That is, partition the treated and control sample in K cells by X
  - Calculate the mean outcome difference in each cell
  - Take a weighted average of the mean differences
  - Using the fraction of the total sample in each cell as the weights

Numerical example

- Suppose that X is binary and $\mathbb{P}[X = 1] = 0.5$, $\mathbb{P}[D_i = 1|X = 0] = 0.9$ and $\mathbb{P}[D_i = 1|X = 1] = 0.5$
- Recall Bayes' rule:

$$\mathbb{P}[X_i = x_k|D_i = 1] = \mathbb{P}[D_i = 1|X_i = x_k]\mathbb{P}[X_i = x_k]/\mathbb{P}[D_i = 1]$$

- Then, we get:

$$\begin{aligned}\beta^{ATT} &= 0.64(\mathbb{E}[Y_i|D_i = 1, X_i = 0] - \mathbb{E}[Y_i|D_i = 0, X_i = 0]) \\ &+ 0.36(\mathbb{E}[Y_i|D_i = 1, X_i = 1] - \mathbb{E}[Y_i|D_i = 0, X_i = 1])\end{aligned}$$

$$\begin{aligned}\beta^{ATE} &= 0.5(\mathbb{E}[Y_i|D_i = 1, X_i = 0] - \mathbb{E}[Y_i|D_i = 0, X_i = 0]) \\ &+ 0.5(\mathbb{E}[Y_i|D_i = 1, X_i = 1] - \mathbb{E}[Y_i|D_i = 0, X_i = 1])\end{aligned}$$

$$\begin{aligned}\beta^{OLS} &= 0.26(\mathbb{E}[Y_i|D_i = 1, X_i = 0] - \mathbb{E}[Y_i|D_i = 0, X_i = 0]) \\ &+ 0.74(\mathbb{E}[Y_i|D_i = 1, X_i = 1] - \mathbb{E}[Y_i|D_i = 0, X_i = 1])\end{aligned}$$

How Propensity Score matching works

- Where before we had troubles estimating

$$\mathbb{E}[Y|D=0,X]$$

- Now we can estimate

$$\mathbb{E}[Y|D=0,p(X)]$$

- This is just a non-parametric regression of Y on 1 dimensional $p(X)$

- There is a little catch: we need to estimate $p(X)$
- This reintroduces the curse of dimensionality in the estimation of $p(X)$

How Propensity Score matching works

- Estimate $p(X)$
  - Choice of X: all the attributes X influencing both D and Y_0, Y_1
    guidance: economic theory, a priori considerations, institutional set-up, previous literature
    plausibility of matching depends on the choice of X
    sensitivity of results to the set of regressors used
  - Estimation: e.g. Probit or Logit
  - Note: we are not interested in the behavioral interpretation, all we want:

$$X \perp\!\!\!\perp D | p(X)$$

- Match treated to controls
- Check Common Support
- Check balancing of X's on CS, if it holds $\rightarrow$ Calculate treatment effect
- Reiterate until balancing condition holds

Matching Estimators Overview

- Pair to each treated individual i some group of "comparable" non - treated individuals
- Associate to the outcome Y of treated individual, a matched outcome $\hat{Y}$ given by the (weighted) outcomes of the "neighbors" in the comparison group:

$$\hat{Y}_0 = \sum_{j \in C^0} w_j Y_j$$

- C^0 - set of neighbours of treated individual in the D=0 group
- w_j - weight of non-treated j in forming a comparison with treated i, where $\sum_{j \in C^0} w_j = 1$
- Weights differ for each treated individual

Classes of Matching Estimators

- Traditional matching estimators
  - one-to-one (nearest neighbour) matching
- Simple smoothed matching estimators
  - K-nearest neighbours
- Weighted smoothed matching estimators
  - kernel-based matching
  - local linear regression-based matching
- Trade-off between matching quality and variance

One-to-one matching

- To each treated unit, match only one non-treated unit:

$$C^0 = \{j : |p(X_i) - p(X_j)| = \min_{k \in \{D=0\}} |p(X_i) - p(X_k)|\}$$

$$\forall m \in \{D = 0\} : w_m = \mathbb{1}[m = j]$$

With replacement

- Many treated units may be matched to the same non-treated unit
- Less bias, more variance

Without replacement

- Once a non-treated unit has been matched, it cannot be used again
- More bias, less variance

Simple smoothed matching

K-nearest neighbours

- C^0= the K units in $D = 0$ with $p(X_k)$ closest to $p(X_i)$

$$w_m = \begin{cases} 1/K & m \in C^0 \\ 0 & o/w \end{cases}$$

- With or without replacement

Radius matching

- $C^0 = \{\text{all } j : |p(X_i) - p(X_j)| < \delta\}$

$$w_m = \begin{cases} 1/|C^0| & m \in C^0 \\ 0 & o/w \end{cases}$$

Kernel matching

- The matched outcome for treated i is a kernel - weighted average of the non - treated outcomes, where the weights are in proportion to the non - treated closeness to i:

$$\hat{Y}_0 = \frac{\sum_{j \in \{D=0\}} K\left(\frac{p(X_i)-p(X_j)}{h}\right) Y_j}{\sum_{j \in \{D=0\}} K\left(\frac{p(X_i)-p(X_j)}{h}\right)} = \sum_{j \in \{D=0\}} w_j Y_j$$

- Non-treated j's outcome Y_j is weighted by

$$w_j = \frac{K\left(\frac{p(X_i)-p(X_j)}{h}\right)}{\sum_{j \in \{D=0\}} K\left(\frac{p(X_i)-p(X_j)}{h}\right)}$$

Non-parametric background

- For each treated i, estimate $\hat{Y}_0 \equiv [Y|D = 0, p(X) = p(X_i)]$ non-parametrically:
  - Fit a constant estimated on a local neighbourhood of $p(X_i)$
  - Apply a weighting scheme with weights K_j, for $p = p(X)$:

$$\min_{\theta_0} \sum_{j \in C^0} (Y_j - \theta_0)^2 K(\frac{p(X_i) - p(X_j)}{h})$$

$\implies$ $\hat{Y}_0$ is approximated locally by a constant:

$$\hat{Y}_0 = \hat{\theta}_0 = \sum_j \frac{K_j}{\sum_j K_j} Y_j$$

Non-parametric background

Kernel choice

- Gaussian

$$K(u) \propto \exp(-\frac{u^2}{2})$$

  uses all the non-treated units

- Epanechnikov

$$K(u) \propto \begin{cases} (1-u^2) & |u| < 1 \\ 0 & o/w \end{cases}$$

  has a moving window within the $D = 0$ group

Bandwidth h selection

- Bias-variance trade-off
- $h \uparrow$, more "tolerant" in terms of closeness of matches

Local linear regression matching

- For each treated i, estimate $\hat{Y}_0 \equiv \mathbb{E}[Y|D=0, p(X)=p(X_i)]$ non-parametrically:
  - Fit a line estimated on a local neighbourhood of $p(X_i)$
  - Apply a weighting scheme with weights:

$$\min_{\theta_0,\theta_1} \sum_{j \in C^0} (Y_j - \theta_0 - \theta_1(p(X_i) - p(X_j)))^2 K(\frac{p(X_i) - p(X_j)}{h})$$

Mahalanobis-metric matching

Alternative to propensity score matching:

- Combine the X's into a distance measure and then match on the resulting scalar:

$$d(i,j) = (X_i - X_j)^T V^{-1} (X_i - X_j)$$

  with V the pooled within-sample covariance matrix of X
- Properties:
  - Unit free
  - Assigns weight to each w in proportion to the inverse of the variance of w
  - reduces differences in X within matched pairs in all directions
  - Uses $d(i,j)$ instead of $|p(X_i) - p(X_j)|$

Matching in practice using LaLonde data

- NSW was a randomized experiment that assigned people to training positions
- mid 1970s:
  - AFDC women, ex-drug addicts, ex-criminal offenders, high school drop-outs of both sexes
  - guaranteed a job for 9 to 18 months
- Baseline earnings and demographics (1975)
- Post treatment earnings demographics (1979)
- LaLonde (AER,1986) compared experimental with non-experimental estimates
- For now, we will just use this data to see how matching can be done in practice
- But get to know paper and data – we will return to this again several times

Estimating the Propensity Score

- Consider treated units from NSW sample - 297 obs., and control units from CPS sample - 15992 obs.

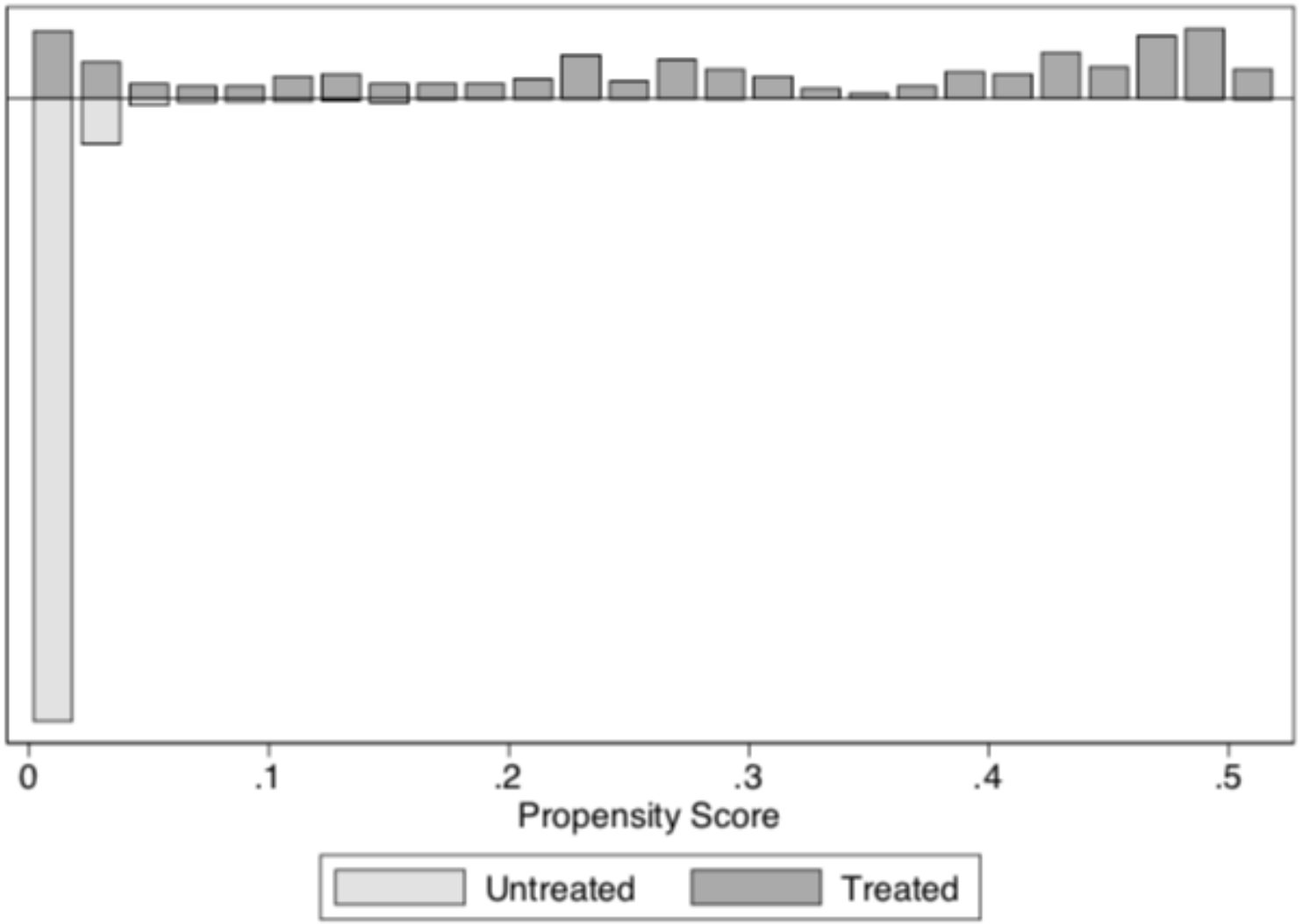
```
. probit treatment age education black hispanic married nodegree earnings75

Probit regression                                 Number of obs   =      16289
                                                  LR chi2(7)      =    1421.68
                                                  Prob > chi2     =     0.0000
Log likelihood = -772.77819                       Pseudo R2       =     0.4791

------------------------------------------------------------------------------
   treatment |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0106229   .0043183    -2.46   0.014    -.0190866   -.0021591
   education |   .0220374   .0194534     1.13   0.257    -.0160905    .0601653
       black |   1.918753   .0834945    22.98   0.000     1.755107     2.0824
    hispanic |   .8917917    .116267     7.67   0.000     .6639125    1.119671
     married |  -.6246106    .091131    -6.85   0.000     -.803224   -.4459972
    nodegree |   .5700337   .1043422     5.46   0.000     .3655266    .7745407
   earnings75 |  -.0000518   6.25e-06    -8.29   0.000     -.000064   -.0000395
       _cons |  -2.513416   .3092365    -8.13   0.000    -3.119508   -1.907324
------------------------------------------------------------------------------
```

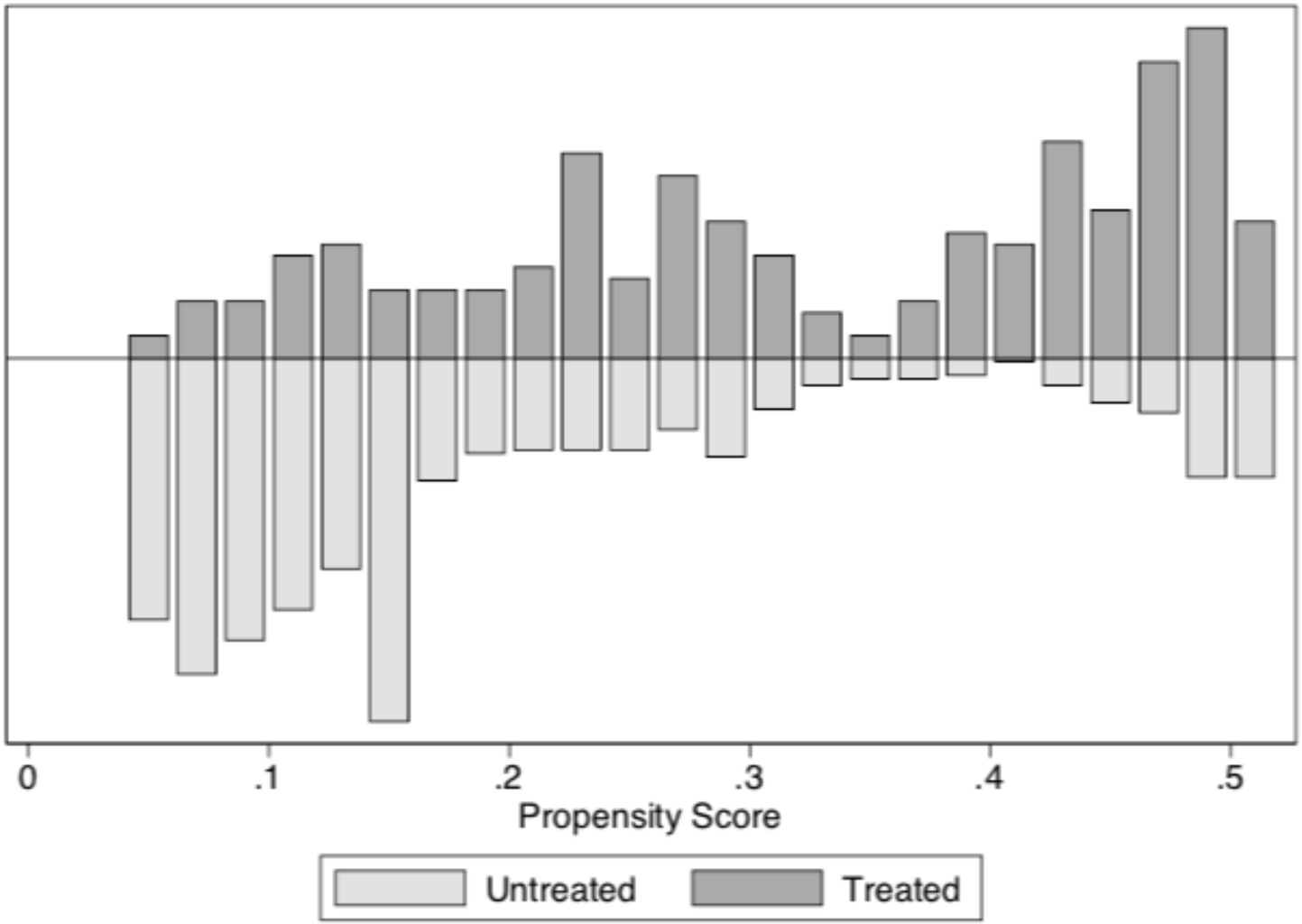
Estimating the Propensity Score

- Checking common support: vertical axis - density



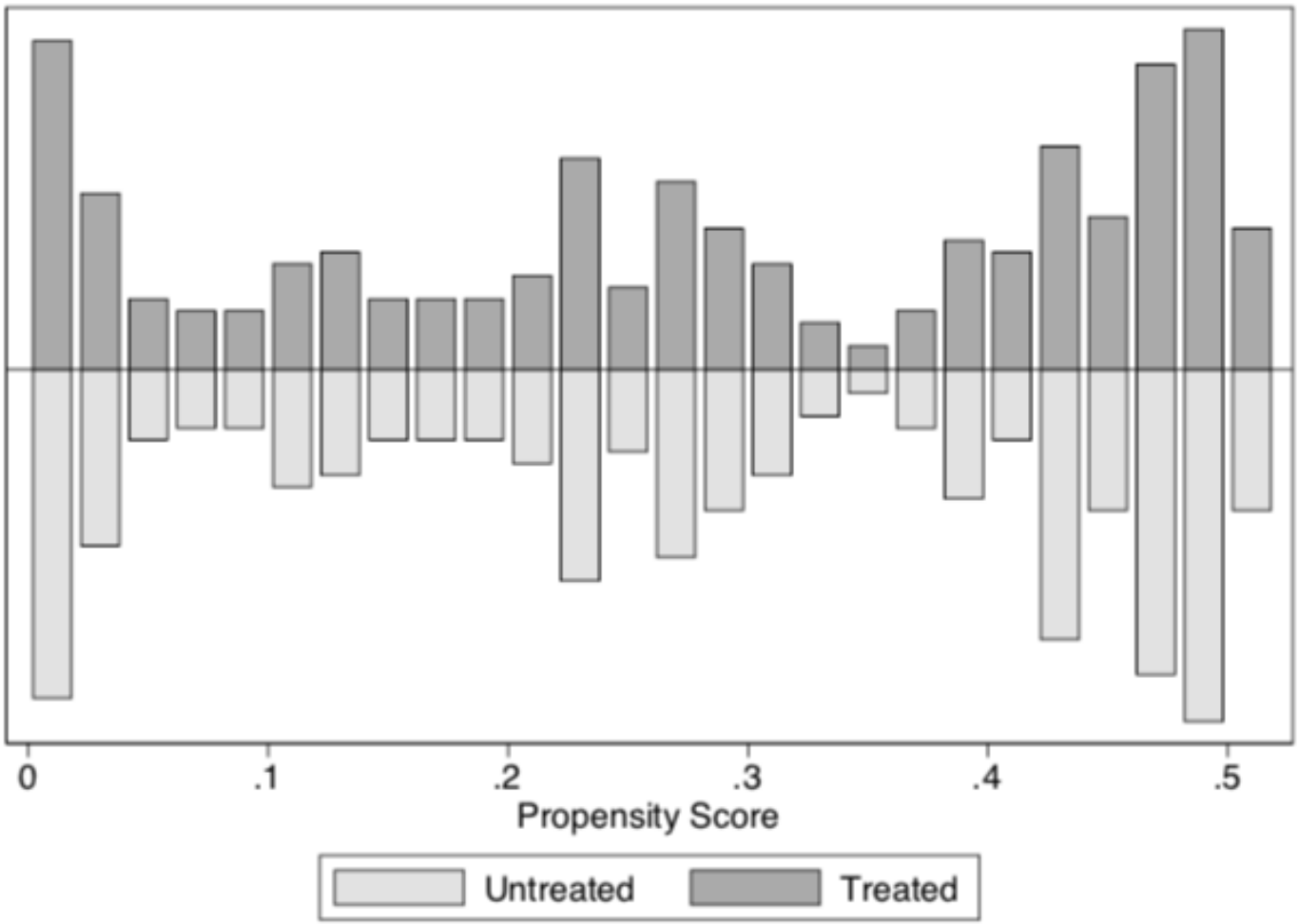
Estimating the Propensity Score

- Checking common support: trimming



One-to-One Matching with Replacement

- Match the treated units with closest (in $p(X)$ metrics) untreated units: 193 out of 15992 untreated units left



Assessing Matching Quality

- Check balancing of observables:

$$D \perp\!\!\!\perp X | p(X)$$

- For each variable compute standardized % bias before and after matching:

$$B_{\text{before}}(X) = 100 \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{(V_1(X) + V_0(X))/2}}$$

$$B_{\text{after}}(X) = 100 \frac{\bar{X}_{1M} - \bar{X}_{0M}}{\sqrt{(V_{1M}(X) + V_{0M}(X))/2}}$$

- Achieved % reduction in |bias|
- t-tests for equality of means before and after matching

Assessing Matching Quality

```
. pstest age education black hispanic married nodegree earnings75

-------------------------------------------------------------------------------
                         |       Mean                      %reduct |     t-test
    Variable      Sample | Treated Control     %bias  |bias| |    t    p>|t|
-------------------------+----------------------------------+-----------------
         age   Unmatched | 24.626   33.225     -94.2          | -13.37  0.000
                 Matched | 24.626   24.923      -3.2    96.6  |  -0.43  0.666
                         |                                    |
   education   Unmatched |  10.38   12.028     -68.5          |  -9.85  0.000
                 Matched |  10.38    10.36       0.8    98.8  |   0.11  0.914
                         |                                    |
       black   Unmatched | .80135   .07354     215.6          |  47.04  0.000
                 Matched | .80135   .83838     -11.0    94.9  |  -1.17  0.241
                         |                                    |
    hispanic   Unmatched | .09428   .07204       8.1          |   1.47  0.143
                 Matched | .09428   .07071       8.5    -6.0  |   1.04  0.297
                         |                                    |
     married   Unmatched | .16835   .71173    -130.7          | -20.54  0.000
                 Matched | .16835   .17172      -0.8    99.4  |  -0.11  0.913
                         |                                    |
    nodegree   Unmatched | .73064   .29584      96.5          |  16.27  0.000
                 Matched | .73064   .68013      11.2    88.4  |   1.35  0.178
                         |                                    |
  earnings75   Unmatched | 3066.1    14017    -144.2          | -19.67  0.000
                 Matched | 3066.1   3362.9      -3.9    97.3  |  -0.76  0.446
                         |                                    |
-------------------------------------------------------------------------------
```

Assessing Matching Quality

Overall measures

- Summary indicators of the distribution of |bias| before and after matching
- Pseudo R^2 from Probit of treatment on covariates before matching and on matched samples
- P-values of the likelihood-ratio test of joint insignificance of covariates before and after matching

Common support

- % of treated lost?
- Compare to those off-the-support: covariates, Y

Assessing Matching Quality

- Probit on matched sample: note the decrease in pseudo R^2 comparing to the unmatched sample

```
. probit treatment age education black hispanic married nodegree earnings75 [iw=_w]

Probit regression                                 Number of obs   =        490
                                                  LR chi2(7)      =       6.61
                                                  Prob > chi2     =     0.4707
Log likelihood = -408.42502                       Pseudo R2       =     0.0080

------------------------------------------------------------------------------
   treatment |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0006298   .0067915     0.09   0.926    -.0126814     .013941
   education |   .0476399   .0324732     1.47   0.142    -.0160065    .1112863
       black |  -.1498752   .1765545    -0.85   0.396    -.4959156    .1961653
    hispanic |   .0943199   .2484845     0.38   0.704    -.3927008    .5813405
     married |  -.0069582   .1480393    -0.05   0.963    -.2971098    .2831935
    nodegree |   .3281476    .161253     2.03   0.042     .0120976    .6441975
  earnings75 |  -9.52e-06   .0000111    -0.86   0.392    -.0000313    .0000123
       _cons |  -.5939974    .516367    -1.15   0.250    -1.606058    .4180634
------------------------------------------------------------------------------

. testparm *

           chi2(  7) =    6.56
         Prob > chi2 =    0.4756
```

Assessing Matching Quality

Iterate the following until balancing is achieved:

- Estimate propensity score
  - Specification, Probit/Logit, probability/index/odds ratio
- Match
  - Metric: $X, \hat{p}(X), \{X, \hat{p}(X)\}$
  - Type of matching
  - Smoothing parameters
- Common support
- Assessment of matching quality

Matching Estimate

```
. psmatch2  treatment age education black hispanic married
> nodegree earnings75, out(earnings78)

------------------------------------------------------------------------------
    Variable      Sample |    Treated     Controls   Difference          S.E.   T-stat
-------------------------+----------------------------------------------------
  earnings78  Unmatched | 5976.35202    13650.8035   -7674.4515   540.709664   -14.19
                    ATT | 5976.35202    4349.83269   1626.51934   677.626206     2.40
-------------------------+----------------------------------------------------
Note: S.E. does not take into account that the propensity score is estimated.

           | psmatch2:
 psmatch2: |   Common
 Treatment |  support
assignment | On suppor |     Total
-----------+-----------+----------
 Untreated |    15,992 |    15,992
   Treated |       297 |       297
-----------+-----------+----------
     Total |    16,289 |    16,289
```

Reducing imbalances

- Although matching should go a long way towards balancing the X between the treated and comparisons
- We can perform a final regression correction for remaining imbalances

$$\min_{\alpha,\beta,\delta} \sum_{i=1}^{N} w_i (Y_i - \alpha - \underbrace{\delta}_{\text{parameter of interest}} D_i - X_i'\beta)^2$$

- Where w_i's are the matching weights

Post-matching regression adjustment

```
. reg earnings78 treatment age education black hispanic married nodegree ///
    earnings75 [iw=_weight]

      Source |       SS       df       MS              Number of obs =     594
-------------+------------------------------           F(  8,   585) =   12.88
       Model |  3.4317e+09     8   428964551           Prob > F      =  0.0000
    Residual |  1.9480e+10   585   33299154.1           R-squared     =  0.1498
-------------+------------------------------           Adj R-squared =  0.1382
       Total |  2.2912e+10   593   38636967.3           Root MSE      =  5770.5

------------------------------------------------------------------------------
  earnings78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |   1686.398   476.1792     3.54   0.000     751.1692    2621.627
         age |   68.52453   31.14949     2.20   0.028     7.346074     129.703
   education |   447.4218   148.3188     3.02   0.003     156.1195     738.724
       black |  -640.7179   812.5813    -0.79   0.431     -2236.65    955.2141
    hispanic |  -115.6389    1135.11    -0.10   0.919    -2345.027    2113.749
     married |    1703.38   679.2145     2.51   0.012     369.3842    3037.376
    nodegree |   830.5569   741.3742     1.12   0.263    -625.5223    2286.636
  earnings75 |   .3565398   .0514142     6.93   0.000      .255561    .4575186
       _cons |  -3504.447   2366.024    -1.48   0.139    -8151.382    1142.488
------------------------------------------------------------------------------
```

Examples of studies

Let's look at some studies assuming selection on obs.

- LaLonde (1986) and the subsequent papers. You should know the arguments of this (heated) debate
- Fagereng et al. (2019). The authors try to use knowledge of treatment assignment to argue selection on observables
- Angrist (1998). Much cited paper for regression vs matching discussion. But why would selection on obs. hold here?

What are we looking for?

- What is the argument for selection on observables?
- How is it implemented?

Selection on observables and information set

Behavioral restrictions

- Recall that selection on observables imply that $(Y_1,\ Y_0)$ or Y_0 do not determine (or correlate with) D conditional on X
- Rules out selection into the program based on unobserved (by the analyst) outcomes (or factors correlate with the outcomes)

Key question:

- What is the information set of the agent making participation decision (taking or assigning treatment)?
- What is the information set of the analyst?
- Is there asymmetry in information? If so, who knows more?

Selection on obs. requires the analyst to know (or assume) the information the agent uses when deciding treatment participation

The Many Ways to Skin a Statistical Cat

What estimator should I choose?

- Many different ways to implement selection on observables. The literature is a bit overwhelming (at least to me)
- Unfortunate truth about statistics is that optimality results are hard
- Even so, some choices may be better than others
  Theory plus judgment and trial and error may take you a long way
  (If you don't want to get your hands dirty, you can always do theory ...)

Standard errors

- We haven't even touched on standard error estimates
- Many of the estimators are quite complicated with multiple steps
- Bootstrap can be invalid for the non-smooth ones (e.g. matching)
- Why? It's a bit complicated but (my) intuition is that matching is like an extreme order statistic – bootstrap doesn't work then

Discrete Nonparametric Estimation

Binning estimator

- Suppose $X \in \{x_1, ..., x_K\}$ is discrete with K small relative to N
- This covers cases where X has multiple discrete components e.g.

$X = \{(male, white), (female, white), (male, nonwhite), (female, nonwhite)\}$

- Then a nonparametric **binning estimator** is very natural: e.g.

$$\hat{\mu}_d(x) = \frac{1}{N_{d,x}} \sum_{i:D_i=d, X_i=x}^{N} Y_i \text{ where } N_{d,x} \equiv \sum_{i=1}^{N} \mathbb{1}[D_i = d, X_i = x]$$

Limitations

- Only works if X is completely discrete, otherwise $N_{d,x} = 0$ or 1!
- If X is continuous, could bin it into discrete chunks
  $\rightarrow$ This can be a bit arbitrary
- Poor finite sample performance if **small bins** (K large relative to N)

Nonparametric Smoothing Regression

Smoothing

- Classical approach to continuous X is to use $X \approx x$ to **smooth**
- Relies on μ_d being a sufficiently smooth function of x
- Old, extremely well-developed, and enormous literature
- Two main approaches considered in economics are very intuitive

Kernel regression

- Take a sample mean of Y over $D_i = d, X_i \in [x - h, x + h]$
- $h > 0$ is a **bandwidth** parameter that needs to be chosen
- Bias goes down as $h \to 0$, variance goes up

Series/sieve approximations

- Write $\mu_d(x) = \sum_{k=1}^{K} \theta_k b_k(x)$ for some **basis functions** b_k
- For example, regressing Y on $1, X, X^2, X^3, ..., X^K$
- Bias goes down as $K \to \infty$, variance goes up

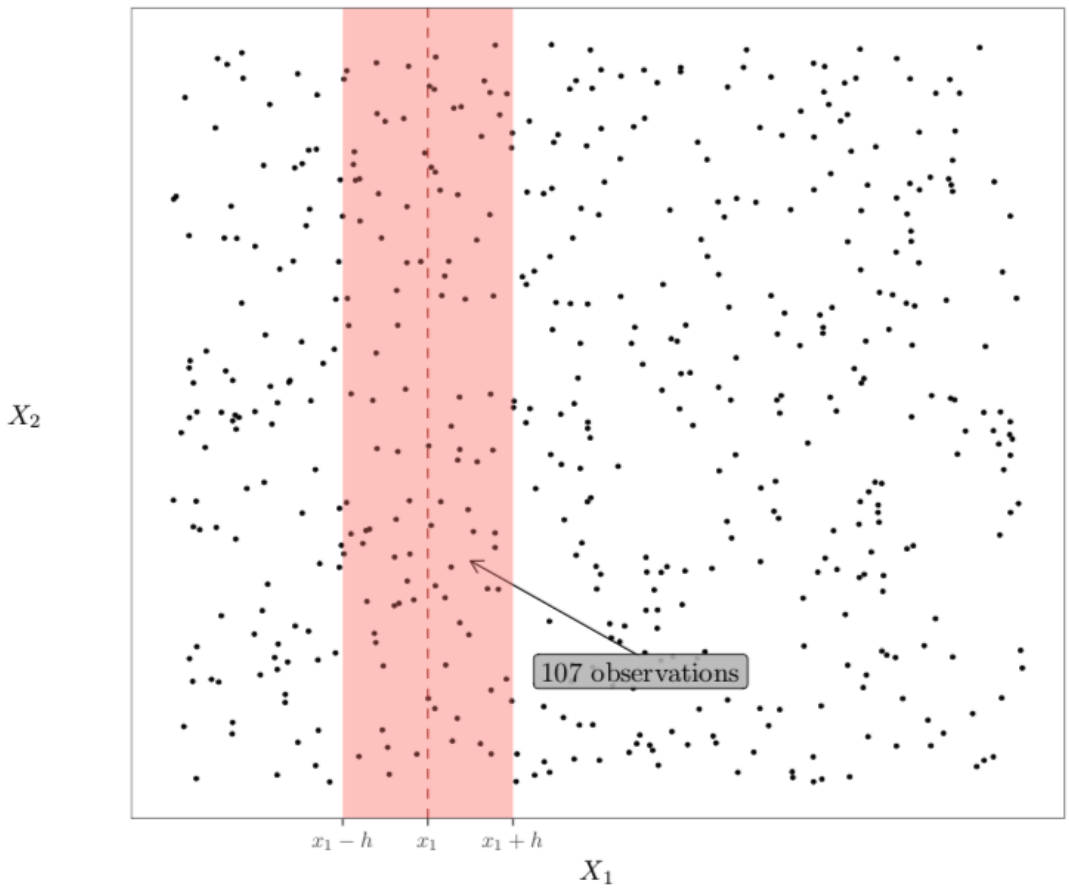
The Curse of Dimensionality

- Smoothing regression is held back by **the curse of dimensionality**
- Estimator quality rapidly deteriorates with the dimension of X
- Affects both kernel and sieve methods (indeed, all rate-optimal methods)

- Formally, the **rate of convergence** of the estimators goes down
- Implication is that each new dimension requires much more data
- Kernel with $X \in [0,1]$ uniformly has $2h \times N$ observations within $\pm h$ of x
- Kernel with $X \in [0,1]^2$ uniformly has only $4h^2 \times N$ observations
- Given h^* need much larger N to maintain the same "effective" N

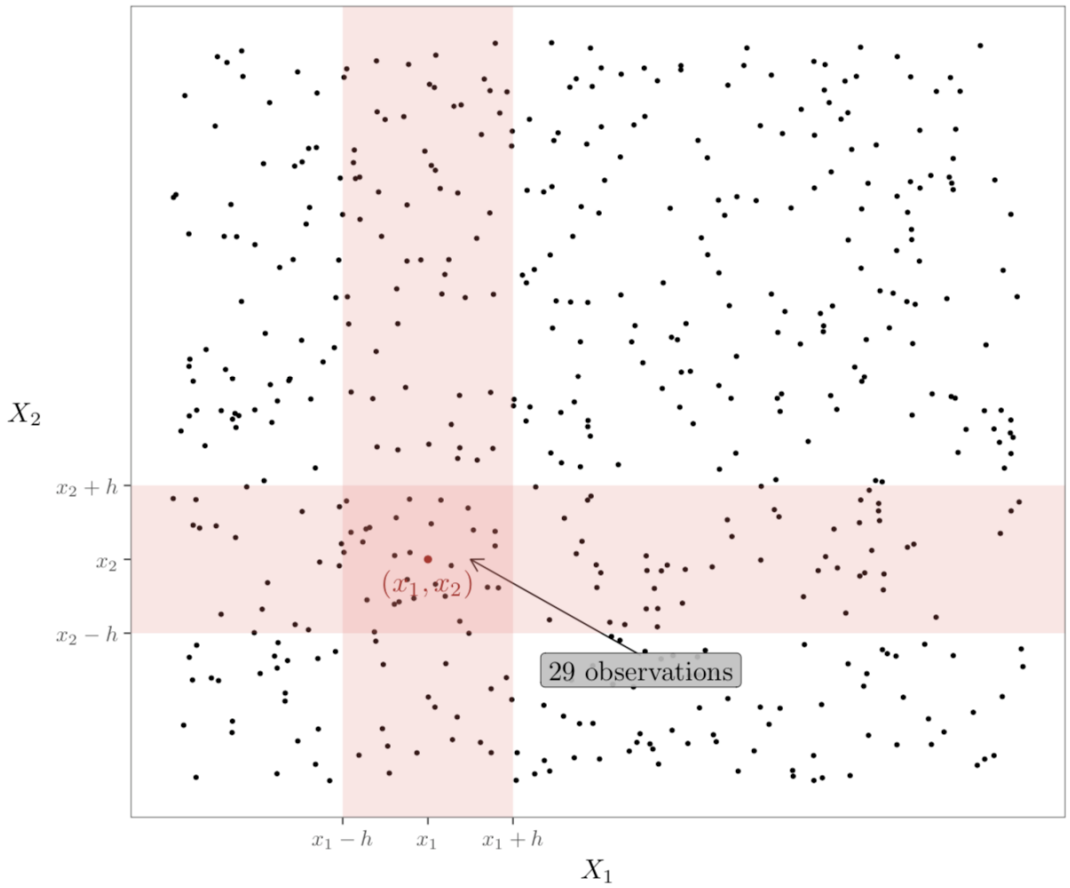
- Each $\hat{\mu}_d$ will have a slow rate of convergence, but $\widehat{ATE}$ can still be $\sqrt{N}$
  $\rightarrow$ Intuition is that by averaging we are again using all N observations
- This can be misleading "asymptopia" - practical value is questionable
- Finite sample performance of $\widehat{ATE}$ will still be poor in high dimensions

Curse of Dimensionality



- Kernel regression with 500 draws from a bivariate uniform $[0, 1]$
- Observations used to estimate $\mathbb{E}[Y|X = x_1]$ with bandwidth h

Curse of Dimensionality



- Observations used to estimate $\mathbb{E}[Y|X = x_1, X = x_2]$ with bandwidth h
- Effective number of observations drops by an order of magnitude

Nonparametric Smoothing with the Propensity Score

Smoothing with the propensity score

- Recall that selection on observables implies $(Y_0, Y_1) \perp\!\!\!\perp D | p(X)$
- So instead of estimating $\mu_d(x)$, we could estimate

$$\nu_d(p) \equiv \mathbb{E}[Y|D = d, P = p]$$

- Appears to break the curse of dimensionality ...

Estimating p

- An immediate practical problem here is p needs to be estimated by $\hat{p}$
- However now we see that dimension reduction was an illusion
  $\rightarrow$ Same curse of dimensionality in nonparametrically estimating p!
- Still, parametric p is arguably better than parametric μ_d
  $\rightarrow$ p is just a matter of fit, whereas μ_d is a counterfactual object
- In practice, usually see a logit (maybe probit) estimator of p

Subclassification (Blocking)

- One could use kernel or sieve methods for $\nu_d(p)$
- Instead, more common to use blocking on the propensity score
- Just a particular type of nonparametric smoothing regression

Blocking

- Divide $[0,1]$ into $\{b_0, b_1, \ldots, b_J\}$ with $b_0 = 0, b_J = 1$
- Define $B_j = 1$ if $p(X) \in (b_{j-1}, b_j)$ as membership in block j
- If $b_j - b_{j-1}$ is small then roughly random assignment within block
- Estimate $\widehat{ATE}_j = Y_{1,j} - Y_{0,j}$ per block, i.e. conditional on $B_j = 1$
- Then average $\widehat{ATE}_j$ by block size into $\widehat{ATE}$

- Key question is how to construct the blocks
- Imbens (2015) suggests an algorithm based on testing $D \perp\!\!\!\perp X | \{B_j\}_{j=1}^J$ $\rightarrow D \perp\!\!\!\perp X|p$ implied by selection on observables - so check within blocks

Blocking with Linear Regression

Combining two approaches

- Imbens (2015) suggests combining blocking with linear regression
- First construct the blocks
- Then within each block, run a linear regression Y on $1, D, X$
- Estimate regression-adjusted ATE for each block, then average up

Why?

- Intuitively, this could potentially reduce both bias and variance
- The variance part is clear if accounting for X reduces variation in Y
- The bias part is less clear (i.e. not necessarily true) ...
- Recall that linear regression extrapolates if $\bar{X}_1 \neq \bar{X}_0$
  $\rightarrow$ However within each block $\bar{X}_1 \approx \bar{X}_0$ - little extrapolation
- Adjusting for X reduces remaining differences within blocks
  $\rightarrow$ But presumably the remaining differences should be small anyway?

Job Training: Background

The National Supported Work (NSW) Demonstration

- Publicly and privately funded training program in the 1975-1977
- Provided 6 - 18 months work experience to disadvantaged individuals
- Applicants were randomly assigned to treatment or control
- Typically focus on post December 1975 $\Longrightarrow$ 1975 earnings predet.

The LaLonde (1986) critique

- Want to know impact of being offered training ($D = 1$) on applicants
- Applying/being offered job training is usually not randomly assigned
- In the NSW it is (for those who apply) $\Longrightarrow$ $\mathbb{E}[Y_1 - Y_0|D = 1]$ identified
- Suppose we didn't have this - then we need to infer $\mathbb{E}[Y_0|D = 1]$
- Could try this with (e.g.) PSID/CPS using a variety of methods
- How close do we get to the experimental benchmark?
- **Answer:** Not very close $\Longrightarrow$ extremely influential paper

$\rightarrow$ Findings were challenged (e.g. Heckman & Hotz, 1987)

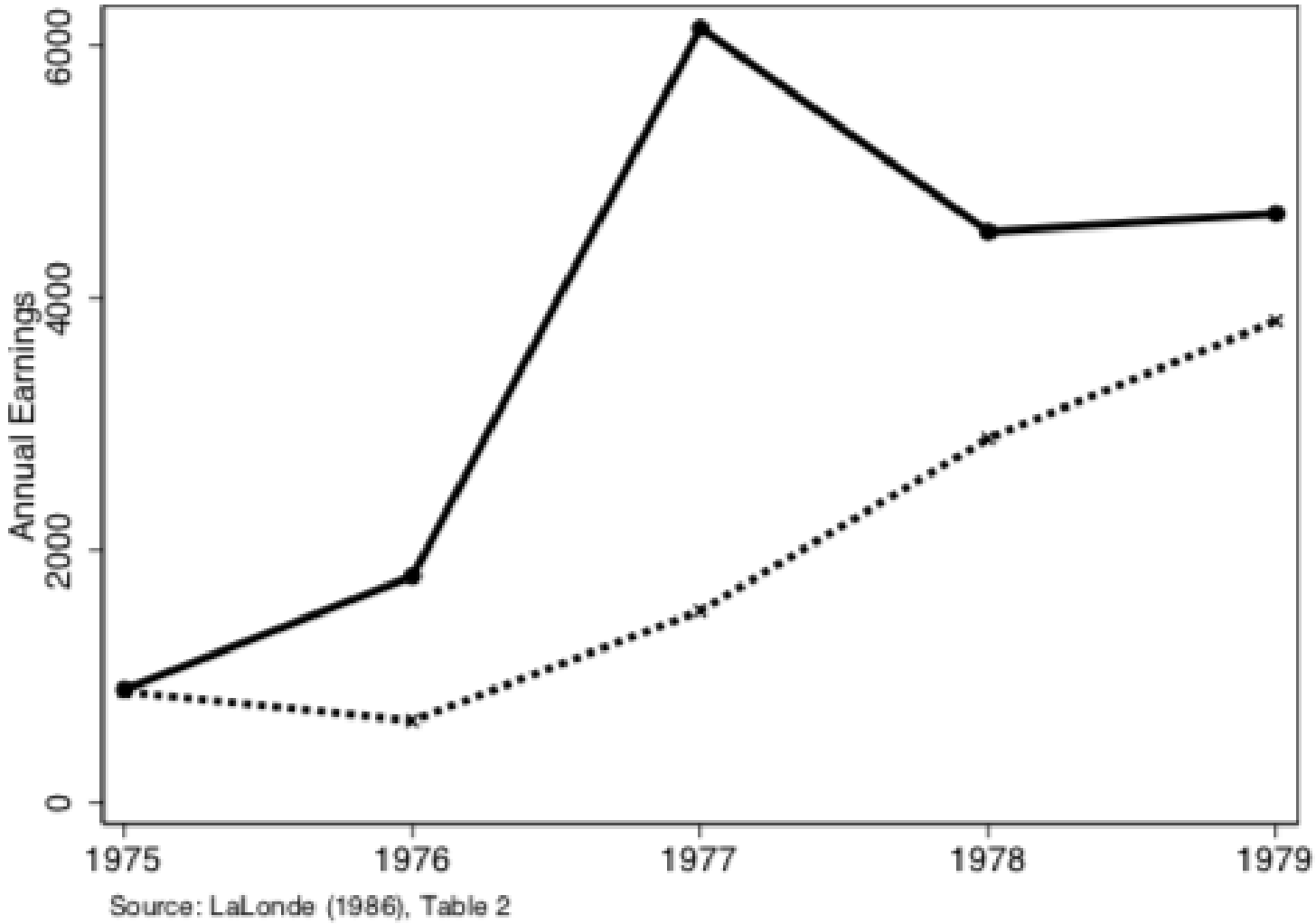
Balancing

TABLE 1—THE SAMPLE MEANS AND STANDARD DEVIATIONS OF PRE-TRAINING EARNINGS AND OTHER CHARACTERISTICS FOR THE NSW AFDC AND MALE PARTICIPANTS

	Full National Supported Work Sample			
	AFDC Participants		Male Participants	
Variable	Treatments	Controls	Treatments	Controls
Age	33.37	33.63	24.49	23.99
	(7.43)	(7.18)	(6.58)	(6.54)
Years of School	10.30	10.27	10.17	10.17
	(1.92)	(2.00)	(1.75)	(1.76)
Proportion	.70	.69	.79	.80
High School Dropouts	(.46)	(.46)	(.41)	(.40)
Proportion Married	.02	.04	.14	.13
	(.15)	(.20)	(.35)	(.35)
Proportion Black	.84	.82	.76	.75
	(.37)	(.39)	(.43)	(.43)
Proportion Hispanic	.12	.13	.12	.14
	(.32)	(.33)	(.33)	(.35)
Real Earnings	$393	$395	1472	1558
1 year Before	(1,203)	(1,149)	(2656)	(2961)
Training	[43]	[41]	[58]	[63]
Real Earnings	$854	$894	2860	3030
2 years Before	(2,087)	(2,240)	(4729)	(5293)
Training	[74]	[79]	[104]	[113]
Hours Worked	90	92	278	274
1 year Before	(251)	(253)	(466)	(458)
Training	[9]	[9]	[10]	[10]
Hours Worked	186	188	458	469
2 years Before	(434)	(450)	(654)	(689)
Training	[15]	[16]	[14]	[15]
Month of Assignment	−12.26	−12.30	−16.08	−15.91
(Jan. 78 = 0)	(4.30)	(4.23)	(5.97)	(5.89)
Number of Observations	800	802	2083	2193

Note: The numbers shown in parentheses are the standard deviations and those in the square brackets are the standard errors.

Annual Earnings - Females



Source: LaLonde (1986), Table 2

Choice of comparison group - Females

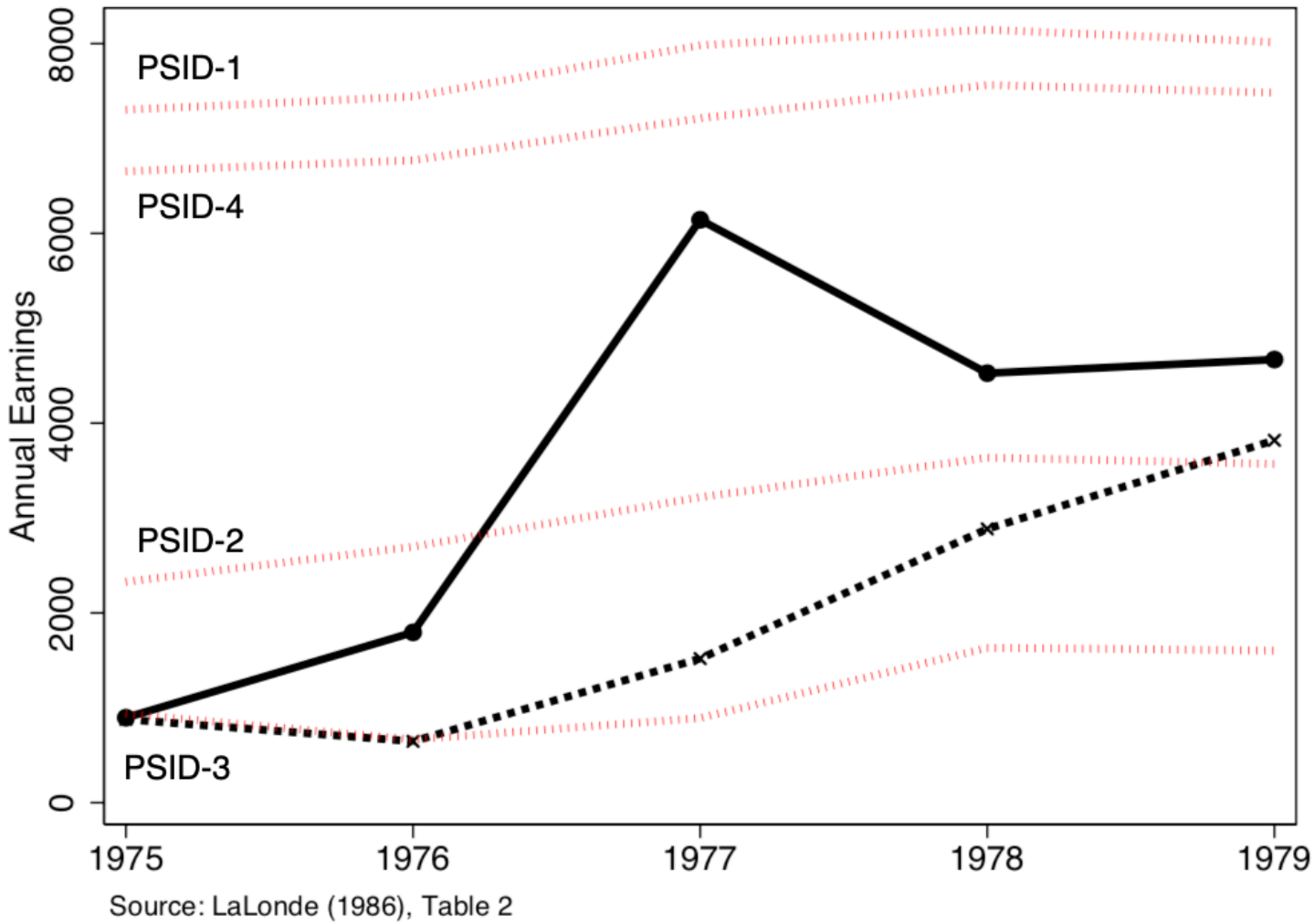
PSID

1. All female household heads aged 20-55
2. (1) + received AFDC in 1975
3. (2) + not working in 1976
4. (1) + no children below 5

CPS SSA

1. Matched sample
2. (1) + received AFDC in 1975
3. (1) + not working in 1976
4. (2) + not working in 1976

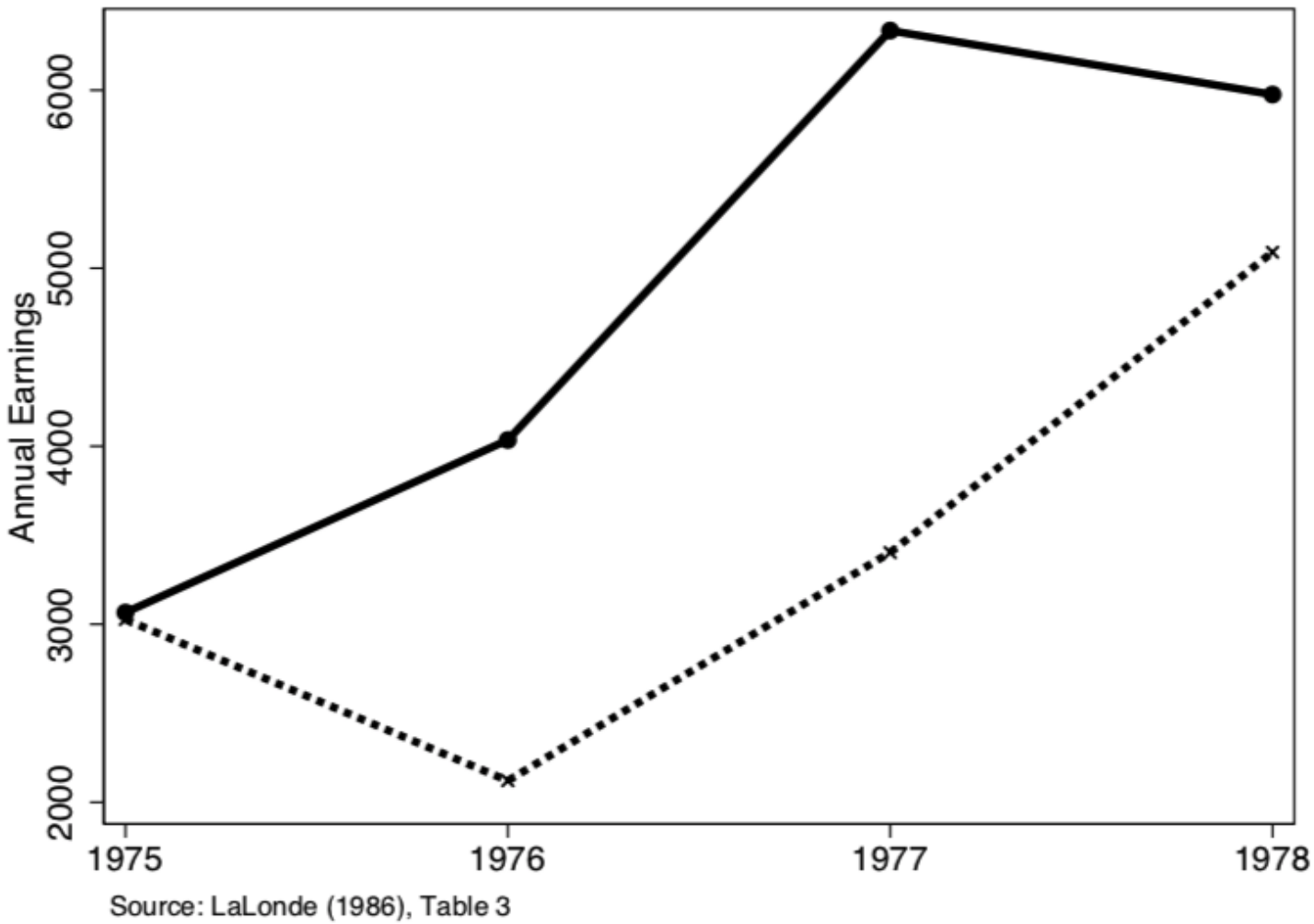
Annual Earnings - Females + PSID Comparisons



Treatment effects Females

Name of Comparison Group[d]	Comparison Group Earnings Growth 1975–79 (1)	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975 Unadjusted (2)	Pre-Training Year, 1975 Adjusted[c] (3)	Post-Training Year, 1979 Unadjusted (4)	Post-Training Year, 1979 Adjusted[c] (5)	Difference in Differences: Difference in Earnings Growth 1975–79 Treatments Less Comparisons Without Age (6)	With Age (7)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–79 Unadjusted (8)	Adjusted[c] (9)	Controlling for All Observed Variables and Pre-Training Earnings: Without AFDC (10)	With AFDC (11)
	2,942	−17	−22	851	861	833	883	843	864	854	–
Controls	(220)	(122)	(122)	(307)	(306)	(323)	(323)	(308)	(306)	(312)	
PSID-1	713	−6,443	−4,882	−3,357	−2,143	3,097	2,657	1746	1,354	1664	2,097
	(210)	(326)	(336)	(403)	(425)	(317)	(333)	(357)	(380)	(409)	(491)
PSID-2	1,242	−1,467	−1,515	1,090	870	2,568	2,392	1,764	1,535	1,826	–
	(314)	(216)	(224)	(468)	(484)	(473)	(481)	(472)	(487)	(537)	
PSID-3	665	−77	−100	3,057	2,915	3,145	3,020	3,070	2,930	2,919	–
	(351)	(202)	(208)	(532)	(543)	(557)	(563)	(531)	(543)	(592)	
PSID-4	928	−5,694	−4,976	−2,822	−2,268	2,883	2,655	1,184	950	1,406	2,146
	(311)	(306)	(323)	(460)	(491)	(417)	(434)	(483)	(503)	(542)	(652)
CPS-SSA-1	233	−6,928	−5,813	−3,363	−2,650	3,578	3,501	1,214	1,127	536	1,041
	(64)	(272)	(309)	(320)	(365)	(280)	(282)	(272)	(309)	(349)	(503)
CPS-SSA-2	1,595	−2,888	−2,332	−683	−240	2,215	2,068	447	620	665	–
	(360)	(204)	(256)	(428)	(536)	(438)	(446)	(468)	(554)	(651)	
CPS-SSA-3	1,207	−3,715	−3,150	−1,122	−812	2,603	2,615	814	784	−99	1,246
	(166)	(226)	(325)	(311)	(452)	(307)	(328)	(305)	(429)	(481)	(720)
CPS-SSA-4	1,684	−1,189	−780	926	756	2,126	1,833	1,222	952	827	–
	(524)	(249)	(283)	(630)	(716)	(654)	(663)	(637)	(717)	(814)	

Annual Earnings - Males



Source: LaLonde (1986), Table 3

Choice of comparison group - Males

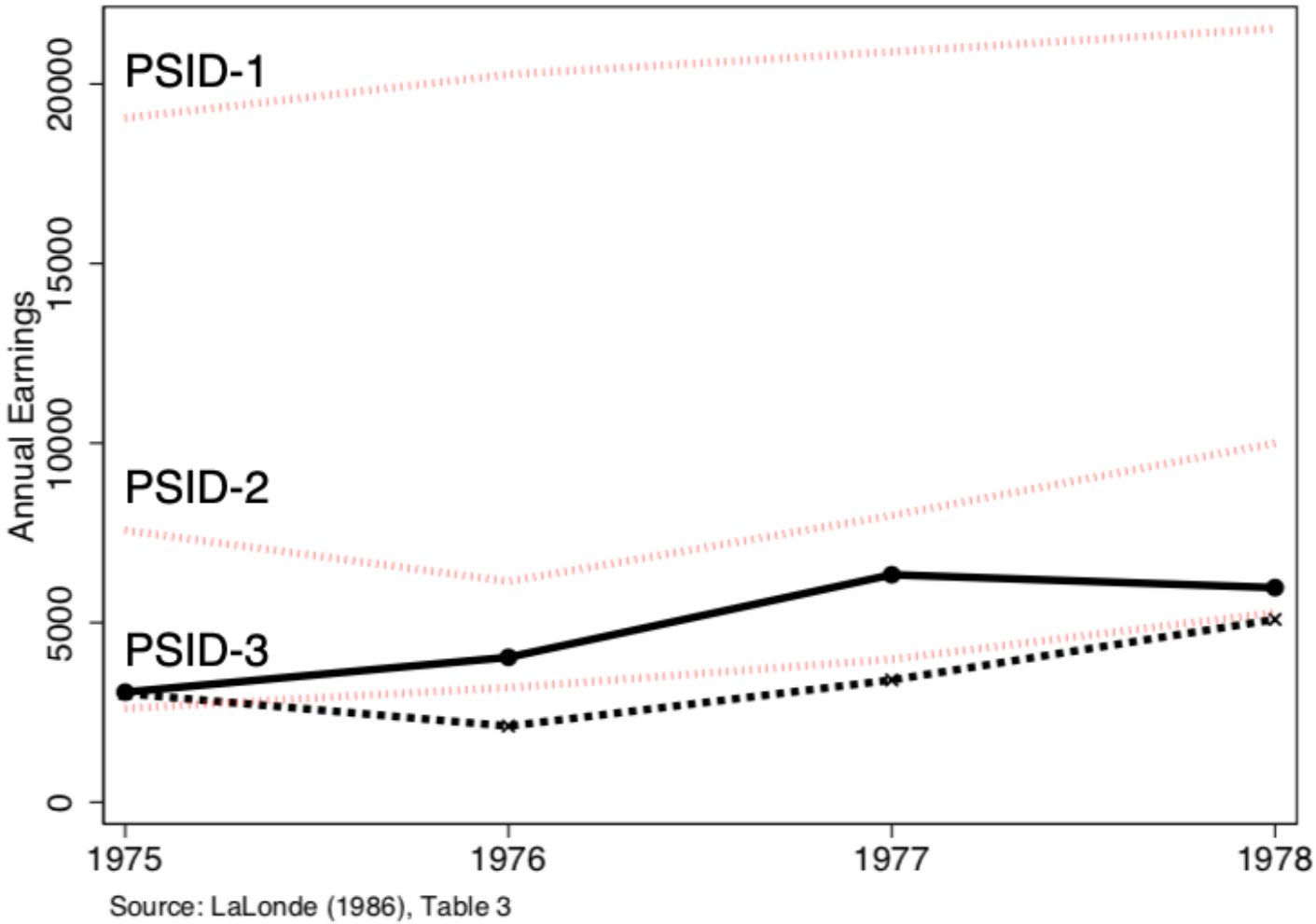
PSID

1. All male household heads not older than 55
2. (1) + not working in 1976
3. (2) + not working in 1975 or 1976

CPS SSA

1. Matched sample not older than 55
2. (1) + not working in 1976
3. (2) + not working in 1976 and income in 1975 below poverty level

Annual Earnings - Males + PSID Comparisons



Source: LaLonde (1986), Table 3

Treatment effects Males

Name of Comparison Group[d]	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975 Unadjusted (2)	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975 Adjusted[c] (3)	NSW Treatment Earnings Less Comparison Group Earnings: Post-Training Year, 1978 Unadjusted (4)	NSW Treatment Earnings Less Comparison Group Earnings: Post-Training Year, 1978 Adjusted[c] (5)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons: Without Age (6)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons: With Age (7)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78: Unadjusted (8)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78: Adjusted[c] (9)	Controlling for All Observed Variables and Pre-Training Earnings (10)
Controls	$2,063	$39	$−21	$886	$798	$847	$856	$897	$802	$662
	(325)	(383)	(378)	(476)	(472)	(560)	(558)	(467)	(467)	(506)
PSID-1	$2,043	−$15,997	−$7,624	−$15,578	−$8,067	$425	−$749	−$2,380	−$2,119	−$1,228
	(237)	(795)	(851)	(913)	(990)	(650)	(692)	(680)	(746)	(896)
PSID-2	$6,071	−$4,503	−$3,669	−$4,020	−$3,482	$484	−$650	−$1,364	−$1,694	−$792
	(637)	(608)	(757)	(781)	(935)	(738)	(850)	(729)	(878)	(1024)
PSID-3	($3,322	($455	$455	$697	−$509	$242	−$1,325	$629	−$552	$397
	(780)	(539)	(704)	(760)	(967)	(884)	(1078)	(757)	(967)	(1103)
CPS-SSA-1	$1,196	−$10,585	−$4,654	−$8,870	−$4,416	$1,714	$195	−$1,543	−$1,102	−$805
	(61)	(539)	(509)	(562)	(557)	(452)	(441)	(426)	(450)	(484)
CPS-SSA-2	$2,684	−$4,321	−$1,824	−$4,095	−$1,675	$226	−$488	−$1,850	−$782	−$319
	(229)	(450)	(535)	(537)	(672)	(539)	(530)	(497)	(621)	(761)
CPS-SSA-3	$4,548	$337	$878	−$1,300	$224	−$1,637	−$1,388	−$1,396	$17	$1,466
	(409)	(343)	(447)	(590)	(766)	(631)	(655)	(582)	(761)	(984)

Dehejia and Wahba (1999, "DW")

Background

- LaLonde (1986) used a variety of methods available at the time
- The huge array of fancy selection on observables methods came later
- DW repeat the exercise using some of these methods
- They find results that suggest selection on observables works well
  $\rightarrow$ "Works well" here means matches the experimental benchmark
- See also Dehejia and Wahba (2002), which is a bit clearer

Impact

- This has also been a hugely influential paper
- Often cited in support of propensity score methods
- Findings have again been controversial (more later)

Data

Table 1. Sample Means of Characteristics for NSW and Comparison Samples

	No. of observations	Age	Education	Black	Hispanic	No degree	Married	RE74 (U.S. $)	RE75 (U.S. $)
NSW/Lalonde:[a]									
Treated	297	24.63	10.38	.80	.09	.73	.17		3,066
		(.32)	(.09)	(.02)	(.01)	(.02)	(.02)		(236)
Control	425	24.45	10.19	.80	.11	.81	.16		3,026
		(.32)	(.08)	(.02)	(.02)	(.02)	(.02)		(252)
RE74 subset:[b]									
Treated	185	25.81	10.35	.84	.059	.71	.19	2,096	1,532
		(.35)	(.10)	(.02)	(.01)	(.02)	(.02)	(237)	(156)
Control	260	25.05	10.09	.83	.1	.83	.15	2,107	1,267
		(.34)	(.08)	(.02)	(.02)	(.02)	(.02)	(276)	(151)
Comparison groups:[c]									
PSID-1	2,490	34.85	12.11	.25	.032	.31	.87	19,429	19,063
		[.78]	[.23]	[.03]	[.01]	[.04]	[.03]	[991]	[1,002]
PSID-2	253	36.10	10.77	.39	.067	.49	.74	11,027	7,569
		[1.00]	[.27]	[.04]	[.02]	[.05]	[.04]	[853]	[695]
PSID-3	128	38.25	10.30	.45	.18	.51	.70	5,566	2,611
		[1.17]	[.29]	[.05]	[.03]	[.05]	[.05]	(686)	[499]
CPS-1	15,992	33.22	12.02	.07	.07	.29	.71	14,016	13,650
		[.81]	[.21]	[.02]	[.02]	[.03]	[.03]	[705]	[682]
CPS-2	2,369	28.25	11.24	.11	.08	.45	.46	8,728	7,397
		[.87]	[.19]	[.02]	[.02]	[.04]	[.04]	[667]	[600]
CPS-3	429	28.03	10.23	.21	.14	.60	.51	5,619	2,467
		[.87]	[.23]	[.03]	[.03]	[.04]	[.04]	[552]	[288]

NOTE: Standard errors are in parentheses. Standard error on difference in means with RE74 subset/treated is given in brackets. Age = age in years; Education = number of years of schooling; Black = 1 if black, 0 otherwise; Hispanic = 1 if Hispanic, 0 otherwise; No degree = 1 if no high school degree, 0 otherwise; Married = 1 if married, 0 otherwise; RE*x* = earnings in calendar year 19*x*.

[a] NSW sample as constructed by Lalonde (1986).

[b] The subset of the Lalonde sample for which RE74 is available.

[c] Definition of comparison groups (Lalonde 1986):

PSID-1: All male household heads under age 55 who did not classify themselves as retired in 1975.

PSID-2: Selects from PSID-1 all men who were not working when surveyed in the spring of 1976.

PSID-3: Selects from PSID-2 all men who were not working in 1975.

CPS-1: All CPS males under age 55.

CPS-2: Selects from CPS-1 all males who were not working when surveyed in March 1976.

CPS-3: Selects from CPS-2 all the unemployed males in 1976 whose income in 1975 was below the poverty level.

PSID1–3 and CPS-1 are identical to those used by Lalonde. CPS2–3 are similar to those used by Lalonde, but Lalonde's original subset could not be recreated.

- NSW treatment and control are quite similar (balance)
- NSW and observational samples (PSID/CPS) are quite different

Propensity Score Overlap

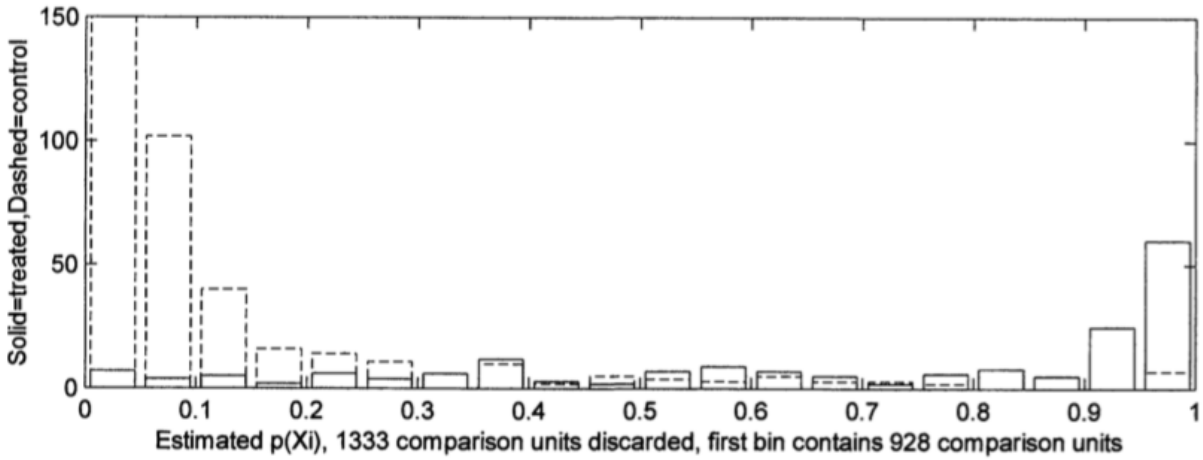


Figure 1. Histogram of the Estimated Propensity Score for NSW Treated Units and PSID Comparison Units. The 1,333 PSID units whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 928 PSID units. There is minimal overlap between the two groups. Three bins (.8–.85, .85–.9, and .9–.95) contain no comparison units. There are 97 treated units with an estimated propensity score greater than .8 and only 7 comparison units.

- Here this is a regression of treatment on X
- Assess both the overlap and fit of the propensity score
- Most comparison group units have very small propensity scores
- Treated units are more evenly spread out, but more close to 1

Results

Table 3. Estimated Training Effects for the NSW Male Participants Using Comparison Groups From PSID and CPS

	NSW earnings less comparison group earnings		NSW treatment earnings less comparison group earnings, conditional on the estimated propensity score					
			Quadratic in score[b]	Stratifying on the score			Matching on the score	
	(1) Unadjusted	(2) Adjusted[a]	(3)	(4) Unadjusted	(5) Adjusted	(6) Observations[c]	(7) Unadjusted	(8) Adjusted[d]
NSW	1,794	1,672						
	(633)	(638)						
PSID-1[e]	−15,205	731	294	1,608	1,494	1,255	1,691	1,473
	(1,154)	(886)	(1,389)	(1,571)	(1,581)		(2,209)	(809)
PSID-2[f]	−3,647	683	496	2,220	2,235	389	1,455	1,480
	(959)	(1,028)	(1,193)	(1,768)	(1,793)		(2,303)	(808)
PSID-3[f]	1,069	825	647	2,321	1,870	247	2,120	1,549
	(899)	(1,104)	(1,383)	(1,994)	(2,002)		(2,335)	(826)
CPS-1[g]	−8,498	972	1,117	1,713	1,774	4,117	1,582	1,616
	(712)	(550)	(747)	(1,115)	(1,152)		(1,069)	(751)
CPS-2[g]	−3,822	790	505	1,543	1,622	1,493	1,788	1,563
	(670)	(658)	(847)	(1,461)	(1,346)		(1,205)	(753)
CPS-3[g]	−635	1,326	556	1,252	2,219	514	587	662
	(657)	(798)	(951)	(1,617)	(2,082)		(1,496)	(776)

[a] Least squares regression: RE78 on a constant, a treatment indicator, age, age^2, education, no degree, black, Hispanic, RE74, RE75.
[b] Least squares regression of RE78 on a quadratic on the estimated propensity score and a treatment indicator, for observations used under stratification; see note (g).
[c] Number of observations refers to the actual number of comparison and treatment units used for (3)–(5); namely, all treatment units and those comparison units whose estimated propensity score is greater than the minimum, and less than the maximum, estimated propensity score for the treatment group.
[d] Weighted least squares: treatment observations weighted as 1, and control observations weighted by the number of times they are matched to a treatment observation [same covariates as (a)]. Propensity scores are estimated using the logistic model, with specifications as follows:
[e] PSID-1: Prob (T_i = 1) = F(age, age^2, education, education^2, married, no degree, black, Hispanic, RE74, RE75, RE74^2, RE75^2, u74*black).
[f] PSID-2 and PSID-3: Prob (T_i = 1) = F(age, age^2, education, education^2, no degree, married, black, Hispanic, RE74, RE74^2, RE75, RE75^2, u74, u75).
[g] CPS-1, CPS-2, and CPS-3: Prob (T_i = 1) = F(age, age^2, education, education^2, no degree, married, black, Hispanic, RE74, RE75, u74, u75, education*RE74, age^3).

- Focus on the first two rows - experimental ATT is about 1800
- They prefer (4)&(5)(blocking), and (7)&(8) (matching)
- Simple regression in column (2) does less well

How much should we trust DW's findings?

Heckman et al. studies

- Dehejia and Wahba's (1999, 2002) finding of low bias from applying propensity score matching to LaLonde's (1986) data is arguably surprising
- Contrasts with results from Heckman et al. (1997a) and Heckman et al. (1996, 1998a)) using the experimental data from the JTPA Study
- They conclude that in order for matching estimators to have low bias, it is important that the data include
  - Variables affecting program part. and labor market outcomes
  - Nonexperimental comparison group be drawn from the same local labor markets as the participants
  - Comparable data for participants and nonparticipants (e.g earnings)

All these conditions fail to hold in the NSW data analyzed by LaLonde (1986) and Dehejia and Wahba (1999, 2002)

Critique and Lessons

Specific critique

- Smith and Todd (2005) show that DW results are not robust
- They argue changing sample from LaLonde (1986) is important
  $\rightarrow$ Sample selection choices as well as specification
- Also point out that income is measured differently in NSW/CPS
  $\rightarrow$Self-reported vs. matched SS earnings data in March CPS
- Indeed, this is a critique of the whole LaLonde (1986) exercise
  $\rightarrow$These points and others: Heckman, Ichimura and Todd (1997)

My reading of the literature

- The preferred nonexperimental evaluation strategy in a given context depends critically on the available data and on the institutions governing selection into the program
- Understanding the sources of bias may help a lot: i) lack of support, ii) quality of match, iii) selection on unobservables

Heckman et al: Understanding the biases

TABLE 2

Decomposition of difference in post-programme mean earnings
Bootstrapped standard errors shown in parentheses‡
Percentage of mean difference attributable to components in square brackets
Earnings measured in average monthly dollars

	Mean difference $\hat{B}$	Non-overlap* $\hat{B}_1$	Density weighting $\hat{B}_2$	Selection bias $\hat{B}_3$	Average bias $(\hat{\bar{B}}_{S_P})$	Selection bias** $(\hat{\bar{B}}_{S_P})$ as a % of treatment impact
Experimental Controls and eligible nonparticipants (ENPs)†						
Adult males	−342	218	−584	23	38	87%
(std. err.) [%]	(47)	(38) [−64%]	(41) [170%]	(33) [−7%]	(63)	
Adult females	33	80	−78	31	38	129%
(std. err.) [%]	(26)	(13) [242%]	(18) [−235%]	(26) [94%]	(33)	
Male youth	20	142	−131	9	14	23%
(std. err.) [%]	(57)	(28) [704%]	(35) [−650%]	(42) [46%]	(64)	
Female youth	42	74	−67	35	49	7239%
(std. err.) [%]	(36)	(17) [177%]	(26) [−161%]	(28) [84%]	(42)	

Heckman et al: Understanding the biases

Experimental Controls and SIPP Eligibles††						
	Mean difference $\hat{B}$	Non-overlap* $\hat{B}_1$	Density weighting $\hat{B}_2$	Selection bias $\hat{B}_3$	Average bias $(\hat{\hat{B}}_{S_P})$	Selection bias** $(\hat{\hat{B}}_{S_P})$ as a % of treatment impact
Adult males	−145	151	−417	121	192	440%
(std. err.) [%]	(56)	(30) [−104%]	(44) [287%]	(33) [−83%]	(57)	
Adult females	47	97	−172	122	198	676%
(std. err.) [%]	(23)	(19) [206%]	(16) [−367%]	(15) [260%]	(26)	
Male youth	−188	65	−263	9	21	36%
(std. err.) [%]	(106)	(108) [−35%]	(53) (139%]	(25) [−5%]	(90)	
Female youth	−88	83	−168	−3	−13	1969%
(std. err.) [%]	(38)	(22) [−94%]	(27) [191%]	(10) [3%]	(58)	

Heckman et al: Understanding the biases

‡ They are based on 50 replications of the data with 100% sampling.
† The best predictor Control-ENP participation models for all the demographic groups include indicator variables for site, age, race, education, marital status, children less than 6 and labour force transitions. In addition to these variables, the adult male model also includes an indicator for vocational training history, the number of household members, earnings in the month of random assignment or eligibility determination (RA or EL) and number of jobs held in 18 months before RA or EL. The adult female model includes an indicator for recent schooling, earnings in the month of RA or EL and number of labour force transitions in the 24 months prior to RA or EL. The male youth model includes average earnings in the 6 months and 12 months prior to RA or EL and average positive earnings in the 6 months before RA or EL. The female youth model includes earnings in the 12 months before RA or EL.
†† The best predictor Control-SIPP model includes indicators for age, race, education, marital status, children age less than 6, labour force transition patterns and levels of earnings in the preceding year. The data used are SIPP eligibles and experimental JTPA controls.
* A 2% trimming rule was used for adult males and females and a 5% trimming rule for youth was used in determining the overlapping support region (see Appendix C for a description of how the support is determined). The proportion of Controls and ENPs falling in the overlap region (S_P) are: 60% and 96% of adult males, 82% and 96% of adult females, 67% and 92% of male youth and 71% and 93% of female youth. The proportion of SIPPs and Controls falling in the overlap region are: 63% and 96% of adult males, 61% and 96% of adult females, 41% and 90% of male youth and 20% and 89% of female youth. A 0·06 fixed bandwidth and a biweight kernel, defined in Appendix A, were used for the nonparametric estimates.
** The final column displays the ratio of the absolute value of $(\hat{\bar{B}}_{S_P})$ to the absolute value of experimental impact estimate.

Fagereng et al. (2019)

Background and motivation

- Many studies document strong intergenerational associations in wealth
- Natural question: Why do wealthy parents have wealthy children?

Several possible explanations

- Selection: Parents genetically pass on abilities and preferences, creating intergen. links in income, savings behavior or financial risk taking
- Causation: Children's accumulation of wealth depends on the actions of their parents
- Several possible causal channels, including direct wealth transfers, parental investment, learning of attitudes and traits
- Distinguishing between selection and causality is key to understand how economic conditions or government policies may shape the persistence of wealth inequality across generations

Research question and design

What they do

- Fagereng et al. investigate the role of family background in determining children's wealth accumulation and investor behavior as adults
- Analysis is made possible by linking Korean-born children who were adopted at infancy by Norwegian parents
- Mechanism by which these Korean-Norwegian adoptees were assigned to adoptive families is known and effectively random

How they argue selection on observables

- First step: Submission of an application to review by case examiners
  Parents were not given the opportunity to specify gender, family background or anything else about their future adoptee
- Next step: Approved files sent to the adoption agency, Holt Korea
  Adoptees assigned to the approved adoptive families in the order the applications arrived
- Thus, as good as random assignment conditional on time of adoption

Testing for Quasi-Random Assignment

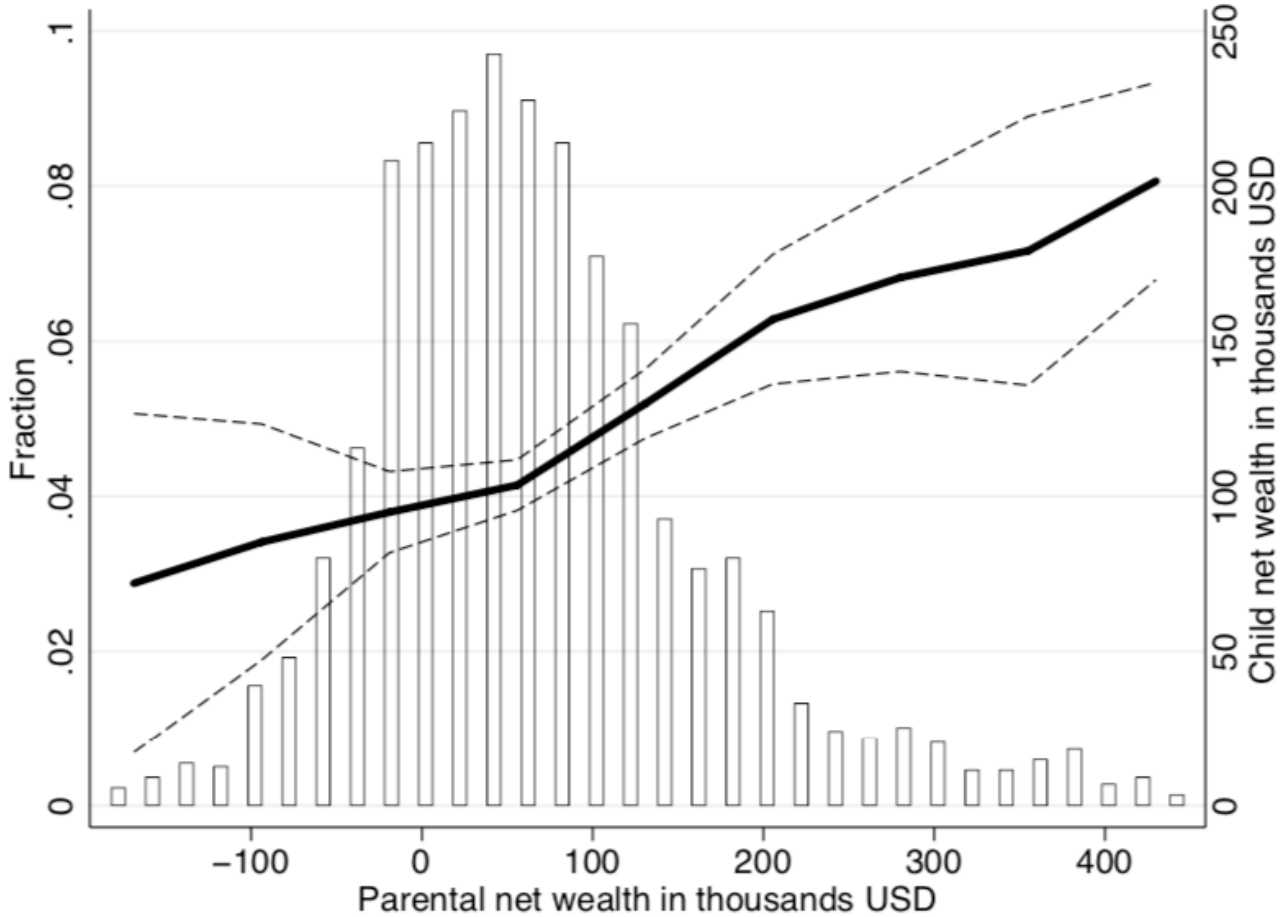
Table 2. Testing for quasi-random assignment of Korean-Norwegian adoptees

	Dependent variable:			
	Age at adoption		Gender	
	Specification:			
Regressors	Bivar. reg.	Multivar. reg.	Bivar. reg.	Multivar. reg.
Parent net wealth	-0.002	-0.002	0.005	0.004
	(0.003)	(0.0037)	(0.004)	(0.004)
Mother's years	0.002	0.003	0.002	0.001
of schooling	(0.002)	(0.003)	(0.003)	(0.004)
Father's years	0.001	-0.000	0.002	-0.000
of schooling	(0.002)	(0.002)	(0.003)	(0.004)
(Log) parent income	0.001	0.007	0.059	0.037
at birth	(0.035)	(0.038)	(0.0488)	(0.054)
Median (log) income	-0.046	-0.047	0.051	0.036
in childhood municipality	(0.034)	(0.035)	(0.0459)	(0.047)
Dependent mean	0.78	0.78	0.75	0.75
F-stat, joint significance of regressors		0.737		0.648
[p-value]		[0.596]		[0.663]

Notes: The table contains estimates from regressions of a pre-determined characteristic of the adoptee (age at adoption or indicator for female) on family background variables such as parental net wealth, education (in years) of the mother and father, the log of parents income and the log the median income in parents' municipality of residence, all measured at the time of birth of the child. In columns 1 and 3, we run separate regressions for each of the family background variables (conditional on a full set of indicators for adoption years of the children). In columns 2 and 4, we run multivariate regressions with all the family characteristics (conditional on a full set of indicators for adoption years of the children). The estimation sample consists of 2,254 Korean-Norwegian adoptees.

Intergenerational Links in Wealth

Figure 4. Association between adoptee's net wealth and adoptive parents' net wealth



Notes: This figure is based on the baseline sample consisting of 2,254 Korean adoptees and their parents. The histogram shows the density of parental wealth (the left y-axis). The solid line shows estimates from a local linear regression of net wealth of the adoptee as an adult (measured as an average of 2012-2014) on the net wealth of her adoptive parents (measured as an average of 1994-1996), conditional on full set of indicators for year of adoption

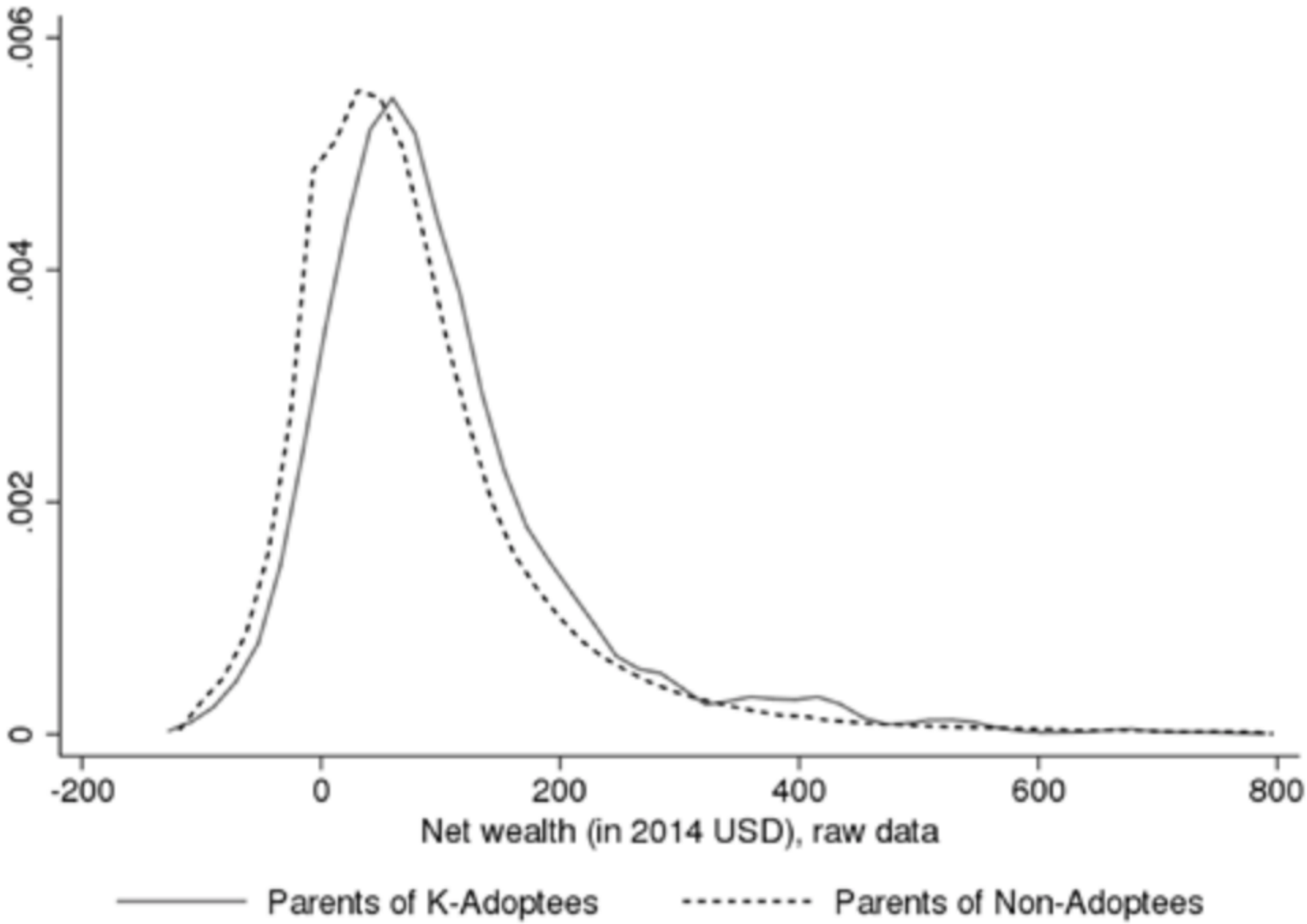
Intergenerational Links in Wealth

Table 3. Intergenerational links in wealth

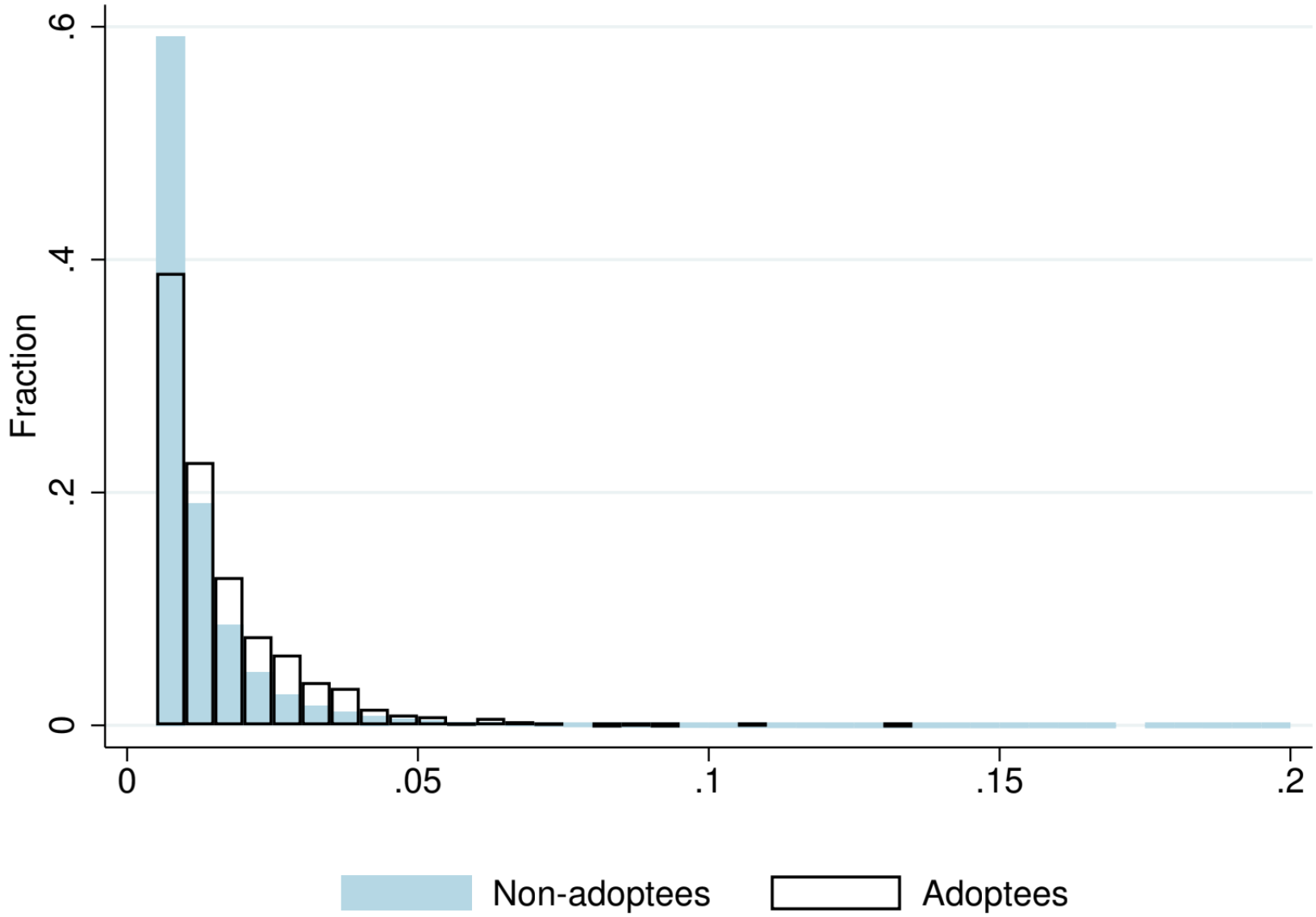
	Korean-Norwegian adoptees			Non-adoptees		
	(1)	(2)	(3)	(4)	(5)	(6)
Child-parent net wealth relation	0.225***	0.225***	0.204***	0.575***	0.547***	0.546***
	(0.041)	(0.041)	(0.042)	(0.011)	(0.011)	(0.026)
Adoption year indicators	Yes	Yes	Yes			
Birth year ind. of child & parents	Yes	Yes	Yes	Yes	Yes	Yes
Gender		Yes	Yes	Yes	Yes	Yes
Adoption age (in days)		Yes	Yes			
Family characteristics			Yes		Yes	Yes
Matched sample (prop. score)						Yes
Observations	2,254			1,206,650		1,174,330

Notes: The Korean-Norwegian adoptees are born in South Korea between 1965 and 1986, and adopted at infancy by Norwegian parents. Family characteristics include education (in years) of the mother and father, the number of siblings, the (log of) parents income and the (log of) the median income in parents' municipality of residence,

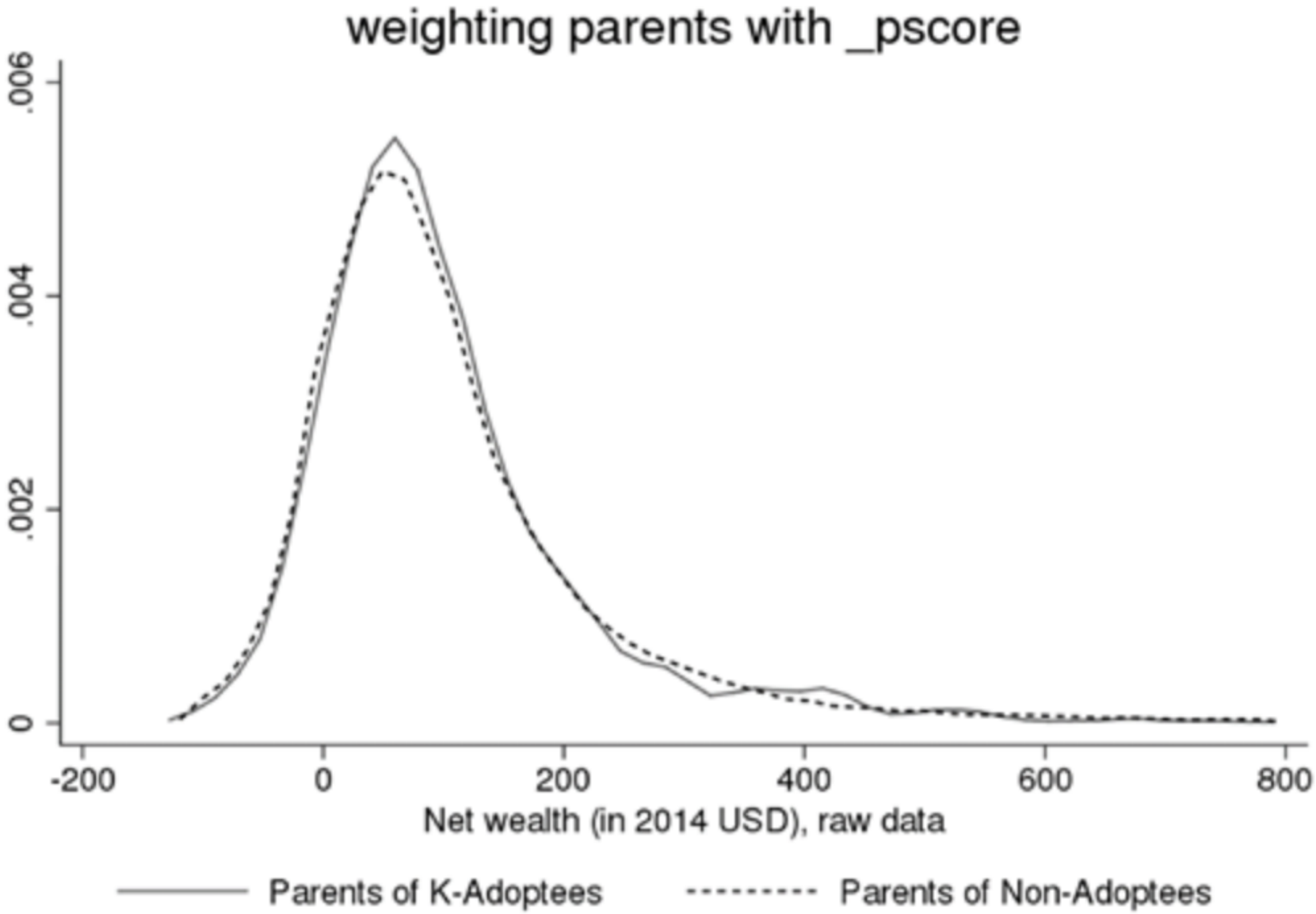
Distribution of Parental Wealth



Propensity Score



Weighted Distribution of Parental Wealth



Intergenerational Links in Wealth

Table 3. Intergenerational links in wealth

	Korean-Norwegian adoptees			Non-adoptees		
	(1)	(2)	(3)	(4)	(5)	(6)
Child-parent net wealth relation	0.225***	0.225***	0.204***	0.575***	0.547***	0.546***
	(0.041)	(0.041)	(0.042)	(0.011)	(0.011)	(0.026)
Adoption year indicators	Yes	Yes	Yes			
Birth year ind. of child & parents	Yes	Yes	Yes	Yes	Yes	Yes
Gender		Yes	Yes	Yes	Yes	Yes
Adoption age (in days)		Yes	Yes			
Family characteristics			Yes		Yes	Yes
Matched sample (prop. score)						Yes
Observations	2,254			1,206,650		1,174,330

Notes: The Korean-Norwegian adoptees are born in South Korea between 1965 and 1986, and adopted at infancy by Norwegian parents. Family characteristics include education (in years) of the mother and father, the number of siblings, the (log of) parents income and the (log of) the median income in parents' municipality of residence,

Criticisms of Assuming Selection on Observables

Inherent unobservables

- Selection on observables can often be difficult to argue
  $\rightarrow$ **Inherent unobservables:** preferences, private info, expectations, ...
- Boils down to arguing i) observationally identical people behave differently due to A, and ii) A is like a coin flip because of B

Controlling for more isn't an (convincing) answer

- Often argued that large X makes selection on observables more likely
  $\rightarrow$ Strictly speaking, this is not necessarily true-we saw this earlier
- Even if it were, still raises an uncomfortable friction with overlap
  $\rightarrow$ If we could perfectly explain D with X then $\mathbb{P}[D = 1|X] = 0$ or 1

Fancy methods for choosing observables will not solve this

- Selection on observables is seeing a resurgence with "machine learning"
- Fancier methods, but the identifying assumption is still the same
- Estimation and bias/variance trade-off is rarely the first-order issue here

Angrist (1998)

A bad use of selection on observables

- Well-known economic application of selection on observables:
  - Y is a labor market outcome (employment/earnings)
  - $D \in \{0, 1\}$ is veteran status (participation in the military)
  - X are socioeconomic variables (race, year, schooling, AFQT, age)
- Assumption is given X, military participation as-if randomly assigned
- **Observationally similar people randomly join the military?!?**
- Ignores first-order issues such as outside employment options
  $\rightarrow$ Also unobservable screening factors (fitness, interpersonal skills)
- These are unobserved and inherently unobservable

Allowing for selection on unobservables

- Most applied microeconomists seem to share this skepticism
- This motivates the other methods that we will discuss in the course
- All use different arguments to allow for **selection on unobservables**

Parametric identification: A cautionary tale

How to deal with selection on unobservables

- One possibility is to find an instrument (next week!)
- Another possibility is to make functional form / parametric assumptions about the relationship between the observed data and the unobservables
- In practice, it is often necessary to do both. Not necessarily anything wrong with that.
- But be wary; with strong enough assumptions you can draw inferences about anything (but learn nothing)

Heckman (2008); first draft 1980

- The Effect of Prayer on God's Attitude Toward Mankind
- Data on intensity of prayer from the National Opinion Research Center's (NORC) survey on religious attitudes
- God's attitude is unobserved to the analyst but why would that stop him?

Parametric assumption

Data and assumption

- $Y \in [0, 1]$ - God's attitude, this is an unobserved variable.
- $X \in [0, 1]$ - the intensity of prayer in the population.
- The population density of prayer is summarized by a univariate density $f(X)$ which has been estimated by Father Greeley (1972).
- Accept on faith that the conditional density of X given Y is of the form

$$g(XY) = a(Y)\exp(XY) \tag{1}$$

where a(Y) is an unknown, continuous, positive, and differentiable function.

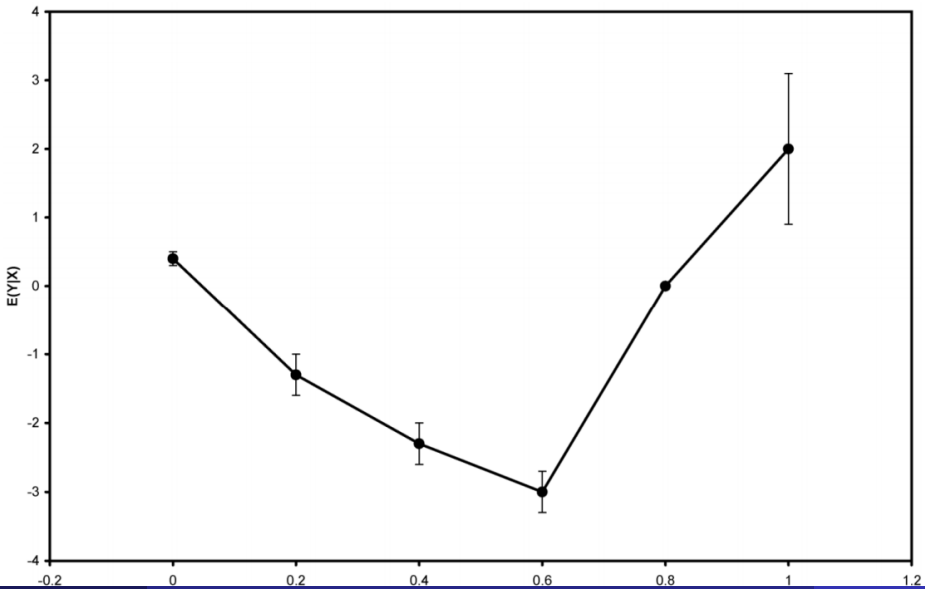
Estimating equation and results

Heckman (2008)

- Singh (1977) demonstrates that:

$$\mathbb{E}[Y|X = x] = \frac{f'(x)}{f(x)}$$

- From the population distribution of prayer we can estimate the population regression function of God's attitude as a function of prayer.



What's the lesson?

- If one is willing to make a parametric assumption, it not necesssary to observe a variable in order to compute its conditional expectation with respect to another variable

- Admittedly an extreme example, but it makes an important point:
  – Think carefully about (the implications of) the assumptions you make
  – Some important questions might not be possible to rigorously study or credibly answer
  – How would you define, identify and estimate parameters answering questions such as: What's the effect of inequality on democracy? Why are some countries rich? What's the meaning of life?