EPE - Lecture 1
Observational Methods of Causal Inference

Sylvain Chabé-Ferret

Toulouse School of Economics, Inra

January 2017

In a nutshell

In this lecture, we are going to study how to estimate the effect of an intervention on an outcome using observational methods, that is methods that try to correct for selection bias by accounting for as many observed confounders as possible.

Outline

Key Assumption: Selection on Observables

Assumption (Selection on Observables)

We assume that there exists a known set of observable covariates X_i such that:

$$\mathbb{E}[Y_i^0 | X_i, D_i = 1] = \mathbb{E}[Y_i^0 | X_i, D_i = 0].$$

Interpretation

Once we account for correlation of X_i with Y_i, there is no selection bias anymore:

$$\begin{aligned}
\Delta^{Y}_{WW}(X_i) &= \mathbb{E}[Y_i|X_i, D_i = 1] - \mathbb{E}[Y_i|X_i, D_i = 0] \\
&= \mathbb{E}[Y_i^1|X_i, D_i = 1] - \mathbb{E}[Y_i^0|X_i, D_i = 0] \\
&= \mathbb{E}[Y_i^1|X_i, D_i = 1] - \mathbb{E}[Y_i^0|X_i, D_i = 1] \\
&= \mathbb{E}[Y_i^1 - Y_i^0|X_i, D_i = 1] \\
&= \Delta^{Y}_{TT}(X_i)
\end{aligned}$$

where the third equality uses Selection on Observables.

Interpretation (cont.)

X_i is assumed to capture all of the relevant confounders. In general, we think that Selection on Observables holds only approximately, so that all remaining influences are "negligible".

$$\mathbb{E}[Y_i^0|X_i, D_i = 1] \approx \mathbb{E}[Y_i^0|X_i, D_i = 0].$$

Whether they truely are negligible is a leap of faith, that has to be verified ex-post comparing methods leveraging on this assumption with more robust methods (e.g. RCTs). The results of such comparisons I call LaLonde Tests.

Implementation

There are several ways to leverage on Selection on Observables to build estimators for TT:

1. Parametric: OLS
2. Nonparametric: Matching
3. Semiparametric: Synthetic Control (SCM)

Outline

Assumption: Parametric Functional Form

Assumption (Parametric functional form)

We assume that there exists a known set of observable covariates X_i such that:

$$\mathbb{E}[Y_i^0 | X_i] = \alpha_0 + \beta_0' X_i.$$

How Can We Estimate TT Using SO and PFF?

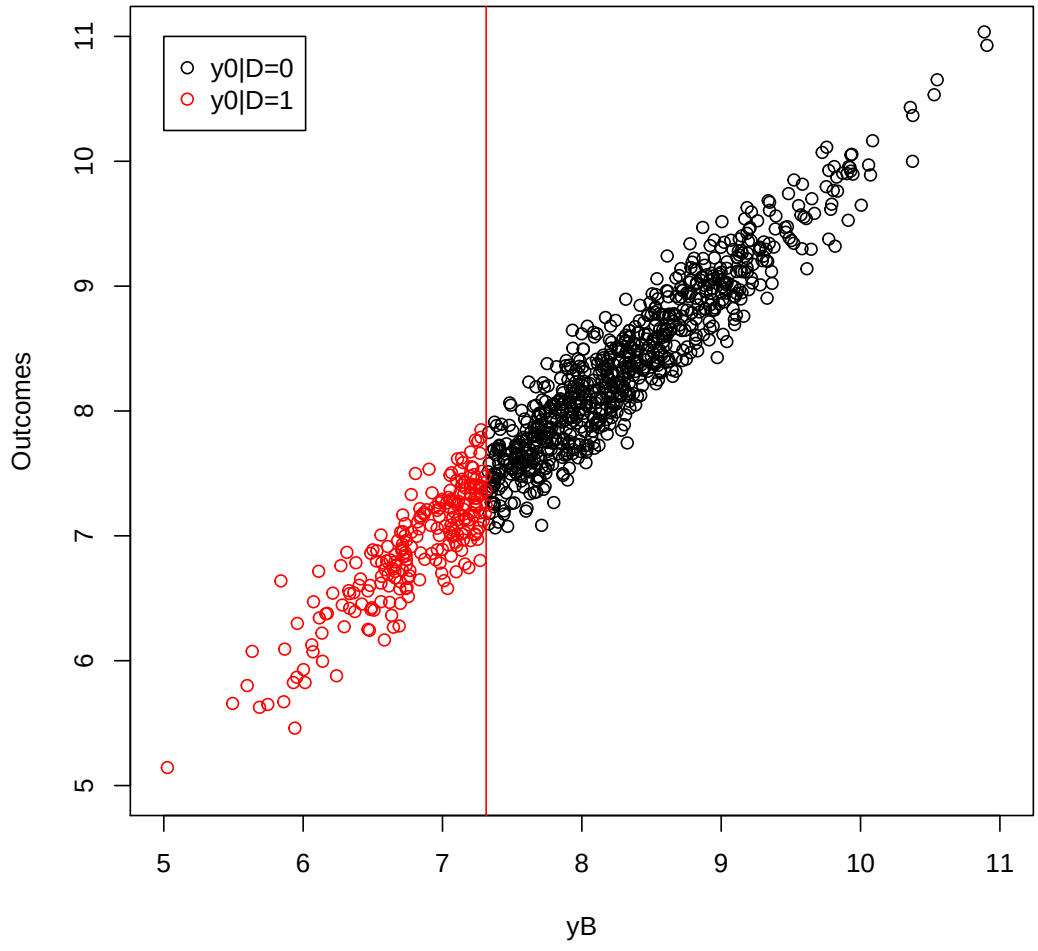


Figure: Step 0

How Can We Estimate TT Using SO and PFF?

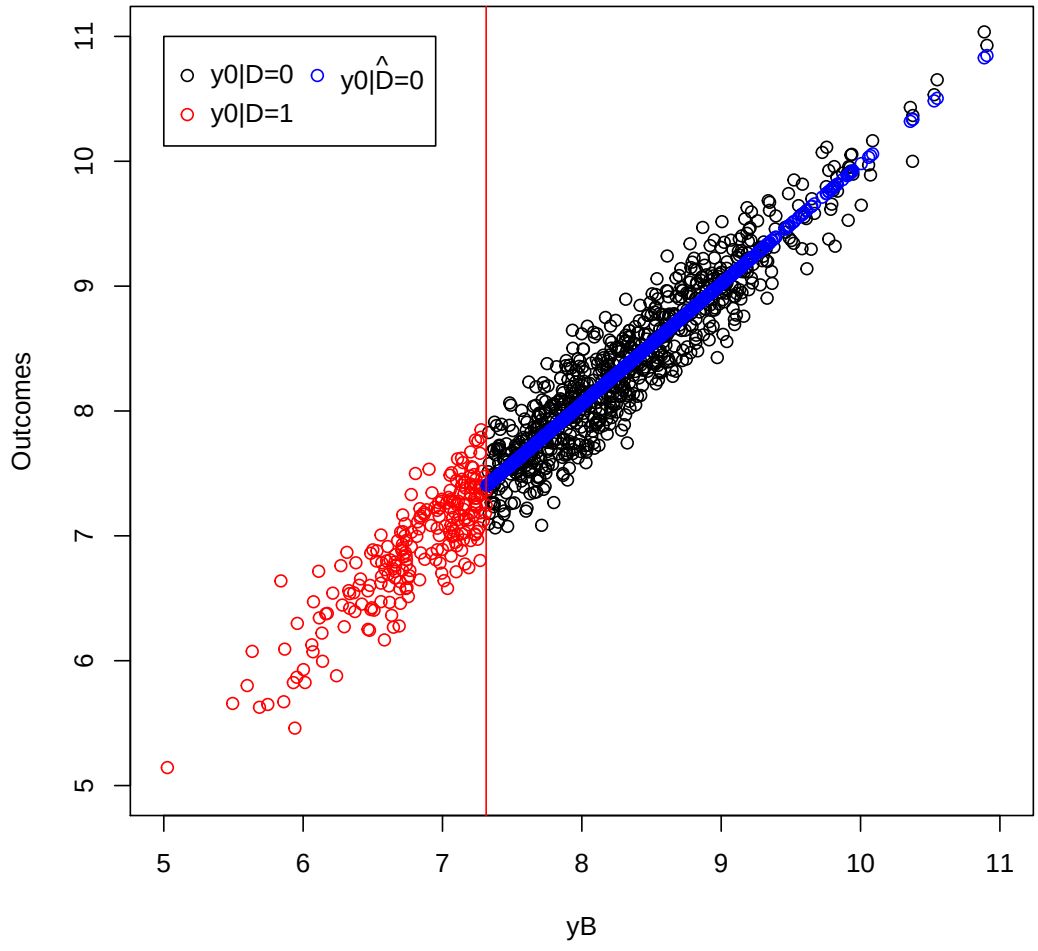


Figure: Step 1

How Can We Estimate TT Using SO and PFF?

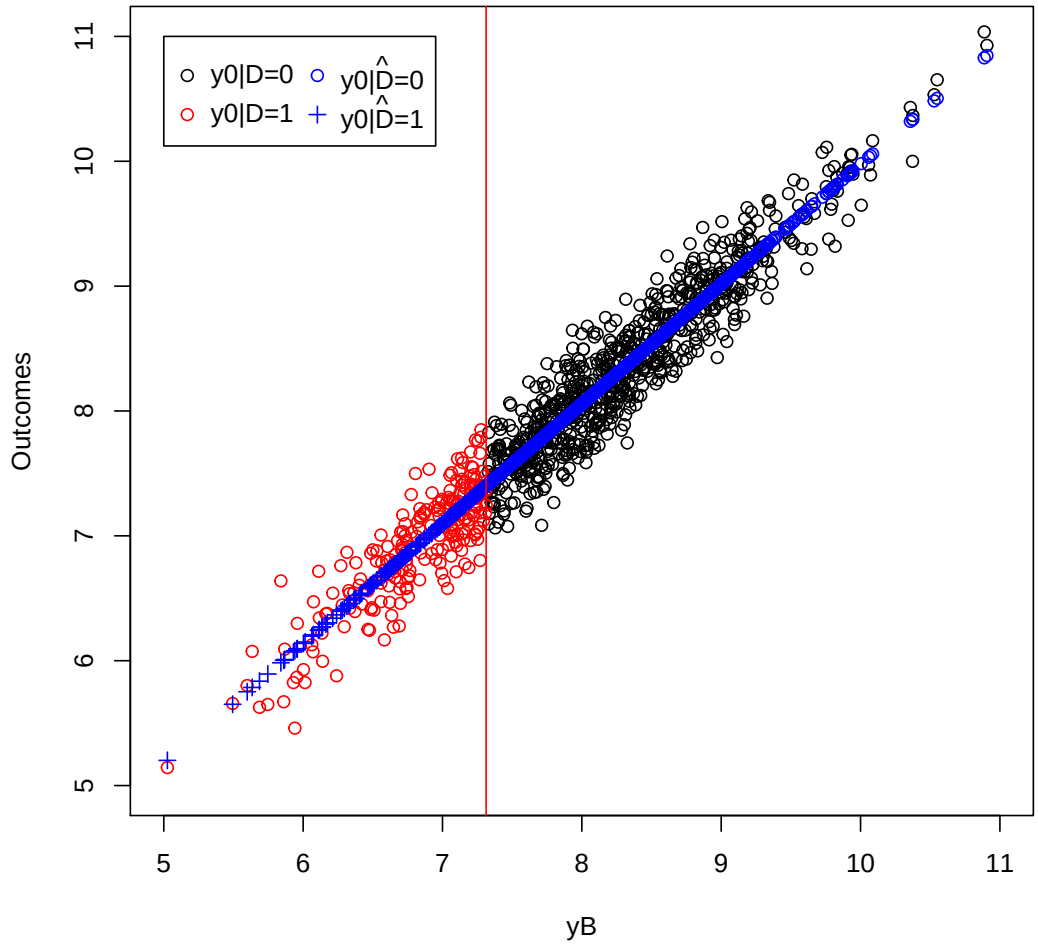


Figure: Step 2

How Can We Estimate TT Using SO and PFF?

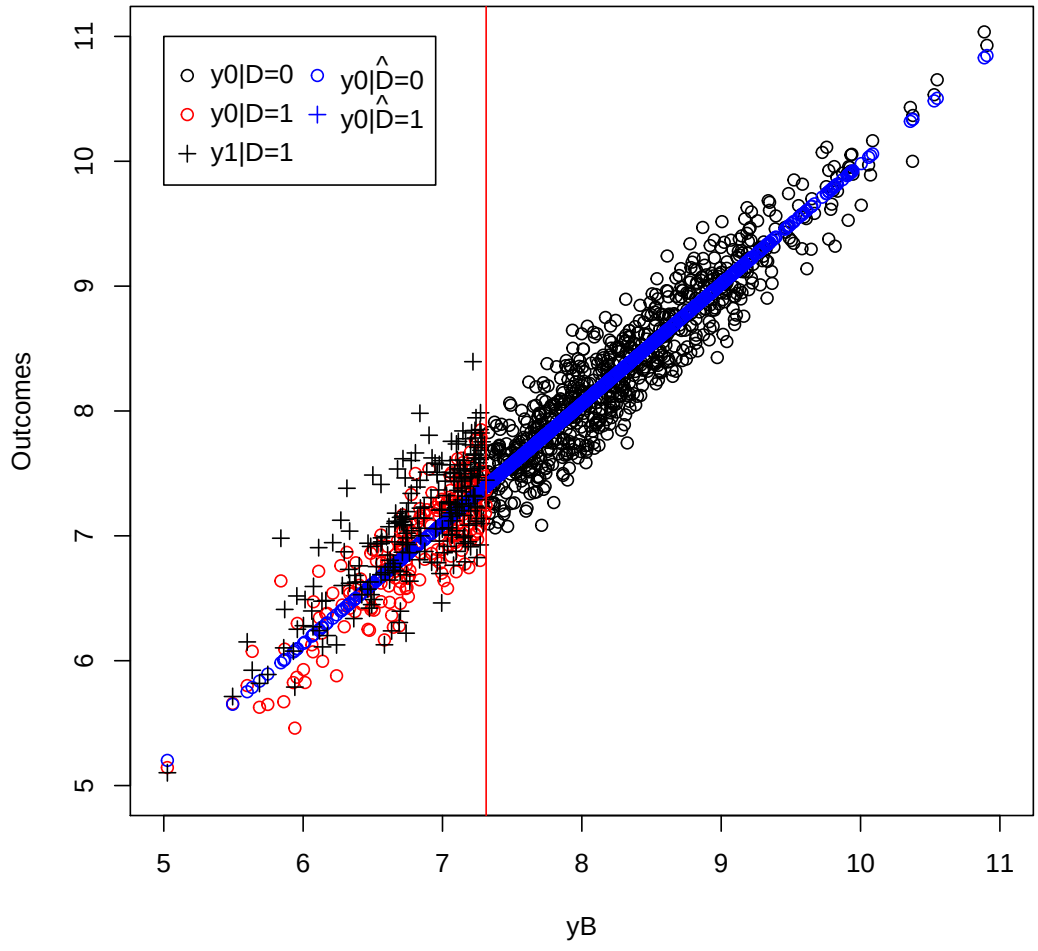


Figure: Step 3

How Can We Estimate TT Using SO and PFF?

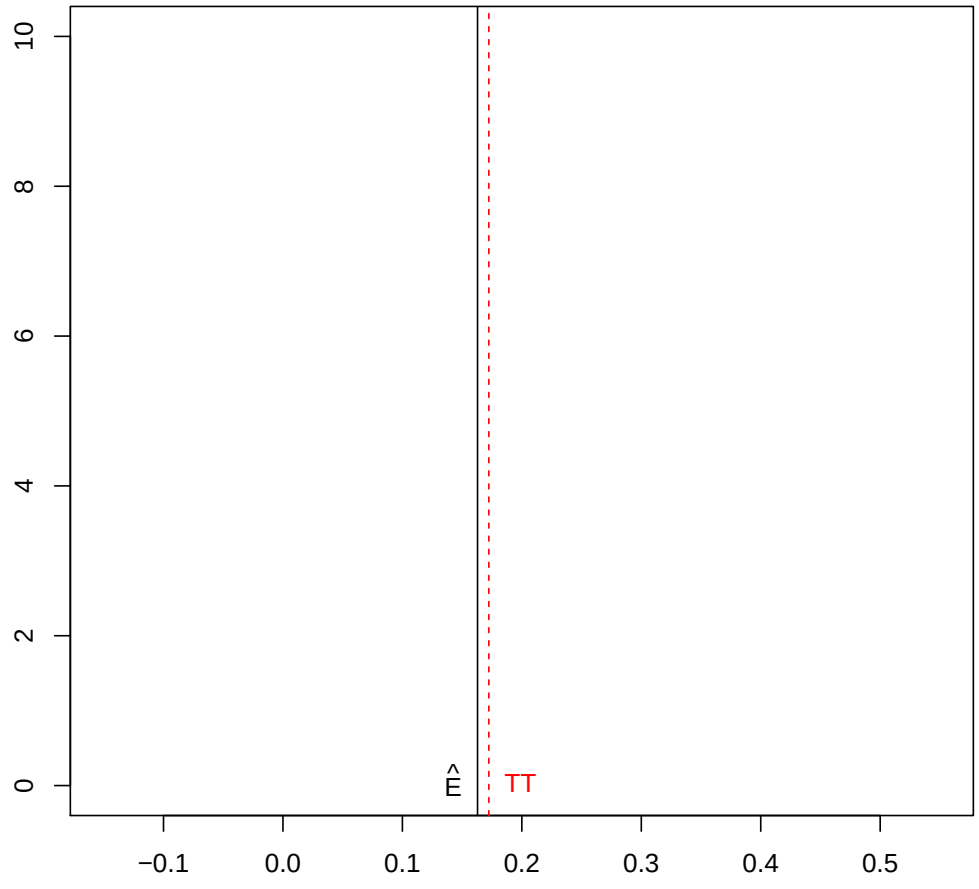


Figure: Step 4

Identification of TT using OLS

Theorem (Identification of TT using OLS)

Under Selection on Observables and Parametric Functional Form, TT is identified using the WW comparison adjusted by the OLS projection.

$$\Delta^{Y}_{WWOLS(X)} = \Delta^{Y}_{TT}.$$

Proof

Under Selection on Observables and Parametric Functional Form, we have:

$$\mathbb{E}[Y_i^0|D_i = 0, X_i] = \mathbb{E}[Y_i^0|X_i] = \alpha_0 + \beta_0' X_i.$$

As a consequence, α_0 and β_0 are identified by using the untreated population, as long as the components of X are not colinear. Then, we have:

$$\begin{aligned}
\Delta^Y_{OLS(X)} &= \mathbb{E}[Y_i - \alpha_0 - \beta_0' X_i | D_i = 1] \\
&= \mathbb{E}[Y_i - \mathbb{E}[Y_i^0|D_i = 1, X_i]|D_i = 1] \\
&= \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[\mathbb{E}[Y_i^0|D_i = 1, X_i]|D_i = 1] \\
&= \mathbb{E}[Y_i^1|D_i = 1] - \mathbb{E}[Y_i^0|D_i = 1] \\
&= \Delta^Y_{TT},
\end{aligned}$$

where the first equality is the definition of the second step of the OLS projection procedure, the second equality is a consequence of Selection on Observables and Parametric Functional Form, the third and fifth equalities use the linearity of conditional expectations and the fourth equality uses the Law of Iterated Expectations.

OLS Adjusted With/Without comparison

Let's denote $\hat{\Delta}^{Y}_{WWOLS(X)}$ the result of the following procedure:

1. Estimate $\hat{\alpha_0}$ and $\hat{\beta_0}$ using an OLS regression of Y_i on X_i on the sample of untreated individuals.
2. Predict the counterfactual values for each treated using $\hat{Y_i^0} = \hat{\alpha_0} + \hat{\beta_0}' X_i$.
3. Compute $\hat{\Delta}^{Y}_{WWOLS(X)} = \frac{1}{\sum_{i=1}^{N} D_i} \sum_{i=1}^{N} D_i (Y_i^1 - \hat{Y_i^0})$.

You can use this estimator directly and derive its precision using the bootstrap. Actually, there is a more direct and simpler approach if you are willing to make another parametric assumption.

Where it OLS Makes Sense (Again!): WWOLS is OLS!

Assumption (Parametric functional form for the Treated)

We assume that there exists a known set of observable covariates X_i such that:

$$\mathbb{E}[Y_i^1|X_i] = \alpha_1 + \beta_1' X_i.$$

Theorem (WWOLS is OLS)

Under No Selection bias and the two Parametric Functional Form Assumptions, the OLS coefficient δ in the following regression:

$$Y_i = \alpha_0 + \beta_0' X_i + (\beta_1 - \beta_0)' \left(X_i - \bar{X}_1\right) D_i + \delta D_i + U_i,$$

with $\bar{X}_1$ the average value of X_i among the treated, is the WWOLS estimator:

$$\hat{\delta}_{OLS} = \hat{\Delta^Y_{WWOLS(X)}}.$$

Proof

To do.

Direct Estimation Using OLS

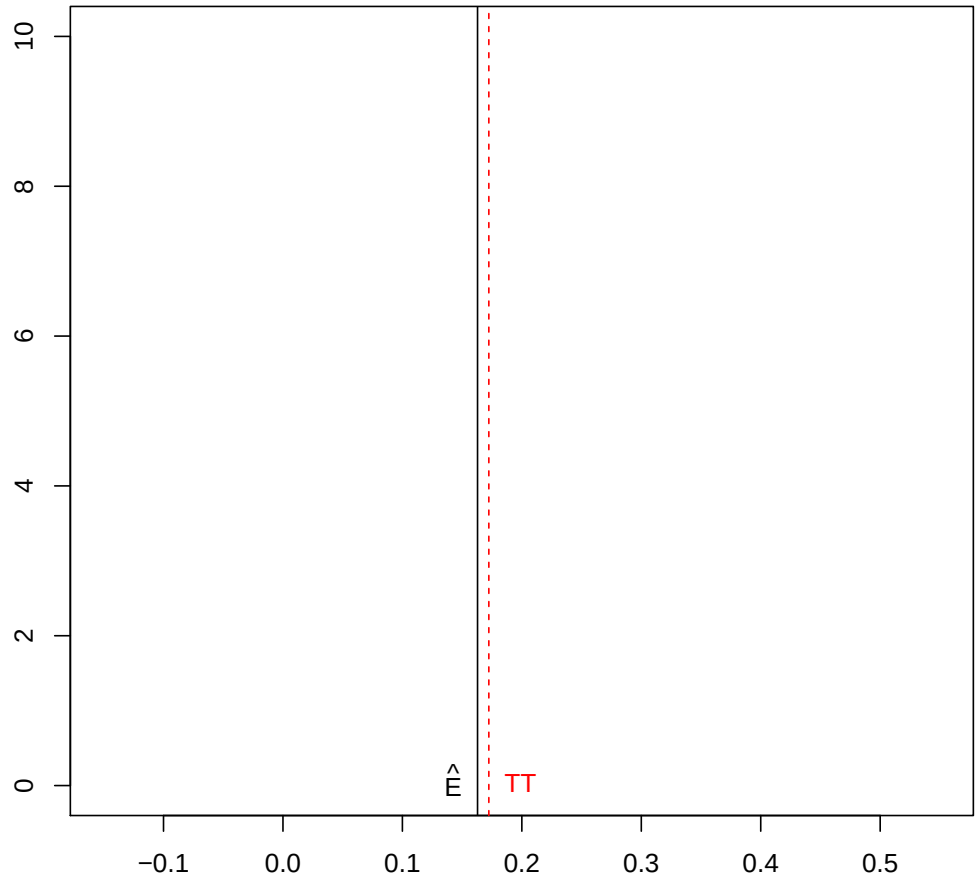


Figure: Direct OLS Estimation

WWOLS is simple OLS when Treatment Effect is Constant

Assumption (Constant Treatment effect)

Assume that the treatment effect is a constant:

$$\forall i,\ \Delta_i^Y = \delta.$$

Theorem (WWOLS is simple OLS)

Under No Selection bias, Parametric Functional Form and Constant Treatment Effect, the OLS coefficient δ in the following regression:

$$Y_i = \alpha_0 + \beta_0' X_i + \delta D_i + U_i$$

is the WWOLS estimator:

$$\hat{\delta}_{OLS} = \hat{\Delta^Y_{WWOLS(X)}}.$$

Proof

To do.

Sampling Noise with OLS: Illustration

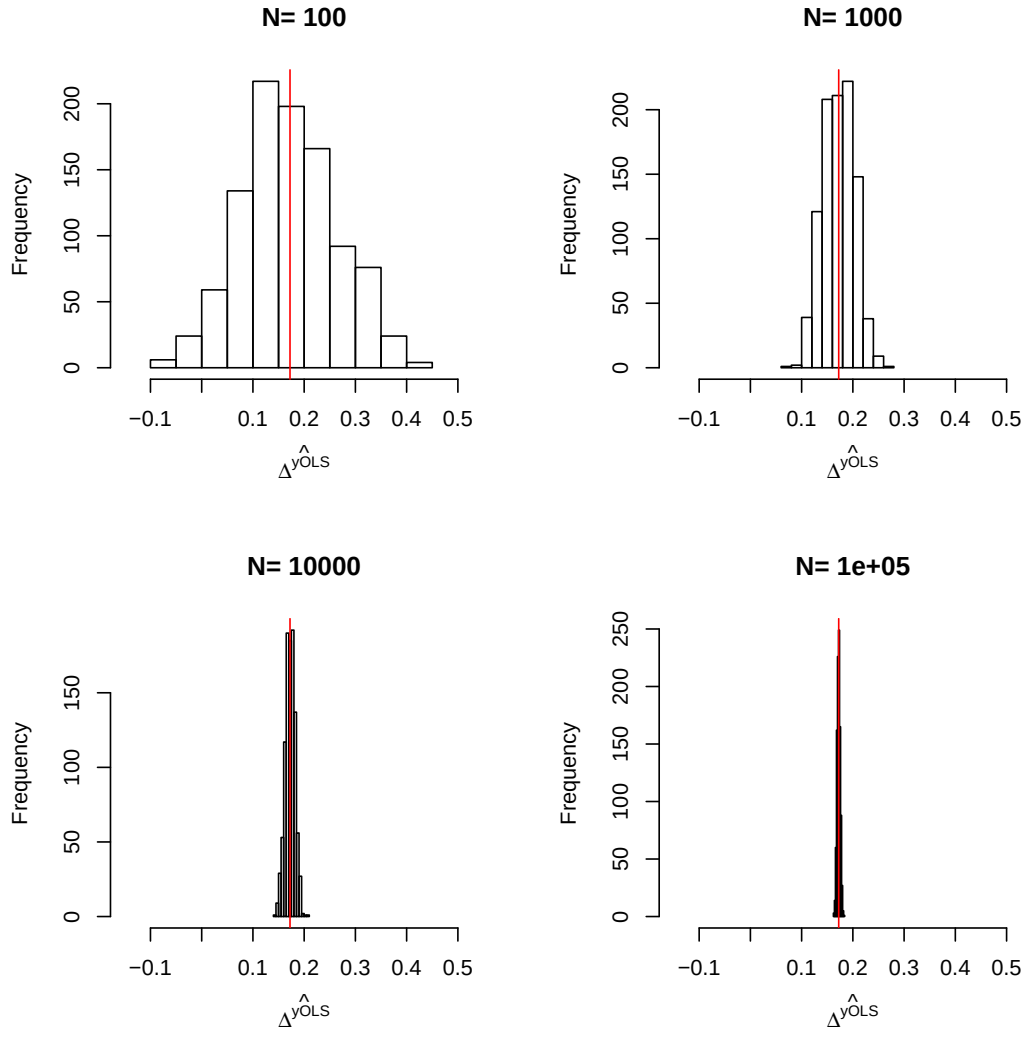


Figure: Distribution of the *OLS* estimator over replications of samples of different sizes

Estimating Precision: Heteroskedasticity

$$U_i = Y_i^0 - \mathbb{E}[Y_i^0|X_i] + D_i(\Delta_i^Y - \Delta_{TT}^Y).$$

The key problem that we face is that of heteroskedasticity. The variance of the error term U_i in the OLS regression depends on:

- D_i: depending on whether or not you are in the treated group, you have a different shock (because the variance of the treatment effect adds to the variance of the outcome in the treated group)
- X_i: the variance in the treatment effect might depend on unobservables correlated with the observables (e.g. in our example, the effect of the treatment depends on μ_i, which is correlated with y_i^B)

Lower Bound on the Precision

Theorem (Efficiency Bound under Unconfoundedness)

Under the assumption that $(Y_i^0, Y_i^1) \perp\!\!\!\perp D_i | X_i$, *we have, for all consistent estimators* $\Delta^{\hat{Y}}_{ATE}$ *and* $\Delta^{\hat{Y}}_{TT}$ *of* Δ^{Y}_{ATE} *and* Δ^{Y}_{TT} *:*

$$\text{Asym}\,\mathbb{V}[\Delta^{\hat{Y}}_{ATE}] \geq \frac{1}{N}\mathbb{E}[\frac{\mathbb{V}[Y_i^1|X_i]}{\Pr(D_i=1|X_i)} + \frac{\mathbb{V}[Y_i^0|X_i]}{1-\Pr(D_i=1|X_i)} + (\Delta^{Y}_{ATE}(X_i) - \Delta^{Y}_{ATE})^2]$$

$$\text{Asym}\,\mathbb{V}[\Delta^{\hat{Y}}_{TT}] \geq \frac{1}{N}\mathbb{E}[\frac{\Pr(D_i=1|X_i)\mathbb{V}[Y_i^1|X_i]}{\Pr(D_i=1)} + \frac{\Pr(D_i=1|X_i)^2\mathbb{V}[Y_i^0|X_i]}{\Pr(D_i=1)(1-\Pr(D_i=1|X_i))} + \frac{(\Delta^{Y}_{TT}(X_i) - \Delta^{Y}_{TT})^2 \Pr(D_i=1|X_i)}{\Pr(D_i=1)}$$

Proof

See Hahn (1998).

Heteroskedasticity Robust Standard Errors using the CLT

Let $\Theta = (\alpha_0, \beta_0, \beta_1, \delta)$ and X be the matrix of all regressors, with a first column of ones and the last column of D_i. Most available heteroskedasticity robust estimators based on the CLT can be written in the following way:

$$\mathbb{V}[\hat{\Theta}_{OLS}] \approx (X'X)^{-1} X' \hat{\Omega} X (X'X)^{-1},$$

where $\hat{\Omega} = \text{diag}(\hat{\sigma}^2_{U_1}, \ldots, \hat{\sigma}^2_{U_N})$ is an estimate the covariance matrix of the residuals U_i.

Heteroskedasticity Robust Standard Errors using the CLT

Here are various classical estimators for $\hat{\Omega}$:

$$\text{HC0:} \qquad \sigma_{\hat{U}_i}^2 = \hat{U}_i^2$$

$$\text{HC1:} \qquad \sigma_{\hat{U}_i}^2 = \frac{N}{N-K}\hat{U}_i^2$$

$$\text{HC2:} \qquad \sigma_{\hat{U}_i}^2 = \frac{\hat{U}_i^2}{1-h_i}$$

$$\text{HC3:} \qquad \sigma_{\hat{U}_i}^2 = \frac{\hat{U}_i^2}{(1-h_i)^2},$$

where $\hat{U}_i$ is the residual from the OLS regression, K is the number of regressors, h_i is the leverage of observation i, and is the i^{th} diagonal element of $H = X(X'X)^{-1}X'$. HC1 is the one reported by Stata when using the 'robust' option.

Heteroskedasticity Robust Standard Errors using R

All of these estimators are programmed in the sandwich package. In practice, it is extremely easy to use:

```
ols.direct.HC1 <- vcovHC(ols.direct,type='HC1')
```

The robust standard error using HC1 is thus 0.0287109, while the default standard error assuming homoskedasticity is 0.0278461.

Precision of OLS

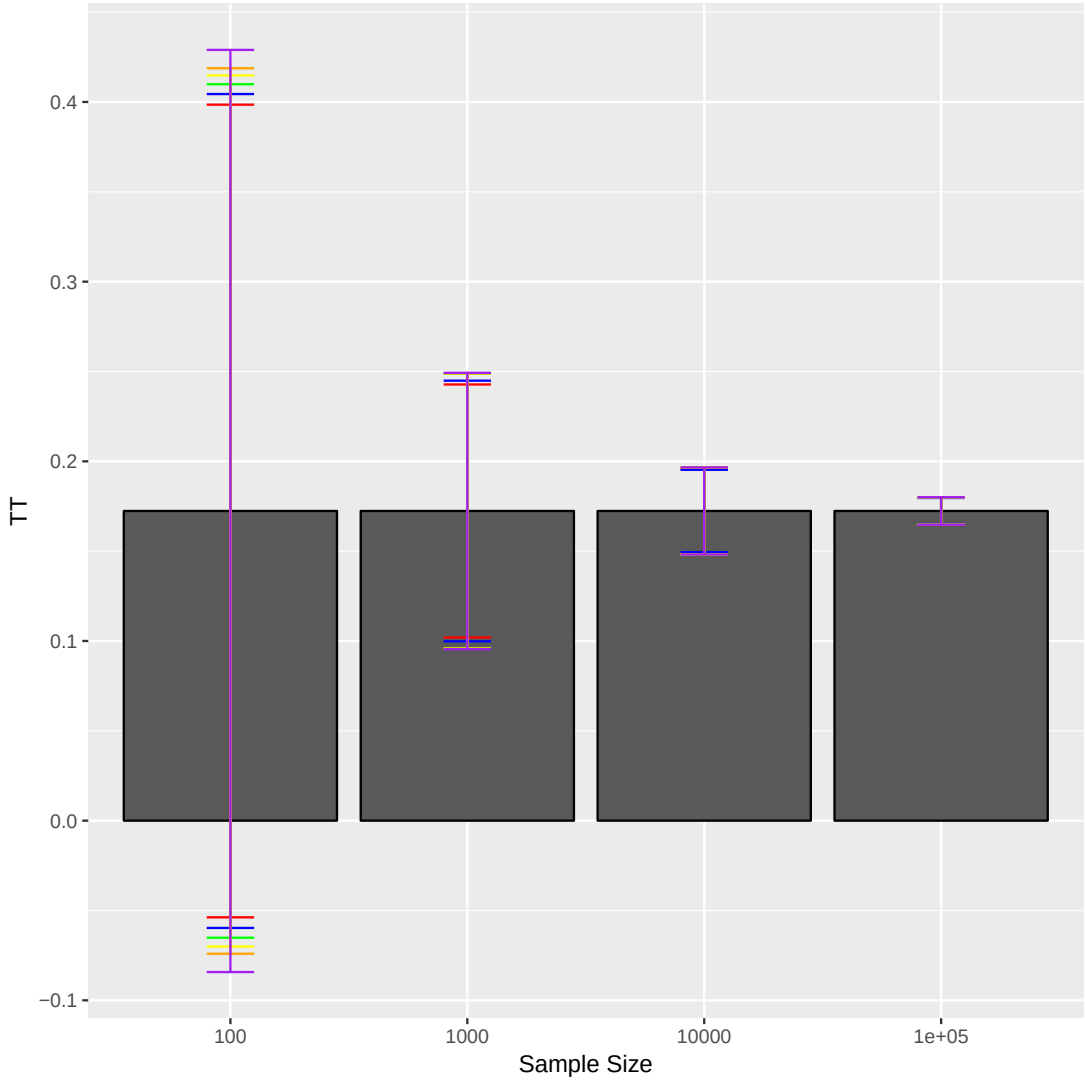


Figure: precision of the OLS estimator with 99% confidence

Problems with OLS

Mostly, the functional form assumption might be false: specification bias.
Illustration:

$$y_i^0 = \mu_i + \delta + \gamma(y_i^B - \bar{y^B}) + U_i^0.$$

Illustration

$$y_i^0 = \mu_i + \delta + \gamma(y_i^B - \bar{y^B})^2 + U_i^0.$$

Nonlinear Relationship Between Outcomes and Confounders

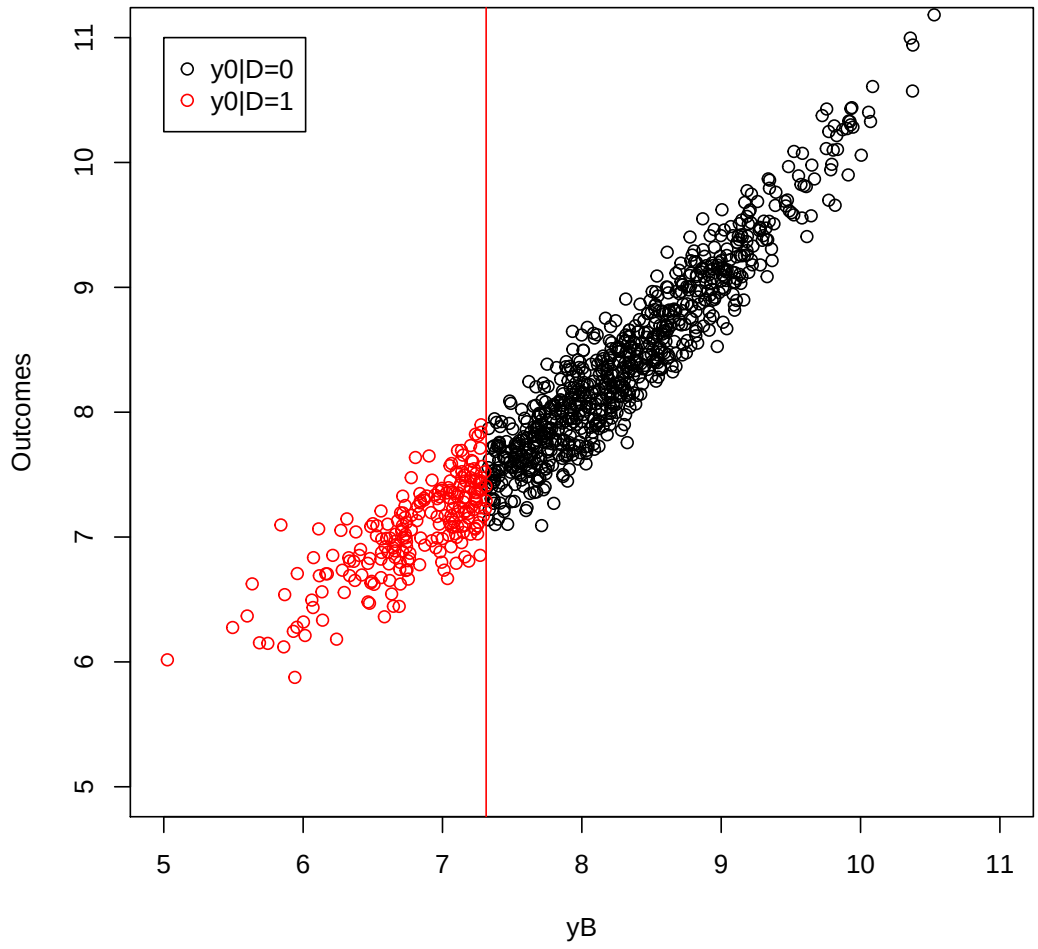


Figure: Non linear

Specification Bias

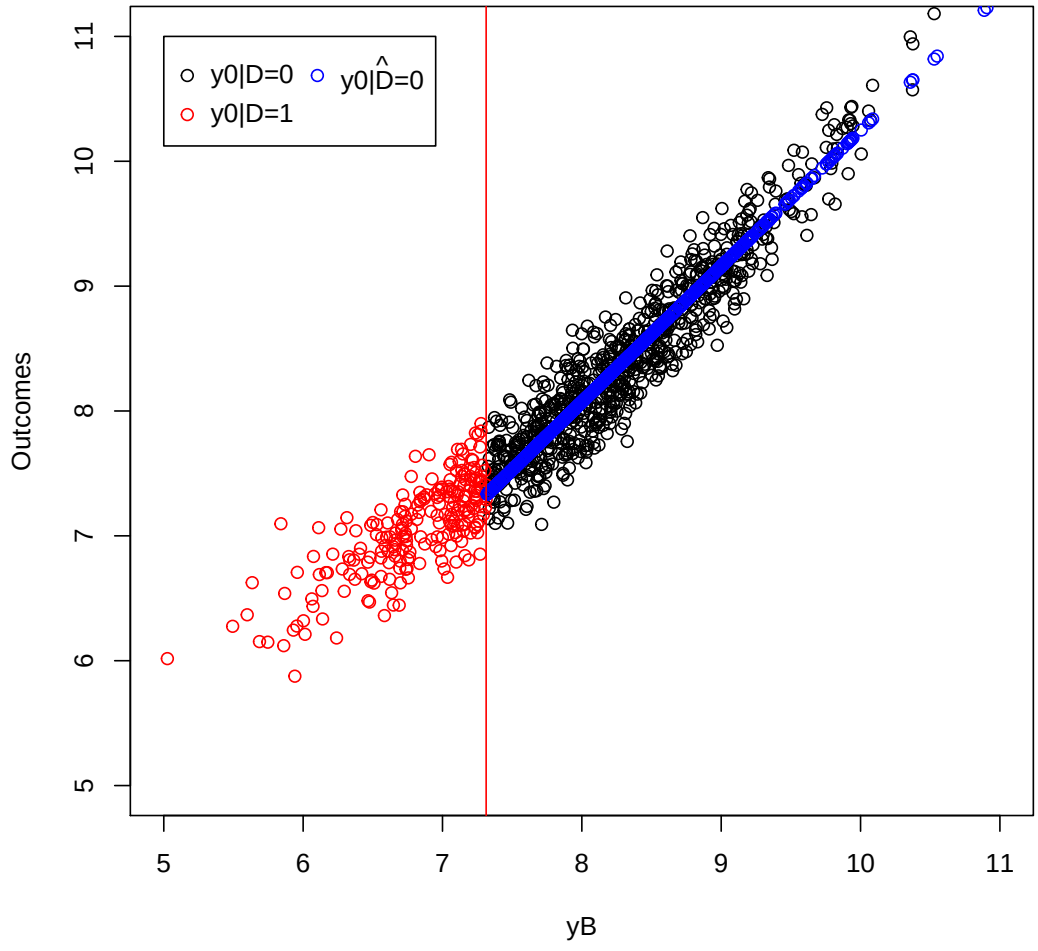


Figure: Step 1

Specification Bias

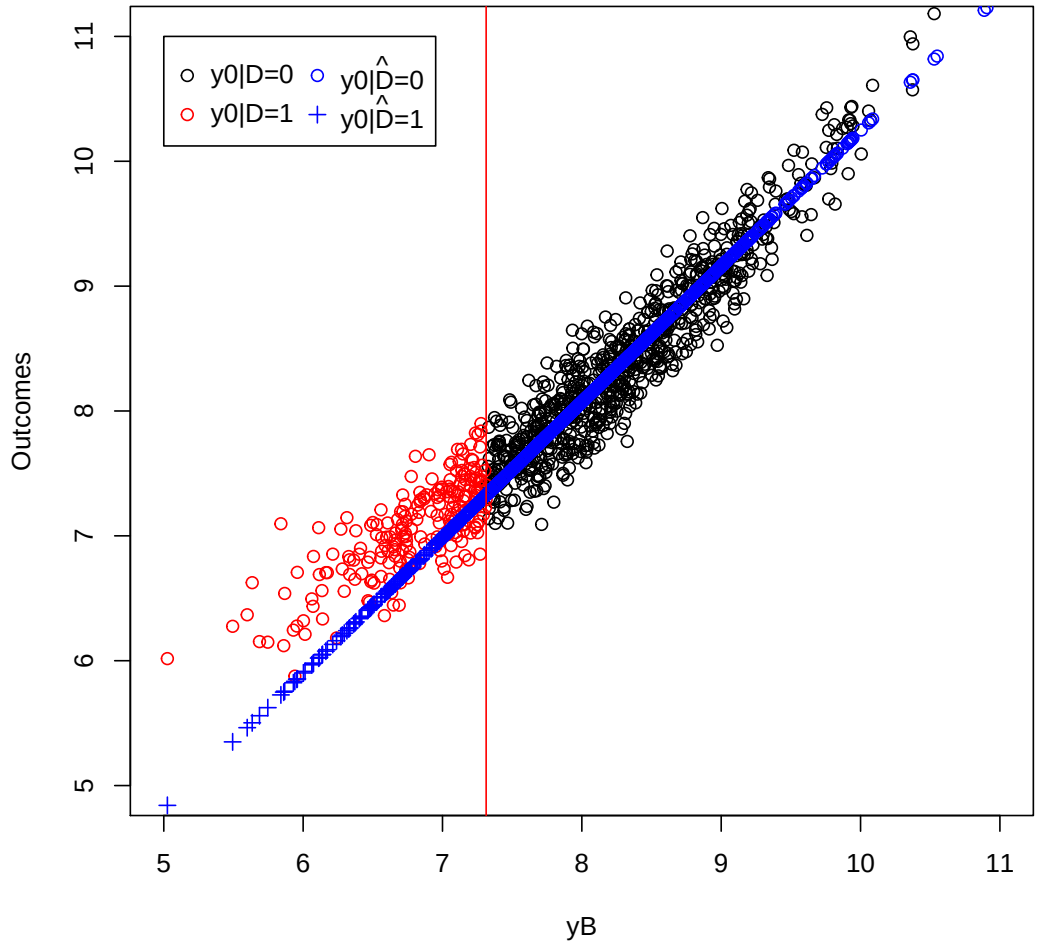


Figure: Step 2

Specification Bias

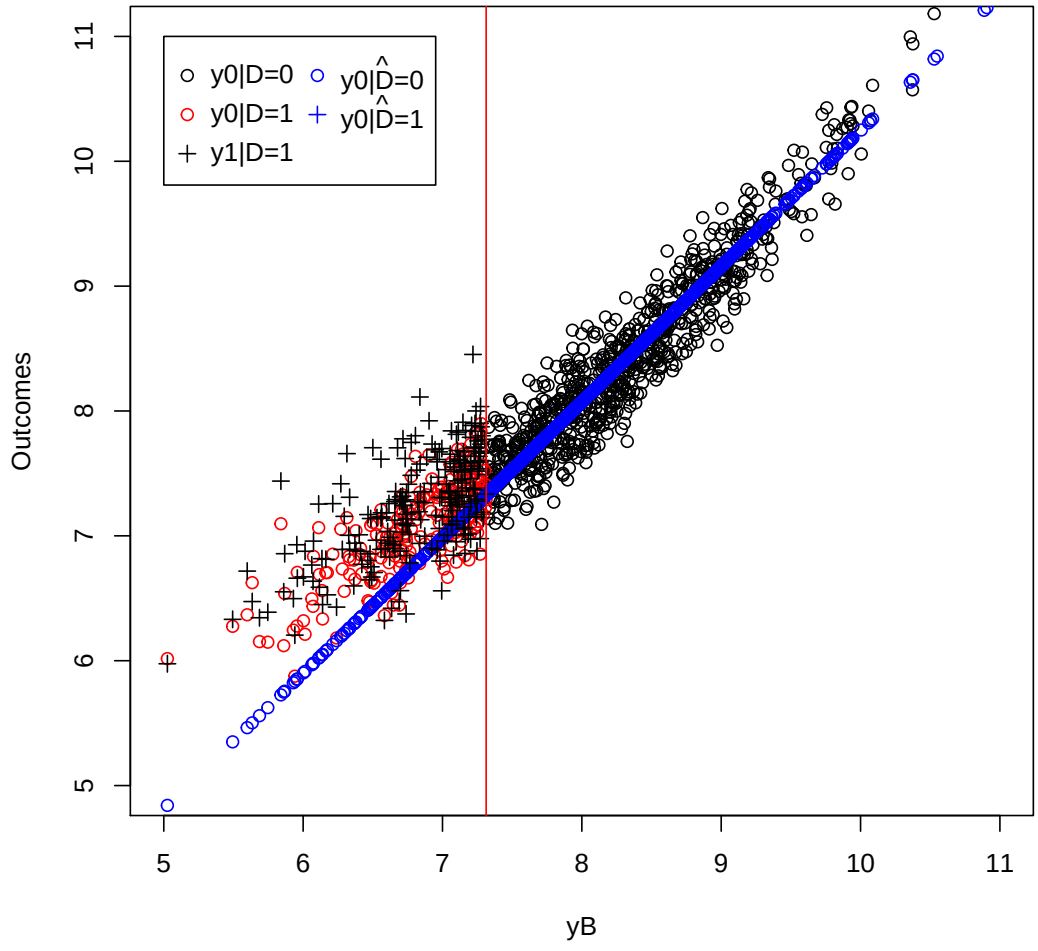


Figure: Step 3

Specification Bias

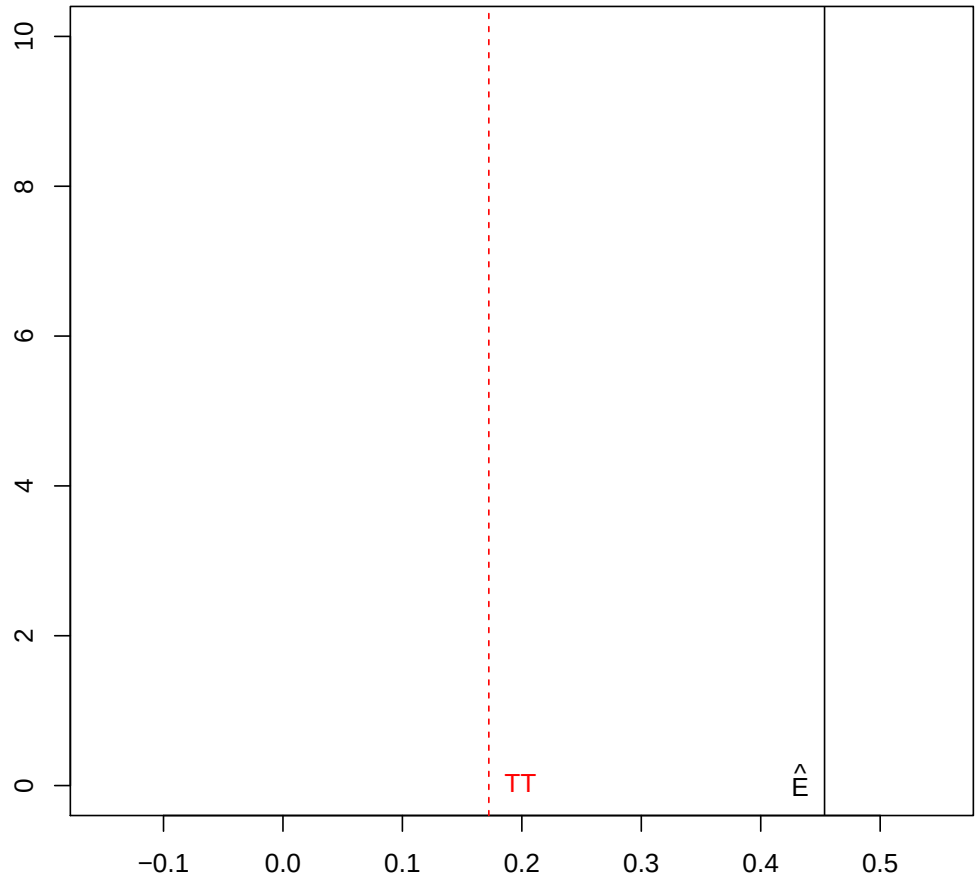


Figure: Step 4

Outline

Nonparametric Approaches in a Nutshell

In order to avoid specification bias, you only rely on local comparisons, that is on the parts of the sample where treated and untreated observations overlap.

Key Assumption: Common Support

Assumption (Common Support)

We assume that, for the known set of observable covariates X_i for which the Selection on Observables Assumption holds, we have:

$$0 < \Pr(D_i = 1 | X_i) < 1.$$

Failure of Common Support: Illustration

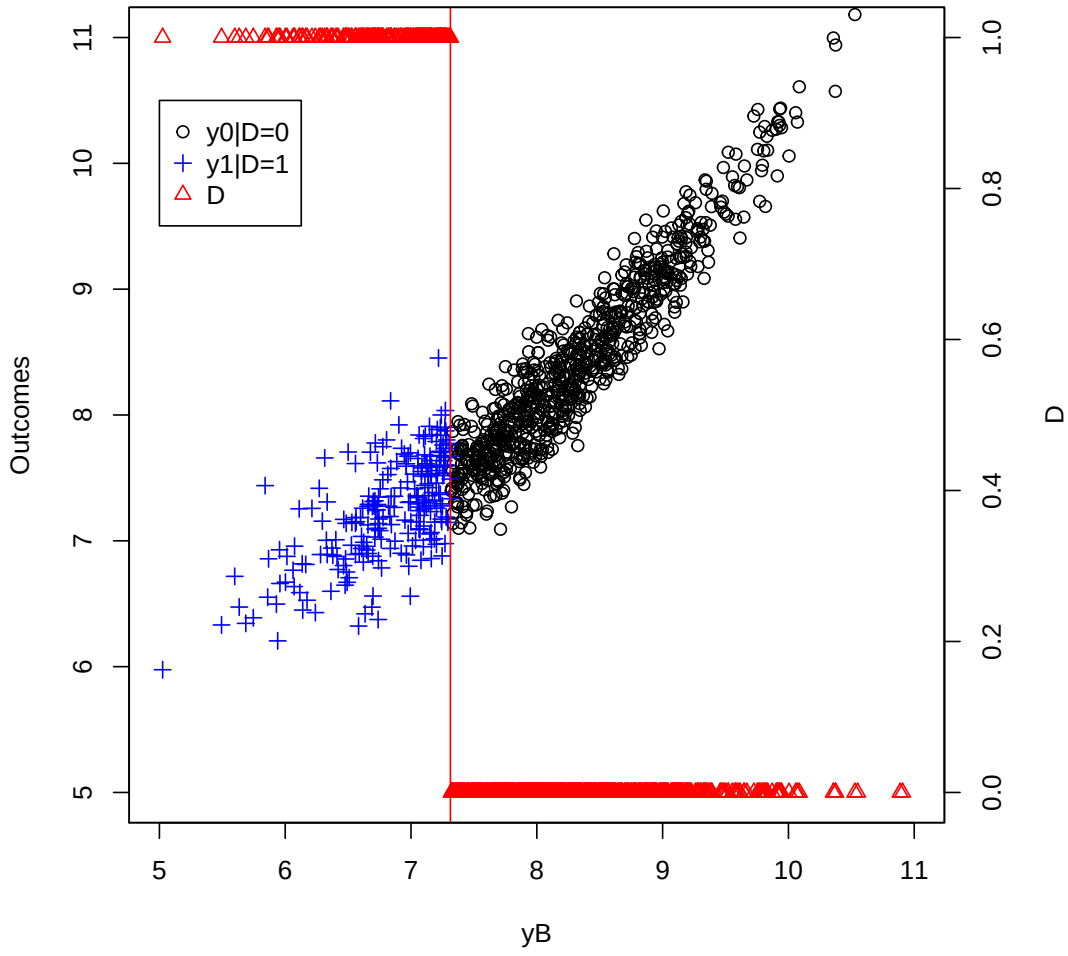


Figure: Failure of Common Support

Generating Common Support

$$D_i = \mathbb{1}[y_i^B + V_i \leq \bar{y}]$$
$$V_i \perp\!\!\!\perp Y_i^0$$
$$V_i \sim \mathcal{N}(0, \sigma_\mu^2 + \sigma_U^2)$$

Common Support: Illustration

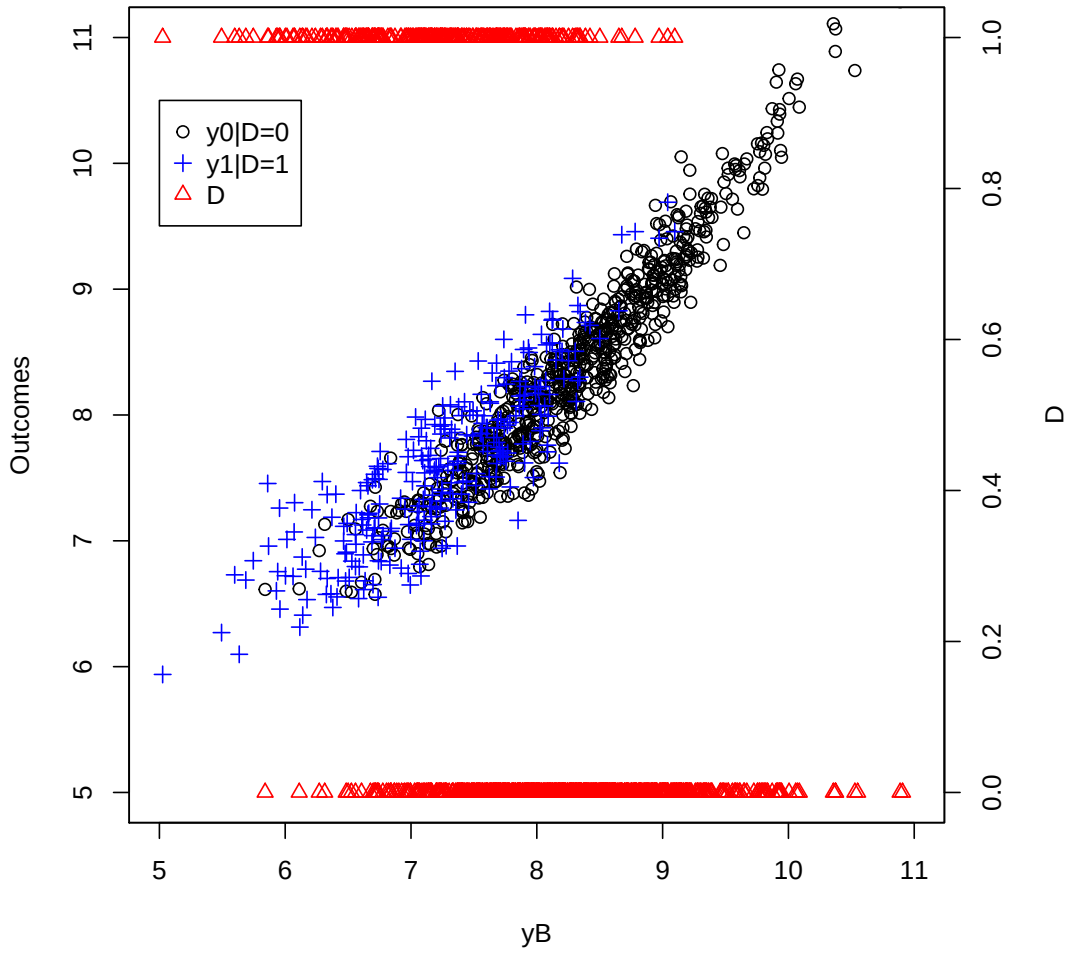


Figure: Common Support Holds

TT in the Nonlinear Example with Common Support

Because we have changed the selection rule, the value of TT in the population has changed also:

$$\Delta^{y}_{TT} = \bar{\alpha} + \theta\bar{\mu} - \theta \frac{\sigma^2_{\mu}}{\sqrt{2(\sigma^2_{\mu} + \sigma^2_{U})}} \frac{\phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{2(\sigma^2_{\mu}+\sigma^2_{U})}}\right)}{\Phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{2(\sigma^2_{\mu}+\sigma^2_{U})}}\right)}.$$

The value of TT is now 0.1752852.

Identification of TT Under SO and CS

Theorem (Identification of TT Under SO and CS)

Under Selection on Observables and Common Support, TT is identified using either a nonparametric regression approach or a reweighting approach:

$$\Delta^Y_{TT} = \Delta^Y_{NPR(X)} = \Delta^Y_{RW(X)},$$

with:

$$\begin{aligned}
\Delta^Y_{NPR(X)} &= \mathbb{E}[Y_i - \mathbb{E}[Y_i|X_i, D_i = 0]|D_i = 1] \\
\Delta^Y_{RW(X)} &= \mathbb{E}[Y_i|D_i = 1] \\
&\quad - \mathbb{E}[Y_i \frac{\Pr(D_i = 1|X_i)}{1 - \Pr(D_i = 1|X_i)} \frac{1 - \Pr(D_i = 1)}{\Pr(D_i = 1)}|D_i = 0].
\end{aligned}$$

Proof (first part)

$$
\begin{aligned}
\Delta^{Y}_{NPR(X)} &= \mathbb{E}[Y_i - \mathbb{E}[Y_i|X_i, D_i = 0]|D_i = 1] \\
&= \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[\mathbb{E}[Y_i|X_i, D_i = 0]|D_i = 1] \\
&= \mathbb{E}[Y_i^1|D_i = 1] - \mathbb{E}[\mathbb{E}[Y_i^0|X_i, D_i = 0]|D_i = 1] \\
&= \mathbb{E}[Y_i^1|D_i = 1] - \mathbb{E}[\mathbb{E}[Y_i^0|X_i, D_i = 1]|D_i = 1] \\
&= \mathbb{E}[Y_i^1|D_i = 1] - \mathbb{E}[Y_i^0|D_i = 1] \\
&= \Delta^{Y}_{TT}.
\end{aligned}
$$

The second equality follows from the linearity of linear expectations, the fourth equality follows from Selection on Observables and the fifth equality follows from the Law of Iterated Expectations and from Common Support. Common Support also allows $\Delta^{Y}_{NPR(X)}$ to be well-defined.

Proof (second part)

$$
\begin{aligned}
\Delta^{Y}_{RW(X)} &= \mathbb{E}[Y_i \frac{\Pr(D_i = 1|X_i)}{1 - \Pr(D_i = 1|X_i)} \frac{1 - \Pr(D_i = 1)}{\Pr(D_i = 1)} | D_i = 0] \\
&= \frac{1}{\Pr(D_i = 1)} \mathbb{E}[Y_i^0 (1 - D_i) \frac{\Pr(D_i = 1|X_i)}{1 - \Pr(D_i = 1|X_i)}] \\
&= \frac{1}{\Pr(D_i = 1)} \mathbb{E}[\mathbb{E}[Y_i^0 (1 - D_i) \frac{\Pr(D_i = 1|X_i)}{1 - \Pr(D_i = 1|X_i)} | X_i]] \\
&= \frac{1}{\Pr(D_i = 1)} \mathbb{E}[\mathbb{E}[Y_i^0 | X_i](1 - \Pr(D_i = 1|X_i)) \frac{\Pr(D_i = 1|X_i)}{1 - \Pr(D_i = 1|X_i)}] \\
&= \frac{1}{\Pr(D_i = 1)} \mathbb{E}[\mathbb{E}[Y_i^0 | X_i] \Pr(D_i = 1|X_i)] \\
&= \frac{1}{\Pr(D_i = 1)} \mathbb{E}[\mathbb{E}[Y_i^0 D_i | X_i]] \\
&= \frac{1}{\Pr(D_i = 1)} \mathbb{E}[Y_i^0 D_i] \\
&= \frac{1}{\Pr(D_i = 1)} \mathbb{E}[Y_i^0 | D_i = 1] \Pr(D_i = 1) \\
&= \mathbb{E}[Y_i^0 | D_i = 1]
\end{aligned}
$$

The third and fifth equalities use Selection on Observables. Common support allows $\Delta^{Y}_{RW(X)}$ to be well-defined.

General formula for matching estimators

$$\hat{\Delta^{Y}_{M}} = \frac{1}{N_1^S} \sum_{i \in \mathcal{I}^1 \cap S} \left(Y_i - \sum_{j \in \mathcal{I}^0} w_{i,j} Y_j \right)$$

with $\mathcal{I}^1$ is the set of treated individuals, $\mathcal{I}^0$ the set of untreated individuals, S the set of individuals lying on the common support, N_1^S the number of treated on the common support

Types of Matching Estimators

- Local Regression Matching
- Local Averaging Matching
- Nearest Neighbor Matching
- Reweighting Matching
- Doubly Robust Matching

Basic Intuition for LLR Matching

- The idea is very simple: run a separate regression for each treated observation using only untreated observations that are close enough.
- Closeness is defined by a bandwidth, or a window around the treated observation of interest.
- In order to be more efficient, more weight is given to untreated observations that lie closer to the treated observation. Weights are given using a kernel function.

Local Linear Regression

In practice, with LLR, $\sum_{j \in \mathcal{I}^0} w_{i,j} Y_j = \widehat{\mathbb{E}[Y_i | X_i, D_i = 0]}$ is equal to $\hat{\theta_0}$ estimated by weighted OLS in the following regression on the sample of untreated individuals

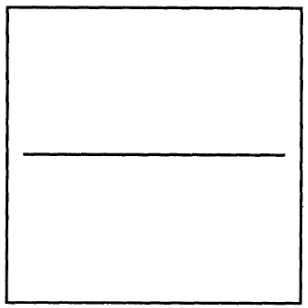
$$Y_j = \theta_0 + (X_i - X_j)\theta_1 + \epsilon_j,$$

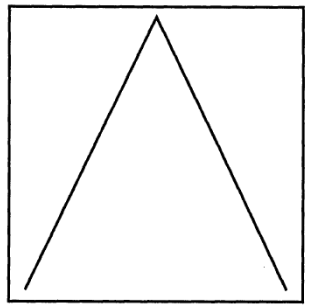
with weights $K\left(\frac{X_i - X_j}{h}\right)$, where h is the bandwidth and K is a kernel function.

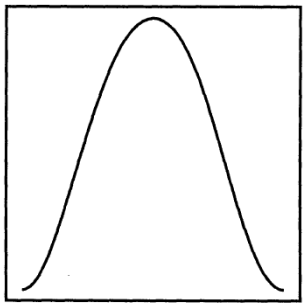
Univariate Kernel Functions

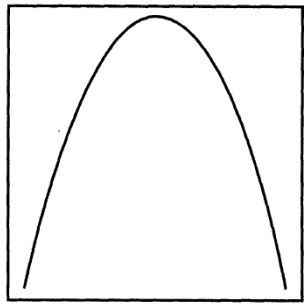
Figure 6. Alternative Kernel Functions

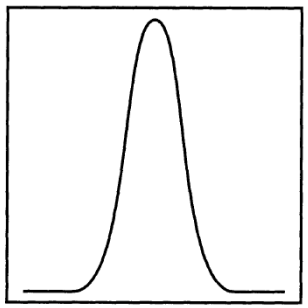
	Kernel	Domain		Kernel	Domain
Uniform°	1	$u \in [-\frac{1}{2},\frac{1}{2}]$	Triangular	$2(1 - 2\lvert u\rvert)$	$u \in [-\frac{1}{2},\frac{1}{2}]$
Quartic	$\frac{15}{8}(1 - 4u^2)^2$	$u \in [-\frac{1}{2},\frac{1}{2}]$	Epanechnikov	$\frac{3}{2}(1 - 4u^2)$	$u \in [-\frac{1}{2},\frac{1}{2}]$
Normal	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} u^2\right)$	$u \in (-\infty,\infty)$			











Weights of the Local Linear Regression

Using some algebra, on can show, in the case of one-dimesional regressor X_i, that the weights of LLR are as follows:

$$w_{i,j} = \frac{K_{i,j} \sum_{k \in \mathcal{I}_0} K_{i,k}(X_k - X_i)^2 - [K_{i,j}(X_j - X_i)][\sum_{k \in \mathcal{I}_0} K_{i,k}(X_k - X_i)]}{\sum_{j \in \mathcal{I}_0} K_{i,j} \sum_{k \in \mathcal{I}_0} K_{i,k}(X_k - X_i)^2 - [\sum_{k \in \mathcal{I}_0} K_{i,k}(X_k - X_i)]^2}$$

with $K_{i,j} = K\left(\frac{X_i - X_j}{h}\right)$.

LLR Matching

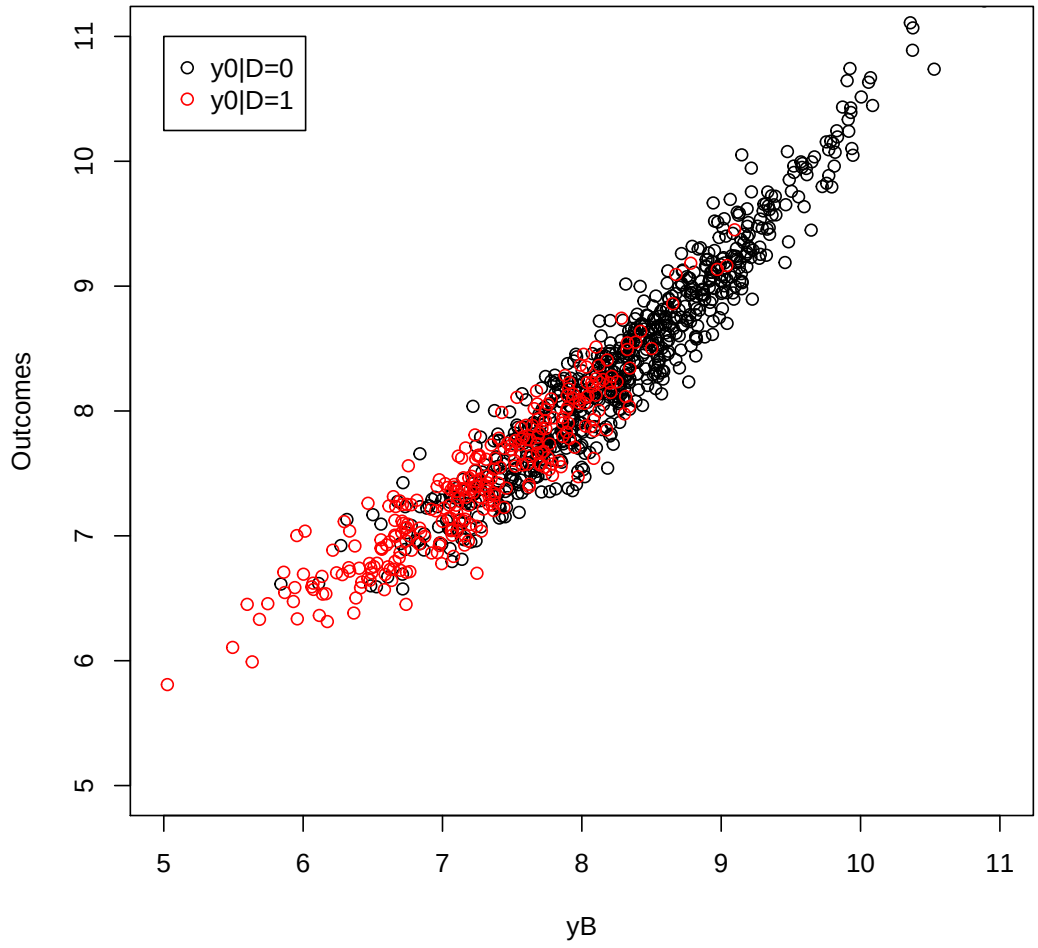


Figure: Step 0

LLR Matching

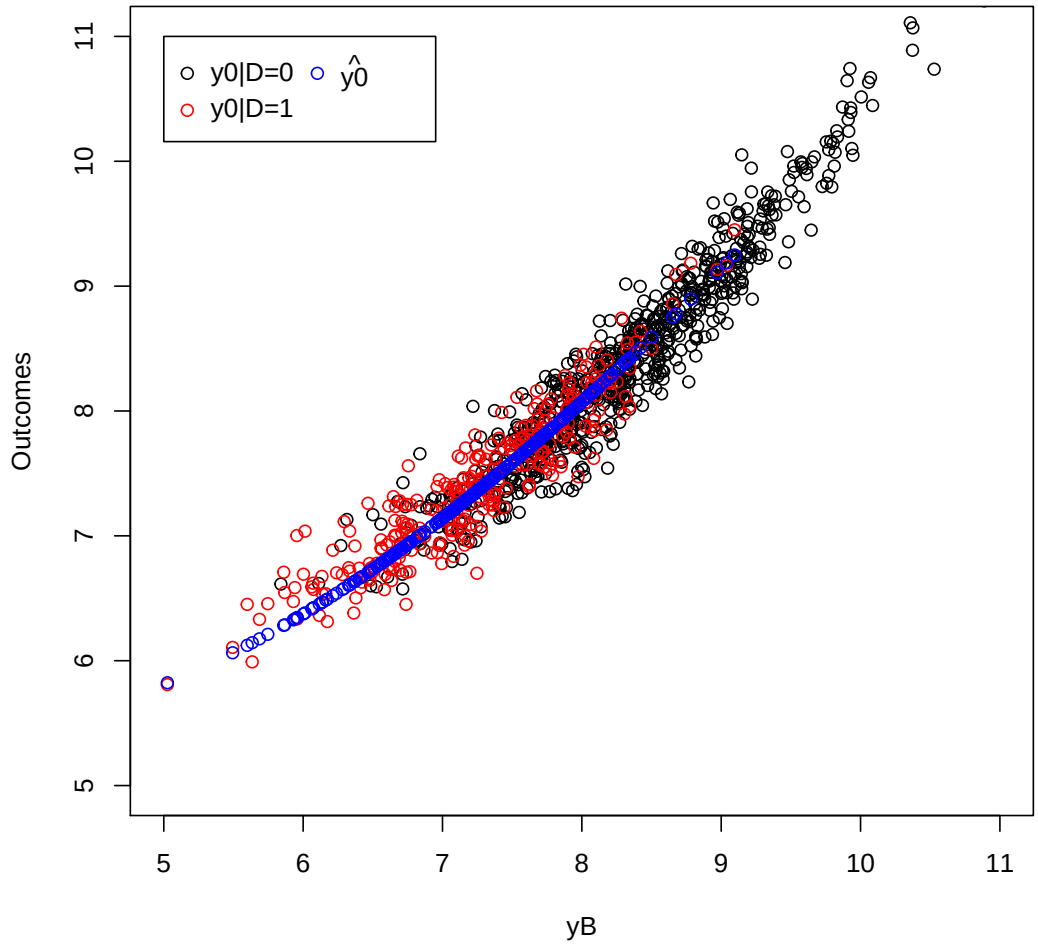


Figure: Step 1

LLR Matching

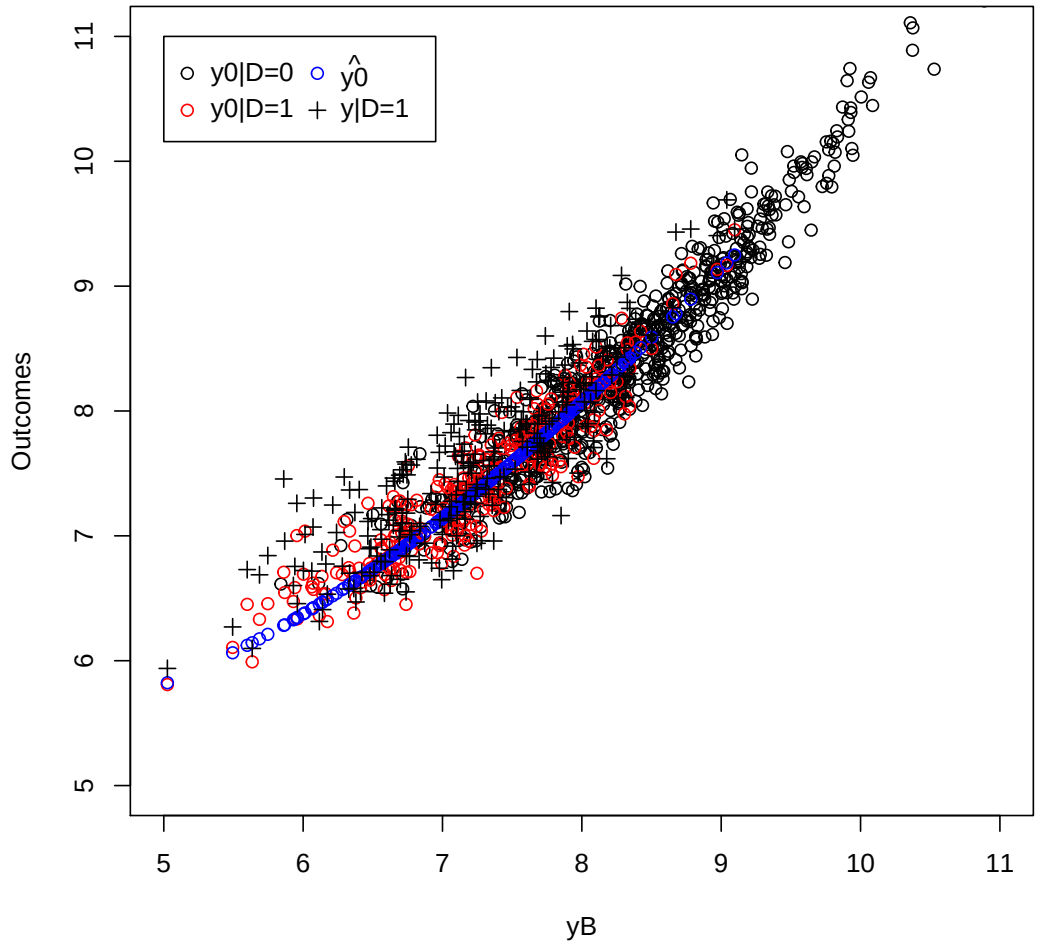


Figure: Step 2

LLR Matching

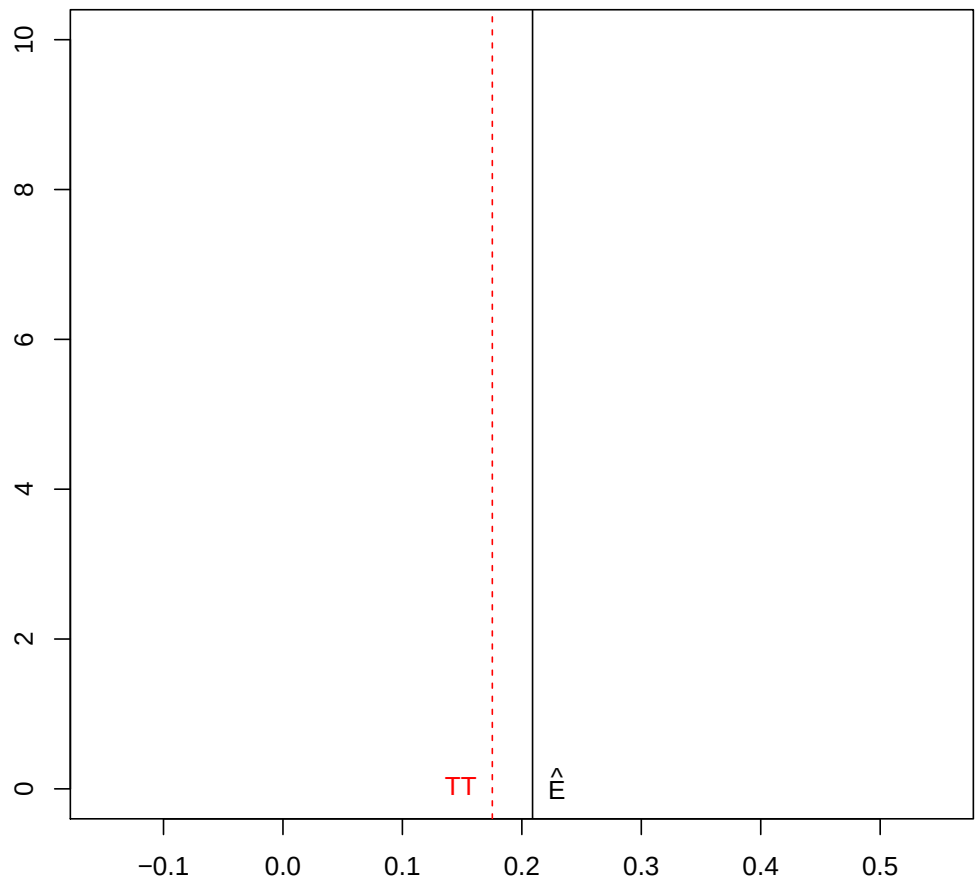


Figure: Step 3

with bandwidth=1 and a gaussian kernel

LLR Matching

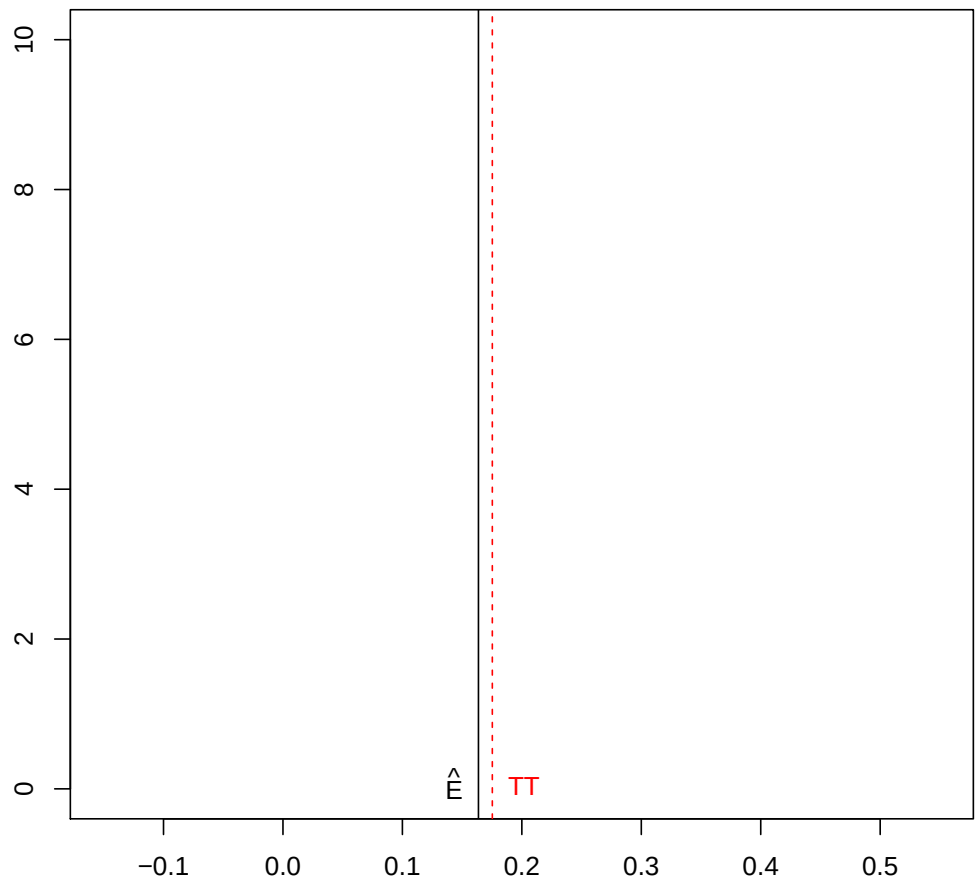


Figure: Step 3

with bandwidth=0.5 and a gaussian kernel

LLR Matching

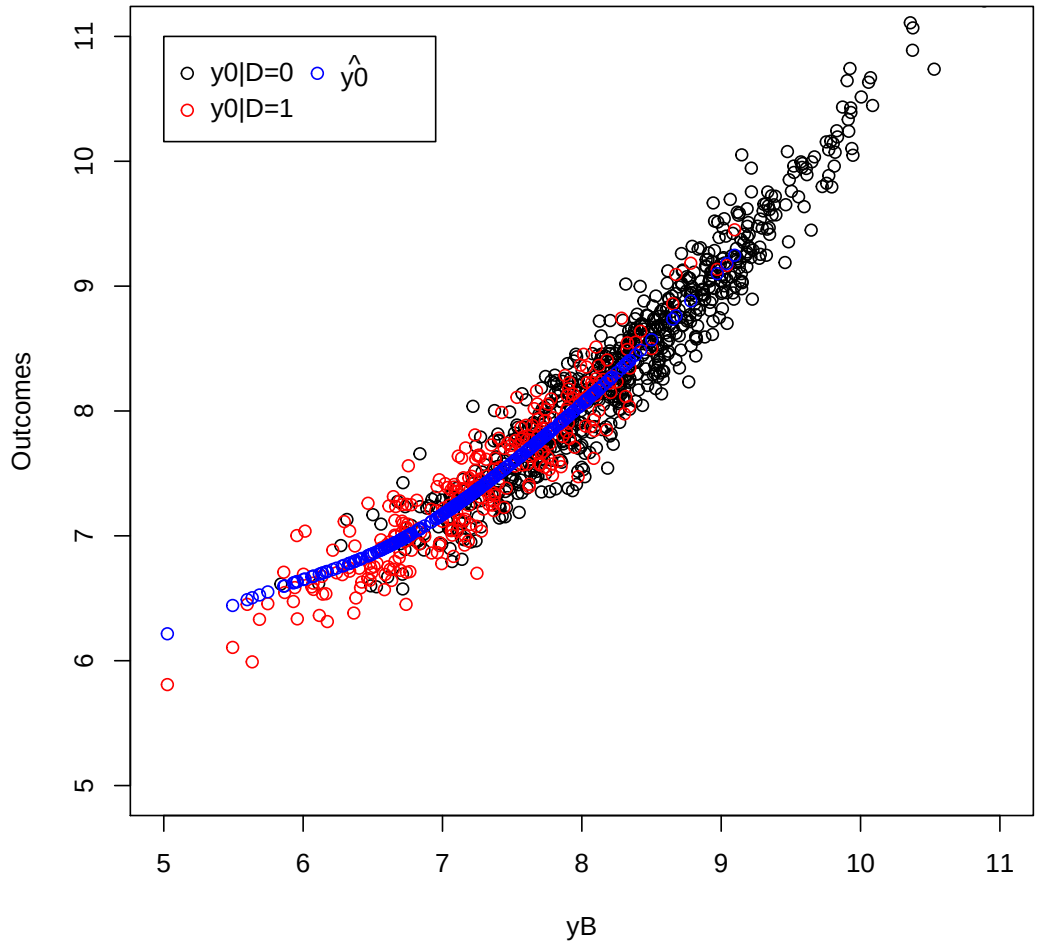


Figure: Step 1

Sampling Noise with LLR: Illustration

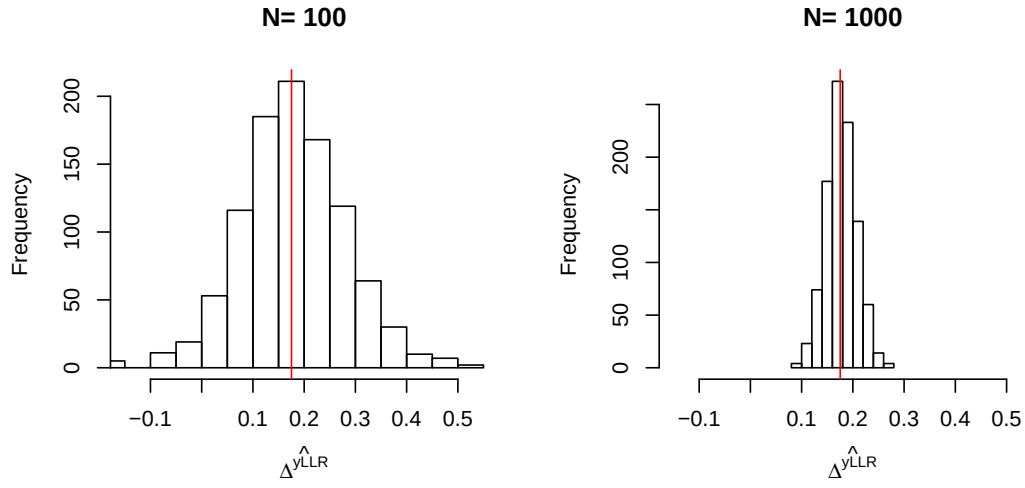


Figure: Distribution of the *LLR* estimator over replications of samples of different sizes

Trimming and Common Support

1. Estimate conditional density $\hat{f}_{X|D=0}$, for $i \in \mathcal{I}_1$
2. Get rid of observations with density lower than the t^{th} percentile of $\hat{f}_{X|D=0}$, where t is the trimming level (5% in general).

$$\hat{f}_{X|D=d}(X_i) = \frac{1}{N_d h} \sum_{j=1}^{N_d} K\left(\frac{X_i - X_j}{h}\right)$$

$$\hat{S} = \begin{cases} 1 & \text{if } \hat{f}_{X|D=0}(X_i) > \text{quantile}_t(\hat{f}_{X|D=1}(X_i)) \\ 0 & \text{if } \hat{f}_{X|D=0}(X_i) \leq \text{quantile}_t(\hat{f}_{X|D=1}(X_i)), \end{cases}$$

where h is a different bandwidth that the one chosen for LLR. You can choose it by using Silverman's rule of thumb, for example.

Trimming: Illustration

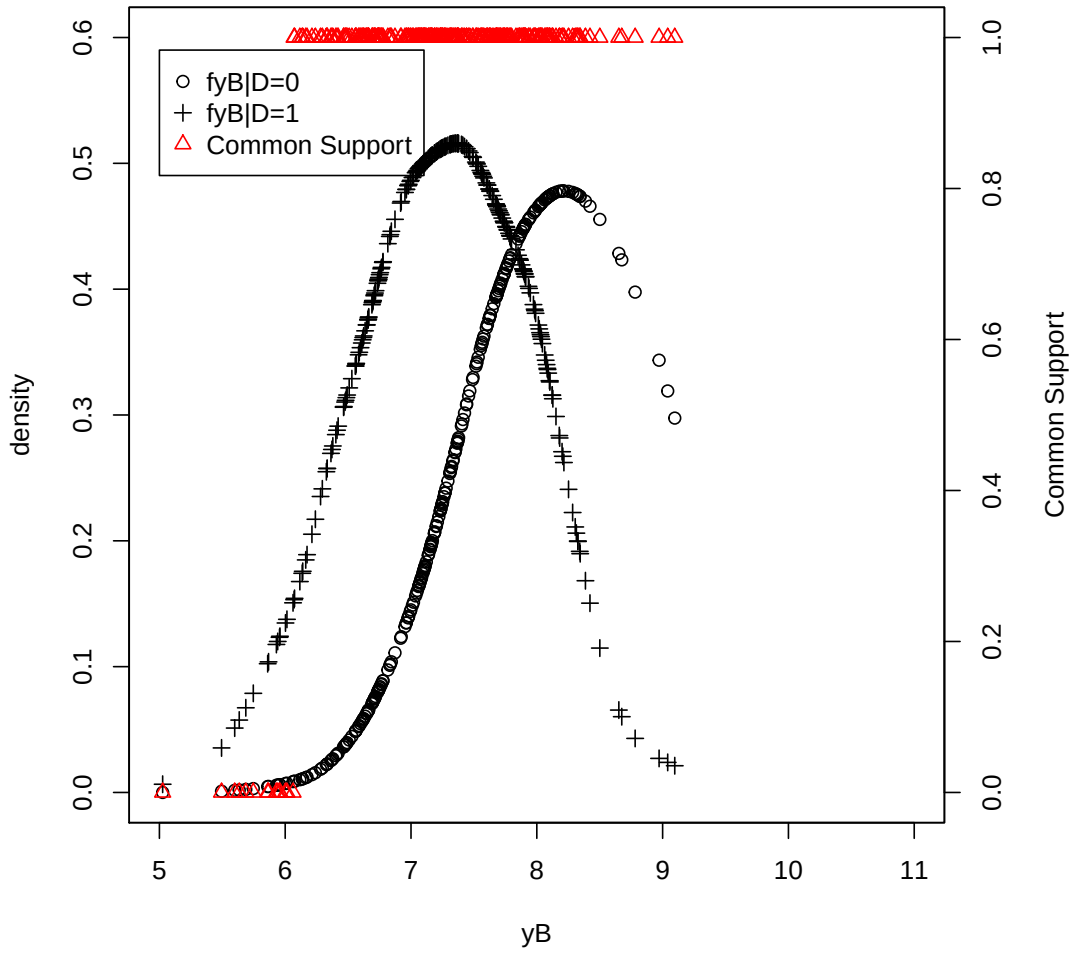


Figure: Trimming and Common Support

LLR Matching with Trimming

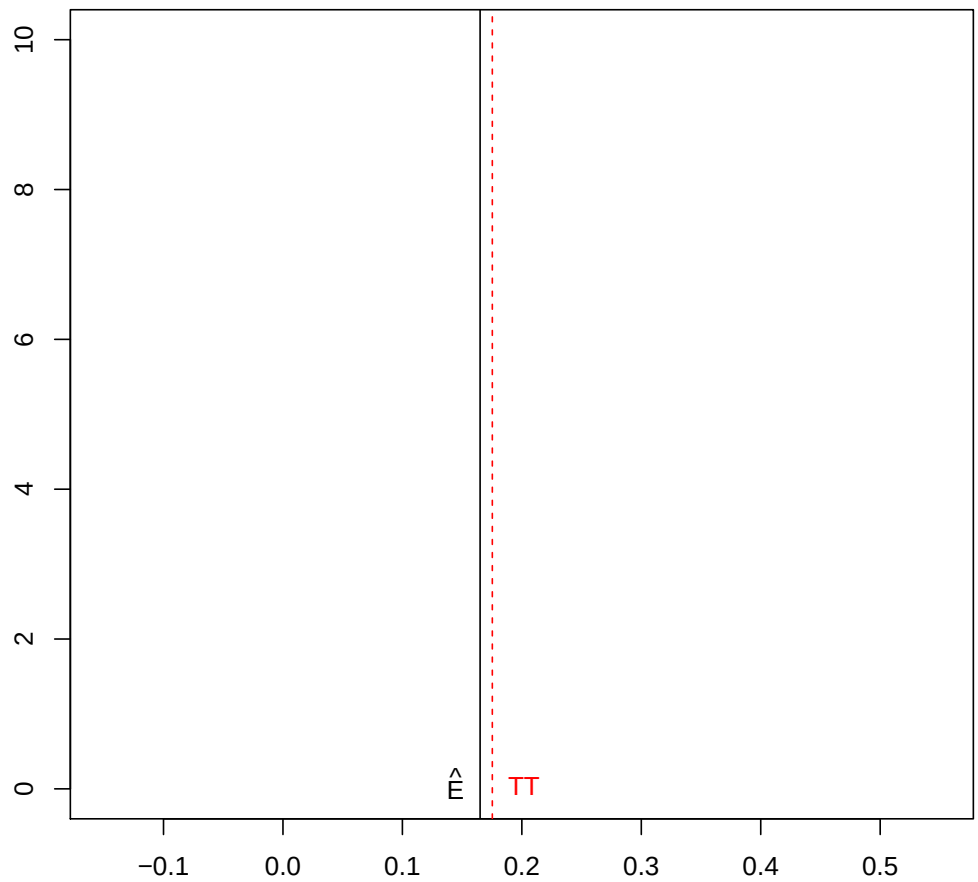


Figure: LLR Matching with Trimming

with bandwidth=0.5, gaussian kernel and trimming level 5%

Sampling Noise with LLR and Trimming: Illustration

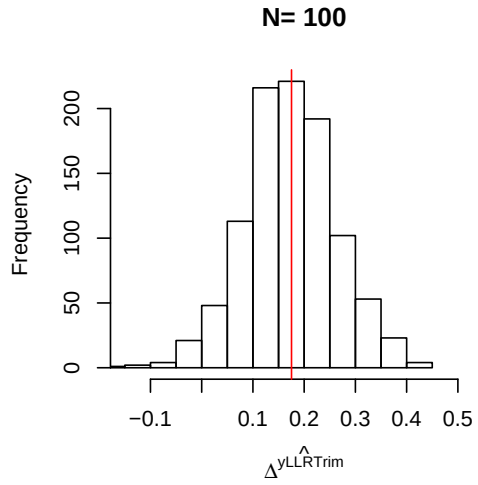


Figure: Distribution of the trimmed *LLR* estimator over replications of samples of different sizes

How to Choose the Bandwidth?

- Usual Bias/Variance trade-off
- Some formulas in Frolich (2004), but complex and flat
- Galdo, Smith and Black (2008): use Cross-Validation (possibly weighted by importance of treated)

Cross validation estimates the MSE using leave-one out estimation and searches for the bandwidth having the lower MSE.

Choosing Bandwidth Using Cross-Validation: illustration

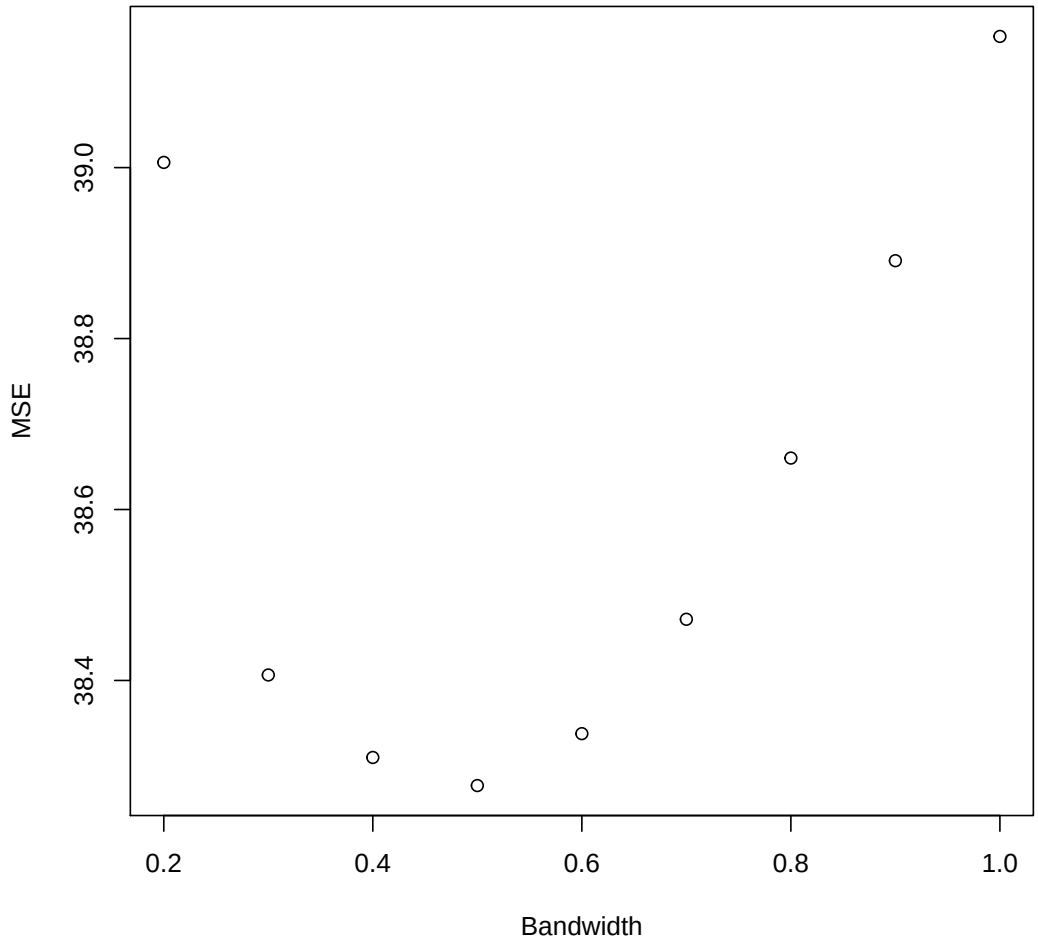


Figure: MSE as a function of Bandwidth for LLR in the untreated sample

What to do When There is More Than One Covariate?

1. Multidimensional kernels
2. Propensity score as dimension reduction device: $\Pr(D_i = 1|X_i) = P(X_i)$.

The Propensity Score as a Dimension Reduction Device

Theorem (Sufficiency of Propensity Score)

$(Y_i^1, Y_i^0) \perp\!\!\!\perp D_i | X_i \Rightarrow (Y_i^1, Y_i^0) \perp\!\!\!\perp D_i | P(X_i)$.

Proof

$$\begin{aligned}\Pr(D_i = 1|Y_i^1, Y_i^0, P(X_i)) &= \mathbb{E}[\Pr(D_i = 1|Y_i^1, Y_i^0, X_i)|Y_i^1, Y_i^0, P(X_i)] \\ &= \mathbb{E}[\Pr(D_i = 1|X_i)|Y_i^1, Y_i^0, P(X_i)] \\ &= P(X_i) = \Pr(D_i = 1|X_i)\end{aligned}$$

Rosenbaum and Rubin (1983)

Propensity Score Matching in Practice

1. Estimate logit or probit: $D_i = \mathbb{1}[\gamma_0 + \gamma_1' X_i + \zeta_i \geq 0]$
2. Compute predicted values: $P(\hat{X}_i)$
3. Use $P(\hat{X}_i)$ as control variable in LLR Matching

Propensity Score: Illustration

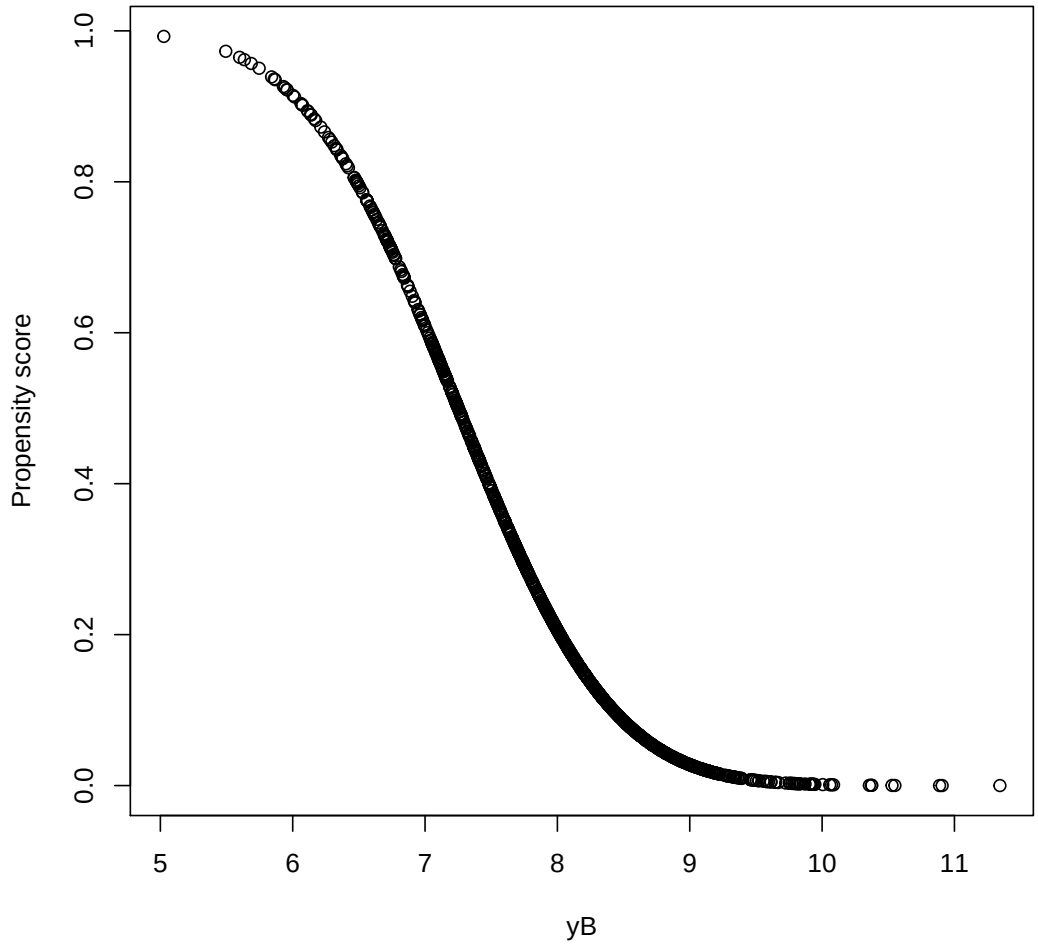


Figure: Propensity score as a function of yB

Propensity Score: Illustration

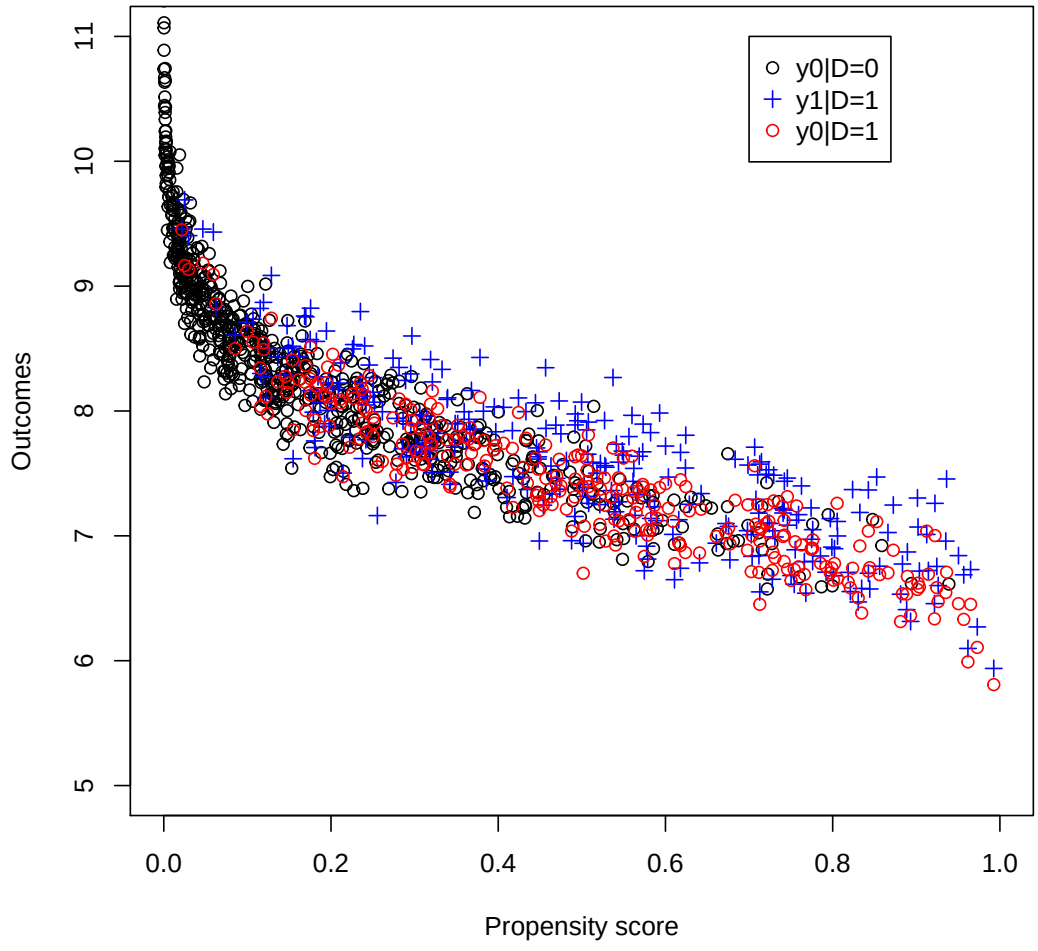


Figure: Propensity score and outcomes

LLR Matching on the Propensity Score

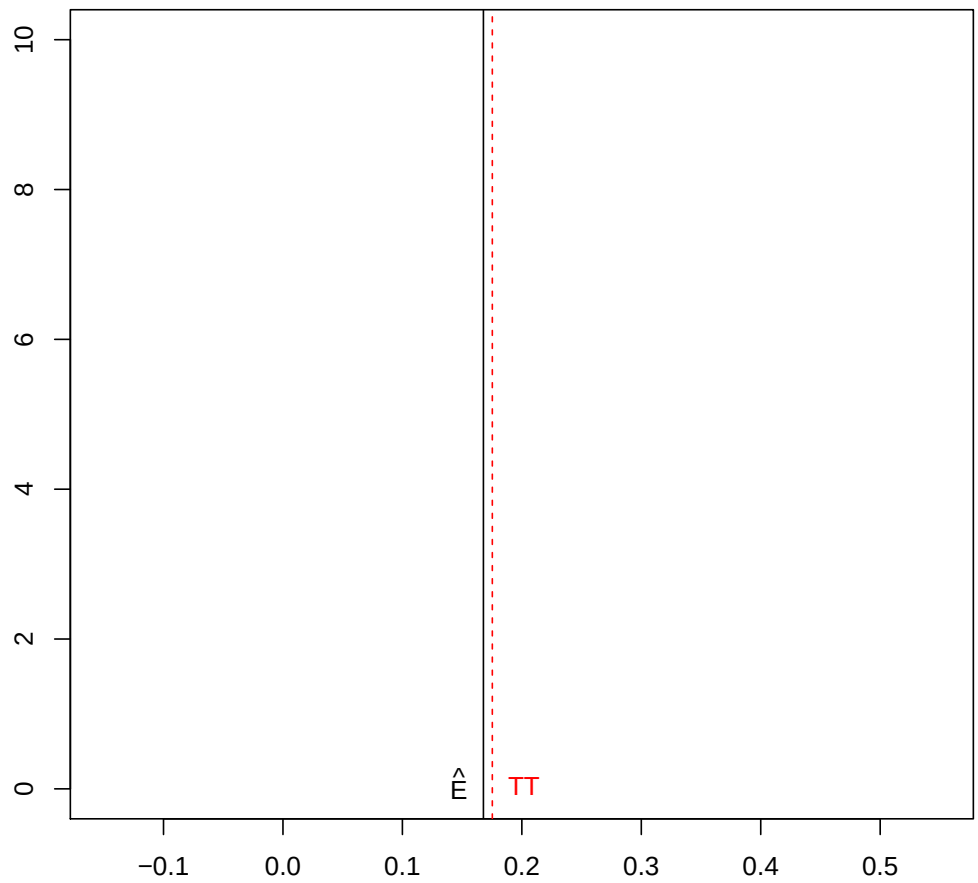


Figure: LLR Matching on the Propensity Score

with bandwidth=0.02, gaussian kernel and trimming level 5%

Local Averaging Matching

$$w_{i,j} = \frac{K\left(\frac{X_i - X_j}{h}\right)}{\sum_{j \in \mathcal{I}^0} K\left(\frac{X_i - X_j}{h}\right)}$$

More simple than LLR: just a weighted average. Problem: bias on boundaries (we will see this again with RDD).

Nearest Neighbor Matching

1. Choose the M closest untreated units for each treated observation i on the common support
2. Compute the average outcome of the M twins: this is the imputed counterfactual mean for i
3. Take the average over all treated i on the common support

Choice of Metric

- Euclidean distance

$$d(i,j) = \sqrt{\sum_k (X_i^k - X_j^k)^2}$$

- Mahalanobis distance

$$d(i,j) = \sqrt{(X - \bar{X})' S^{-1} (X - \bar{X})}$$

with S the covariance matrix of X

- Normalized Euclidean distance

$$d(i,j) = \sqrt{\sum_k \frac{(X_i^k - X_j^k)^2}{\mathbb{V}[X^k]}}$$

- Propensity score distance

$$d(i,j) = \sqrt{(P(X_i) - P(X_j))^2} = |P(X_i) - P(X_j)|$$

- Genetic distance (Sekhon)

Choice of Neighbors

- The m^{th} closest neighbor

$$j_m(i) = \left\{ j : D_j = 1 - D_i \;\&\; \sum_{k:D_k=1-D_i} \mathbb{1}[d(i,k) \leq d(i,j)] = m \right\}$$

- The set of M closest neighbors

$$\mathcal{J}_M(i) = \{j_1(i), \ldots, j_M(i)\}$$

- Number of times observation j is used as a neighbor (matching with replacement)

$$\mathcal{K}_M(j) = \sum_{i=1}^{N} \mathbb{1}[j \in \mathcal{J}_M(i)]$$

Weights

- M nearest-neighbors pair matching with replacement

$$w_{i,j} = \begin{cases} \frac{1}{M} & \text{if } j \in \mathcal{J}_M(i) \\ 0 & \text{if } j \notin \mathcal{J}_M(i) \end{cases}$$

- As a consequence, we can rewrite the estimator of M nearest-neighbors matching with replacement

$$\begin{aligned} \hat{\Delta}^Y_{NNM} &= \frac{1}{N_1^S} \sum_{i \in \mathcal{I}^1 \cap S} \left(Y_i - \sum_{j \in \mathcal{I}^0} w_{i,j} Y_j \right) \\ &= \frac{1}{N_1^S} \sum_{i \in S} \left(D_i - (1 - D_i) \frac{\mathcal{K}_M(i)}{M} \right) Y_i \end{aligned}$$

Weighted matching: in practice

$$w_{i,j} = \frac{\frac{\hat{P}(X_j)}{1-\hat{P}(X_j)}}{\sum_{j \in \mathcal{I}_0} \frac{\hat{P}(X_j)}{1-\hat{P}(X_j)}} = w_j$$

- Weights do not depend on i, we can thus write:

$$\begin{aligned}\hat{\Delta}^{Y}_{TT} &= \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} \left(Y_i - \sum_{j \in \mathcal{I}_0} w_{i,j} Y_j \right) \\ &= \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} Y_i - \sum_{j \in \mathcal{I}_0} w_j Y_j\end{aligned}$$

- Very simple and quick estimator: needs only two separate summations instead of two nested summations

Estimating Precision

- LLR, Local Averaging and Reweighting Matching reach the semi-parametric efficiency bound (while Nearest Neighbour does NOT): can use formula from Hahn to compute precision.
- Bootstrap is calid for LLR, Local Averaging and Reweighting Matching, but NOT for Nearest Neighbor (Abadie and Imbens, 2008)
- Subsampling is valid for all estimators
- Abadie and Imbens (2006) propose a variance estimator for Nearest Neighbor Matching based on the CLT

Precision of LLR Matching: Illustration

- with 100 obs, the true precision of LLR Matching is 0.6072477, while it is of 0.4527886 with trimming.
- With 1000 observation, the true precision of LLR Matching without trimming is 0.151929 and the precision estimated using the bootstrap with 100 replications is 0.1364327.

Outline

Outline

Exercises with generated data: OLS

- With the sharp eligibility design, use the OLS Adjusted With/Without comparison to recover the TT
- Use linear OLS conditioning on y^B to obtain TT. Compare with the previous result.
- Use linear OLS conditioning on $y^B - \bar{y^B}$ to obtain TT. Compare to the previous results.
- Estimate sampling noise for linear OLS using heteroskedasticity robust CLT based approximation.
- Estimate sampling noise for the OLS Adjusted With/Without comparison using the bootstrap. Compare with sampling noise in the previous case.
- With the fuzzy eligibility design and quadratic link between outcomes and y^B, repeat all previous estimates.

Exercises with generated data: Matching

- Generate data with the fuzzy eligibility design and quadratic link between outcomes and y^B.
- Estimate the propensity score using probit.
- Compute the density of the propensity score conditional on $D_i = 1$ and $D_i = 0$
- Compute the trimming dummy with trimming level 0.05.
- Estimate TT using LLR on the propensity score
- Estimate sampling noise for the LLR estimator using the bootstrap.
- Advanced: Estimate sampling noise using Hahn's efficiency bound
- Advanced: Estimate TT using NNM (Matching package in R)
- Advanced: Estimate sampling noise using Abadie and Imbens CLT based approximation (Matching package)
- Advanced: Estimate TT using Reweighting matching (no package) with and without trimming
- Advanced: Estimate sampling noise for this estimator by bootstrap

Exercises with real data: OLS

For eligibility design, on the arm where $R_i = 1$ and for both arms for encouragement designs

- Compute the effect of the treatment on at least one outcome using the OLS Adjusted With/Without comparison
- Do the same with OLS conditioning on X
- Do the same with OLS conditioning on $X - \bar{X}$
- Estimate sampling noise using heteroskedasticity robust CLT-based approximation for OLS
- Estimate sampling noise using the bootstrap for the OLS Adjusted With/Without comparison

Exercises with real data: Matching

- Choose one treatment arm where the proportion of treated is large but not too large.
- Estimate the propensity score
- Compute the density of the propensity score conditional on $D_i = 1$ and $D_i = 0$
- Compute the trimming dummy with trimming level 0.05.
- Estimate TT using LLR on the propensity score
- Estimate sampling noise for the LLR estimator using the bootstrap.
- Advanced: Estimate sampling noise using Hahn's efficiency bound
- Advanced: Estimate TT using NNM (Matching package in R)
- Advanced: Estimate sampling noise using Abadie and Imbens CLT based approximation (Matching package)
- Advanced: Estimate TT using Reweighting matching (no package) with and without trimming
- Advanced: Estimate sampling noise for this estimator by bootstrap