# Applied Microeconomics - Pset 1

Jeanne Sorin *

October 20, 2020

## Problem 1

Let's first get things straight about what we're looking at here.
Writing $Y$ in terms of potential outcomes (and constant treatment effect)

$$\begin{aligned} Y &= DY(1) + (1-D)Y(0) \\ &= Y(0) + D(Y(1) - Y(0)) \\ &= Y(0) + \alpha D \qquad\qquad \alpha \equiv Y(1) - Y(0) \end{aligned}$$

Regression 1: Y on D and a constant:

$$\begin{aligned} Y &= a_1 + \beta D + \varepsilon_1 \\ \hat{D} &\equiv D - BLP(D|1) = D - \mathbb{E}(D) \end{aligned}$$

Regression 2: Y on D, X and a constant:

$$\begin{aligned} Y &= a_2 + \beta D + \zeta X + \varepsilon_2 \\ \tilde{D} &\equiv D - BLP(D|1, X) = D - \mathbb{E}(D) - \frac{cov(D,X)}{var(X)}X \end{aligned}$$

In the most general case, we get the standard regression formula for $\beta$

$$\begin{aligned} \beta &= \frac{cov(Y, \hat{D})}{var(\hat{D})} \\ &= \alpha + \frac{cov(Y(0), \hat{D})}{var(\hat{D})} && \text{using the potential outcome def of Y} \\ &= \alpha + \frac{cov(D, Y(0))}{var(D)} && \text{using } var(\hat{D}) = var(D) \\ & && \text{using } cov(\hat{D}, D) = var(D) \end{aligned}$$

and for $\gamma$

$$\begin{aligned} \gamma &= \frac{cov(\tilde{D}, Y)}{var(\tilde{D})} && \tilde{D} \equiv D - a_2 - \zeta X \\ &= \frac{cov(\tilde{D}, Y(0) + \alpha D)}{var(\tilde{D})} && \text{using potential outcome def (and constant TE)} \\ &= \frac{cov(\tilde{D}, Y(0) + \alpha\tilde{D} + \alpha BLP(D|1, X))}{var(\tilde{D})} \\ &= \alpha + \frac{cov(\tilde{D}, Y(0)}{var(\tilde{D})} && \text{using } \tilde{D} \perp BLP(D|1, X) \end{aligned}$$

*I thank Tom Hierons, Tanya Rajan, Sun Yixin, Camilla Schneier and George Vojta, for comments and discussions.

Moreover define $BLP(D|1) = D - E(D)$, $BLP(D|1, X) = \zeta_1 + \zeta_2 X$, where $\zeta_2 = \frac{cov(D,X)}{var(X)}$

Therefore, we get

$$|\alpha - \beta| = |\frac{cov(D, Y(0))}{var(D)}| = |\frac{cov(\hat{D}, Y(0)}{var(\hat{D})}|$$

$$|\alpha - \gamma| = |\frac{cov(\tilde{D}, Y(0)}{var(\tilde{D})}|$$

(a) *Is it true that $|\alpha - \gamma| \leq |\alpha - \beta|$? If so, prove it. If not, find a counterexample.*
No it is not necessarily true as it depends of the relative covariances between $X, D, Y(0)$ and $X, D$. Indeed, we could have the case where

$$|\alpha - \beta| < |\alpha - \gamma|$$

$$|\frac{Cov(Y(0), D)}{var(D)}| < |\frac{Cov(Y(0), D - \frac{cov(X,D)}{var(X)}X)}{var(D - BLP(D|1, X))}|$$

(b) *Suppose that D and X are uncorrelated. Does this change the answer to a)?*
Yes in this case the inequality holds as $cov(D, X) = 0$, so $var(\tilde{(D)}) = var(D)$ and $cov(Y(0), D - \frac{cov(X,D)}{var(X)}X) = cov(Y(0), D)$, so $|\alpha - \beta| = |\alpha - \gamma|$.

(c) *Suppose that X is uncorrelated with $Y(0)$ and $Y(1)$. Does this change the answer to a)?*
This doesn't change the answer to a), as we are still unable to get rid of $cov(D, X)$ induced by $BLP(D, X)$.
Naively, one could think that including a regressor $X$ that is uncorrelated with the outcome variable $Y(0), Y(1)$ would not change anything. However, including unnecessary variables could bias the estimate in one way or another as they could be correlated with D. In other words, by adding X, we potentially estimate a coefficient for X that captures how X and D co-move and therefore wrongly captures some of the effects of D on Y.
In the equation for $|\alpha - \gamma|$ the numerators become equal, but not the denominators (could go both ways).

(d) *Suppose that $\mathbb{E}[Y(0)|D = d, X = x] = \mathbb{E}[Y(0)|X = x]$. Does this change the answer to a)?*
Unfortunately we don't get the inequality yet: the assumption gives us that $cov(Y(0), D|X) = 0$, but as we don't know anything about the form of $E(Y(0)|X = x)$ (in particular we don't know whether $Y(0)$ is linear in X), and the conditional mean independence doesn't imply the mean independence of $Y(0)$ and $D$,we cannot simplify $\gamma$.

(e) *Suppose that $\mathbb{E}[Y(0)|X = x]$ is a linear function of x. Does this change the answer to a)?*
This doesn't change the answer to a) as it doesn't get rid of the $cov(D, X)$ which is causing our bias.

# Problem 2

We have

$$\begin{aligned} Y &= WX + (1-W)Z \\ W &\perp (X, Z) \\ W &\in \{0, 1\} \\ F(y) &= P(Z \leq y) \\ G(y) &= P(Z \leq y) \end{aligned}$$

(a) Let's play with the definition of Y for a bit

$$\begin{aligned} & Y = WX + (1-W)Z \\ \Rightarrow\ & P(Y \leq y) = P(WX + (1-W)Z \leq y) \\ & G(y) = P(W=1)P(X \leq y) + (1 - P(W=1))P(Z \leq y) && \text{using } W \in \{0,1\} \\ & && \text{and } (X,Z)| \perp W \\ & G(y) = \pi P(X \leq y) + (1-\pi)F(y) \\ & \frac{G(y) - \pi P(X \leq y)}{(1-\pi)} = F(y) \end{aligned}$$

On the above expression, notice that $0 \leq P(X \leq y) \leq 1$ so

$$\begin{aligned} min(\frac{G(y) - \pi P(X \leq y)}{(1-\pi)}) &= \frac{G(y) - \pi}{(1-\pi)} && P(X \leq y) = 1 \\ max(\frac{G(y) - \pi P(X \leq y)}{(1-\pi)}) &= \frac{G(y)}{(1-\pi)} && P(X \leq y) = 0 \end{aligned}$$

Moreover, noticing that $0 \leq F(y) \leq 1$ as F(y) is a CDF, and without further assumption on X, this gives us (strict) bounds on F(y):

$$max(0, \frac{G(y) - \pi}{(1-\pi)}) \leq F(y) \leq min(1, \frac{G(y)}{(1-\pi)})$$

In order to make the lower bound sharp, fix $X < y$ such that $P(X \leq y) = 1$, so $F(y) = \frac{G(y)-\pi}{1-\pi}$. Similarly, to make the upper bound sharp, fix $X > y$ such that $P(X \leq y) = 0$, so $F(y) = \frac{G(y)}{1-\pi}$.

(b) We now assume that Y, X, Z are all continuously distributed, and we let $G^{-1}(q)$ denote the $q^{th}$ quantile of Y. We show that

$$\mathbb{E}[Y|Y \leq G^{-1}(1-\pi)] \leq \mathbb{E}[Z] \leq \mathbb{E}[Y|Y \geq G^{-1}(\pi)]$$

Let's look at each bound successively. First the lower bound

$$\begin{aligned} \mathbb{E}[Y|Y \leq G^{-1}(1-\pi)] &= \mathbb{E}[Y|Y \leq \gamma] && \gamma \equiv G^{-1}(1-\pi)] \\ &= \mathbb{E}[WX + (1-W)Y|Y \leq \gamma] \\ &= P(W=1|Y \leq \gamma)E[X|X \leq \gamma] + P(W=0|Y \leq \gamma)E[Z|Z \leq \gamma] && LTE \\ &= \frac{P(Y \leq \gamma|W=1)\pi}{P(Y \leq \gamma)}E[X|X \leq \gamma] + \frac{P(Y \leq \gamma|W=0)(1-\pi)}{P(Y \leq \gamma)}E[Z|Z \leq \gamma] && \text{Bayes Rule} \\ &= \frac{P(X \leq \gamma)\pi}{P(Y \leq \gamma)}E[X|X \leq \gamma] + \frac{P(Z \leq \gamma)(1-\pi)}{P(Y \leq \gamma)}E[Z|Z \leq \gamma] \end{aligned}$$

Noticing that $P(Y \leq \gamma) = \pi P(X \leq \gamma) + (1-\pi)P(Z \leq \gamma)$, we can replace for $P(X \leq \gamma)$ $pi = P(Y \leq \gamma) - (1-\pi)P(Z \leq \gamma)$, and that $P(Y \leq \gamma) \equiv 1 - \pi$

$$\begin{aligned}
\mathbb{E}[Y|Y \leq \gamma] &= \frac{1}{1-\pi}(E[X|X \leq \gamma](1-\pi) - E[X|X \leq \gamma](1-\pi)P(Z \leq \gamma) + (1-\pi)P(Z \leq \gamma)E[Z|Z \leq \gamma]) \\
&= \frac{1}{1-\pi}(E[X|X \leq \gamma](1-\pi)(1-P(Z \leq \gamma)) + (1-\pi)P(Z \leq \gamma)E[Z|Z \leq \gamma]) \\
&= E[X|X \leq \gamma](1-P(Z \leq \gamma)) + P(Z \leq \gamma)E[Z|Z \leq \gamma] \\
&\leq E[Z|Z \geq \gamma]P(Z \geq \gamma) + P(Z \leq \gamma)E[Z|Z \leq \gamma] \\
&= E(Z)
\end{aligned}$$

The lower bound is sharp if $\pi = 1 W = 1$ with probability 1, $Y = max(Y)$

Similarly for the upper bound with $W = 0$ with probability 1, $\pi = 0$, $Y = min(Y)$.

(c) This question refers to David Lee's 2005 NBER WP[1] and uses insights / lemmas from questions a and b.

Using the independence assumption, and the definition of $Y$ we get that

$$\begin{aligned}
\mu &= E(Y(1)|S(0) = 1, S(1) = 1) - E(Y(0)|S(0) = 1, S(1) = 1) \\
&= E(Y(1)|S(0) = 1, S(1) = 1) - E(Y|S = 1, D = 0)
\end{aligned}$$

We can therefore rewrite the WTS expression as

$$E(Y|D = 1, S = 1, Y \leq \bar{G}^{-1}(1-\pi)) \leq E(Y(1)|S(0) = 1, S(1) = 1) \leq E(Y|D = 1, S = 1, Y \geq \bar{G}^{-1}(\pi))$$

Noticing that $\bar{G}(y)$ being the CDF of Y conditional on $D = 1, S(0) = 1, S(1) = 1$ implies that

$$\begin{aligned}
\bar{G}(y) &= P(Y \leq y|S = 1, D = 1) \\
&= P(Y(1) \leq y|S = 1, D = 1) \\
&= \pi P(Y(1) \leq y|D = 1, S(1) = 1, S(0) = 0) + (1-\pi)P(Y(1) \leq y|S(1) = 1, S(0) = 1)
\end{aligned}$$

Where the last equality uses the law of total expectations, Bayes rule and $Y(0) \perp D$ As, using monotonicity and independence we can rewrite

$$\begin{aligned}
\pi &= \frac{P(S = 1|D = 1) - P(S = 1|D = 0)}{P(S = 1|D = 1)} \\
&= \frac{P(S(1) = 1, S(0) = 0|D = 1)}{P(S = 1|D = 1)}
\end{aligned}$$

Looking at the upper bounds (RHS), and defining $P(Y(1) \leq y|S(1) = 1, S(0) = 1)$ to be $F(y)$ from before, we can apply part b and we obtain

$$E(Y(1)|S(0) = 1, S(1)) \leq E(Y|D = 1, S = 1, Y \geq \bar{G}^{-1}(\pi))$$

Similarly for the lower bound. We therefore get the desired expression.
Sharpness of the bounds follow the same argument as in the previous question.

[1] https://www.nber.org/papers/w11721.pdf

# Problem 3

**Part 1:** *Show, for any jointly supported $d \in D$ and $p \in (0,1)$:*

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

Let's define $Y$ in terms of potential outcomes, with $D$ discrete:

$$Y = \sum_{d'} Y(d')\mathbf{1}\{D=d'\}$$

Taking the conditional expectation on both sides

$$\begin{aligned}
\mathbb{E}[Y|D=d,P_d=p] &= \sum_{d'} \mathbb{E}[Y(d')\mathbf{1}\{D=d'\}|D=d,P_d=p] && \text{linearity of the expectation operator}\\
&= \sum_{d'} \mathbb{E}[Y(d')|D=d,P_d=p]\mathbb{E}[\mathbf{1}\{D=d'\}|D=d,P_d=p] && \mathbf{1}\{D=d'\}|D=d \text{ deterministic}\\
&= \mathbb{E}[Y(d)|D=d,P_d=p] + \sum_{d'\neq d} \mathbb{E}[Y(d')|D=d,P_d=p]\cdot 0 &&\\
&= \mathbb{E}[Y(d)|D=d,P_d=p] &&\\
&= \mathbb{E}[\mathbb{E}[Y(d)|D=d,P_d=p,X=x]|D=d,P_d=p] && \text{LIE}\\
&= \mathbb{E}[\mathbb{E}[Y(d)|D=d,X=x]|D=d,P_d=p] && \text{redundant}\\
&= \mathbb{E}[\mathbb{E}[Y(d)|D=d,X=x]|P_d=p] && \text{redundant}\\
&= \mathbb{E}[\mathbb{E}[Y(d)|X=x]|P_d=p] && F(Y(d)|X=x)=F(Y(d)|X=x,D=d)
\end{aligned}$$

where $F(.)$ is the distribution of $Y(d)$.

$$\begin{aligned}
\mathbb{E}[Y|D=d,P_d=p] &= \mathbb{E}[\mathbb{E}[Y(d)|P_d=d,X=x]|P_d=p] && \text{LIE}\\
\mathbb{E}[Y|D=d,P_d=p] &= \mathbb{E}[\mathbb{E}[Y(d)|P_d=d]|P_d=p] && \text{redundant}\\
\mathbb{E}[Y|D=d,P_d=p] &= \mathbb{E}[Y(d)|P_d=d]
\end{aligned}$$

*Explain what this result shows and why it is significant for empirical practice.*
This is an important result in practice because it shows that conditonal on the propensity score $P(D=d|X=x)=P_d$, the expected value of $Y|D,P_d=p$ is the same as the expected value of $Y(d)|P_d=p$. This shifts the curse of dimensionality to estimating the propensity score.

**Part 2**:
*Let $P=p(D,X)$ and show that (where we assume $P>0$ with probability 1)*

$$\mathbb{E}[Y(d)] = \mathbb{E}[\frac{Y\mathbf{1}[D=d]}{p}]$$

**Way 1:**
Let's define $Y$ in terms of potential outcomes, with $D$ discrete:

$$\begin{aligned}
Y &= \sum_{d'} Y(d')\mathbf{1}\{D=d'\}\\
Y\mathbf{1}\{D=d\} &= \sum_{d'} Y(d')\mathbf{1}\{D=d'\}\mathbf{1}\{D=d\}\\
Y\mathbf{1}\{D=d\} &= Y(d)\mathbf{1}\{D=d\} \text{ as for all } d'\neq d: \mathbf{1}\{D=d'\}\mathbf{1}\{D=d\}=0
\end{aligned}$$

Dividing by $P = P(D = d|X = x)$ on both sides and taking the expectation

$$\mathbb{E}[\frac{Y\mathbf{1}\{D = d\}}{P}] = \mathbb{E}[\frac{Y(d)\mathbf{1}\{D = d\}}{P}]$$

Focusing on the RHS

$$\begin{aligned}
\mathbb{E}[\frac{Y(d)\mathbf{1}\{D = d\}}{P}] &= \mathbb{E}[\mathbb{E}[\frac{Y(d)\mathbf{1}\{D = d\}}{P}|D = d, X = x]] && \text{LIE} \\
&= \mathbb{E}[\frac{\mathbb{E}[Y(d)|X = x, D = d]\mathbb{E}[\mathbf{1}\{D = d\}|X = x, D = d]}{\mathbb{E}[P(D = d|X = x)|X = x, D = d}] && P(D = d\,|X = x) \perp Y(d)|X = x, D = d \\
& && \mathbf{1}\{D = d\} \perp Y(d)|D = d, X = x \\
&= \mathbb{E}[\frac{\mathbb{E}[Y(d)|X = x, D = d]P(D = d|X = x)}{P(D = d|X = x)}] && D \text{ binary and LIE} \\
&= \mathbb{E}(Y(d)) && \text{LIE}
\end{aligned}$$

Replacing the RHS from before by $\mathbb{E}(Y(d))$ we obtain

$$\mathbb{E}[\frac{Y\mathbf{1}\{D = d\}}{P}] = \mathbb{E}(Y(d))$$

**Way 2**:

$$\begin{aligned}
\mathbb{E}[Y(d)] &= \mathbb{E}[\mathbb{E}[Y(d)|X]] \\
&= \mathbb{E}[\mathbb{E}[Y(d)|X, D = d]] && \text{independence} \\
&= \mathbb{E}[\mathbb{E}[Y|X, D = d]] \\
&= \mathbb{E}[\frac{\mathbb{E}[Y|X, D = d]P}{P}] && P \equiv P(D = d|X = x) \\
&= \mathbb{E}[\frac{\mathbb{E}[Y\mathbf{1}\{D = d\}|X]}{P}] && \text{Conditional expectation} \\
&= \mathbb{E}[\mathbb{E}[\frac{Y\mathbf{1}\{D = d\}}{P}]] && \text{redundance} \\
&= \mathbb{E}[\frac{Y\mathbf{1}\{D = d\}}{P}]
\end{aligned}$$

# Problem 4

$$
\begin{aligned}
Y &= sin(2X) + 2\exp(-16X^2) + U \\
U &\sim \mathcal{N}(0, .3) \\
X &\sim U[-2, 2] \\
m(x) &= \mathbb{E}[Y|X = x]
\end{aligned}
$$

We want to estimate $m(x)$ nonparametrically by conducting a Monte Carlo simulation that demonstrates the bias-variance trade-off in the context of nonparametric regression.
As one can show below, we face a trade off between low bias - high variance (left panel) approach and high bias - low variance (right panel) approach. Interestingly enough, this trade off seems to be more or less pronounced depending on the different non-parametric approaches.
The graphs below plot the true distribution of $Y$ (black line) versus the average projected $Y$ using the non-parametric method (red dots) in a 500 iterations Monte Carlo simulation, and the one-standard deviation bounds (blue dots) of these projected $Y$. All formula are taken from A. Torgovitsky's slides.

- Local constant (kernel) regression: I use the uniform kernel.

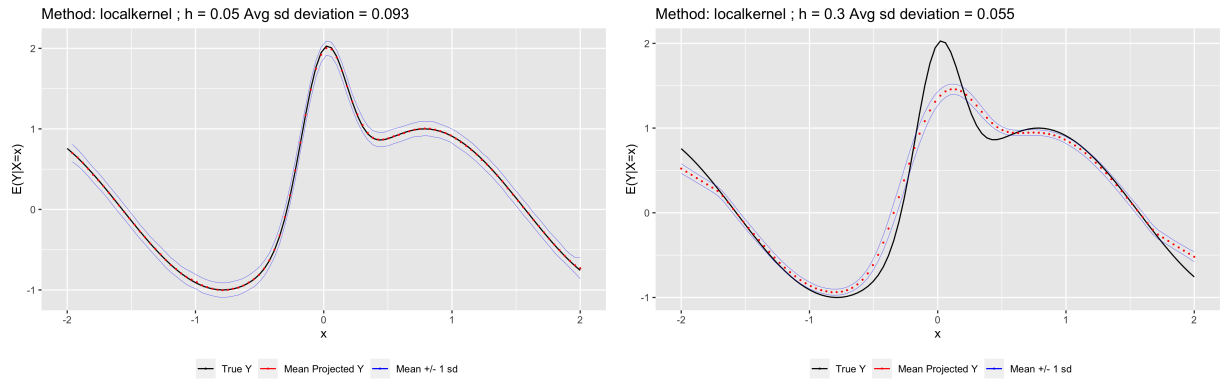


Figure 1: Trade-off bias-variance for the Local Kernel Approach

- Local linear regression

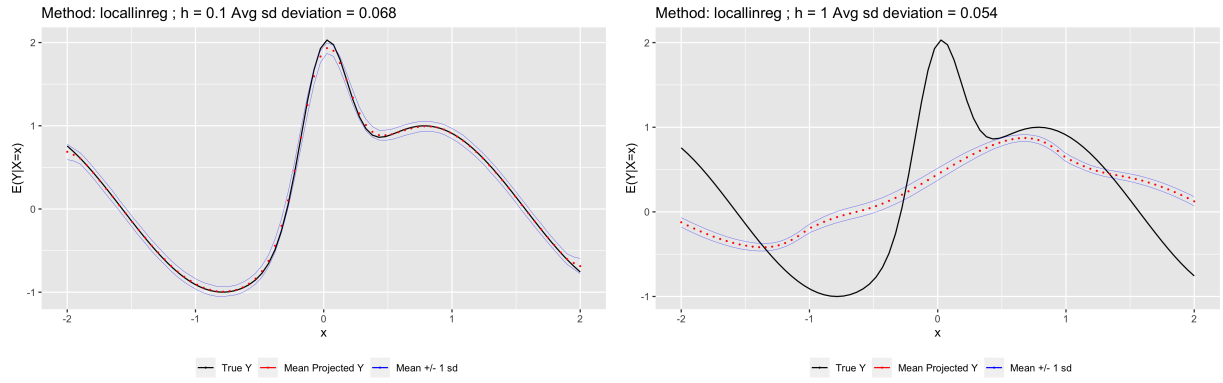


Figure 2: Trade-off bias-variance for the Local Linear Regression Approach

- A sieve approximation using the standard polynomial basis

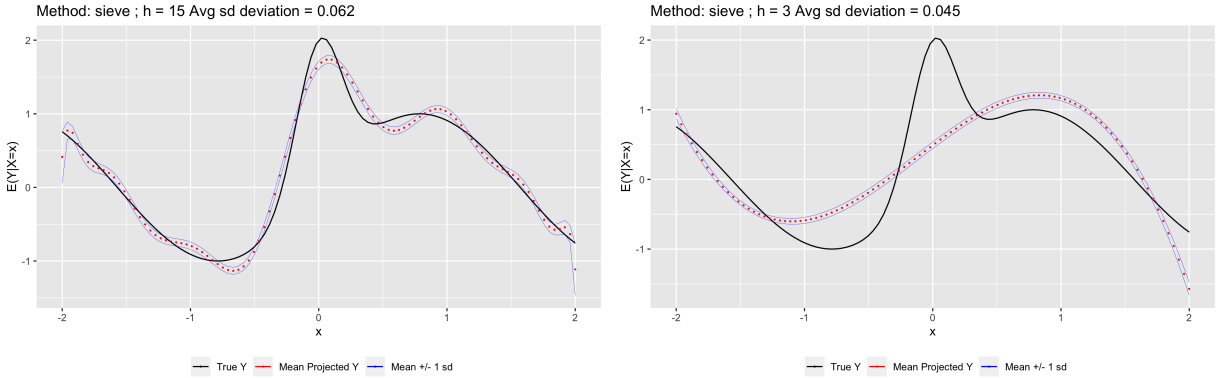


Figure 3: Trade-off bias-variance for the Sieve with Standard Polynomial Basis Approach

- The nearest neighbors estimator

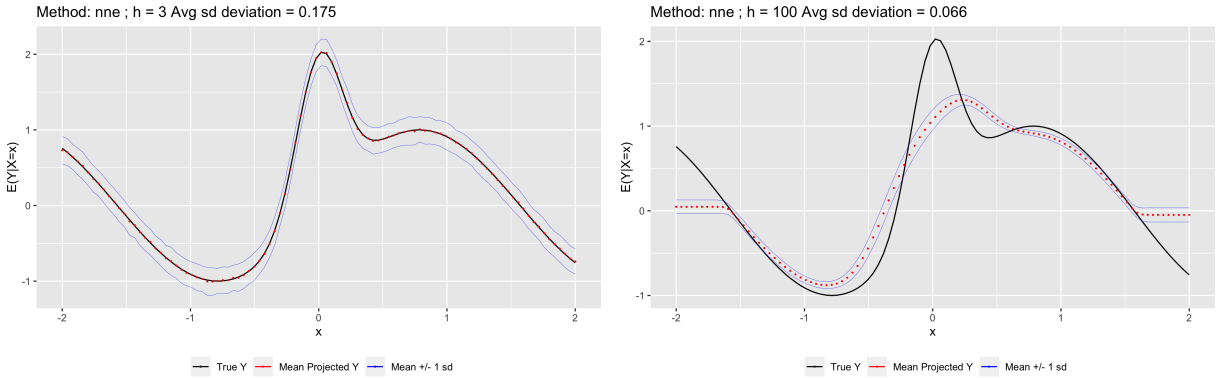


Figure 4: Trade-off bias-variance for the Nearest Neighbors Estimator Approach

- A sieve approximation using the Bernstein polynomial basis

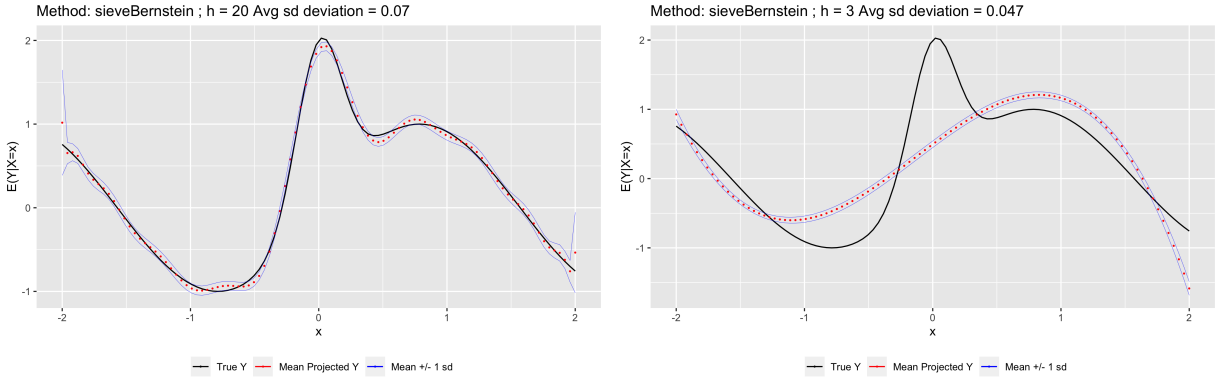


Figure 5: Trade-off bias-variance for the Bernstein Polynomial Basis Approach

- A sieve approximation using linear splines represented in the truncated power basis

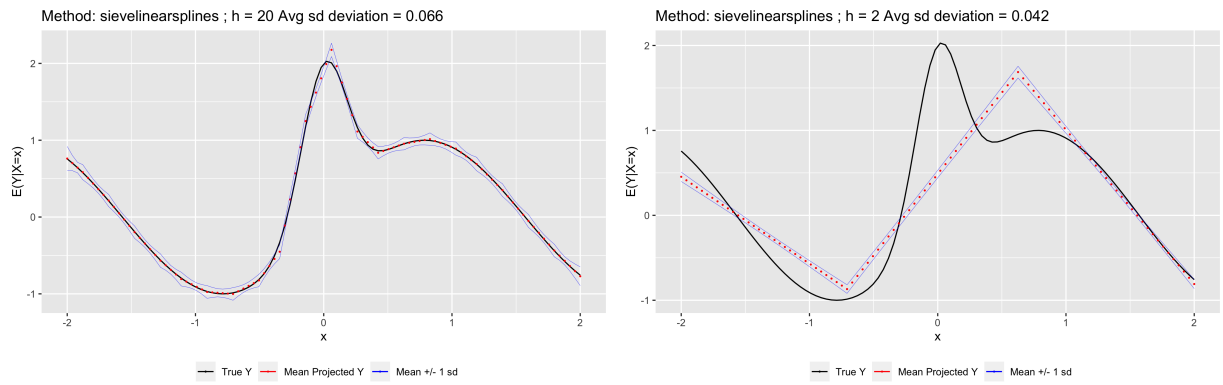


Figure 6: Trade-off bias-variance for the Linear Splines Approach

# Problem 5

a) *The authors seem to argue that conditioning on covariates is important. Given their argument, why is the set of covariates that they use in (1) a bit odd?*

$$AS_i = \alpha + \beta \cdot POG_i^{1349} + \gamma X_i + \varepsilon_i \tag{1}$$

Where $X_i$ contains

- City population
- The percentage of the population that is Jewish
- The percentage of the population that is Protestant

Moreover, footnote (37): *"Protestants were more prone to vote for the Nazi Party (Falter 1991). City population and the share of Jewish population are measured for the year closest to each outcome variable; for the latter variable, data are available for 1925 and 1933. The share of Protestants is available only for 1925. In cases where we do not have city- or town-level observations for control variables, we use county- (Kreis-) level data. Standard errors are clustered at the county level."*

By including the percentage of protestants as a control, while acknowledging that protestants are more prone to vote for the Nazi Party, the authors potentially capture part of the treatment effect through their control.
In other words, one could think of a mechanism where the percentage of protestants / jews in 1925 is somehow affected by the treatment variable: anti-semitism in 1349 may have had an impact on the share of jews in a city in the 20th century (and mechanically share of protestants).

b) *Replicate Comlumn 1 of Table VI.*

Panel A:

$$\text{1920s pogroms} = \hat{\alpha} + \hat{\beta} POG^{1349} + \hat{\gamma_1}\% \text{ Jewish} + \hat{\gamma_2}\% \text{ Protestant} + u_i$$

c) *Implement propensity score matching estimators of both the ATE and ATT, using the same covariates as the authors do. I leave the specifics up to you, but you might consider nearest neighbor matching on the propensity score, and/or a blocking approach. Compare your estimates to the authors' estimates. You may use the bootstrap to compute standard errors.*

Table 1: Voigtlander & Voth (2012), Table VI (Column 1) - Panel A

| Dep. Var | Pogroms 1920s |
|---|---|
| Pogrom 1349 | 0.0607<br>(0.0226) |
| ln(Pop) | 0.039<br>(0.0152) |
| Percent. Jewish | 0.0135<br>(0.0114) |
| Percent. Protestant | 0.00034<br>(0.00042) |
| N | 320 |

Note : The dependent variable is a dummy for whether or not the town had pogroms in the 1920s. Standard errors are clustered at the *kreis* level and are reported in parenthesis below the estimates. The table presents the result for an OLS specification with a constant (non-displayed).

Table 2: Voigtlander & Voth (2012), Table VI (Column 1) - Panel B

| Dep. Var | Pogroms 1920s |
|---|---|
| Pogrom 1349 | 0.0711<br>(0.0214) |
| N | 320 |

Note : The dependent variable is a dummy for whether or not the town had pogroms in the 1920s. The table displays the ATT. Observations are matched using the Mahalanobis distance and the 4 closest neighbors using the same covariates as in Panel A. Standard errors are displayed below estimates in parenthesis and computed using bootstrapping for B=100. Estimates and standard errors are really closed to the authors'. Potential sources of discrepancies are number of bootstrap iterations and measure of distance used.

Table 3: Voigtlander & Voth (2012), Table VI (Column 1) - Panel C

| Dep. Var | Pogroms 1920s |
|---|---|
| Pogrom 1349 | 0.0819<br>(0.0186) |
| N | 320 |

Note : The dependent variable is a dummy for whether or not the town had pogroms in the 1920s. The table displays the ATT. Observations are matched using the Mahalanobis distance and the 2 closest neighbors using coordinates. Note that in order to match the paper's estimates, I perform matching on the coordinates using the mahalanobis distance. However, it is worth emphasizing that the estimates are slightly different if one uses instead the actual geographical distance implied by these estimates (*geodist* in R). Standard errors are displayed below estimates in parenthesis and computed using bootstrapping for B=100. Estimates and standard errors are really closed to the authors'. Potential sources of discrepancies are number of bootstrap iterations and measure of distance used.

Table 4: Voigtlander & Voth (2012), Table VI (Column 1) - Propensity score Matching

| Dep. Var | Pogroms 1920s |
|---|---|
| Pogrom 1349 - Pscore ATT | 0.0733<br>(0.0217) |
| Pogrom 1349 - Pscore ATE | 0.0648<br>(0.0179) |
| N | 320 |

Note : The dependent variable is a dummy for whether or not the town had pogroms in the 1920s. The table displays both the ATT and ATE. Observations are matched on their propensity score, using a logit model, and the 4 closest neighbors. Standard errors are displayed below estimates in parenthesis and computed using bootstrapping for B=100. Estimates and standard errors are really closed to the authors'. Potential sources of discrepancies are number of bootstrap iterations and measure of distance used.