IFAU – INSTITUTE FOR
LABOUR MARKET POLICY
EVALUATION

Non-parametric adjustment for covariates when estimating a treatment effect

Eva Cantoni
Xavier de Luna

WORKING PAPER 2004:9

The Institute for Labour Market Policy Evaluation (IFAU) is a research institute under the Swedish Ministry of Industry, Employment and Communications, situated in Uppsala. IFAU's objective is to promote, support and carry out: evaluations of the effects of labour market policies, studies of the functioning of the labour market and evaluations of the labour market effects of measures within the educational system. Besides research, IFAU also works on: spreading knowledge about the activities of the institute through publications, seminars, courses, workshops and conferences; creating a library of Swedish evaluational studies; influencing the collection of data and making data easily available to researchers all over the country.

IFAU also provides funding for research projects within its areas of interest. There are two fixed dates for applications every year: April 1 and November 1. Since the researchers at IFAU are mainly economists, researchers from other disciplines are encouraged to apply for funding.

IFAU is run by a Director-General. The authority has a traditional board, consisting of a chairman, the Director-General and eight other members. The tasks of the board are, among other things, to make decisions about external grants and give its views on the activities at IFAU. A reference group including representatives for employers and employees as well as the ministries and authorities concerned is also connected to the institute.

Postal address: P.O. Box 513, 751 20 Uppsala
Visiting address: Kyrkogårdsgatan 6, Uppsala
Phone: +46 18 471 70 70
Fax: +46 18 471 70 71
ifau@ifau.uu.se
www.ifau.se

Papers published in the Working Paper Series should, according to the IFAU policy, have been discussed at seminars held at IFAU and at least one other academic forum, and have been read by one external and one internal referee. They need not, however, have undergone the standard scrutiny for publication in a scientific journal. The purpose of the Working Paper Series is to provide a factual basis for public policy and the public policy discussion.

ISSN 1651-1166

Non-parametric adjustment for covariates when estimating a treatment effect *

Eva Cantoni† and Xavier de Luna‡

June 16, 2004

Abstract

We consider a non-parametric model for estimating the effect of a binary treatment on an outcome variable while adjusting for an observed covariate. A naive procedure consists in performing two separate non-parametric regression of the response on the covariate: one with the treated individuals and the other with the untreated. The treatment effect is then obtained by taking the difference between the two fitted regression functions. This paper proposes a backfitting algorithm which uses all the data for the two above-mentioned non-parametric regression. We give theoretical results showing that the resulting estimator of the treatment effect can have lower finite sample variance. This improvement may be achieved at the cost of a larger bias. However, in a simulation study we observe that mean squared error is lowest for the proposed backfitting estimator. When more than one covariate is observed our backfitting estimator can still be applied by using the propensity score (probability of being treated for a given setup of the covariates). We illustrate the use of the backfitting estimator in a several covariate situation with data on a training program for individuals having faced social and economic problems.

Keywords: Analysis of covariance, Backfitting algorithm, Linear smoothers, Propensity score.

JEL: C14

*We are grateful to Markus Frölich and Per Johansson for comments that helped in improving the paper. The work reported in this article was financially supported by the Swedish Institute for Labour Market Policy Evaluation (IFAU), the Swedish Council for Working Life and Social Research, and by the Wikström Foundation at the Umeå School of Business Administration and Economics. Eva Cantoni was, moreover, partially supported by CUS (Conférence Universitaire Suisse) and its partners, through the IRIS program of the University of Lausanne, of Geneva and the Swiss Federal Institute of Technology, Lausanne.

†Department of Econometrics, University of Geneva, 40, Bd du Pont d'Arve,CH-1211 Geneva 4, Switzerland. Email: Eva.Cantoni@metri.unige.ch

‡Department of Statistics, Umeå University, Umeå S-90187, Sweden. Email: Xavier.deLuna@stat.umu.se

1 Introduction

The estimation of the effect of a binary treatment w on an outcome variable y is often performed with the classical linear analysis of covariance[1] when a covariate x must be adjusted for. A more general non-parametric analysis of covariance can be performed by considering the model, for a random sample of size n,

$$y_i = \beta_0(x_i) + w_i\tau(x_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where ϵ_i is the usual regression error term with mean zero, and $\tau(x_i)$ is the conditional treatment effect which is often of main interest. The use of this model can be illustrated with a dataset previously analysed in Young and Bowman (1995) and Ratkowsky (1983), where the logarithm of the yield, y (g/plant) of a variety of Spanish Onion is explained by the covariate density, x (plants/m^2), and what may be called a treatment, that is a binary indicator, w, for two different regions in South Australia. The data consists in 42 observations for each of the two regions Virginia ($w = 0$) and Purnong Landing ($w = 1$), and is displayed in Figure 1 (top left panel) together with a non-parametric fit of the functions $\beta_0(x)$ and $\beta_1(x) = \tau(x) + \beta_0(x)$. Different inferential purposes may be sought with such a fit. For instance, different hypotheses (e.g., $\tau(x)$, is a constant function) may be formally tested, see, e.g., Young and Bowman (1995), Akritas and Van Keilegom (2001) and Neumeyer and Dette (2003). It is also common to provide pointwise confidence bands around non-parametric fits as shown in Figure 1 (dotted lines). Such confidence bands are ± 2 times the standard error of the fitted value at a given design point. They correspond to pointwise 95% confidence intervals for the true curve if bias in estimation is negligible, see Bowman and Azzalini (1997, Sec. 4.4), and Hastie and Tibshirani (1990, Sec. 3.8).

Model (1) is more general than it appears because when more than one covariate must be adjusted for the model can be used by replacing the univariate variable x_i with the propensity score, $\Pr(w_i = 1|x_{1i}, \ldots, x_{pi})$ if p covariates are available, see Rosenbaum and Rubin (1983). This can be done under certain conditions given in Section 3, where an application is also presented.

Model (1) is usually fitted by considering separately the treated ($w_i = 1$) and untreated ($w_i = 0$) individuals. A non-parametric regression technique (e.g. kernels, smoothing splines, etc.; see, for example, Härdle, 1990, Fan and Gijbels, 1996) is used to fit the function $\beta_0(x)$ based solely on the untreated and to fit the function $\beta_1(x)$ based solely on the treated.

In this paper we propose a backfitting algorithm which improves on the above naive estimation of the functions $\beta_0(x)$, $\beta_1(x)$ and $\tau(x)$. This is achieved by using the information contained in both the treated and untreated individuals when estimating both $\beta_0(x)$ and $\beta_1(x)$ non-parametrically. For linear smoothers (e.g. smoothing

[1] That is an additive separable linear regression model.

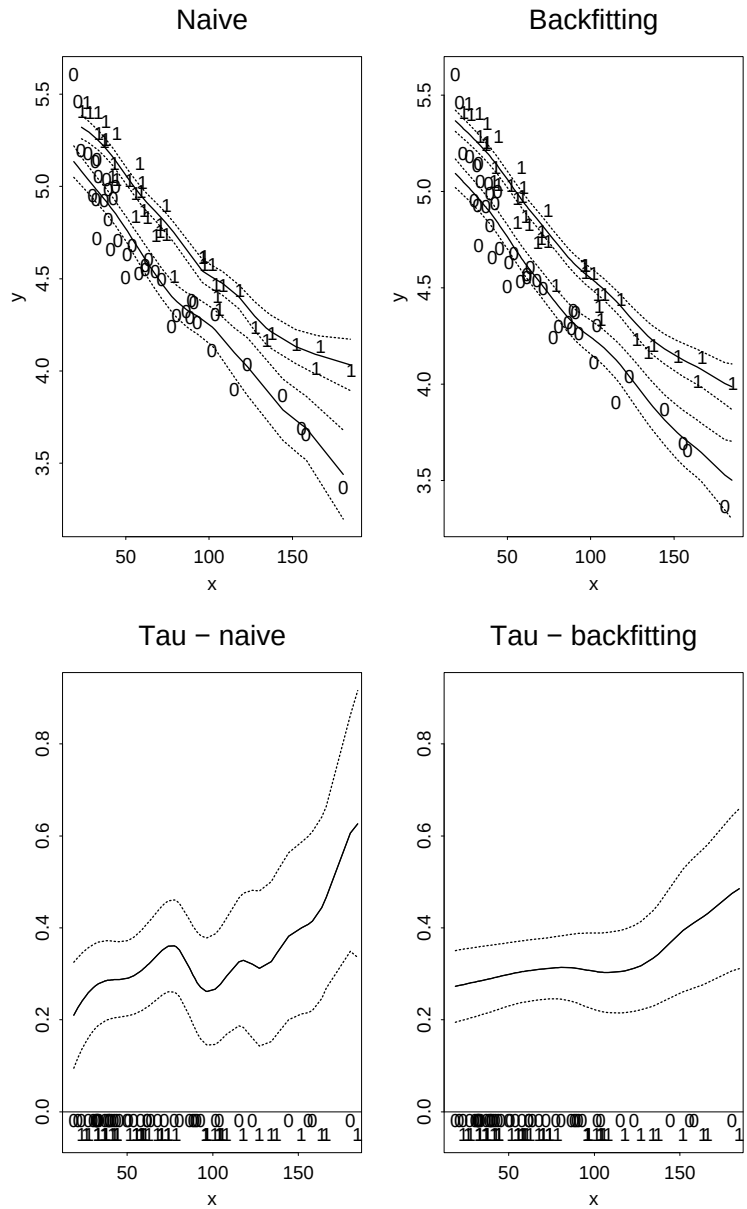

Figure 1: White Spanish Onions dataset. Top panels: Non-parametric fits (gaussian kernel with smoothing parameter $h = 12$) of the functions $\beta_0(x)$ and $\beta_1(x)$ (plain lines). The top left panel displays the fits obtained separately for data on location $w = 0$ and $w = 1$. The top right panel shows the fits obtained with the backfitting algorithm. The bottom panels display the corresponding fits of the function $\tau(x)$. Dotted lines are confidence bands for the fitted functions.

splines, kernels), we show that our algorithm provides an estimator with lower variance under certain conditions. The improvement is illustrated in Figure 1 where both the naive estimators and the backfitting estimators are displayed together with their respective confidence bands. Bias may increase with the backfitting estimator, although, in a simulation study we observe that the decrease in variance is large enough to imply a decrease in mean squared error.

The paper is organised as follow. In the next section we briefly introduce linear smoothers. The backfitting estimator is then presented, followed by finite sample theoretical and simulation results showing the difference in terms of variance and bias between the naive and the novel estimator. Section 3 presents an application where the propensity score of Rosenbaum and Rubin (1983) is utilized to adjust for several covariates. Section 4 concludes the paper.

2 Estimators and properties

2.1 Linear smoothers

Various methods (see, e.g., Hastie and Tibshirani, 1990, Sec. 9.5, Härdle, 1990, Fan and Gijbels, 1996 and Young and Bowman, 1995) can be used to estimate the involved functions in model (1) without making stringent parametric assumptions. Linear smoothers are such methods, including smoothing splines, kernels and local polynomials. They are called linear because the implied fitted values at the design points are linear in the outcome. That is, for a model $y_i = f(x_i) + \varepsilon_i$ for $i = 1, \ldots, n$, the estimation of f at the design points, $\mathbf{x} = (x_1, \ldots, x_n)^T$, is given by

$$\hat{\mathbf{f}}(\mathbf{x}) = S^h[\mathbf{x}]\mathbf{y}, \tag{2}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ contains the observed outcomes, $\hat{\mathbf{f}}(\mathbf{x})$ is the vector containing the fitted values at each x_i, and $S^h[\mathbf{x}]$ is a matrix of weights not depending on the y_i's and depending on a smoothing parameter h.

The results derived in this paper focus on linear smoothers, and, in particular, their kernel representation. A kernel estimator is defined at a generic point z based on a sample (x_i, y_i) for $i = 1, \ldots, n$ by

$$\hat{f}(z) = \frac{\sum_{i=1}^{n} K(\frac{z-x_i}{h}) y_i}{\sum_{i=1}^{n} K(\frac{z-x_i}{h})}, \tag{3}$$

where K satisfies the following conditions $\int K(u)du = 1$ and $\int uK(u)du = 0$. A commonly used kernel smoother is the gaussian kernel, which utilizes the standard normal density as function K. Linear smoothers have an equivalent kernel representation, see, e.g., Hastie and Tibshirani (1990, Sec. 2.8).

In the setting of model (1), naive kernel estimators of the functions β_0 and β_1 are obtained by considering two separate subsamples consisting in the untreated

and treated individuals respectively. Denote by $\mathbf{y}_0 = (y_{01}, \ldots, y_{0n_0})^T$ and $\mathbf{x}_0 = (x_{01}, \ldots, x_{0n_0})^T$ the observed response and covariate values for the n_0 non-treated individuals, and similarly $\mathbf{y}_1 = (y_{11}, \ldots, y_{1n_1})^T$ and $\mathbf{x}_1 = (x_{11}, \ldots, x_{1n_1})^T$ for the n_1 treated units. Then, the fitted values at $\mathbf{x}_0$ and $\mathbf{x}_1$ are $\hat{\beta}_0^{naive}(\mathbf{x}_0) = S_0^{h_0}[\mathbf{x}_0]\mathbf{y}_0$ and $\hat{\beta}_1^{naive}(\mathbf{x}_1) = S_1^{h_1}[\mathbf{x}_1]\mathbf{y}_1$. More generally, for a vector $\mathbf{z}$ of size n_z, we have the predictions $\hat{\beta}_j^{naive}(\mathbf{z}) = S_j^{h_j}[\mathbf{z}]\mathbf{y}_j$, $j = 0, 1$, where the $n_z \times n_j$ matrix $S_j^h[\mathbf{z}]$ is defined as

$$S_j^h[\mathbf{z}] = \begin{pmatrix} \frac{K(\frac{z_1 - x_{j1}}{h})}{\sum_{i=1}^{n_j} K(\frac{z_1 - x_{ji}}{h})} & \frac{K(\frac{z_1 - x_{j2}}{h})}{\sum_{i=1}^{n_j} K(\frac{z_1 - x_{ji}}{h})} & \cdots & \frac{K(\frac{z_1 - x_{jn_j}}{h})}{\sum_{i=1}^{n_j} K(\frac{z_1 - x_{ji}}{h})} \\ \frac{K(\frac{z_2 - x_{j1}}{h})}{\sum_{i=1}^{n_j} K(\frac{z_2 - x_{ji}}{h})} & \frac{K(\frac{z_2 - x_{j2}}{h})}{\sum_{i=1}^{n_j} K(\frac{z_2 - x_{ji}}{h})} & \cdots & \frac{K(\frac{z_2 - x_{jn_j}}{h})}{\sum_{i=1}^{n_j} K(\frac{z_2 - x_{ji}}{h})} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{K(\frac{z_{n_z} - x_{j1}}{h})}{\sum_{i=1}^{n_j} K(\frac{z_{n_z} - x_{ji}}{h})} & \cdots & \cdots & \frac{K(\frac{z_{n_z} - x_{jn_j}}{h})}{\sum_{i=1}^{n_j} K(\frac{z_{n_z} - x_{ji}}{h})} \end{pmatrix}. \quad (4)$$

From the above fits/predictions of the functions β_0 and β_1, we obtain the naive estimator of τ as $\hat{\tau}^{naive}(\mathbf{z}) = \hat{\beta}_1^{naive}(\mathbf{z}) - \hat{\beta}_0^{naive}(\mathbf{z})$.

2.2 A backfitting procedure

To improve the quality of the naive fit, we propose the backfitting procedure described in Algorithm 1.

Algorithm 1 The backfitting algorithm to estimate β_0, β_1 and τ.

1. $\hat{\beta}_0(\mathbf{x}_0) = S_0^{h_0}[\mathbf{x}_0]\ \mathbf{y}_0 = \hat{\beta}_0^{naive}(\mathbf{x}_0)$. Predict $\hat{\beta}_0(\mathbf{x}_1) = S_0^{h_0}[\mathbf{x}_1]\ \mathbf{y}_0$.
2. $\hat{\tau}(\mathbf{x}_1) = S_1^{h_\tau}[\mathbf{x}_1]\ (\mathbf{y}_1 - \hat{\beta}_0(\mathbf{x}_1))$.
3. $\hat{\beta}_0^{backfit}((\mathbf{x}_0^T, \mathbf{x}_1^T)^T) = S_{0,1}^{h_0}[(\mathbf{x}_0^T, \mathbf{x}_1^T)^T](\mathbf{y}_0^T, (\mathbf{y}_1 - \hat{\tau}(\mathbf{x}_1))^T)^T$.
4. $\hat{\beta}_1(\mathbf{x}_1) = S_1^{h_1}[\mathbf{x}_1]\ \mathbf{y}_1 = \hat{\beta}_1^{naive}(\mathbf{x}_1)$. Predict $\hat{\beta}_1(\mathbf{x}_0) = S_1^{h_1}[\mathbf{x}_0]\ \mathbf{y}_1$.
5. $\hat{\tau}(\mathbf{x}_0) = S_0^{h_\tau}[\mathbf{x}_0]\ (-\mathbf{y}_0 + \hat{\beta}_1(\mathbf{x}_0))$.
6. $\hat{\beta}_1^{\ backfit}((\mathbf{x}_0^T, \mathbf{x}_1^T)^T) = S_{0,1}^{h_1}[(\mathbf{x}_0^T, \mathbf{x}_1^T)^T]((\mathbf{y}_0 + \hat{\tau}(\mathbf{x}_0))^T, \mathbf{y}_1^T)^T$.
7. $\hat{\tau}^{backfit}((\mathbf{x}_0^T, \mathbf{x}_1^T)^T) = \hat{\beta}_1^{\ backfit}((\mathbf{x}_0^T, \mathbf{x}_1^T)^T) - \hat{\beta}_0^{backfit}((\mathbf{x}_0^T, \mathbf{x}_1^T)^T)$.

In this algorithm, predictions at $\mathbf{z}$ are given by

$$\hat{\beta}_0^{backfit}(\mathbf{z}) = S_{0,1}^{h_0}[\mathbf{z}](\mathbf{y}_0^T, (\mathbf{y}_1 - \hat{\tau}(\mathbf{x}_1))^T)^T$$

and

$$\hat{\beta}_1^{backfit}(\mathbf{z}) = S_{0,1}^{h_1}[\mathbf{z}]((\mathbf{y}_0 + \hat{\tau}(\mathbf{x}_0))^T, \mathbf{y}_1^T)^T.$$

For a kernel estimator, the (k,l) element of the $n_z \times (n_0 + n_1)$ matrix $S_{0,1}^h[\mathbf{z}]$ is

$$(S_{0,1}^h[\mathbf{z}])_{k,l} = \begin{cases} \frac{K(\frac{z_k - x_{0l}}{h})}{\sum_{i=1}^{n_0} K(\frac{z_k - x_{0i}}{h}) + \sum_{i=1}^{n_1} K(\frac{z_k - x_{1i}}{h})} & 1 \leq l \leq n_0 \\ \frac{K(\frac{z_k - x_{1(l-n_0)}}{h})}{\sum_{i=1}^{n_0} K(\frac{z_k - x_{0i}}{h}) + \sum_{i=1}^{n_1} K(\frac{z_k - x_{1i}}{h})} & n_0 + 1 \leq l \leq n_0 + n_1, \end{cases}$$

for $k = 1, \ldots, n_z$.

In Figure 2 we illustrate Steps 1. and 3. of the backfitting algorithm on the Spanish Onion dataset. The algorithm starts by computing a first estimate of β_0 based on the untreated individuals only. The fit produced (panel (a) of Figure 2) is the naive estimator $\hat{\beta}_0^{naive}$ of Section 2.1. This fit is used to obtain predicted values of β_0 at the design points $\mathbf{x}_1$. Note that $\hat{\beta}_0(\mathbf{x}_1)$ does not depend on $\mathbf{y}_1$. In Step 2. we use the fact that $\tau(\mathbf{x}_1) = \beta_1(\mathbf{x}_1) - \beta_0(\mathbf{x}_1)$ to obtain a first estimate of $\tau(\mathbf{x}_1)$ by smoothing the prediction errors $\mathbf{y}_1 - \hat{\beta}_0(\mathbf{x}_1)$ on $\mathbf{x}_1$. With this estimate $\hat{\tau}(\mathbf{x}_1)$ we impute values of the response for pseudo non-treated individuals at the design points $\mathbf{x}_1$ as $\mathbf{y}_1 - \hat{\tau}(\mathbf{x}_1)$. Step 3. re-estimates β_0 by smoothing $(\mathbf{y}_0^T, (\mathbf{y}_1 - \hat{\tau}(\mathbf{x}_1)^T)^T$ on $(\mathbf{x}_0^T, \mathbf{x}_1^T)^T$ as it appears on panel (b) of Figure 2. Intuitively, this refitting based on a larger sample should improve the finite sample properties of the final fit. This is studied in Sections 2.3 to 2.5. The algorithm is fully symmetric for both groups and Steps 4. to 6. mimic Steps 1. to 3. for the estimation of β_1. Finally, Step 7. produces the estimation of τ by taking the difference between $\hat{\beta}_1$ and $\hat{\beta}_0$.

It is possible to iterate Steps 1. to 3. (4. to 6. respectively) by using $\hat{\beta}_0(\mathbf{x}_0) = \hat{\beta}_0^{backfit}(\mathbf{x}_0)$ (and $\hat{\beta}_1(\mathbf{x}_1) = \hat{\beta}_1^{backfit}(\mathbf{x}_1)$ respectively). The gain of precision by iterating these steps is, however, negligible.

The smoothing parameter h_0, h_τ and h_1 are usually unknown in practice. They must be estimated and cross-validation is often used in this setting, see, e.g., Hastie and Tibshirani (1990, Sec. 3.4).

Note that the algorithm implicitly assumes that the distribution of $x_i|w_i = 0$ and $x_i|w_i = 1$ have common support. The practical counterpart of this assumption is that we want the prediction $\hat{\beta}_0(\mathbf{x}_1)$ and $\hat{\beta}_1(\mathbf{x}_0)$ to be made within (or at least very close) to the design space where the functions are fitted. This is because extrapolation does not make sense in non-parametric regression unless some restrictive parametric assumptions at the border are made. When the common support assumption does not hold, a solution consists in applying the backfitting algorithm selectively. That is by making predictions only where the functions β_0 and β_1 are fitted in Step 1. and 4. respectively.

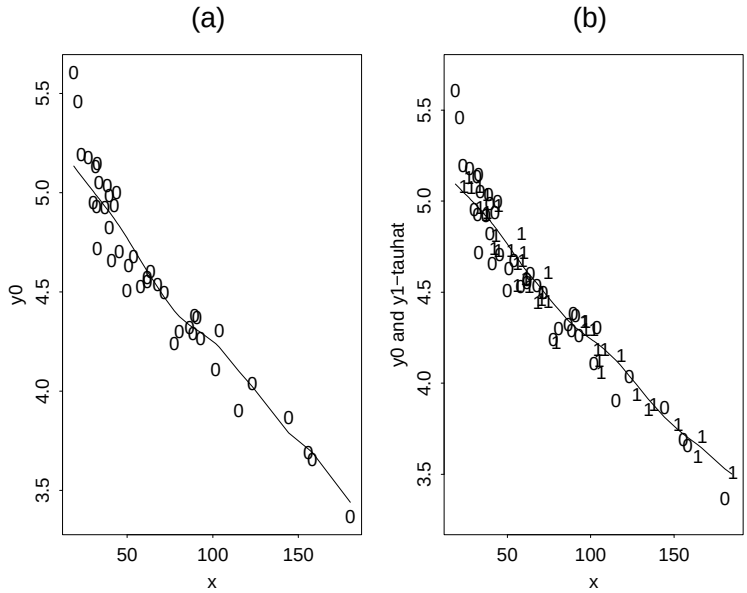

Figure 2: Illustration of Algorithm 1: Panel (a) displays the fit of $\beta_0(\mathbf{x}_0)$ at Step 1 of the algorithm; panel (b) displays the fit of $\beta_0(\mathbf{x}_0, \mathbf{x}_1)$ at Step 3.

2.3 Variance

In this section we present the exact variances of the naive and backfitting estimators introduced earlier. We, moreover, give theoretical results describing situations where the backfitting estimators have lower variance. The design points are throughout considered as fixed.

For the naive estimators we have

$$Var(\hat{\beta}_0^{naive}(z)) = \sigma^2 S_0^{h_0}[z] S_0^{h_0}[z]^T \text{ and } Var(\hat{\beta}_1^{naive}(z)) = \sigma^2 S_1^{h_1}[z] S_1^{h_1}[z]^T. \quad (5)$$

Furthermore, because these estimates are obtained with two independent sub-samples we obtain

$$Var(\hat{\tau}^{naive}(z)) = Var(\hat{\beta}_0^{naive}(z)) + Var(\hat{\beta}_1^{naive}(z)). \quad (6)$$

The variance of the backfitting estimators are deduced in Appendix A.1. We find

$$Var\big(\hat{\beta}_0^{backfit}(z)\big) = \sigma^2 S_{0,1}^{h_0}[z] V S_{0,1}^{h_0}[z]^T \quad (7)$$

$$Var\big(\hat{\beta}_1^{backfit}(z)\big) = \sigma^2 S_{0,1}^{h_1}[z] W S_{0,1}^{h_1}[z]^T, \quad (8)$$

where

$$V = \begin{pmatrix} I_{n_0} & C \\ C^T & B \end{pmatrix} \text{ and } W = \begin{pmatrix} D & E \\ E^T & I_{n_1} \end{pmatrix}.$$

with

$$C = S_0^{h_0}[\mathbf{x}_1]^T S_1^{h_\tau}[\mathbf{x}_1]^T, \quad E = S_0^{h_\tau}[\mathbf{x}_0] S_1^{h_1}[\mathbf{x}_0],$$
$$B = I_{n_1} + S_1^{h_\tau}[\mathbf{x}_1]\big[I_{n_1} + S_0^{h_0}[\mathbf{x}_1] S_0^{h_0}[\mathbf{x}_1]^T\big] S_1^{h_\tau}[\mathbf{x}_1]^T - S_1^{h_\tau}[\mathbf{x}_1] - S_1^{h_\tau}[\mathbf{x}_1]^T,$$

and

$$D = I_{n_0} + S_0^{h_\tau}[\mathbf{x}_0]\big[I_{n_0} + S_1^{h_1}[\mathbf{x}_0] S_1^{h_1}[\mathbf{x}_0]^T] S_0^{h_\tau}[\mathbf{x}_0]^T - S_0^{h_\tau}[\mathbf{x}_0\big] - S_0^{h_\tau}[\mathbf{x}_0]^T.$$

Moreover,

$$\begin{aligned} Var(\hat{\tau}^{backfit}(z)) &= Var(\hat{\beta}_0^{backfit}(z)) + Var(\hat{\beta}_1^{backfit}(z)) \\ &\quad -2Cov\big(\hat{\beta}_1^{backfit}(z), \hat{\beta}_0^{backfit}(z)\big), \end{aligned} \tag{9}$$

where the latter covariance is deduced in Appendix A.2.

We now give two results giving conditions ensuring that the backfitting estimators have lower variance than the naive estimators.

Proposition 1 *Assume that Algorithm 1 of Section 2.2 is used with symmetric matrices $S_0[\mathbf{x}_0]$ and $S_1[\mathbf{x}_1]$, whose eigenvalues are within $[0,1]$. If, moreover, treated and non-treated have the same design points ($\mathbf{x}_0 \equiv \mathbf{x}_1$) we have that*

$$Var\big(\hat{\beta}_0^{backfit}(z)\big) \leq Var\big(\hat{\beta}_0^{naive}(z)\big), \tag{10}$$
$$Var\big(\hat{\beta}_1^{backfit}(z)\big) \leq Var\big(\hat{\beta}_1^{naive}(z)\big). \tag{11}$$

Proof. We give the proof for (10) only. We need to prove that

$$S_{0,1}^{h_0}[z] V S_{0,1}^{h_0}[z]^T \leq S_0^{h_0}[z] S_0^{h_0}[z]^T.$$

Because $\mathbf{x}_0 \equiv \mathbf{x}_1$, we have that the $(n_0+n_0)\times 1$ matrix $S_{0,1}^{h_0}[z] = \frac{1}{2}(S_0^{h_0}[z], S_0^{h_0}[z])$. Hence,

$$\begin{aligned} S_{0,1}^{h_0}[z] V S_{0,1}^{h_0}[z]^T &= \frac{1}{4}(S_0^{h_0}[z], S_0^{h_0}[z]) \begin{pmatrix} I_{n_0} & C \\ C^T & B \end{pmatrix} (S_0^{h_0}[z], S_0^{h_0}[z])^T \\ &= \frac{1}{4} S_0^{h_0}[z](I_{n_0} + C^T + C + B) S_0^{h_0}[z]^T. \end{aligned}$$

Further, because $\mathbf{x}_1 \equiv \mathbf{x}_0$ and we work with a symmetric smoother we have $C = S_0^{h_0}[\mathbf{x}_0] S_0^{h_\tau}[\mathbf{x}_0]$. Also, by the condition on the eigenvalues of the smoothing matrix we can write $C \leq S_0^{h_0}[\mathbf{x}_0] \leq I_{n_0}$.

The matrix B can also be bounded above by I_{n_0}. Indeed, by the same arguments than above we have $S_1^{h_\tau}[\mathbf{x}_1] S_1^{h_\tau}[\mathbf{x}_1] \leq S_1^{h_\tau}[\mathbf{x}_1]$, and $S_0^{h_0}[\mathbf{x}_1] S_0^{h_0}[\mathbf{x}_1] = S_0^{h_0}[\mathbf{x}_0] S_0^{h_0}[\mathbf{x}_0] \leq I_{n_0}$. Therefore, we can write

$$\begin{aligned} B &= I_{n_0} + S_1^{h_\tau}[\mathbf{x}_1]\big[I_{n_0} + S_0^{h_0}[\mathbf{x}_1] S_0^{h_0}[\mathbf{x}_1]\big] S_1^{h_\tau}[\mathbf{x}_1] - S_1^{h_\tau}[\mathbf{x}_1] - S_1^{h_\tau}[\mathbf{x}_1] \\ &\leq I_{n_0} + S_1^{h_\tau}[\mathbf{x}_1] S_1^{h_\tau}[\mathbf{x}_1] + S_1^{h_\tau}[\mathbf{x}_1] S_1^{h_\tau}[\mathbf{x}_1] - S_1^{h_\tau}[\mathbf{x}_1] - S_1^{h_\tau}[\mathbf{x}_1] \leq I_{n_0}. \end{aligned}$$

Finally, we have the desired result

$$\begin{aligned} S_{0,1}^{h_0}[z]VS_{0,1}^{h_0}[z]^T &= \frac{1}{4}S_0^{h_0}[z](I_{n_0}+C^T+C+B)S_0^{h_0}[z]^T \\ &\leq \frac{1}{4}S_0^{h_0}[z](4I_{n_0})S_0^{h_0}[z]^T = S_0^{h_0}[z]S_0^{h_0}[z]^T. \end{aligned}$$

■

Proposition 2 *Assume that Algorithm 1 of Section 2.2 is used with symmetric matrices $S_0[\mathbf{x}_0]$ and $S_1[\mathbf{x}_1]$, whose eigenvalues are within $[0,1]$. If $Var(\hat{\beta}_j^{naive}(z)) > Var(\hat{\beta}_j^{backfit}(z))$, for $j=0,1$, and $h_0=h_1$ then*

$$Var(\hat{\tau}^{backfit}(z)) < Var(\hat{\tau}^{naive}(z)).$$

Proof. The proposition is shown by noting that

$$Cov(\hat{\beta}_1^{backfit}(z), \hat{\beta}_0^{backfit}(z)) \geq 0.$$

This covariance was deduced in Appendix A.2 and has the form

$$Cov(\hat{\beta}_1^{backfit}(z), \hat{\beta}_0^{backfit}(z)) = \sigma^2 S_{0,1}^{h_0}[z]US_{0,1}^{h_1}[z]^T,$$

where

$$U = \begin{pmatrix} I_{n_0} - S_0^{h_\tau}[\mathbf{x}_0] & F \\ 0 & I_{n_1} - S_1^{h_\tau}[\mathbf{x}_1] \end{pmatrix}.$$

Because the eigenvalues of the smoothing matrix are between 0 and 1 the two diagonal blocks of U are positive definite, and so is U itself because it is triangular by block. Hence, $S_{0,1}^{h_0}[z]US_{0,1}^{h_1}[z]^T \geq 0$ if $S_{0,1}^{h_0}[z] = S_{0,1}^{h_1}[z]$. The latter equality holds by the assumption $h_0 = h_1$, thereby completing the proof. ■

Note that the assumption $h_0 = h_1$ will rarely be needed in practical applications. Indeed, $S_{0,1}^{h_0}[z]US_{0,1}^{h_1}[z]^T \geq 0$ hold with $h_0 \neq h_1$, for instance, if the vectors $S_{0,1}^{h_j}[z]$, $j=0,1$, contain only positive values. Even when negative weights are allowed, those are typically very close to zero, making the condition $h_0 = h_1$ superfluous in practice.

The condition on the eigenvalues has been used before in the literature, see, e.g., Hastie and Tibshirani (1990, Sec. 5.3.7). Smoothers fulfilling it include cubic splines, regression splines and linear regression. With asymmetric smoothers, empirical evidence shows that the conclusion of the propositions hold often; see Section 2.5.

The condition asking for equality of design points for treated and untreated in Proposition 1 is not either a necessary assumption. In real applications, we often have design points that are not too different from each other for treated and untreated, in which case we will often observe an improvement of the variance when using the backfitting estimator.

We stress here that, in any particular case, the improvement in variance can be checked by computing the exact variances of the naive and backfitting estimators with the explicit formulas given in (5)–(9). In Figure 1 the confidence bands provided are based on these exact variances. In this example we observe narrower confidence bands for the backfitting estimators, even if the assumptions of the Propositions 1 and 2 do not hold exactly.

2.4 Bias

For an estimator $\hat{f}(z)$ of $f(z)$, we define its bias at z as $Bias(\hat{f}(z)) = E(\hat{f}(z)) - f(z)$. It depends on the unknown function f. Notice first that the bias at a given design point z can both be decreased or increased by adding information/observations at other design points. It is, therefore, not possible to give a general statement when comparing the bias of the naive and backfitting estimators at a given design point.

The naive estimators have biases

$$Bias(\hat{\tau}^{naive}(z)) = Bias(\hat{\beta}_1^{naive}(z)) - Bias(\hat{\beta}_0^{naive}(z)) \tag{12}$$

with $Bias(\hat{\beta}_0^{naive}(z)) = S_0^{h_0}[z]\beta_0(\mathbf{x}_0) - \beta_0(z)$ and $Bias(\hat{\beta}_1^{naive}(z)) = S_1^{h_1}[z]\beta_1(\mathbf{x}_1) - \beta_1(z)$.

The backfitting estimators have biases

$$Bias(\hat{\tau}^{backfit}(z)) = Bias(\hat{\beta}_1^{backfit}(z)) - Bias(\hat{\beta}_0^{backfit}(z)), \tag{13}$$

with

$$\begin{aligned} Bias(\hat{\beta}_0^{backfit}(z)) &= S_{0,1}^{h_0}[z]E\left[\begin{pmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 - \hat{\tau}(\mathbf{x}_1) \end{pmatrix}\right] - \beta_0(z) = \\ &= S_{0,1}^{h_0}[z]\begin{pmatrix} \beta_0(\mathbf{x}_0) \\ \beta_1(\mathbf{x}_1) - S_1^{h_\tau}[\mathbf{x}_1]\beta_1(\mathbf{x}_1) + S_1^{h_\tau}[\mathbf{x}_1]S_0^{h_0}[\mathbf{x}_1]\beta_0(\mathbf{x}_0) \end{pmatrix} - \beta_0(z) \end{aligned}$$

and

$$\begin{aligned} &Bias(\hat{\beta}_1^{backfit}(z)) = \\ &= S_{0,1}^{h_1}[z]\begin{pmatrix} \beta_0(\mathbf{x}_0) - S_0^{h_\tau}[\mathbf{x}_0]\beta_0(\mathbf{x}_0) + S_0^{h_\tau}[\mathbf{x}_0]S_1^{h_1}[\mathbf{x}_0]\beta_1(\mathbf{x}_1) \\ \beta_1(\mathbf{x}_1) \end{pmatrix} - \beta_1(z). \end{aligned}$$

Looking at $Bias(\hat{\beta}_0^{backfit}(z))$, we see that the extra observations $y_1 - \hat{\tau}(\mathbf{x}_1)$ that are utilized for $\hat{\beta}_0^{backfit}(z)$ are themselves biased (as estimators of $y_1 - \tau(\mathbf{x}_1)$). The simulations performed in Section 2.5 indicate that using such biased observations tend to increase the bias in estimating $\beta_0(\mathbf{x}_0)$, but may have the reverse effect when estimating $\beta_0(\mathbf{x}_1)$. The latter effect may be explained by the fact that extra information at $\mathbf{x}_1$, even biased, is beneficial to $\hat{\beta}_0^{backfit}(\mathbf{x}_1)$. Note that to obtain a fit of $\tau(\mathbf{x}_1)$, for instance, both $\beta_1(\mathbf{x}_1)$ and $\beta_0(\mathbf{x}_1)$ are needed.

2.5 Simulation study

In this section we aim at studying two main issues: (1) Compare the variance for the naive and backfitting estimators in cases where Propositions 1 and 2 do not apply exactly, and (2) study the bias and mean squared error (MSE) of these two estimators.

We simulate data inspired from the white Spanish Onions dataset. Ratkowsky (1983) fitted two parametric curves $\beta_j(x) = (\alpha_{j0} + \alpha_{j1}x + \alpha_{j2}x^2)^{-1}$, $j = 0, 1$ to the data with least squares. We use their estimate and simulate

$$y_i = \beta_0(x_i) + (\beta_1(x_i) - \beta_0(x_i))w_i + \epsilon_i, \tag{14}$$

with $\alpha_{00} = 0.002054$, $\alpha_{01} = 0.8571 * 10^{-4}$, $\alpha_{02} = 0.3808 * 10^{-7}$, $\alpha_{10} = 0.002084$, $\alpha_{11} = 0.1311 * 10^{-3}$, $\alpha_{12} = 0.7796 * 10^{-7}$, and where $\epsilon_i \sim N(0, 0.01)$. We simulate design points x_i from a uniform distribution with support $(18.78, 184.75)$ which is the range of the design points available for the Spanish Onions dataset. In the first experiment (Experiment 1 in the sequel) we simulate 42 design points and use them to simulate the outcome y_i both with $w_i = 0$ and $w_i = 1$. That is we simulate data where treated and untreated have the same design points. The second experiment (Experiment 2) is obtained by simulating different design points for 42 treated and 42 untreated individuals. All computations are performed with Splus. Each experiment is replicated 1000 times.

For all the nonparametric fit performed we fix the value of h_0, h_1 and h_τ. It is indeed not the purpose of this paper to study the estimation of the smoothing parameters on which there exists an extensive literature. Instead of choosing arbitrary values for the smoothing parameters we used cross-validation to estimate h_0 and h_1 on the untreated and treated individuals respectively. This was done on the 1000 replicated data sets and the median of the 1000 estimated parameters is used in the sequel, namely $h_0 = 7.5$ and $h_1 = 8.2$ for gaussian kernels, and $h_0 = 0.005$ and $h_1 = 0.006$ for the cubic smoothing splines. The backfitting algorithm was then run on the 1000 replicates and cross-validation was used to estimate h_τ in Step 2 of the algorithm. The median of the 1000 estimates is used in the sequel, namely $h_\tau = 17.6$ for kernels and $h_\tau = 0.109$ for splines. Note that even though it is used to pursue the same goal, the definition of the smoothing parameter is particular to each smoother, therefore explaining the difference in magnitude we observe for kernels and splines.

To compare estimators we compute on each replicate an average (over the design) bias (using (12) and (13)), variance (using (6) and (9)) and MSE. We present results on the estimation of τ since it is the curve of main interest. In Experiment 1 we used both smoothing splines and kernels. When using smoothing splines, we are in a situation covered by Proposition 1. With kernels, the assumptions of the proposition are violated since the smoother is not symmetric. In Experiment 2 we only used kernels.

The results summarized as boxplots in Figure 3 and Figure 4 show that the backfitting estimator lead to a decrease in average variance in all 1000 cases of

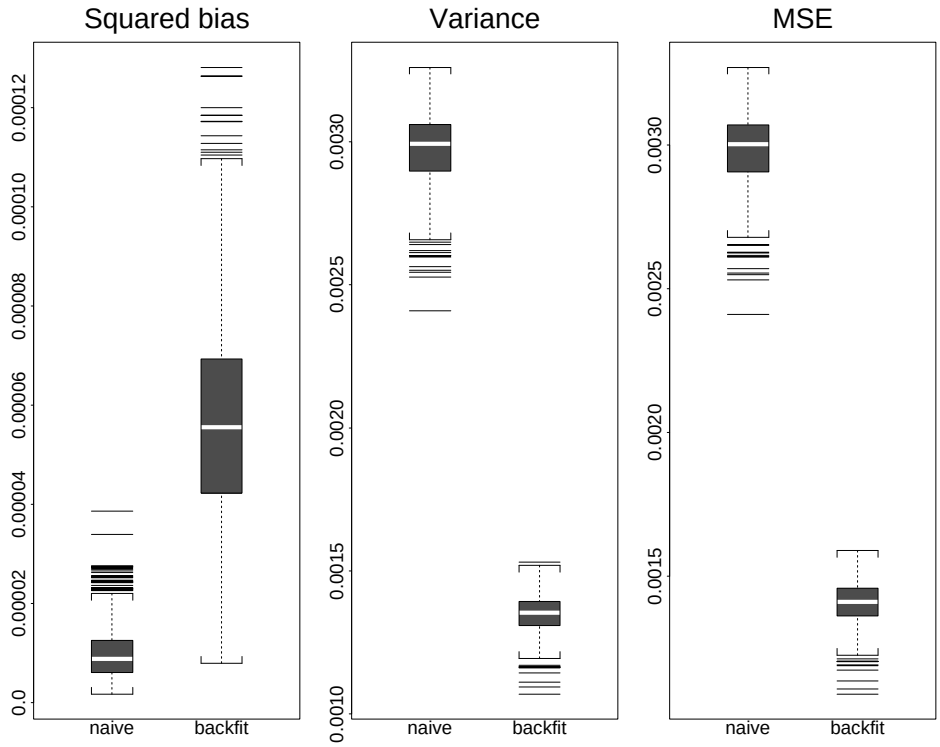

Figure 3: Results for Experiment 1 with gaussian kernels. Boxplots of the average (over the design) squared bias, variance and MSE of the naive and backfitting estimators of τ for 1000 simulated cases.

Experiment 1, both with splines and kernels. Averaged squared bias is increased as expected. However, the average MSE is always improved by the backfitting estimator.

In Experiment 2 where the design points are not identical for treated and untreated, we see (Figure 5) that the average variance is still decreased by the backfitting estimator. Moreover, the average squared biases are here slightly lower for the backfitting estimator. This decrease in bias can be explained as follows. When estimating τ, for instance, at $\mathbf{x}_0$, $\hat{\beta}_1$ must be evaluated at $\mathbf{x}_0$, which is a prediction in the case of the naive estimator. These predictions have large bias, as can be noticed by comparing the bias of the naive estimator in Experiment 1 $-$where $\mathbf{x}_0 \equiv \mathbf{x}_1$ (Figure 3) and no predictions are therefore made to estimate $\tau-$ and Experiment 2 (Figure 5). In contrast, the backfitting estimator uses information at $\mathbf{x}_0$ when fitting β_1. This yields a less biased estimator of $\beta_1(\mathbf{x}_0)$, thereby explaining the pattern of squared biases observed in Figure 5.

Finally, the average MSE is decreased with the backfitting estimator. Although this is not apparent from the last boxplot in Figure 5, the decrease in MSE takes

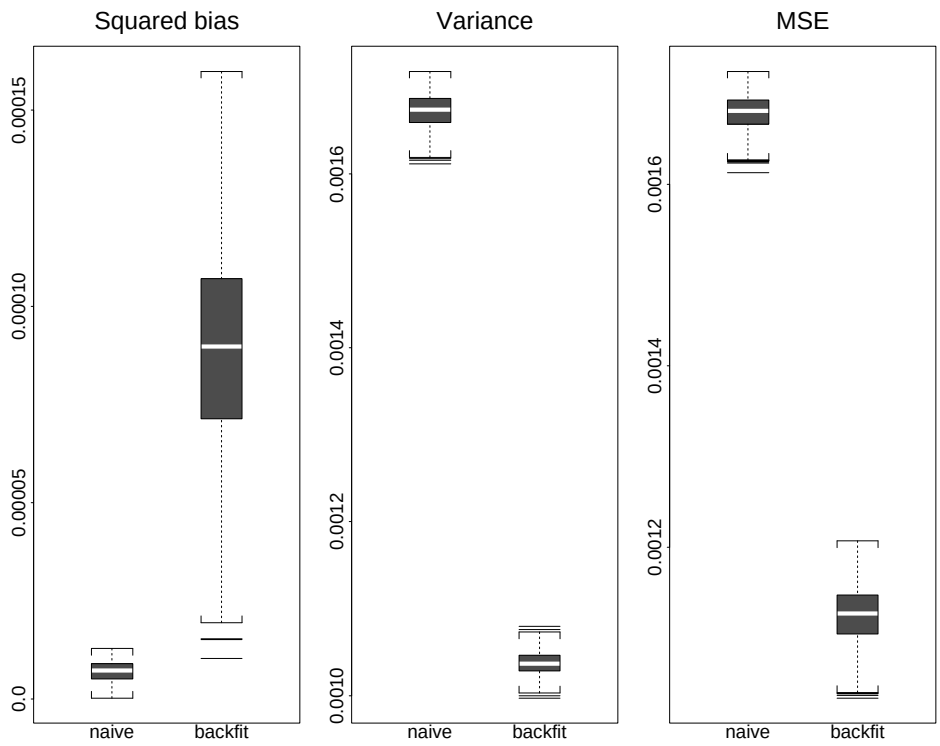

Figure 4: Results for Experiment 1 with cubic smoothing splines. Boxplots of the average (over the design) squared bias, variance and MSE of the naive and backfitting estimators of τ for 1000 simulated cases.

place in all 1000 simulated cases.

3 Covariance adjustment with the propensity score: an application

The results presented in this paper are not restricted to a single covariate situation thanks to the results obtained by Rosenbaum and Rubin (1983). We now briefly introduce the potential outcome framework for non-randomized experiments (Rubin, 1974) and how it leads to model (1) where a single covariate x_i is replaced by a scalar valued function of a vector of p covariates, denoted $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})^T$. The use of the backfitting estimator in this general context is then illustrated with a data set on training program evaluation.

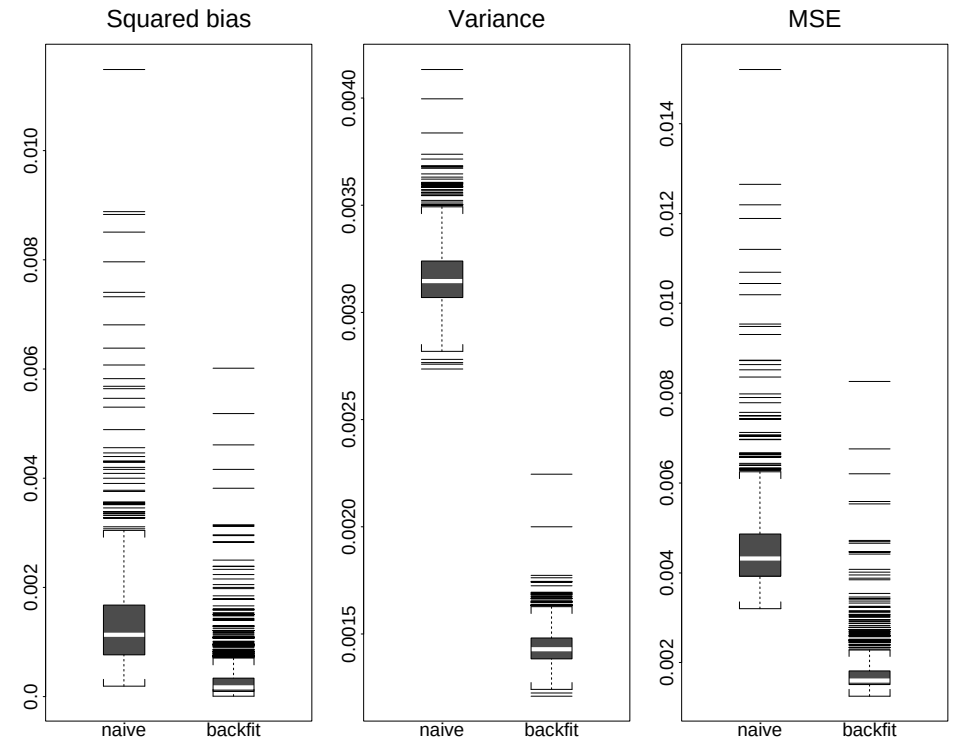

Figure 5: Results for Experiment 2 with gaussian kernels. Boxplots of the average (over the design) squared bias, variance and MSE of the naive and backfitting estimators of τ for 1000 simulated cases.

3.1 Potential outcomes and non-parametric covariance adjustment

Let y_i^0 and y_i^1 be the response for individual i had he been not treated or treated, respectively. Assume that, for all $\mathbf{x}_i$, *i)* y_i^0 and y_i^1 are independent of the treatment w_i when conditioning on $\mathbf{x}_i$ (denoted $y_i^0, y_i^1 \perp\!\!\!\perp w_i|\mathbf{x}_i$) and *ii)* $0 < \Pr(w_i = 1|\mathbf{x}_i) = p(\mathbf{x}_i) < 1$, then (Rosenbaum and Rubin, 1983, Theorem 3)

$$y_i^0, y_i^1 \perp\!\!\!\perp w_i|p(\mathbf{x}_i), \tag{15}$$

where $p(\mathbf{x}_i)$ is called the propensity score. Assumption *ii)* is equivalent to the common support assumption discussed at the end of Section 2.2.

Note that y_i^0, y_i^1 cannot be both observed for a given individual i. On the other hand, we observe always the response $y_i = y_i^0(1 - w_i) + y_i^1 w_i$, whose expectation we want to model. A direct consequence of (15) is that

$$E(y_i|p(\mathbf{x}_i), w_i) = E(y_i^0|p(\mathbf{x}_i))(1 - w_i) + E(y_i^1|p(\mathbf{x}_i))w_i. \tag{16}$$

Rearranging we have

$$E(y_i|p(\mathbf{x}_i), w_i) = E(y_i^0|p(\mathbf{x}_i)) + w_i(E(y_i^1|p(\mathbf{x}_i)) - E(y_i^0|p(\mathbf{x}_i))).$$

We, therefore, retrieve model (1)

$$y_i = \beta_0(p(\mathbf{x}_i)) + w_i\tau(p(\mathbf{x}_i)) + \varepsilon_i,$$

where $\beta_0 = E(y_i^0|p(\mathbf{x}_i))$ and $\tau(x_i) = E(y_i^1|p(\mathbf{x}_i)) - E(y_i^0|p(\mathbf{x}_i))$. The functions involved are functions of scalars and the backfitting algorithm may be applied as described in Algorithm 1. In practice the propensity score is not known and must be estimated.[2]

Finally, note that assumptions *i)* and *ii)* are natural since the former is essential for $\tau(p(\mathbf{x}_i))$ to have causal content, and the latter guarantees $\tau(p(\mathbf{x}_i))$ to be well defined on the support of $\mathbf{x}_i$.

3.2 Training program: estimation of a conditional training effect

We consider data on a training program implemented in the mid-1970's for individuals having faced economic and social problems prior to enrollment (Lalonde, 1986). Because both a randomized and several non-randomized control (untreated) groups are available, this data was used by Dehejia and Wahba (1999) to validate the use of result (15) to estimate average treatment effects on wage based on non-randomized data. They performed their analysis on various subsets of individuals obtained by stratification in order to make treated and untreated more homogeneous in their covariate values. We refer the reader to Dehejia and Wahba (1999) for a detailed description of the data. We consider in the sequel a subset of the data where the control group was obtained from the Westat's Matched Current Population Survey-Social Security Administration File. The data set, called CPS3 in Dehejia and Wahba (1999), consists in 185 treated and 429 controls, on which ten covariates are measured: Age (x_1), Education (x_2), Black (x_3), Hispanic (x_4), No degree (x_5), Married (x_6), Unemployed in 1974 (x_7), Unemployed in 1975 (x_8), Earnings in 1974 (x_9, U.S. \$) and Earnings in 1975 (x_{10}, U.S. \$). The outcome of interest is Earnings 1978 (y, U.S. \$). The data is available at `http://www.columbia.edu/~rd247/nswdata.html`.

We focus on the conditional treatment effect $\tau(p(\mathbf{x}_i))$ defined in the previous section, and illustrate the use of the backfitting estimator proposed earlier. For this purpose, we need to estimate $p(\mathbf{x}_i)$. We follow Dehejia and Wahba (1999) and use a logistic regression model with the following linear predictor

[2]It is common practice to use the estimated propensity score since the true one is generally unknown. Although Theorem 5 in Rosenbaum and Rubin (1983) shows that an estimated propensity score can have the desired property (15), this is not guaranteed.

$$\log\left(\frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)}\right) = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \theta_3 x_{3i} + \theta_4 x_{4i} + \theta_5 x_{5i} + \theta_6 x_{6i} + \theta_7 x_{7i} + \theta_8 x_{8i} + \theta_9 x_{9i} + \theta_{10} x_{10i} + \theta_{11} x_{1i}^2 + \theta_{12} x_{1i}^3 + \theta_{13} x_{2i}^2 + \theta_{14} x_{2i} x_{9i}.$$

Based on the estimated propensity scores we estimate $\tau(\hat{p}(\mathbf{x_i}))$ with the naive and backfitting estimator. The fits and their confidence bands are displayed in Figure 6 (bottom panels). We observe that the backfitting estimator decrease significantly the variability on the right hand side of the range of the propensity score, while on the left hand side the variability is slightly increased. This asymmetry is due to the distribution of the treated and untreated along this propensity score axis. Most controls are found on the lower values of the propensity score, making the naive estimator of the β_0 function highly variable for large values of the propensity score. This large variability is corrected by the backfitting estimator of β_0 using information from the treated group, thereby improving on the variability of the estimation of τ. On the other hand, there is no evidence of a treatment effect even with the backfitting estimators since the value zero is overlapped by the confidence bands.

4 Discussion

We have considered a model for estimating a conditional treatment effect while adjusting for covariates without making strong parametric assumptions. A backfitting algorithm has been proposed to estimate non-parametrically the functions of the covariates involved. This new estimator has been shown to improve on the naive procedure which consists in estimating separately a function of the covariates for the treated individuals and for the controls. The variance calculated are for finite samples, thereby allowing us to avoid the use of asymptotic arguments when comparing estimators.

We have implicitly assumed (see model (1)) that treated and controls have identical residual variances: $Var(y_i|x_i, w_i = 1) = Var(y_i|x_i, w_i = 0) = \sigma^2$. Diverging variances do not affect the implementation of the backfitting estimator. The expressions deduced for the variances of the different estimators must, however, be adapted. For both the real data sets used in this paper the estimated variances were close enough to be assumed equal (an F-test can be carried out to test for equality).

We have focused our work on linear smoothers because of their analytical tractability. Similar results are, however, expected to hold with more complex non-parametric regression methods, such as neural networks and wavelets. Another natural generalization would be to consider discrete responses through generalized additive models.

A series of papers have recently appeared on tests of constant treatment effect, that is for the null hypothesis $\tau(x) = c$ for all x, see Young and Bowman (1995),

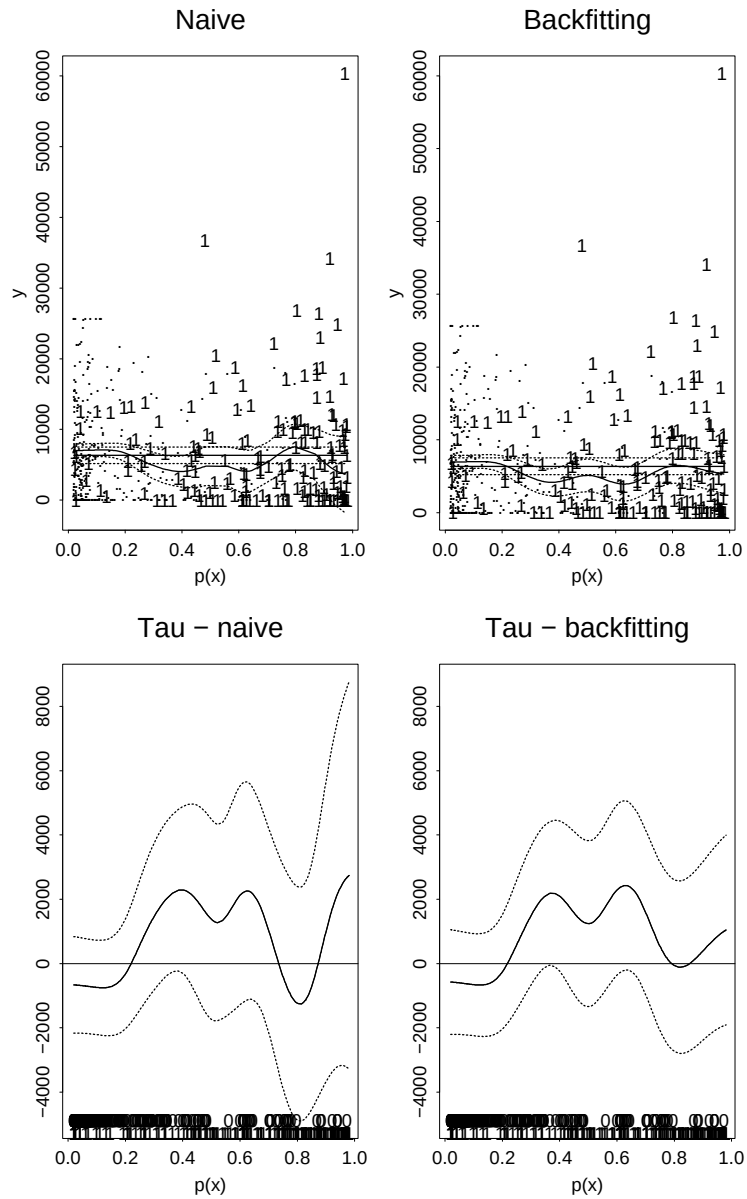

Figure 6: Non-parametric fits (gaussian kernels with smoothing parameter chosen with cross-validation) for the CPS3 data set. Treated are marked with 1's and controls with dots. Top panels: fits of the functions $\beta_0(\hat{p}(\mathbf{x}))$ and $\beta_1(\hat{p}(\mathbf{x}))$ (plain lines) with confidence bands (dotted lines). Bottom panels: Corresponding fits of the function $\tau(\hat{p}(\mathbf{x}))$ (plain lines) with confidence bands (dotted lines).

Akritas and Van Keilegom (2001) and Neumeyer and Dette (2003), and the references therein. The tests proposed build on the naive estimator. A more powerful test could result based on the backfitting estimator.

Finally, another area of possible application of the backfitting algorithm is the estimation of an average treatment effect $E_x(\tau(x))$. The non-parametric estimation of this parameter in non-randomized experiments has been largely discussed in the literature building on the work of Rosenbaum and Rubin (1983). In particular, Heckman, Ichimura, and Todd (1998) and Abadie and Imbens (2004) consider a regression imputation estimator which uses an estimate of the conditional treatment effect $\tau(x)$.

A Variances

All the variances and covariances computed here are conditional on $(\mathbf{x}_0^T, \mathbf{x}_1^T)^T$.

A.1 Variance of $\hat{\beta}_0^{backfit}$ and $\hat{\beta}_1^{backfit}$

$$
\begin{aligned}
&Var\big(\hat{\beta}_0^{backfit}(z)\big) = \\
&\quad = S_{0,1}^{h_0}[z]\; Var\big((\mathbf{y}_0^T, (\mathbf{y}_1 - \hat{\tau}(\mathbf{x}_1))^T)^T\big)\; S_{0,1}^{h_0}[z]^T \\
&\quad = S_{0,1}^{h_0}[z] \begin{pmatrix} Var(\mathbf{y_0}) & Cov(\mathbf{y}_0, \mathbf{y}_1 - \hat{\tau}(\mathbf{x}_1)) \\ Cov(\mathbf{y}_1 - \hat{\tau}(\mathbf{x}_1), \mathbf{y}_0) & Var(\mathbf{y}_1 - \hat{\tau}(\mathbf{x}_1)) \end{pmatrix} S_{0,1}^{h_0}[z]^T
\end{aligned} \tag{17}
$$

Moreover, we have that

$$
\begin{aligned}
&Var(\mathbf{y}_1 - \hat{\tau}(\mathbf{x}_1)) = \\
&\quad = Var(\mathbf{y_1}) + Var(\hat{\tau}(\mathbf{x}_1)) - Cov(\mathbf{y_1}, \hat{\tau}(\mathbf{x}_1)) - Cov(\hat{\tau}(\mathbf{x}_1), \mathbf{y_1}) \\
&\quad = \sigma^2 I_{n_1} + S_1^{h_\tau}[\mathbf{x}_1] Var(\mathbf{y}_1 - \hat{\beta}_0(\mathbf{x}_1)) S_1^{h_\tau}[\mathbf{x}_1]^T \\
&\qquad -Cov(\mathbf{y}_1, \mathbf{y}_1 - \hat{\beta}_0(\mathbf{x}_1)) S_1^{h_\tau}[\mathbf{x}_1]^T - S_1^{h_\tau}[\mathbf{x}_1] Cov(\mathbf{y}_1 - \hat{\beta}_0(\mathbf{x}_1), \mathbf{y}_1) \\
&\quad = \sigma^2 I_{n_1} + S_1^{h_\tau}[\mathbf{x}_1] \big[\sigma^2 I_{n_1} + Var(\hat{\beta}_0(\mathbf{x}_1))\big] S_1^{h_\tau}[\mathbf{x}_1]^T - \sigma^2 S_1^{h_\tau}[\mathbf{x}_1] - \sigma^2 S_1^{h_\tau}[\mathbf{x}_1]^T \\
&\quad = \sigma^2 I_{n_1} + S_1^{h_\tau}[\mathbf{x}_1] \big[\sigma^2 I_{n_1} + Var(S_0^{h_0}[\mathbf{x}_1]\ \mathbf{y}_0)\big] S_1^{h_\tau}[\mathbf{x}_1]^T - \sigma^2 [S_1^{h_\tau}[\mathbf{x}_1] + S_1^{h_\tau}[\mathbf{x}_1]^T] \\
&\quad = \sigma^2 \Big(I_{n_1} + S_1^{h_\tau}[\mathbf{x}_1] \big[I_{n_1} + S_0^{h_0}[\mathbf{x}_1] S_0^{h_0}[\mathbf{x}_1]^T\big] S_1^{h_\tau}[\mathbf{x}_1]^T - S_1^{h_\tau}[\mathbf{x}_1] - S_1^{h_\tau}[\mathbf{x}_1]^T \Big) \\
&\quad = \sigma^2 B
\end{aligned}
$$

where we have used the fact that $Cov(\mathbf{y}_1, \hat{\beta}_0(\mathbf{x}_1)) = 0$.

We also have that

$$\begin{aligned} Cov(\mathbf{y}_0, \mathbf{y}_1 - \hat{\tau}(\mathbf{x}_1)) = \\ &= -Cov(\mathbf{y}_0, \hat{\tau}(\mathbf{x}_1)) \\ &= -Cov(\mathbf{y}_0, S_1^{h_\tau}[\mathbf{x}_1](\mathbf{y}_1 - \hat{\beta}_0[\mathbf{x}_1]) \\ &= Cov(\mathbf{y}_0, S_1^{h_\tau}[\mathbf{x}_1]S_0^{h_0}[\mathbf{x}_1]\ \mathbf{y}_0) \\ &= \sigma^2 S_0^{h_0}[\mathbf{x}_1]^T S_1^{h_\tau}[\mathbf{x}_1]^T = \sigma^2 C \end{aligned}$$

Finally, $Var(\mathbf{y}_0) = \sigma^2 I_{n_0}$, and we write

$$Var\big(\hat{\beta}_0^{backfit}(z)\big) = \sigma^2 S_{0,1}^{h_0}[z] V S_{0,1}^{h_0}[z]^T, \tag{18}$$

where

$$V = \begin{pmatrix} I_{n_0} & C \\ C^T & B \end{pmatrix}.$$

Similarly we have

$$Var\big(\hat{\beta}_1^{backfit}(z)\big) = \sigma^2 S_{0,1}^{h_1}[\mathbf{z}] W S_{0,1}^{h_1}[z]^T \tag{19}$$

where

$$W = \begin{pmatrix} D & E \\ E^T & I_{n_1} \end{pmatrix},$$

where $D = I_{n_0} + S_0^{h_\tau}[\mathbf{x}_0]\big[I_{n_0} + S_1^{h_1}[\mathbf{x}_0]S_1^{h_1}[\mathbf{x}_0]^T]S_0^{h_\tau}[\mathbf{x}_0]^T - S_0^{h_\tau}[\mathbf{x}_0]\big] - S_0^{h_\tau}[\mathbf{x}_0]^T$ and $E = S_0^{h_\tau}[\mathbf{x}_0]S_1^{h_1}[\mathbf{x}_0]$.

A.2 Variance of $\hat{\tau}^{backfit}$

$$\begin{aligned} Var\big(\hat{\tau}^{backfit}(z)\big) &= Var\big(\hat{\beta}_1^{backfit}(z) - \hat{\beta}_0^{backfit}(z)\big) \\ &= Var\big(\hat{\beta}_1^{backfit}(z)\big) + Var\big(\hat{\beta}_0^{backfit}(z)\big) \\ &\quad -Cov\big(\hat{\beta}_1^{backfit}(z), \hat{\beta}_0^{backfit}(z)\big) - Cov\big(\hat{\beta}_1^{backfit}(z), \hat{\beta}_0^{backfit}(z)\big)^T \end{aligned} \tag{20}$$

We are left to evaluate the covariance between $\hat{\beta}_1^{backfit}(z)$ and $\hat{\beta}_0^{backfit}(z)$. We have

$$Cov\big(\hat{\beta}_1^{backfit}(z), \hat{\beta}_0^{backfit}(z)\big) = S_{0,1}^{h_0}[z] U S_{0,1}^{h_1}[z]^T,$$

where

$$\begin{aligned} U &= Cov\big((y_0 + S_0^{h_\tau}[\mathbf{x}_0](S_1^{h_1}[\mathbf{x}_0]y_1 - y_0), y_1)^T, (y_0, y_1 - S_1^{h_\tau}[\mathbf{x}_1](y_1 - S_0^{h_0}[\mathbf{x}_1]y_0))^T\big) \\ &= \sigma^2 \begin{pmatrix} I_{n_0} - S_0^{h_\tau}[\mathbf{x}_0] & F \\ 0 & I_{n_1} - S_1^{h_\tau}[\mathbf{x}_1] \end{pmatrix}, \end{aligned}$$

and

$$F = (I_{n_0} - S_0^{h_\tau}[\mathbf{x}_0])S_0^{h_0}[\mathbf{x}_1]^T S_1^{h_\tau}[\mathbf{x}_1]^T + S_0^{h_\tau}[\mathbf{x}_0]S_1^{h_1}[\mathbf{x}_0](I_{n_1} - S_1^{h_\tau}[\mathbf{x}_1])^T.$$

References

Abadie, A. and Imbens, G. (2004). Large sample properties of matching estimators for average treatment effects. Working paper, Berkeley University.

Akritas, M. G. and Van Keilegom, I. (2001). Nonparametric ancova methods for heteroscedastic nonparametric regression models. *Journal of the American Statistical Association*, **96**, 220–232.

Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-plus Illustrations*. Cambridge: Cambridge University Press.

Dehejia, R. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, **94**, 1053–1062.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. London: Chapman & Hall.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. London: Chapman & Hall.

Heckman, J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies*, **65**, 261–294.

Lalonde, R. (1986). Evaluating the econometric evaluations of training programs. *American Economic Review*, **76**, 604–620.

Neumeyer, N. and Dette, H. (2003). Nonparametric comparison of regression curves: an empirical process approach. *Annals of Statistics*, **31**, 880–920.

Ratkowsky, D. A. (1983). *Nonlinear regression modeling: A unified practical approach*. Marcel Dekker Inc.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688–701.

Young, S. G. and Bowman, A. W. (1995). Non-parametric analysis of covariance. *Biometrics*, **51**, 920–931.

Publication series published by the Institute for Labour Market Policy Evaluation (IFAU) – latest issues

Rapporter/Reports

2004:1 Björklund Anders, Per-Anders Edin, Peter Fredriksson & Alan Krueger "Education, equality, and efficiency – An analysis of Swedish school reforms during the 1990s"

2004:2 Lindell Mats "Erfarenheter av utbildningsreformen Kvalificerad yrkesutbildning: ett arbetsmarknadsperspektiv"

2004:3 Eriksson Stefan & Jonas Lagerström "Väljer företag bort arbetslösa jobbsökande?

2004:4 Forslund Anders, Daniela Fröberg & Linus Lindqvist "The Swedish activity guarantee"

2004:5 Franzén Elsie C & Lennart Johansson "Föreställningar om praktik som åtgärd för invandrares integration och socialisation i arbetslivet"

2004:6 Lindqvist Linus "Deltagare och arbetsgivare i friårsförsöket"

2004:7 Larsson Laura "Samspel mellan arbetslöshets- och sjukförsäkringen"

2004:8 Ericson Thomas "Personalutbildning: en teoretisk och empirisk översikt"

Working Papers

2004:1 Frölich Markus, Michael Lechner & Heidi Steiger "Statistically assisted programme selection – International experiences and potential benefits for Switzerland"

2004:2 Eriksson Stefan & Jonas Lagerström "Competition between employed and unemployed job applicants: Swedish evidence

2004:3 Forslund Anders & Thomas Lindh "Decentralisation of bargaining and manufacturing employment: Sweden 1970–96"

2004:4 Kolm Ann-Sofie & Birthe Larsen "Does tax evasion affect unemployment and educational choice?

2004:5 Schröder Lena "The role of youth programmes in the transition from school to work"

2004:6 Nilsson Anna "Income inequality and crime: The case of Sweden"

2004:7 Larsson Laura & Oskar Nordström Skans "Early indication of program performance: The case of a Swedish temporary employment program"

2004:8 Larsson Laura "Harmonizing unemployment and sickness insurance: Why (not)?"

2004:9 Cantoni Eva & Xavier de Luna "Non-parametric adjustment for covariates when estimating a treatment effect"

Dissertation Series

2002:1 Larsson Laura "Evaluating social programs: active labor market policies and social insurance"

2002:2 Nordström Skans Oskar "Labour market effects of working time reductions and demographic changes"

2002:3 Sianesi Barbara "Essays on the evaluation of social programmes and educational qualifications"

2002:4 Eriksson Stefan "The persistence of unemployment: Does competition between employed and unemployed job applicants matter?"

2003:1 Andersson Fredrik "Causes and labor market consequences of producer heterogeneity"

2003:2 Ekström Erika "Essays on inequality and education"