Propensity score matching without conditional independence assumption - with an application to the gender wage gap in the UK

Markus Frölich

Department of Economics, University of St. Gallen

Last changes: October 22, 2003

Abstract:

Propensity score matching is frequently used for estimating average treatment effects. Its applicability, however, is not confined to treatment evaluation. In this paper it is shown that propensity score matching does not hinge on a selection on observables assumption and can be used to estimate not only adjusted means but also their distributions, even with non-iid sampling. Propensity score matching is used to analyze the gender wage gap among graduates in the UK. It is found that subject of degree contributes substantially to explaining the gender wage gap, particularly at higher quantiles of the wage distribution.

Keywords: Selection on unobservables, covariate-adjustment, discrimination, wage gap, field of major, subject of degree

JEL classification: C13, C14, J16, J31, J71

The author is also affiliated with the Institute for the Study of Labor (IZA), Bonn. This research was supported by the Swiss National Science Foundation (project NSF 4043-058311) and the Grundlagenforschungsfonds HSG (project G02110112). I am grateful to the UK Data Archive at the University of Essex for providing the data and to Michael Lechner, Stephen Machin, Blaise Melly, Ruth Miquel, Patrick Puhani, Conny Wunsch and two anonymous referees for helpful comments and suggestions. Address for correspondence: Markus Frölich, Swiss Institute for International Economics and Applied Economic Research (SIAW), University of St.Gallen, Dufourstrasse 48, CH-9000 St.Gallen, Switzerland; markus.froelich@unisg.ch, www.siaw.unisg.ch/froelich

1 Introduction

Propensity score matching is widely used in evaluation to estimate average treatment effects when selection is on observables. The propensity score plays such a fundamental role since it enables using *one-dimensional* nonparametric regression techniques for estimating treatment effects, even with many confounding variables. Rosenbaum and Rubin (1983) showed that, if treatment assignment is ignorable (i.e. selection is on observables), matching on the probability of treatment receipt estimates the average treatment effect consistently. Hence, instead of matching on all covariates X, matching on this one-dimensional probability is sufficient, and this dimension reduction has been used in many evaluations.[1]

In this paper, it is argued that the nonparametric technique of propensity score matching is more versatile and not restricted to treatment evaluation only. It is shown that propensity score matching can be used to estimate a *covariate-distribution adjusted mean* even if selection is on unobservables. The justification of propensity score matching does not hinge on a conditional independence assumption. Propensity score matching can not only be used to estimate the mean, but also to estimate the density and distribution function of the counterfactual outcome variable. These results extend also to the case of non-random sampling (stratified sampling and/or choice-based sampling), when using *weighted* propensity score matching with an appropriate estimator of the propensity score. With stratified sampling, the propensity score estimator generally must take account of the sampling weights. Nevertheless, under certain conditions on the stratification scheme, sampling weights can be neglected for the propensity score, but need to be accounted for in the weighted matching estimator. Choice-based sampling, on the other hand, can be neglected in any case.

These results are useful in at least three areas. First, propensity score matching can be used in a wide range of applications to disentangle effects due to observables and due to unobservables. For example, in the analysis of discrimination, propensity score matching can be employed to distinguish between income differences due to different human capital compositions and differences due to unobservables or discrimination. Such a nonparametric approach avoids

[1] See for example, Brodaty, Crépon, and Fougère (2001), Dehejia and Wahba (1999), Frölich, Heshmati, and Lechner (2004), Gerfin and Lechner (2002), Heckman, Ichimura, and Todd (1997), Heckman, Ichimura, Smith, and Todd (1998), Jalan and Ravallion (2002), Larsson (2003), Lechner (1999), Puhani (1999), Rosenbaum and Rubin (1983, 1985) , Rubin and Thomas (1992, 1996) and Smith and Todd (2000).

the parametric assumptions of, for example, a Blinder (1973)-Oaxaca (1973) decomposition. It also allows to examine the effects of observables and unobservables on the whole distribution.

Second, propensity score matching can be used to predict the counterfactual distribution of Y if the distribution of X were changed but the relationship between Y and X remained unchanged. Juhn, Murphy, and Pierce (1993), DiNardo, Fortin, and Lemieux (1996) and Gosling, Machin, and Meghir (2000) analyzed the increasing wage inequality in the USA and UK. They simulated the counterfactual wage distribution that would have prevailed if the distribution of workers' characteristics and other covariates such as the real value of the minimum wage had changed over time but the relationship between these covariates had remained unchanged (or vice versa). To simulate this counterfactual wage inequality, propensity score matching could be used as well.

Third, propensity score matching is also applicable with difference-in-difference matching, even when the selection bias does not remain constant over time for every value of X, as assumed for example in Heckman, Ichimura, and Todd (1997) and Smith and Todd (2000). It suffices if the average selection bias remains constant and can be estimated in a baseline period. In fact, when multiple baseline periods are available, the average selection bias does not even need to remain constant. It suffices if it can be predicted consistently from the estimated average selection biases in the previous periods. E.g. if the selection bias grows at a constant rate over time and two baseline periods are available, the growth rate of the average selection bias can be estimated from the average selection biases estimated in the two baseline periods. For example, in the analysis of longitudinal data, attrition is a frequent problem. If interest is in the evolution of a cohorts' characteristics over time (e.g. income, employment rate, health), non-random attrition has to be taken into account. If data is missing at random conditional on some covariates X (Little and Rubin 1987), adjusting for the distribution of X in the baseline period (e.g. by propensity score matching) gives an unbiased estimate of the cohorts' characteristics that would have been observed in the absence of any attrition. However, if data are not missing at random (conditional on X), this estimate would be biased. If all individuals responded in the baseline period, the magnitude of the bias from using the matching estimator could be estimated by pretending that all individuals attriting up to a time t were also non-responding in the baseline period. Comparing the simulated baseline-period characteristics for the attriters with the observed characteristics gives an estimate of the bias for the baseline

period. If this bias can be assumed to be constant over time, it can be subtracted from the matching estimator at time t. If multiple periods without attrition are available, the average selection bias can be estimated for all these periods and then be extrapolated to later periods. Usually imputation methods or inverse probability weighting are used to deal with attrition and non-response, but propensity score matching could be used as well.

In Section 2 the applicability of propensity score matching without assuming selection on observables is shown. Section 3 analyzes the finite-sample properties of propensity score matching and multivariate matching estimators. In Section 4, propensity score matching is applied to analyze the gender wage gap of college graduates in the UK. It is examined to which extent the gender wage gap can be explained by observed characteristics. By simulating the whole counterfactual wage distribution, not only the mean wage gap is analyzed but also the gender wage gap at different quantiles. Of particular interest is the contribution of 'subject of degree' (= field of major) to explaining the gender wage differences; a variable that is not available in most datasets. It turns out that subject of degree contributes substantially to reducing the unexplained wage gap, particularly in the upper tail of the wage distribution. The huge wage differential between high-earning men and high-earning women is thus to a large extent the result of men and women choosing different subjects in university. Section 5 concludes.

2 Propensity score matching without CIA

Let $f_{x|s}$ be the density of some covariates X in a particular population (denoted source population) and let $f_{x|t}$ be the density of X in a target population. In many situations, one is interested in the mean of an outcome variable Y in the source population if its covariates were distributed as in the target population:

$$\int m_s(x) \cdot f_{x|t}(x)dx, \tag{1}$$

where $m_s(x) = E_s[Y|X = x]$ is the conditional mean of Y in the source population. A nonparametric estimator of this *covariate-distribution adjusted mean* is the matching estimator

$$\frac{1}{n_t} \sum_{i:D_i=t} \hat{m}_s\left(X_i\right),$$

where $D_i \in \{s, t\}$ indicates whether an observation i is drawn from the source or the target population, n_t is the number of observations in the target population and $\hat{m}_s(x)$ is a nonparametric estimator of $E_s[Y|X = x] = E[Y|X = x, D = s]$.

For example, let the source population be women and the target population be men, and let Y being wages and the variables X describing human capital endowment, then the adjusted mean (1) would represent the counterfactual mean wage that women would receive if they had the human capital endowment of men.

Estimation of the covariate-distribution adjusted mean plays a prominent role in treatment evaluation for estimating average treatment effects when selection is on observables. For treatment evaluation, one wants to compare the expected outcome with and without a particular treatment.[2] Define Y_i^0, Y_i^1 as the *potential outcomes* of individual i. Let $D_i \in \{0, 1\}$ indicate whether an observation received treatment or not. The average impact of the treatment is often measured by the average treatment effect on the treated $E[Y^1 - Y^0|D = 1]$, which is not identified (unless treatment was randomly assigned) since the counterfactual mean $E[Y^0|D = 1]$ is unidentified. If all variables that influenced treatment assignment as well as the potential outcomes are observed, then conditional on these confounding variables X, the potential outcomes are stochastically independent ($\perp\!\!\!\perp$) of D

$$Y^0 \perp\!\!\!\perp D \,|X. \tag{2}$$

With this assumption, known as *selection on observables* (Barnow, Cain, and Goldberger 1981, Heckman and Robb 1985), *ignorable treatment assignment* (Rosenbaum and Rubin 1983) or *conditional independence assumption* (Lechner 1999), the counterfactual mean outcome $E[Y^0|D = 1]$ is identified by adjusting the mean of Y among the non-participants ($D_i = 0$) for the distribution of X among the participants ($D_i = 1$)

$$E[Y^0|D = 1] = \int m_0(x) \cdot f_{x|D=1}(x)dx,$$

where $m_0(x) \equiv E[Y|X = x, D = 0]$. This expression is identical to (1), with the source population being the non-participants and the target population being the participants.

Rosenbaum and Rubin (1983) showed that, if the conditional independence assumption (2) is valid, then the potential outcomes are also independent of treatment assignment, conditional

[2] For example, a drug treatment, college attendance or participation in an active labour market programme.

on the *propensity score:*

$$Y^0 \perp\!\!\!\perp D \,|p(X), \tag{3}$$

where $p(x) = \Pr(D = 1|X = x)$. Hence, the conditional mean $E[Y^0|D = 1]$ can also be estimated by matching on the propensity score

$$\frac{1}{n_t} \sum_{i:D_i=t} \hat{\mathfrak{m}}_0 (p_i)$$

where $\hat{\mathfrak{m}}_0 (\rho)$ is an estimator of $\mathfrak{m}_0 (\rho) = E[Y^0|p(X) = \rho, D = 0]$ and $p_i = p(X_i)$. Imbens (2000) and Lechner (2001) extended this result to the evaluation of multiple treatments (e.g. varieties of treatments). Since the propensity score $p(x)$ is *one-dimensional*, in finite samples nonparametric regression of Y_i on p_i can be much less demanding than nonparametric regression on X_i, particularly when the number of confounding variables X is large. Because of this 'dimension reduction' property (Hahn 1998), *propensity score matching* is used in many evaluation studies.[3]

However, the use of propensity score matching is not confined to treatment evaluation only, and can be used in many other situations as well. This versatility is demonstrated in Theorem 1, which shows that matching on the propensity score is justified as an estimator of the adjusted mean even when the conditional independence assumption (2) is not valid and selection is on unobservables. Theorem 1 also shows that propensity score matching can not only be used for estimating the adjusted mean but also its pdf and cdf.

Theorem 1:

Let $(Y, X, D) \in \Re \times \Re^k \times \Re$ be random variables with $\Pr(D = s) > 0$ and $\Pr(D = t) > 0$. Let $\xi_{y|x,d}(x, d)$ be a functional of the distribution of Y given $X = x$ and $D = d$. Let $\xi_{y|p,d}(p, d)$ be the corresponding functional given the propensity score $P = p$ and $D = d$, where the propensity score is defined as $p(x) = \Pr(D = t|X = x, D \in \{s, t\})$. Denote by $f_{x|s}(x)$ the density of X given $D = s$ and let $\mathcal{S} \equiv \{x : f_{x|s}(x) > 0\} = \{x : p(x) < 1\}$ denote the support of X in the source population. Suppose that $\xi_{y|x,d}(x, s)$ and $\xi_{y|p,d}(p, s)$ are both finite for all values of

[3] The properties of (propensity score) matching estimators of average treatment effects under the selection-on-observables assumption (2) have been studied by Abadie and Imbens (2001), Hahn (1998), Heckman, Ichimura, and Todd (1998), Hirano, Imbens, and Ridder (2003) and Lechner (2001). For a recent overview see Imbens (2003).

$x \in \mathcal{S}$ and $p < 1$. If

$$E\left[\xi_{y|x,d}(X,s) \mid p(X) = \rho, D = s\right] = \xi_{y|p,d}(\rho, s) \tag{4}$$

then

$$\underset{\mathcal{S}}{E}\left[\xi_{y|x,d}(X,s) \mid D = t\right] \quad = \quad \underset{\mathcal{S}}{E}\left[\xi_{y|p,d}(P,s) \mid D = t\right], \tag{5}$$

where the expectation is taken with respect to the support $\mathcal{S}$ in the source population. (Proof in appendix.)

Hence, matching on X

$$\frac{1}{n_t^{\mathcal{S}}} \sum_{\substack{i:D_i=t \\ X_i \in \mathcal{S}}} \hat{\xi}_{y|x,d}(X_i, s)$$

and matching on the propensity score

$$\frac{1}{n_t^{\mathcal{S}}} \sum_{\substack{i:D_i=t \\ p_i<1}} \hat{\xi}_{y|p,d}(p_i, s)$$

converge to the same limit, provided $\hat{\xi}_{y|x,d}$ and $\hat{\xi}_{y|p,d}$ are consistent estimators. $n_t^{\mathcal{S}}$ is the number of observations with $D_i = t$ and $X_i \in \mathcal{S}$.

The condition (4) is satisfied, for example, for the *conditional mean* $\xi_{y|x,d}(x,d) = E[Y|X = x, D = d]$ and $\xi_{y|p,d}(\rho,d) = E[Y|P = \rho, D = d]$. It is also satisfied for the *conditional density function* $\xi_{y|x,d} = f_{Y|X,D=d}$ and $\xi_{y|p,d} = f_{Y|P,D=d}$, and for the *conditional distribution function* $\xi_{y|x,d} = F_{Y|X,D=d}$ and $\xi_{y|p,d} = F_{Y|P,D=d}$.

On the other hand, (4) is *not* satisfied for the conditional variance and for the conditional quantiles, e.g. the median. Hence, if one were interested in the adjusted variance

$$\int Var\left[Y|X = x, D = s\right] \cdot f_{x|t}(x)dx$$

or the adjusted median

$$\int Median\left[Y|X = x, D = s\right] \cdot f_{x|t}(x)dx,$$

matching on X would be consistent, but propensity score matching would not be.

The results of Theorem 1 are relevant, since it is often not only of interest to estimate the mean of Y in a particular population if its covariates X were distributed as in an other population, but also its whole distribution. The density function of Y in the adjusted population is

given by

$$\int f_{y|x,s}(a,x) \cdot f_{x|t}(x)dx \tag{6}$$

and its distribution function is

$$\int F_{y|x,s}(a,x) \cdot f_{x|t}(x)dx, \tag{7}$$

where $f_{y|x,s}(a,x)$ is the conditional density function of Y in the source population evaluated at x and $F_{y|x,s}(a,x)$ is the conditional distribution function. Since nonparametric estimation of the conditional pdf $f_{y|x,s}$ and conditional cdf $F_{y|x,s}$ suffers from a similar curse of dimensionality problem as the estimation of the conditional mean $m_s(x)$, matching on the propensity score may help here as well to reduce the dimensionality of the estimation problem. The propensity score matching estimators for the adjusted mean, pdf and cdf are

$$\begin{aligned}
&\frac{1}{n_t^{\mathcal{S}}} \sum_{\substack{i:D_i=t \\ p_i<1}} \hat{\mathfrak{m}}_s(p_i) \\
&\frac{1}{n_t^{\mathcal{S}}} \sum_{\substack{i:D_i=t \\ p_i<1}} \hat{f}_{y|p,s}(a,p_i) \\
&\frac{1}{n_t^{\mathcal{S}}} \sum_{\substack{i:D_i=t \\ p_i<1}} \hat{F}_{y|p,s}(a,p_i),
\end{aligned}$$

respectively, where $\hat{\mathfrak{m}}_s(\rho)$ is a nonparametric estimator of $\mathfrak{m}_s(\rho) = E[Y|p(X) = \rho, D = s]$, $\hat{f}_{y|p,s}(a,\rho)$ is a nonparametric estimator of the pdf of Y at a in the source population given the propensity score, and $\hat{F}_{y|p,s}(a,\rho)$ is a nonparametric estimator of the cdf of Y given the propensity score. $F_{y|p,s}(a,\rho)$ could be estimated, for example, by nonparametric regression of $1(Y \leq a)$ on the propensity score in the source sample.

Hence, propensity score matching may be useful in a variety of circumstances. In Section 4, propensity score matching will be used to analyze the gender wage gap. In the analysis of race or gender discrimination, the Blinder (1973)-Oaxaca (1973)-decomposition has frequently been used to decompose the mean wage gap. The underlying idea is to estimate which part of the wage gap is due to observables and which part is due to unobservables (e.g. differences in returns to human capital, differences in unobserved ability or motivation). The nonparametric counterpart to the Blinder-Oaxaca decomposition is to simulate the counterfactual mean wage

$$\int m_m(x) \cdot f_{x|w}(x)dx,$$

where $m_m(x)$ is the mean wage for men with characteristics x and $f_{x|w}$ is the density of these characteristics among women. This counterfactual represents the average wage women would earn if they received the same returns to their skills as men. It also represents the wage that men would earn on average if they had the human capital endowment of women.

However, it might not only be interesting to simulate the counterfactual mean, but also the whole counterfactual distribution, i.e. the distribution of wages women would earn if they were remunerated as men. Estimating the counterfactual wage distribution (6) is more informative than only estimating the counterfactual mean. As Theorem 1 shows, propensity score matching can be used for this purpose.

Another context where propensity score matching could be used is in accounting for survey non-response and attrition in panel studies. Attrition is a serious problem in many longitudinal surveys, where after several waves often only a fraction of the original respondents may be located and willing to participate in the interview. The remaining sample may no longer be representative of the population, since attrition is usually non-random. If one is interested in how the population mean of some outcome variable Y, e.g. the employment rate, evolves over time, one needs to control for non-random attrition to distinguish changes in the population mean from changes in the sample composition. Consider a sample of observations (Y_t, X_t), where t indexes the survey wave. In the baseline period $t = 0$, the sample is representative of the population of interest. In the subsequent periods, (Y_t, X_t) is only observed for individuals that did not attrit. Denote by $D_t = 1$ if an individual remains in the survey, i.e. such that (Y_t, X_t) are observed. In many surveys, a substantial fraction attrits after the first interview, such that only (Y_0, X_0) is observed for them. For estimating the population mean $E[Y_1]$ in period 1, it is necessary to identify the mean $E[Y_1|D_1 = 0]$ in the attriting subpopulation. If data is missing at random (Little and Rubin 1987) conditional on the X_0 values in the base period:[4] $Y_1 \perp\!\!\!\perp D_1 \,|X_0$, the mean among the attriters is identified by the adjusted mean

$$E\left[E\left[Y_1|X_0, D_1 = 1\right] \,|D_1 = 0\right] = \int E\left[Y_1|X_0, D_1 = 1\right] f_{X_0|D_1=0}(x)dx, \tag{8}$$

where $f_{X_0|D_1=0}$ is the density of X_0 in the attriting subpopulation. However, even after controlling for the characteristics X_0, attrition may not be random. Attrition is often related to

[4] If $E[Y_t]$ at time $t > 1$ is of interest, attrition may depend as well on the values X_0 as on $X_1, .., X_{t-1}$. Then the sequential nature of attrition in subsequent periods needs to be taken into account. This, however, is beyond the scope of this paper. See for instance Lechner and Miquel (2002).

an individual's attitude or propensity to moving, which may also affect the variable of interest Y, for example earnings. An individual who does not feel bound to a certain neighbourhood may have higher earnings, since he is more likely to take advantage of better job offers in other geographic regions. In this case, the adjusted mean (8) would lead to a biased estimate of the mean $E[Y_1|D_1 = 0]$ in the attriting subpopulation. The individual's attitude to moving, however, is likely to having affected Y also in period 0, implying that $E[Y_0|X_0, D_1 = 1] \neq E[Y_0|X_0, D_1 = 0]$. Accordingly, the adjusted mean in period 0 would also be a biased estimate of $E[Y_0|D_1 = 0]$:

$$\int E\left[Y_0|X_0, D_1 = 1\right] f_{X_0|D_1=0}(x)dx - E\left[Y_0|D_1 = 0\right] \;=\; \mathcal{B} \;\neq 0. \tag{9}$$

Since (Y, X) are observed for all individuals in the base period $t = 0$, all terms in (9) are identified and the bias $\mathcal{B}$ can be estimated. If the unobserved characteristics have a similar effect in period 0 and 1, e.g. the effect of individuals' attitude to moving on earnings, the bias may be similar in period 0 and 1, and the mean $E[Y_1|D_1 = 0]$ in the attriting subpopulation can be estimated via

$$E\left[Y_1|D_1 = 0\right] = \int E\left[Y_1|X_0, D_1 = 1\right] f_{X_0|D_1=0}(x)dx - \mathcal{B}, \tag{10}$$

where the bias is estimated from (9). This is a type of difference-in-difference estimator after adjusting for some covariates X_0. An analogous expression can be obtained, if interest lies in the estimation of the distribution function $F_{Y_1|D_1=0}$ of Y_1 in the attriting subpopulation. If multiple baseline periods without attrition are available, the assumption that the bias is identical in period 0 and 1 can be relaxed. It suffices if the bias in the period of interest can be predicted consistently from the estimated biases in the baseline periods. This is further discussed below.

The most frequent approach used to estimate the adjusted mean in (8) or in (9) and (10) is weighting by the inverse of the probability to attrit. This inverse probability weighting is analyzed for example in Horvitz and Thompson (1952), Horowitz and Manski (1998), Robins, Rotnitzky, and Zhao (1995), Wooldridge (2002) and others. According to Theorem 1 also propensity score matching could be used to estimate the adjusted means, where the propensity score is the probability to attrit given X_0. Propensity score matching may have the advantage over inverse probability weighting in small samples that it is less susceptible to values of the

propensity score close to 1, which may lead to almost zero denominators in the weighting estimator.

This difference-in-difference matching is also often used in treatment evaluation. In evaluation settings where not all confounding variables can be observed, the conditional independence assumption (2) is invalid and selection bias is introduced due to unobservables. Nevertheless, propensity score matching can still be used to estimate the adjusted mean. After adjusting for the differences in the X distributions, the remaining selection bias due to unobservables at time t is

$$\mathcal{B}_t = \int E[Y_t^0|X=x, D=0] \cdot f_{X|D=1}(x)dx - E\left[Y_t^0|D=1\right].$$

If data on the outcome variable are available before treatment starts ($t<0$), the bias due to unobservables can be estimated for these pre-treatment periods since all individuals are in the untreated state: $E[Y_\tau^0|X, D=1]$ and $E[Y_\tau^0|X, D=0]$ are identified for $\tau<0$. If the changes in the average bias over time are smooth (e.g. bias grows at a constant rate) such that the bias $\mathcal{B}_t$ in period t can be predicted consistently from $\hat{\mathcal{B}}_{-1}$, $\hat{\mathcal{B}}_{-2}$, $\hat{\mathcal{B}}_{-3}$... (*predictable bias assumption*), the average treatment effect on the treated is identified. For example, in the evaluation of active labour market programmes, administrative data on earnings and employment status are often available for several years before treatment.

If pre-treatment data are available only for a single period τ, the treatment effect is identified only if the bias is constant over time: $\mathcal{B}_t = \mathcal{B}_\tau$. (It is not necessary to assume that for every value of x, the bias remains constant $E\left[Y_t^0 - Y_\tau \,|X, D=1\right] = E\left[Y_t^0 - Y_\tau \,|X, D=0\right]$, as it is done in Heckman, Ichimura, and Todd (1997), Heckman, Ichimura, Smith, and Todd (1998) and Smith and Todd (2000).)

Two further issues are relevant for using propensity score matching to estimate the adjusted mean/pdf/cdf. First, the adjusted mean/pdf/cdf is only well defined with respect to the support $\mathcal{S}$ of X in the source population. In the literature, it has often been assumed that

$$0 < p(x) < 1 \qquad \forall x. \tag{11}$$

However, as Heckman, Ichimura, and Todd (1998) point out, in many empirical applications this assumption is not appropriate. The support of X in the target population may often not

be a subset of $\mathcal{S}$, such that $\xi_{y|x,d}(x,s)$ is undefined for all $x \notin \mathcal{S}$. In this case, the adjusted mean/pdf/cdf is only identified with respect to the subpopulation of the target population with $x \in \mathcal{S}$. With a (propensity score) matching estimator, this restriction can easily be implemented by eliminating all target sample observations with $X_i \notin \mathcal{S}$. Since the supports of X in source and target population are unknown, $\mathcal{S}$ needs to be estimated. Heckman, Ichimura, and Todd (1997) and Heckman, Ichimura, Smith, and Todd (1998) incorporate this support restriction through a trimming rule. The density $f_{x|s}(X_i)$ is nonparametrically estimated for all observations i of the target sample. All observations with estimated density $\hat{f}_{x|s}(X_i) = 0$ are eliminated[5] along with another 2% of the target sample observations with the smallest estimated positive density values, see Heckman, Ichimura, Smith, and Todd (1998, p.1081). This procedure, however, would also trim 2% or more of the observations even if the assumption (11) were valid. An alternative approach, used for example in Nopo (2002), is to estimate $\mathcal{S}$ by the empirical support of X in the source population: x is considered as being in $\mathcal{S}$ if there is at least one observation j in the source sample with $X_j = x$. Besides being inapplicable if x contains any continuous variables, this approach may lead to a very conservative estimate of the support. The number of target sample observations that are classified as being out of the support $\mathcal{S}$ strongly depends on the coarseness of the definition of the X variables and decreases with the number of observations in the source sample. Also the shape of the estimated support may be very unreasonable: A 36-year old woman may be classified as being out of support, while a 35-year and a 37-year old are not, simply because no 36-year old woman is observed in the source sample. If the estimate of the adjusted mean with respect to $\mathcal{S}$ is interpreted as being indicative of the adjusted mean in the whole population (as it is often done), a very restrictive estimate of $\mathcal{S}$ reduces considerably the size of the subpopulation for which the adjusted mean is estimated and thus reduces the usefulness of this estimate for interpretation and policy conclusions.

A more formal approach to incorporating the support restriction consists in estimating $f_{x|s}(X_i)$, e.g. by kernel density estimation, and to test $f_{x|s}(X_i) = 0$ versus $f_{x|s}(X_i) > 0$ for every observation i of the target sample. The significance level α of the test determines the probability that an out-of-support observation of the target sample is classified as within-support. If the supports in source and target population have empty overlap, then $100(1-\alpha)\%$

[5] This can only occur when a bounded kernel is used.

of the target sample observations would be deleted. In the case of complete overlap, the fraction deleted depends on the power of the test. Since by definition $f_{x|s}(X_j) > 0$ for every observation j of the source sample, the power of the test can be examined by testing $f_{x|s}(X_j) = 0$ also for all the source sample observations. If for a reasonable number of source observations the test does not reject, increasing the α-level might be appropriate. Since $p(x) < 1$ if and only if $x \in \mathcal{S}$, testing $p(X_i) = 1$ versus $p(X_i) < 1$ for every target observation, is an alternative to implementing the common support restriction. For calibrating the α-level, it can be used that $p(X_j) < 1$ by definition for every source sample observation.

The second issue concerns the estimation of the propensity score, which is usually unknown in most applications. Since the estimation of the propensity score is by itself a highly dimensional estimation problem, the practical value of the dimension-reduction property of the propensity score is somewhat diminished. In practice, it may be that inference is "less sensitive to specifications of the propensity score than to specification of the conditional expectation of the potential outcomes" (Imbens 2000). In addition, estimating the propensity score by a logit or probit model may lead to very similar results as when using more sophisticated nonparametric or semiparametric estimation techniques for the propensity score (Heckman, Ichimura, and Todd 1997). Nevertheless, it is theoretically ambiguous whether propensity score matching is superior to direct matching on the X covariates, since even when the propensity score is known, matching on the X can be more efficient than propensity score matching for estimating the treatment effect on the treated, see Heckman, Ichimura, and Todd (1998). Hahn (1998) derived the semiparametric efficiency bounds for estimating average treatment effects and showed that matching on the propensity score is not necessary to achieve this bound.[6] Because of this ambiguity, a small Monte Carlo simulation is conducted in Section 3, to compare the finite-sample properties of propensity score matching and multivariate matching on X.

[6] Imbens (2003) points out an advantage of propensity score matching as regards the accuracy of asymptotic variance approximations, arguing that the asymptotic approximations may be more accurate with one-dimensional nonparametric regression (as for using the propensity score) compared to higher-dimensional nonparametric regression (as with multivariate matching).

2.1 Propensity score matching with non-iid data

The previous discussion implicitly assumed the availability of random samples. In practice, however, the available data is rarely randomly sampled. Survey data is often collected through multi-stage stratified sampling or variable probability sampling procedures (see e.g. Imbens and Lancaster (1996) or Wooldridge (1999)), leading to samples that are not directly representative of the population(s) of interest. Certain groups, such as treatment participants, foreigners, low-income individuals or residents of particular regions, are often deliberately oversampled. Usually, sampling is non-random across strata but random within strata, and stratification may be with respect to Y, X or D. Often sampling weights (proportional to the inverse of the sampling probability) are provided, from which the true sampling probabilities can be recovered, see e.g. Wooldridge (1999).

Stratified sampling also includes *choice-based sampling*, where non-random sampling is with respect to D. Choice-based sampling occurs very frequently, for example, in treatment evaluation where the treatment participants are usually oversampled relative to the number of non-participants. It will also quite likely occur, when the source sample and the target sample stem from separate surveys or data sources. With choice-based sampling, the ratio of the number of observations in the source and the target sample does not correspond to the relative size of the source and the target populations. When sampling is choice-based, estimation of the propensity score will usually be biased (Manski and Lerman 1977). However, as will be seen below, this is irrelevant for estimating the adjusted mean/pdf/cdf by propensity score matching. Weighted propensity score matching is consistent even with this biased propensity score.

Stratification may be with respect to Y, X or D. Let $f^0_{yx|s}$ be the joint population density of Y and X in the source population and $f^0_{yx|t}$ be the population density in the target population. Let $f^*_{yx|s}$ and $f^*_{yx|t}$ be the *sampling* densities in the respective populations, and suppose that the sampling densities are positive wherever the population densities are:

$$supp\left(f^*_{yx|d}\right) = supp\left(f^0_{yx|d}\right) \qquad \text{for } d \in \{s,t\}\,.$$

Define the sampling weights $w_{yx|d}(y,x|d)$ as

$$w_{yx|d} = \frac{f^0_{yx|d}}{f^*_{yx|d}}, \tag{12}$$

which are assumed to be known. These sampling weights are inversely proportional to the fraction of the population with characteristics X, Y that is sampled. With variable probability sampling, for example, an individual is drawn randomly from the population and some preliminary characteristics (age, gender, income) are gathered. With a probability $\Pr^0(\text{in sample}|X,Y)$ depending on these characteristics, the individual is kept in the sample and is further interviewed. Otherwise it is dropped from the sample. These probabilities are inversely proportional to $w_{yx|d}$

$$\Pr{}^0(\text{in sample}|X,Y,D=d) \propto \frac{1}{w_{yx|d}}.$$

Besides stratification on Y and X, there may also be *choice-based sampling* (stratification on D). Choice-based sampling occurs frequently in treatment evaluation where the treatment participants are usually oversampled relative to the number of non-participants. Let $P_t^0 = \Pr^0(D=t|D\in\{s,t\})$ be the true population measure of the target population (relative to the source population). Let $P_t^* = \Pr^*(D=t|D\in\{s,t\})$ be the measure of the target population under choice-based sampling. Let α characterize the odds-ratio:

$$\alpha = \frac{P_t^*}{P_s^*}\frac{P_s^0}{P_t^0} - 1. \tag{13}$$

If $\alpha > 0$, the target population is oversampled. If $\alpha = 0$, there is no choice-based sampling.

The object of interest is, as before, the adjusted functional

$$\underset{\mathcal{S}}{E^0}\left[\xi^0_{y|x,d}(X,s)\ |D=t\right] \tag{14}$$

where the superscript ^{0} in E^0 and in $\xi^0_{y|x,d}$ symbolizes that the expectation and the conditional functional are with respect to the true population distribution. To make the notation somewhat less cumbersome, the following derivation refers only to the adjusted *mean* and abstracts from the support problem, i.e. it supposes that the supports of X in the source and the target population are identical. The necessary changes for the adjustment of other functionals, e.g. the pdf or cdf, and for the incorporation of the support $\mathcal{S}$ are straightforward and would only complicate the notation.

Theorem 2 shows that the adjusted functional (14) can be estimated by propensity score matching from data sampled under the sampling distribution, provided the matching estimator is properly redefined. Three modifications are required to obtain an estimator consistent under

the population distribution. The *weighted propensity score matching* estimator

$$\frac{\sum_{i:D_i=t} w_i \cdot \hat{\mathfrak{m}}_s^0(p_i)}{\sum_{i:D_i=t} w_i}, \tag{15}$$

takes account of the sampling weights $w_i = w_{yx|d}(Y_i, X_i|D_i)$ in the target population. Second, to take into account the sampling weights in the source population, a weighted estimator $\hat{\mathfrak{m}}_s^0$ is required, for example weighted kernel regression[7]

$$\hat{\mathfrak{m}}_s^0(\rho) = \frac{\sum_{j:D_j=s} w_j \cdot Y_j \cdot K\left(\frac{p_j-\rho}{h}\right)}{\sum_{j:D_j=s} w_j \cdot K\left(\frac{p_j-\rho}{h}\right)}. \tag{16}$$

Third, the non-identical sampling needs to be taken into account in the estimation of the propensity score, which could be estimated e.g. by weighted nonparametric regression or by weighted Maximum Likelihood (Manski and Lerman 1977). The weights can be accounted for in different ways. On the one hand, one could take account of the sampling weights *and* of choice-based sampling. The resulting propensity score $p^0(x)$ would satisfy

$$p^0(x) = \frac{f_{x|t}^0(x) \cdot P_t^0}{f_{x|t}^0(x) \cdot P_t^0 + f_{x|s}^0(x) \cdot P_s^0}.$$

On the other hand, one could take account of sampling weights but *neglect* choice-based sampling weights, yielding the propensity score

$$p^{0*}(x) = \frac{f_{x|t}^0(x) \cdot P_t^*}{f_{x|t}^0(x) \cdot P_t^* + f_{x|s}^0(x) \cdot P_s^*}.$$

Completely neglecting both kinds of weighting would give the propensity score

$$p^*(x) = \frac{f_{x|t}^*(x) \cdot P_t^*}{f_{x|t}^*(x) \cdot P_t^* + f_{x|s}^*(x) \cdot P_s^*},$$

which would be estimated under the sampling distribution when weights are ignored. Under choice-based sampling ($\alpha \neq 0$), both propensity scores p^{0*} and p^* are biased (with respect to the population distribution).

Theorem 2 shows under which conditions propensity score matching with $p^0(x)$, $p^{0*}(x)$ or $p^*(x)$ as propensity score can be used to estimate the adjusted mean. The first part of Theorem

[7] With nearest-neighbour regression, as discussed in the next section, incorporating the weights would be less straightforward. This holds particularly for pair-matching.

2 shows that the limit expression of the weighted propensity score matching estimator based on the propensity score $p^0(x)$ is identical to the object of interest (14). The second part of Theorem 2 shows that this expression is identical when $p^{0*}(x)$ is used instead of $p^0(x)$. Hence, from this perspective, *choice-based sampling* is irrelevant for consistency of the matching estimator and can be neglected in the estimation of the propensity score.

The third part of Theorem 2 shows that in particular situations, also the propensity score matching estimator with $p^*(x)$ as propensity score has the same limit expression. Then it is irrelevant whether the propensity score under the population distribution or under the sampling distribution is used. In this case, the propensity score can be estimated by conventional (unweighted) estimation methods under the sampling distribution, completely neglecting *stratification and choice-based sampling.* The relevant condition for this equivalence result is (20)

$$\Pr{}^0(\text{in sample}|X, D = t) \propto \Pr{}^0(\text{in sample}|X, D = s)$$

as a function of x, or equivalently

$$w_{x|t}(x) \propto w_{x|s}(x),$$

where $w_{x|d} = f^0_{x,d}/f^*_{x,d}$ is the sampling weight with respect to X.[8] This condition requires that stratification with respect to x in the source sample is proportional to the stratification in the target sample. This condition is satisfied, for example, when the same *stratification design* was used for sampling the source sample and the target sample, e.g. when source and target sample are sampled in the same survey.[9] Hence, if the stratification design is the same in source and

[8] The sampling weights can be expressed as $w_{yx|d} = w_{y|x,d} \cdot w_{x|d}$, where $w_{y|x,d} = f^0_{y|x,d}/f^*_{y|x,d}$.

[9] Notice, however, that this proportionality condition refers to the marginal sampling probability with respect to X only. Even if the sampling weights with respect to Y and X are the same in the source and target population, they may differ with respect to X only. Consider the following example: Let Y be wages (high, low) and X education (high, low) and source and target population are men and women, respectively. A variable probability sampling scheme is employed, where first an individual is drawn and its wage is enquired. If the individual is a low-wage earner, it is kept with 100% probability in the sample and is further surveyed. Otherwise it is kept in the sample only with a probability of 20%. Hence, low wage earners are oversampled in the same way in the male and female population:

$$\Pr{}^0(\text{in sample}|X, Y, D = t) = \Pr{}^0(\text{in sample}|X, Y, D = s).$$

However, the marginal sampling probabilities with respect to X

$$\int \Pr{}^0(\text{in sample}|X, Y, D = d) \cdot dF_{y|x,d}$$

target sample, any weights can be neglected in the estimation of the propensity score. (I.e. a simple logit model instead of a weighted logit model could be used.) The presence ($\alpha \neq 0$) or absence ($\alpha = 0$) of choice-based sampling is completely irrelevant, only the stratification with respect to X and Y is relevant.

Similar to Theorem 1, the first part of Theorem 2 shows that the probability limit of the weighted propensity score matching estimator, which is based on the sampling distribution, is identical to the adjusted mean under the *population distribution*.[10] The left-hand side of (18) is the adjusted mean, which is the object of interest. The right-hand side is the probability limit of the weighted propensity score matching estimator. To see this, note that the weighted kernel regression $\hat{\mathfrak{m}}_s^0(\rho)$ in (16) converges to

$$\mathfrak{m}_s^0(\rho) = \frac{E^*\left[Y \cdot W_{yx|s}|p(X) = \rho, D = s\right]}{E^*\left[W_{yx|s}|p(X) = \rho, D = s\right]}$$

under appropriate conditions on the bandwidth sequence. With $\hat{\mathfrak{m}}_s^0(\rho)$ converging to $\mathfrak{m}_s^0(\rho)$, the limit of the weighted propensity score matching estimator (15) is

$$\int E^*\left[W_{yx|t}|p(X) = \rho, D = t\right] \cdot \mathfrak{m}_s^0(\rho) \cdot f_{p|t}^*(\rho)d\rho. \tag{17}$$

This is just the expression on the right-hand side in (18) of Theorem 2, when $p^0(x)$ is used as propensity score.

Theorem 2:

Let $f_{yx|d}^0$ be the population density of Y and X in the $D = d$ population. Let $f_{yx|d}^*$ be the *sampling* density, with $supp(f_{yx|d}^*) = supp(f_{yx|d}^0)$. Let $w_{yx|d} = f_{yx|d}^0/f_{yx|d}^*$ be the known sampling weights. Let $p^0(x) = \Pr^0(D = t|X = x, D \in \{s, t\})$ denote the propensity score under the population distribution and let $p^*(x) = \Pr^*(D = t|X = x, D \in \{s, t\})$ be the propensity

may differ if education and wages are related. Suppose that half of the men and half of the women are highly educated. High-educated men earn low wages with 10% probability, while low-educated men earn low wages with 50% probability. Women are paid lower wages than men, and the prevalence of low wage earners is 20% among high-educated and 60% among low-educated women. The probabilities of being kept in the sample are: 28% (60%) for high (low)-educated men and 36% (68%) for high (low)-educated women. Since the sampling probability ratio between source and target population for X=high-educated (36/28) is different from the ratio for X=low-educated (68/60), the proportionality assumption is violated.

[10] Where for ease of exposition, it has been assumed that the supports of X are identical in source and target population, such that the conditioning on $\mathcal{S}$ is no longer necessary.

score under the sampling distribution. Let $p^{0*}(x)$ be the propensity score under the population distribution but with choice-based sampling neglected. Let $f^0_{p^0}$, $f^0_{p^{0*}}$ and $f^0_{p^*}$ denote the densities of p^0, p^{0*} and p^*, respectively, under the population distribution, and let $f^*_{p^0}$, $f^*_{p^{0*}}$ and $f^*_{p^*}$ denote the corresponding densities under the sampling distribution. The expectation operator E^0 refers to the population distribution and E^* to the sampling distribution. Suppose that the expected values of Y, X and the sampling weights W exist under the population and under the sampling distribution. The following equivalences hold

a) for $p^0(x)$

$$E^0\left[E^0\left[Y|X, D=s\right] \ |D=t\right]$$
$$= \int E^*\left[W_{yx|t}|p^0(X)=\rho, D=t\right] \cdot \frac{E^*\left[Y \cdot W_{yx|s}|p^0(X)=\rho, D=s\right]}{E^*\left[W_{yx|s}|p^0(X)=\rho, D=s\right]} \cdot f^*_{p^0|t}(\rho)d\rho. \quad (18)$$

b) for $p^{0*}(x)$

$$= \int E^*\left[W_{yx|t}|p^{0*}(X)=\rho, D=t\right] \cdot \frac{E^*\left[Y \cdot W_{yx|s}|p^{0*}(X)=\rho, D=s\right]}{E^*\left[W_{yx|s}|p^{0*}(X)=\rho, D=s\right]} \cdot f^*_{p^{0*}|t}(\rho)d\rho. \quad (19)$$

c) and if the stratification design is identical for source and target sample in the sense that

$$\frac{f^*_{x|t}(x)}{f^0_{x|t}(x)} \frac{f^0_{x|s}(x)}{f^*_{x|s}(x)} = \gamma \quad (20)$$

is constant, then for $p^*(x)$

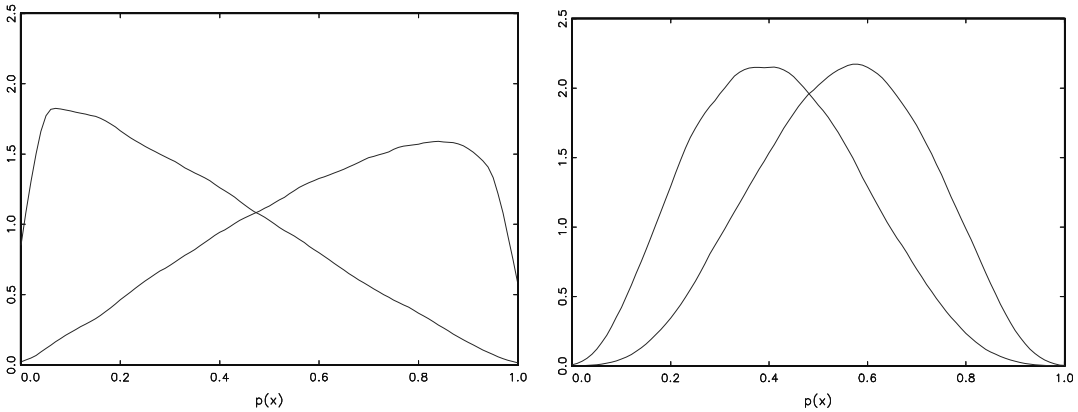
$$= \int E^*\left[W_{yx|t}|p^*(X)=\rho, D=t\right] \cdot \frac{E^*\left[Y \cdot W_{yx|s}|p^*(X)=\rho, D=s\right]}{E^*\left[W_{yx|s}|p^*(X)=\rho, D=s\right]} \cdot f^*_{p^*|t}(\rho)d\rho. \quad (21)$$

Proof in appendix.

3 Monte Carlo study

To analyze the properties of propensity score matching in finite samples, a small Monte Carlo experiment is conducted, where different estimators of the counterfactual mean are compared. Two simulation experiments are conducted: In the first experiment, the adjustment proceeds with respect to 5 continuous X variables. In the second experiment, X consists of 10 binary variables. The source and target populations are specified indirectly by first drawing the X covariates and assigning them subsequently to the source ($D = 0$) and target ($D = 1$)

Figure 3.1: Distribution of true propensity score in source and target population



Note: Left figure: $\sigma_\varepsilon^2 = 30$. Right figure: $\sigma_\varepsilon^2 = 100$.

population. This process allows to analyze the properties for correctly specified and misspecified propensity scores conveniently.

In the first experiment, the X covariates consist of 5 continuous variables. The first three variables are jointly normal (with variances 2, 1, 1, respectively, and covariances 1, -1, -0.5), X_4 is uniformly distributed in $(-3, 3)$ and X_5 is $\chi^2_{(1)}$ distributed. The source ($D = 0$) and target ($D = 1$) sample are formed by assigning each observation according to

$$D = 1\,(\; X_1 + 2X_2 - 2X_3 - X_4 - 0.5X_5 + \varepsilon \;> 0)\,,$$

where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. Two different values of σ_ε^2 are examined (30 and 100) and Tables B.1 and B.2 describe the distribution of X in the source and target population for these two values. In the design with $\sigma_\varepsilon^2 = 30$, the source and target population are very different, whereas they are more similar when $\sigma_\varepsilon^2 = 100$. This can also be seen from Figure 3.1, which plots the density of the propensity score in source and target population. In the left graph ($\sigma_\varepsilon^2 = 30$), the densities of the propensity score are more dissimilar than in the right graph ($\sigma_\varepsilon^2 = 100$). Hence, covariate adjustment will be more difficult in the design with $\sigma_\varepsilon^2 = 30$.

For the outcome variable Y (which is observed only in the $D = 0$ population), three designs are examined:

$$\begin{aligned} Y design\ 1\text{:} \quad & Y = X_1 + X_2 + X_3 - X_4 + X_5 + u \\ Y design\ 2\text{:} \quad & Y = X_1 + X_2 + 0.2X_3X_4 - \sqrt{X_5} + u \\ Y design\ 3\text{:} \quad & Y = (X_1 + X_2 + X_5)^2 + u, \end{aligned}$$

where $u \sim N(0, 1)$.

For these different designs, samples of size 100, 400 and 1600, respectively,[11] are drawn and the adjusted mean

$$\int m_0(x) \cdot f_{x|1}(x)dx$$

is estimated. Three different types of estimators are considered: covariate matching estimators, propensity score matching estimators and weighting estimators. Covariate matching estimates the adjusted mean as

$$\frac{1}{n_1} \sum_{i:D=1} \hat{m}_0(X_i),$$

where $\hat{m}_0(x)$ is an estimator of $E[Y|X = x, D = 0]$ obtained from the $D = 0$ sample. *Mahalanobis matching* estimates $\hat{m}_0(x)$ by first-nearest neighbour regression, where the distance between two observations X_1 and X_2 is defined as

$$(X_1 - X_2)'\hat{\Omega}^{-1}(X_1 - X_2)$$

and $\hat{\Omega}$ is an estimate of the covariance matrix of X.

Multivariate kernel matching estimates $\hat{m}_0(x)$ by nonparametric kernel regression

$$\hat{m}_0(x) = \frac{\sum_{j:D=0} w_j Y_j}{\sum_{j:D=0} w_j}$$

where the weights $w_j = w(X_j, x)$ are obtained through a multiplicative kernel

$$w(X_j, x) = \prod_{k=1}^{5} K\left(\frac{X_{j,k} - x_k}{h}\right),$$

where K is a kernel function, x_k denotes the k-th component of x and h is a bandwidth value. Two different kernel functions are examined in the simulations: Gaussian and Epanechnikov. Since only a single bandwidth value is used, the X data are scaled beforehand to mean zero and unit variance. The bandwidth is chosen by leave-one-out cross-validation in the $D = 0$ sample from a grid of 20 bandwidth values.[12]

The propensity score matching estimators require first an estimate of the propensity score for all observations. The propensity score is estimated by Probit regression of D on a set of regressors. Table 3.1 shows the three different regressor specifications that are considered in

[11] This is the total sample size. About half are in the source and half in the target sample, respectively.

[12] 0.001, $0.001 \cdot 1.7$, $0.001 \cdot 1.7^2$, ..., $0.001 \cdot 1.7^{18}$, ∞.

Table 3.1: Specification of the propensity score model

	Regressors	$Corr(\hat{p}, p)$
Specification 1:	$const, X_1, X_2, X_3, X_4, X_5$	1.00
Specification 2:	$const, X_1, X_2^2, X_3, X_4, X_5$	0.96
Specification 3:	$const, X_1^2, X_2^2, X_3, X_4^2, X_5^2$	0.80

$Corr(\hat{p}, p)$ is correlation of estimated propensity score with true propensity score (based on 10^6 observations).

the simulations. In the first specification, the model is correctly specified and the propensity score is estimated consistently for each observation. The second and third specifications are misspecified: the squares of some variables instead of the variables themselves are used as regressors.[13] These latter two models will give some indication about the sensitivity of the propensity score matching estimators to misspecification of the propensity score model. As a rough indicator of the extent of misspecification, the last column of Table 3.1 shows the correlation of the true propensity score with the propensity score estimated by the above probit specifications. As specification 1 is correct, the propensity scores are consistently estimated. In specification 2, only one regressor is misspecified and the estimated propensity scores are highly correlated with the true propensity scores. In specification 3, the misspecification is more severe and the correlation between the true and the estimated propensity score is only 0.80.

With the propensity score p_i estimated for all observations, the propensity score matching estimators estimate the adjusted mean as

$$\frac{1}{n_1} \sum_{i:D_i=1} \hat{\mathfrak{m}}_0(\hat{p}_i),$$

and differ in how the conditional mean $\mathfrak{m}_0(\rho) = E[Y|p(X) = \rho, D = 0]$ is estimated from the $D = 0$ observations. *Propensity score pair matching* estimates $\mathfrak{m}_0(\rho)$ by the outcome Y of the nearest neighbour. *Propensity score kernel matching* estimates $\mathfrak{m}_0(\rho)$ as

$$\hat{\mathfrak{m}}_0(\rho) = \frac{\sum_{j:D_j=0} Y_j \cdot K\left(\frac{p_j-\rho}{h}\right)}{\sum_{j:D_j=0} K\left(\frac{p_j-\rho}{h}\right)},$$

where the Gaussian and the Epanechnikov kernel are examined in the simulations. The bandwidth h is chosen by leave-one-out cross validation in the $D = 0$ sample, using the same grid

[13] This corresponds to a situation where the regressors enter linearly in the model, but enter with its square root in the true propensity score.

as with multivariate kernel matching.

Propensity score ridge matching, which was examined in Frölich (2004) and is based on the ridge regression estimator of Seifert and Gasser (1996, 2000), estimates $\hat{\mathfrak{m}}_0(\rho)$ as

$$\hat{\mathfrak{m}}_0(\rho) = \frac{T_0}{S_0} + \frac{T_1 \cdot (\rho - \bar{p})}{S_2 + rh\,|\rho - \bar{p}|},$$

where $S_a(\rho) = \sum_{j:D_j=0} (p_j - \bar{p})^a K\left(\frac{p_j-\rho}{h}\right)$ and $T_a(\rho) = \sum_{j:D_j=0} Y_j (p_j - \bar{p})^a K\left(\frac{p_j-\rho}{h}\right)$ and $\bar{p} = \sum_{j:D_j=0} p_j K\left(\frac{p_j-\rho}{h}\right) / \sum_{j:D_j=0} K\left(\frac{p_j-\rho}{h}\right)$. The ridge parameter r is set to 5/16 for the Epanechnikov kernel and 0.3535 for the Gaussian kernel, according to the 'rule-of-thumb' of Seifert and Gasser (2000).

Finally, a weighting estimator as suggested by Horvitz and Thompson (1952) and Hirano, Imbens, and Ridder (2003) is examined. *Propensity score weighting* estimates the adjusted mean as

$$\frac{1}{n_1} \sum_{j:D_j=0} Y_j \cdot \frac{\hat{p}_j}{1-\hat{p}_j}.$$

Since the ratio $\hat{p}_j/(1-\hat{p}_j)$ might be very large for some observations with very large estimated propensity scores, also a trimmed version of the weighting estimator is considered:

$$\frac{1}{n_1} \sum_{j:D_j=0} Y_j \cdot \min\left\{\frac{\hat{p}_j}{1-\hat{p}_j}, c\right\},$$

where the constant c provides a ceiling to the ratio. Five different values for c (2, 5, 10, 20, 50) are examined in the simulations.

Tables 3.2 to 3.4 provide the simulation results for the Ydesigns 1 to 3. The upper panel gives the results for different sample sizes for $\sigma_\varepsilon^2 = 30$, the lower panel refers to $\sigma_\varepsilon^2 = 100$. For each estimator, the simulated bias, mean squared error (MSE), median squared error (MedSE), mean absolute error (MAE) and median absolute error (MedAE) are given. The first column refers to the Mahalanobis matching estimator, the second and the third column to the multivariate kernel matching estimator with, respectively, Gaussian and Epanechnikov kernel. The following columns represent various propensity score estimators. The first block gives the results with a correctly specified propensity score. The other two blocks refer to the misspecified propensity scores (see Table 3.1).

The propensity score based estimators are pair matching, kernel matching, ridge matching, and weighting without trimming and with a ceiling of 5 and 20, respectively. The results for the

propensity score kernel and ridge matching estimators are given for the Epanechnikov kernel. The simulations were also performed with Gaussian kernel, but are not presented in the tables. The results were similar, with the Epanechnikov kernel in general leading to slightly better results. Also not presented in the tables are the results for the weighting estimator with ceiling level 2, 10 and 50, respectively. A ceiling level of 2 often led to a very high bias. A ceiling level of 10 or 50 usually gave results similar to those presented for a ceiling of 5 or 20. (The additional results are available on request.)

The simulation results for the estimators which are not based on the propensity score show that the multivariate kernel matching estimator with Gaussian kernel clearly dominates Mahalanobis matching and multivariate matching with Epanechnikov kernel. The particularly poor performance of the Epanechnikov kernel may be indicating that a more careful bandwidth choice is necessary when using a bounded kernel in higher dimensional regression.

Comparing the different propensity score estimators with each other, under the correct specification ridge matching most often performs best and clearly dominates pair matching. It is also often more precise and less biased than kernel matching. The weighting estimator is sensitive to the trimming level chosen and without trimming it often performs poorly. With appropriate trimming[14] it can be better than pair matching, but is almost always inferior to ridge matching. This relative ordering of the estimators is largely preserved when the propensity score is misspecified, with the differences being less pronounced, though, since the bias is becoming the dominating component.

Finally, comparing the multivariate matching estimator (with Gaussian kernel) with propensity score ridge matching, the sensitivity to the specification of the propensity score is of main interest. Propensity score ridge matching clearly dominates when the propensity score is correctly specified. Also with a limited misspecification (Specification 2), propensity score ridge matching is often superior to multivariate matching (particularly with respect to the median error measures). When the propensity score is heavily misspecified (Specification 3), however, propensity score matching can be much worse. Nevertheless, in particular situations, propensity score matching can work well even with a misspecified propensity score, as evidenced from Table 3.4 (Ydesign 3).

[14] However, so far no rule has been developed in the literature on how to select the optimal trimming level for a given data set.

Table 3.2: Simulation results with 5 continuous variables for Ydesign 1

	Multivariate			Propensity Score Specification 1						Propensity Score Specification 2						Propensity Score Specification 3					
	Mahal	Gauss	Epa	Pair	Kernel	Ridge	W	W 5	W 20	Pair	Kernel	Ridge	W	W 5	W 20	Pair	Kernel	Ridge	W	W 5	W 20
										σ_ε^2 =30											
n=100																					
Bias	-81.0	-67.8	-112.3	-21.1	-50.5	-29.5	-15.8	-61.8	-28.3	-45.8	-75.9	-57.6	-44.1	-77.0	-52.5	-187.7	-183.0	-185.9	-188.9	-187.8	-188.7
MSE	86.2	66.7	147.4	78.1	63.8	54.4	179.4	62.9	70.9	90.8	92.6	75.3	144.4	84.4	81.5	421.5	359.8	378.0	396.4	378.7	393.2
MedSE	63.2	45.4	125.3	29.6	33.9	24.1	40.3	41.0	37.3	38.4	63.2	40.7	51.1	61.6	48.1	362.2	335.3	345.9	360.3	356.1	360.2
MAE	82.4	70.2	112.6	68.4	65.3	58.8	80.4	67.4	68.6	75.0	82.6	71.5	83.6	80.4	76.0	189.0	183.0	186.3	189.3	187.8	188.9
MedAE	79.5	67.3	112.0	54.4	58.2	49.1	63.5	64.0	61.1	61.9	79.5	63.8	71.5	78.5	69.3	190.3	183.1	186.0	189.8	188.7	189.8
n=400																					
Bias	-56.1	-52.0	-108.0	-7.8	-26.1	-12.6	-2.0	-52.8	-17.5	-35.3	-53.8	-39.9	-32.7	-69.0	-41.6	-183.9	-181.9	-183.6	-185.0	-184.8	-185.0
MSE	36.8	32.0	128.2	21.6	17.6	12.5	68.5	34.1	19.8	31.8	38.6	26.2	61.8	53.7	31.4	353.8	336.8	345.2	351.2	347.8	350.4
MedSE	30.4	28.1	123.9	9.0	9.2	5.3	12.8	29.7	10.9	17.3	30.1	16.8	24.6	49.4	22.6	345.3	330.2	340.4	348.4	344.9	347.2
MAE	56.2	52.1	108.0	36.4	34.6	28.0	49.1	53.2	36.8	46.3	55.2	43.3	53.9	69.1	48.7	183.9	181.9	183.6	185.0	184.8	185.0
MedAE	55.1	53.0	111.3	30.0	30.4	23.1	35.7	54.5	33.0	41.6	54.9	41.0	49.6	70.3	47.6	185.8	181.7	184.5	186.7	185.7	186.3
n=1600																					
Bias	-37.9	-30.1	-109.0	-1.5	-11.8	-4.4	-1.7	-49.8	-13.5	-31.9	-42.3	-34.5	-30.6	-66.3	-38.1	-183.4	-181.2	-182.5	-183.2	-182.9	-183.2
MSE	15.7	10.7	129.8	5.3	4.5	3.5	12.6	26.4	6.5	15.4	20.7	14.8	17.6	45.6	18.6	341.0	330.5	335.5	338.2	336.6	338.1
MedSE	14.9	9.3	135.7	2.6	2.3	1.6	5.2	25.6	3.3	9.3	18.5	11.9	13.2	45.0	15.6	337.4	327.1	332.9	331.8	333.4	332.7
MAE	37.9	30.1	109.0	18.4	17.3	14.9	26.4	49.8	21.1	33.5	42.4	35.0	37.1	66.3	39.2	183.4	181.2	182.5	183.2	182.9	183.2
MedAE	38.6	30.4	116.5	16.0	15.2	12.6	22.9	50.6	18.3	30.5	43.0	34.4	36.4	67.1	39.5	183.7	180.9	182.5	182.2	182.6	182.4
										σ_ε^2 =100											
n=100																					
Bias	-55.1	-46.4	-77.5	-10.1	-33.6	-15.4	-5.1	-18.1	-5.7	-24.8	-47.8	-31.6	-23.5	-32.3	-23.9	-114.8	-117.2	-115.6	-115.1	-116.2	-115.1
MSE	49.9	38.0	78.0	49.6	35.6	30.4	45.0	28.5	41.8	54.5	45.6	38.1	42.2	34.9	40.4	184.6	158.5	159.7	159.8	159.1	159.7
MedSE	32.0	23.3	61.4	21.2	18.9	13.8	16.3	14.1	16.2	23.7	25.3	18.0	18.8	17.5	18.8	129.1	137.6	133.2	134.6	135.7	135.1
MAE	59.5	51.6	78.5	55.4	48.7	43.9	50.2	43.4	49.5	58.5	56.1	49.6	50.8	48.1	50.4	119.0	117.5	116.4	116.2	116.7	116.1
MedAE	56.6	48.3	78.4	46.0	43.5	37.2	40.4	37.5	40.2	48.7	50.3	42.4	43.4	41.9	43.4	113.6	117.3	115.4	116.0	116.5	116.2
n=400																					
Bias	-34.6	-32.8	-71.7	-1.3	-16.8	-3.9	-1.1	-8.7	-1.3	-17.8	-32.3	-19.7	-18.2	-23.5	-18.3	-105.9	-111.4	-106.7	-107.7	-108.3	-107.7
MSE	16.8	14.7	59.5	11.3	8.9	6.5	9.4	6.9	9.0	15.9	16.6	10.3	11.5	11.5	11.2	126.7	129.4	120.3	122.4	123.2	122.3
MedSE	12.2	11.1	53.8	4.5	4.5	3.2	3.8	3.1	3.9	6.9	10.9	5.5	6.3	6.8	6.2	113.3	125.3	114.2	116.8	117.2	116.8
MAE	35.8	33.8	71.8	26.4	24.4	20.4	23.8	21.0	23.6	31.9	35.0	26.2	27.6	28.3	27.5	106.1	111.4	106.7	107.7	108.3	107.7
MedAE	34.9	33.4	73.3	21.2	21.1	18.0	19.5	17.6	19.7	26.2	33.1	23.4	25.0	26.1	25.0	106.4	112.0	106.9	108.1	108.3	108.1
n=1600																					
Bias	-22.4	-18.2	-69.6	-0.2	-9.0	-1.4	-0.1	-6.0	-0.2	-17.5	-24.7	-17.0	-16.1	-20.1	-16.1	-105.0	-108.2	-103.8	-105.1	-105.4	-105.1
MSE	6.2	4.3	54.1	2.5	2.1	1.4	2.2	1.9	2.1	5.8	7.5	4.2	4.4	5.5	4.4	113.7	118.5	109.1	112.0	112.6	112.0
MedSE	5.1	3.3	55.9	1.1	1.1	0.6	1.0	0.8	1.0	3.1	6.2	2.9	2.7	4.1	2.7	109.9	117.2	106.7	110.4	110.8	110.4
MAE	22.5	18.4	69.6	12.6	11.8	9.4	11.8	11.1	11.6	19.8	24.9	17.7	17.9	20.7	17.9	105.0	108.2	103.8	105.1	105.4	105.1
MedAE	22.7	18.1	74.8	10.3	10.7	7.7	10.0	9.1	9.9	17.6	24.8	16.9	16.5	20.2	16.5	104.8	108.2	103.3	105.1	105.3	105.1

Note: Simulation results for sample sizes n=100, 400, 1600 and σ_ε^2 =30, 100, respectively. Mahal is the Mahalanobis matching estimator. Multivariate Gauss and Epa are the multivariate matching estimators using the Gaussian and Epanechnikov kernel, respectively. Pair, kernel and ridge are propensity score pair-matching, propensity score kernel matching and propensity score ridge matching (both with Epanechnikov kernel). W, W5 and W20 are weighting estimators without and with trimming.

Table 3.3: Simulation results with 5 continuous variables for Ydesign 2

	Multivariate			Propensity Score Specification 1						Propensity Score Specification 2						Propensity Score Specification 3					
	Mahal	Gauss	Epa	Pair	Kernel	Ridge	W	W 5	W 20	Pair	Kernel	Ridge	W	W 5	W 20	Pair	Kernel	Ridge	W	W 5	W 20
										σ_ε^2 =30											
n=100																					
Bias	-85.4	-64.5	-110.2	-23.6	-46.7	-27.7	-15.4	-45.7	-23.5	-49.1	-73.7	-53.9	-43.1	-64.9	-47.9	-122.0	-141.7	-127.6	-120.5	-123.1	-120.9
MSE	88.1	56.4	141.8	39.0	44.9	31.5	88.9	33.3	40.5	62.2	76.8	53.5	75.0	55.0	54.8	186.7	219.8	184.9	166.7	166.2	166.4
MedSE	73.2	41.7	122.8	17.2	24.6	13.6	24.1	23.8	23.1	29.4	56.6	29.8	42.4	43.2	39.0	150.5	204.3	163.3	152.2	152.8	151.7
MAE	85.9	66.2	110.6	49.5	55.2	44.5	60.6	49.5	52.6	63.5	76.8	60.9	68.5	66.1	63.7	123.3	141.8	127.9	121.6	123.1	121.5
MedAE	85.5	64.6	110.8	41.5	49.6	36.9	49.1	48.8	48.1	54.3	75.3	54.6	65.1	65.7	62.4	122.7	142.9	127.8	123.4	123.6	123.2
n=400																					
Bias	-58.5	-41.8	-103.9	-10.1	-22.1	-13.7	-2.9	-41.3	-15.2	-37.5	-49.9	-40.8	-32.7	-60.6	-39.9	-118.8	-128.4	-120.0	-118.2	-121.9	-118.8
MSE	38.4	21.0	122.9	9.8	10.7	7.9	39.1	20.5	13.3	23.1	30.7	22.6	33.6	40.4	25.6	149.8	169.7	149.4	145.9	152.7	146.6
MedSE	33.8	16.7	115.9	3.8	5.1	3.4	9.1	16.5	7.3	12.4	23.6	15.4	21.3	37.0	20.6	137.8	164.9	142.5	140.3	147.4	140.3
MAE	58.6	41.9	103.9	24.3	26.1	21.9	40.2	41.5	29.9	39.9	50.3	41.7	47.9	60.6	44.2	118.8	128.4	120.0	118.3	121.9	118.8
MedAE	58.1	40.9	107.7	19.5	22.7	18.3	30.1	40.7	27.0	35.3	48.6	39.2	46.2	60.8	45.3	117.4	128.4	119.4	118.4	121.4	118.4
n=1600																					
Bias	-39.7	-35.4	-103.0	-2.3	-8.3	-3.4	-0.5	-39.4	-11.8	-32.5	-38.6	-33.2	-31.0	-58.8	-37.1	-118.4	-122.9	-119.5	-118.6	-122.1	-118.8
MSE	16.7	13.4	123.7	2.5	2.4	1.9	8.4	16.4	4.0	12.9	16.5	12.7	14.3	35.5	16.1	142.6	152.4	144.4	142.3	150.1	142.7
MedSE	15.6	12.4	123.7	1.1	1.1	0.9	3.5	16.1	2.1	10.5	14.7	10.4	12.0	35.7	14.0	138.5	150.7	142.2	142.5	149.7	142.5
MAE	39.7	35.4	103.0	12.7	12.3	11.0	21.5	39.4	16.2	32.8	38.7	33.4	34.1	58.8	37.3	118.4	122.9	119.5	118.6	122.1	118.8
MedAE	39.5	35.3	111.2	10.6	10.5	9.3	18.7	40.1	14.3	32.4	38.3	32.2	34.7	59.8	37.4	117.7	122.7	119.3	119.4	122.3	119.4
										σ_ε^2 =100											
n=100																					
Bias	-44.5	-32.9	-66.2	-8.9	-28.2	-10.0	-2.0	-11.7	-2.6	-25.3	-44.5	-26.1	-20.4	-27.0	-20.6	-68.9	-85.9	-73.6	-67.9	-69.3	-68.0
MSE	34.2	22.9	58.3	28.8	24.8	19.0	31.2	17.8	28.4	35.5	36.9	26.1	28.6	23.7	28.0	80.1	90.6	73.1	65.0	64.0	64.8
MedSE	20.8	12.5	45.9	11.8	12.3	8.2	10.4	7.9	10.4	16.2	22.3	10.9	14.0	11.7	14.0	48.8	77.6	54.4	50.4	50.0	50.3
MAE	48.8	39.6	67.5	42.1	40.5	34.5	40.6	33.8	40.1	47.5	50.7	40.5	42.8	39.7	42.7	74.5	86.5	75.2	71.4	71.0	71.3
MedAE	45.6	35.4	67.7	34.3	35.1	28.7	32.3	28.2	32.2	40.2	47.3	33.0	37.4	34.2	37.4	69.9	88.1	73.7	71.0	70.7	71.0
n=400																					
Bias	-29.5	-21.1	-64.6	-3.2	-14.3	-3.9	-0.8	-7.6	-1.1	-19.3	-30.5	-20.0	-17.3	-22.6	-17.5	-66.7	-77.1	-68.8	-67.6	-68.7	-67.6
MSE	12.4	7.3	50.6	5.9	6.0	3.9	7.2	4.6	6.5	10.1	13.3	7.9	9.4	9.1	9.1	51.8	63.4	51.3	49.9	50.9	49.9
MedSE	8.6	4.6	41.5	2.8	2.9	1.8	2.7	2.2	2.7	5.3	9.4	4.5	5.0	5.7	5.0	44.3	59.0	47.1	45.9	46.8	45.9
MAE	30.5	22.9	64.6	19.4	19.8	15.8	20.2	17.2	19.9	26.0	31.8	23.4	25.1	25.3	24.9	66.9	77.1	68.8	67.7	68.7	67.6
MedAE	29.4	21.4	64.4	16.7	17.2	13.2	16.5	14.7	16.5	23.0	30.7	21.3	22.4	23.8	22.3	66.6	76.8	68.6	67.8	68.4	67.8
n=1600																					
Bias	-19.1	-18.5	-64.6	-1.5	-7.1	-1.0	0.0	-6.0	-0.1	-17.5	-23.1	-16.9	-16.1	-20.5	-16.3	-67.6	-73.4	-68.5	-67.8	-68.4	-67.8
MSE	4.5	4.1	50.0	1.3	1.4	0.9	1.6	1.4	1.5	4.4	6.2	3.7	3.9	5.2	3.9	47.4	54.8	47.8	46.8	47.7	46.8
MedSE	3.7	3.3	48.8	0.6	0.7	0.4	0.7	0.7	0.7	3.0	5.2	2.8	2.9	4.2	2.8	45.8	53.8	46.4	45.7	46.5	45.7
MAE	19.3	18.6	64.6	9.1	9.4	7.5	10.0	9.4	9.8	18.1	23.1	17.1	17.1	20.6	17.1	67.6	73.4	68.5	67.8	68.4	67.8
MedAE	19.2	18.3	69.8	7.7	8.1	6.3	8.3	8.1	8.3	17.2	22.9	16.8	16.9	20.4	16.7	67.7	73.3	68.1	67.6	68.2	67.6

See Note below Table 3.2.

Table 3.4: Simulation results with 5 continuous variables for Ydesign 3

	Multivariate			Propensity Score Specification 1						Propensity Score Specification 2						Propensity Score Specification 3					
	Mahal	Gauss	Epa	Pair	Kernel	Ridge	W	W 5	W 20	Pair	Kernel	Ridge	W	W 5	W 20	Pair	Kernel	Ridge	W	W 5	W 20
										σ_ε^2 =30											
n=100																					
Bias	-49.1	-41.7	-45.4	-16.9	-26.8	-22.4	-13.3	-35.1	-18.9	-23.8	-30.6	-28.5	-23.1	-37.4	-26.1	-37.9	-38.7	-37.9	-38.5	-43.1	-38.7
MSE	33.4	24.7	25.8	40.5	19.8	24.0	74.8	23.5	33.7	39.5	19.4	22.3	48.4	24.3	30.4	32.9	21.7	23.8	30.7	26.2	28.3
MedSE	23.8	19.3	21.6	13.7	11.1	10.8	15.1	15.7	14.7	16.5	11.5	12.0	16.0	16.6	15.5	19.2	16.5	17.3	19.9	20.1	19.9
MAE	50.9	43.7	46.2	47.4	36.4	38.6	49.8	41.3	44.8	48.9	36.8	38.9	47.1	42.4	44.1	48.1	41.0	42.2	46.4	45.3	45.7
MedAE	48.8	44.0	46.4	37.1	33.4	32.9	38.9	39.7	38.3	40.6	33.9	34.6	40.0	40.8	39.3	43.9	40.6	41.6	44.7	44.8	44.7
n=400																					
Bias	-39.7	-33.5	-42.8	-7.7	-18.4	-9.0	1.0	-30.2	-11.7	-17.9	-25.6	-18.7	-14.8	-33.2	-19.8	-34.1	-36.8	-33.7	-33.7	-38.7	-34.2
MSE	18.4	13.7	19.9	13.2	8.5	8.2	122.6	12.3	10.2	13.1	10.2	9.3	22.9	13.9	10.6	16.8	15.4	14.2	15.0	17.0	14.8
MedSE	15.9	11.8	19.1	5.0	5.3	3.7	6.1	10.0	5.2	7.2	7.9	5.6	7.1	11.8	6.7	13.0	14.0	12.3	13.1	15.5	13.1
MAE	39.8	33.8	42.9	27.7	24.4	22.7	37.0	31.5	25.8	29.6	28.3	25.4	30.8	34.0	27.5	36.2	36.9	34.4	35.4	38.8	35.3
MedAE	39.9	34.3	43.8	22.3	23.1	19.4	24.7	31.6	22.9	26.9	28.1	23.6	26.6	34.3	25.9	36.1	37.5	35.1	36.2	39.4	36.1
n=1600																					
Bias	-30.1	-25.3	-41.6	-4.7	-9.7	-3.5	-1.8	-28.9	-10.6	-14.1	-19.0	-13.1	-12.1	-31.8	-17.8	-31.5	-34.0	-30.4	-31.7	-36.7	-32.4
MSE	9.7	7.0	18.1	3.7	2.6	2.5	13.0	9.1	3.1	5.0	5.1	3.8	8.4	10.8	4.8	11.5	12.1	10.2	11.1	14.0	11.2
MedSE	9.4	6.5	18.8	1.7	1.5	1.1	1.9	8.5	1.7	3.1	3.7	2.2	3.3	10.3	3.7	10.2	12.1	9.4	10.4	13.8	10.7
MAE	30.1	25.3	41.6	15.1	13.4	12.5	18.8	28.9	14.5	18.4	20.1	16.0	20.4	31.8	19.4	31.9	34.0	30.5	32.0	36.7	32.4
MedAE	30.7	25.4	43.3	12.9	12.2	10.4	13.7	29.2	13.1	17.5	19.1	14.7	18.3	32.0	19.3	31.9	34.8	30.7	32.3	37.2	32.7
										σ_ε^2 =100											
n=100																					
Bias	-36.1	-31.3	-34.5	-7.1	-17.2	-9.9	-2.4	-11.7	-3.7	-11.1	-19.8	-13.9	-10.4	-15.7	-10.7	-23.4	-26.0	-22.3	-22.5	-24.4	-22.5
MSE	20.6	14.8	16.0	20.4	10.3	11.6	30.0	12.1	22.3	22.8	10.1	10.7	15.5	11.2	14.8	20.8	11.9	11.6	13.0	12.4	13.0
MedSE	13.8	10.2	12.7	7.8	5.7	5.3	6.8	5.9	6.7	8.9	5.8	5.4	5.9	5.7	5.9	10.2	7.6	6.7	7.7	7.8	7.7
MAE	38.9	33.4	35.6	34.6	26.4	26.9	33.9	28.1	32.7	36.0	26.5	26.5	29.4	27.3	29.2	36.4	29.3	28.4	30.0	29.7	30.0
MedAE	37.1	31.9	35.7	28.0	23.9	23.0	26.1	24.3	25.9	29.9	24.1	23.3	24.3	23.9	24.2	32.0	27.6	25.9	27.7	27.8	27.7
n=400																					
Bias	-27.3	-23.7	-32.0	-2.3	-10.1	1.4	-0.1	-5.6	-0.5	-6.7	-14.2	-4.0	-6.1	-9.7	-6.3	-17.9	-22.3	-15.3	-17.6	-18.6	-17.7
MSE	9.5	7.2	11.4	6.1	3.5	3.5	6.8	3.4	4.9	6.1	4.1	3.3	4.1	3.4	3.9	7.4	6.5	4.6	5.2	5.3	5.2
MedSE	7.6	5.8	10.4	2.6	1.9	1.4	1.7	1.5	1.7	2.7	2.6	1.4	1.8	1.9	1.8	4.4	5.3	3.0	3.7	3.8	3.7
MAE	27.7	24.1	32.0	19.3	15.4	14.6	17.0	14.6	16.6	19.4	17.1	14.2	15.7	15.1	15.6	22.8	22.9	18.2	19.8	20.0	19.8
MedAE	27.5	24.2	32.3	16.0	13.9	12.0	13.2	12.4	13.2	16.3	16.1	11.9	13.5	13.6	13.5	20.9	23.0	17.3	19.1	19.5	19.1
n=1600																					
Bias	-18.9	-17.9	-30.7	-0.1	-5.3	3.8	0.5	-3.7	0.4	-6.4	-9.8	-1.4	-4.9	-7.7	-5.0	-15.8	-19.2	-12.3	-16.1	-16.6	-16.1
MSE	4.1	3.5	9.9	1.5	0.9	0.9	1.1	0.9	1.1	1.7	1.5	0.7	1.0	1.2	1.0	3.6	4.1	2.1	3.1	3.2	3.1
MedSE	3.6	3.3	9.8	0.6	0.5	0.4	0.4	0.4	0.4	0.8	1.0	0.4	0.5	0.7	0.5	2.6	3.6	1.4	2.6	2.7	2.6
MAE	18.9	17.9	30.7	9.4	7.8	7.7	8.3	7.6	8.2	10.6	10.4	6.9	8.3	9.0	8.2	16.4	19.2	12.7	16.1	16.6	16.1
MedAE	19.0	18.1	31.3	7.8	6.9	6.4	6.7	6.5	6.7	9.2	9.9	6.2	7.4	8.2	7.4	16.0	19.1	12.0	16.1	16.5	16.1

See Note below Table 3.2.

In the second Monte Carlo experiment, X consists of 10 binary variables. With X being discrete, it is now also possible to match exactly on all covariates, since the number of cells spanned by X is finite. *Exact matching* proceeds by estimating $m_0(x)$ by the average of the Y values in the cell defined by X. If the cell is empty in the $D = 0$ sample, the estimate is undefined. Then, in the $D = 1$, sample the average of $\hat{m}_0$ is taken for all values of X_i where the estimate is defined.

The multivariate kernel matching estimator is also adopted to this design. A kernel that smooths over the dummy variables is used, as suggested in Racine and Li (2000). The kernel weights are defined as

$$w(X_j, x) = \prod_{k=1}^{10} \lambda^{1(X_{j,k} \neq x_k)},$$

where $\lambda \in [0, 1]$ is the bandwidth value.[15] For $\lambda = 0$, no smoothing over the discrete regressors takes place and multivariate kernel matching is equivalent to exact matching. For $\lambda = 1$, the weights are all equal to one and the estimate corresponds to the sample mean.

The X covariates consist of 10 binary variables, taking the values -0.5 and 0.5 with equal probability. $X_1..X_{10}$ are generated as the sum of a common component and an individual component (both uniform random variables), taking the value 0.5 if this sum is greater than one, and -0.5 otherwise. The correlation between $X_1..X_{10}$ is 0.34. The observations are assigned to the source ($D = 0$) or the target ($D = 1$) sample by

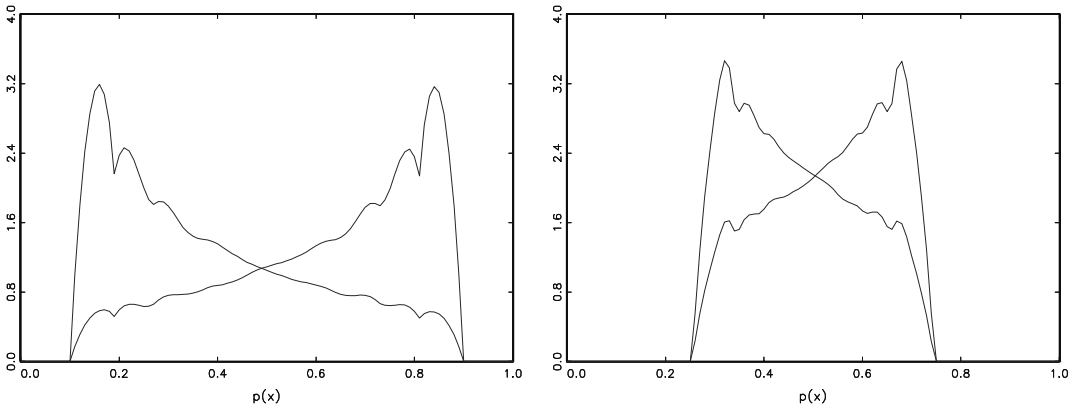
$$D = 1\left(X'\beta + \varepsilon \ > 0 \right),$$

where $\beta = (\sqrt{0.5}, \sqrt{1.5}, \sqrt{2.5}, .., \sqrt{9.5})'$ and $\varepsilon \sim N(0, \sigma_\varepsilon^2)$.[16] Two different values of σ_ε^2 are examined (100 and 400), and Table B.3 shows the means of X in source and target population for these two scenarios. Figure 3.2 displays the corresponding distributions of the propensity score. In the scenario with the larger σ_ε^2, the source and target population are more similar.

[15] The value λ is chosen by leave-one-out cross-validation from the grid: 0, 0.05, 0.10, ..., 1.

[16] This β vector was chosen to obtain a large cardinality of the support of the propensity score. Since with X discrete, the propensity score $\Phi\left(\frac{X'\beta}{\sigma_\varepsilon}\right)$ is also discrete. Were β chosen, for example, as a vector with identical elements, $X'\beta$ could take only 11 different values and the support of the propensity score would thus contain only 11 points; this would not be a very realistic scenario.

Figure 3.2: Distribution of true propensity score in source and target population



Note: Left figure: $\sigma_\varepsilon^2 = 100$. Right figure: $\sigma_\varepsilon^2 = 400$.

The outcome variable Y is generated as

$$
\begin{aligned}
Ydesign\ 1: \quad & Y = \sum_{k=1}^{10} X_k + u \\
Ydesign\ 2: \quad & Y = \sum_{k=1}^{5} X_k - \sum_{k=6}^{10} X_k + \sum_{k=1}^{5} X_k X_{k+5} + u \\
Ydesign\ 3: \quad & Y = \ln\left(\sum_{k=1}^{10} (X_k + 0.5) + 0.1 \right) + u,
\end{aligned}
$$

where $u \sim N(0,1)$.

Three different specifications are also considered for the estimation of the propensity score, shown in Table 3.5. Whereas the first specification is correct, the latter two are misspecified. In specification 2, the interaction term X_9X_{10} instead of the variable X_{10} is included in the Probit model. In specification 3, the interaction term X_1X_6 instead of the variable X_6 is used as regressor, and also X_2X_7 instead of X_7 and so forth. The informational content is the same in all 3 specifications, since X_{10} could be recovered from the regressors X_9 and X_9X_{10} in specification 2 (and analogously in specification 3).[17] The last column in Table 3.5 shows the correlation of the true with the estimated propensity score. Whereas the bias in the estimated propensity scores is rather limited in specification 2, the estimated propensity scores are more seriously biased in specification 3.

Tables 3.6 to 3.8 give the simulation results for the three different Ydesigns. These tables are similar in structure to the Tables 3.2 to 3.4, with the only differences being that there is only one multivariate kernel matching estimator (without variants in the kernel function) and

[17] Specification 2 corresponds to a model where two variables A and B are included in a Probit model, but where the true propensity score is of the type $\Phi(\beta_1 A + \beta_2 \frac{B}{A})$, i.e. depending on the ratio B/A.

Table 3.5: Specification of the propensity score model

	Regressors	$Corr(\hat{p}, p)$
Specification 1:	$const, X_1, X_2, .., X_{10}$	1.00
Specification 2:	$const, X_1, X_2, .., X_9,\ X_9X_{10}$	0.98
Specification 3:	$const, X_1, X_2, .., X_5,\ X_1X_6,\ X_2X_7,\ X_3X_8,\ X_4X_9,\ X_5X_{10}$	0.87

Note: See note below Table 3.1.

that exact multivariate matching enters as an additional estimator. (As before, the results for propensity score kernel and ridge matching with Gaussian kernel and for the weighting estimator with ceiling level 2, 10 and 50 are not presented and available upon request.)

Among the estimators not based on the propensity score, Mahalanobis matching is usually, but not always, worse than multivariate matching. Exact matching performs much worse than the other two estimators, particularly in terms of bias. Since exact matching discards all observations for which no exact match can be found, in small samples it might occur that the estimate is undefined as no matches might be found at all. If the bias and the average absolute/squared error are presented in brackets in Tables 3.6 to 3.8, only those Monte Carlo replications where the estimate was defined were used for calculating them.

Among the propensity score based estimators, the patterns are less clear cut as in the Monte Carlo experiment with the continuous regressors. Although pair matching is still clearly dominated by kernel and ridge matching, ridge matching is now slightly inferior to kernel matching (although the differences are not large). For the weighting estimator, trimming is not so crucial anymore since only few estimated propensity scores are close to one, see Figure 3.2. Nevertheless, kernel matching still dominates weighting in most situations. These findings roughly hold through with a misspecified propensity score. For the grossly misspecified specification 3, however, the results do not show a consistent pattern (with pair matching often performing very well).

Comparing multivariate matching with propensity score matching, we find that for the correctly and the slightly misspecified propensity score model (Specifications 1 and 2) propensity score kernel matching always (and ridge matching almost always) dominates multivariate matching. For specification 3, on the other hand, multivariate matching is usually superior to propensity score matching. However, that this is not generally the case is demonstrated by Table 3.8. For this design, propensity score matching with misspecified propensity score often

performs better than multivariate matching.

The results of this (limited) Monte Carlo study do not allow for very general conclusions. Nevertheless, some patterns regarding the ordering among the various estimators can be extracted. Among the propensity score based estimators, pair matching is usually inferior to kernel matching and ridge matching. Ridge matching performed slightly better than kernel matching in the simulations with the five continuous variables, while this relationship was reversed in the simulations with the ten binary variables. The weighting estimator was susceptible to the trimming level and did not dominate any of the other estimators. Among the estimators not based on the propensity score, Mahalanobis matching and particularly exact matching were clearly inferior to multivariate kernel matching.

The relationship between multivariate kernel matching and propensity score matching, unsurprisingly, depends on the degree of misspecification of the propensity score. With a correctly or a slightly misspecified propensity score, propensity score matching was superior. Even with a grossly misspecified propensity score, it was not always worse than multivariate matching. From this perspective, propensity score matching provides a convenient and practical approach to estimating covariate-distribution adjusted means. Nevertheless, the simulation results also showed that in certain situations, propensity score matching can be much worse than multivariate matching, particularly when recognizing that only a rather simple nonparametric multivariate regression estimator (with a single bandwidth) has been used in the simulation study. Employing more sophisticated multivariate nonparametric estimators might be worthwhile, but is beyond the scope of this paper. Compared to propensity score matching, implementing such a more sophisticated nonparametric estimator with data-driven bandwidth choice for multiple bandwidths will be more demanding, though.

Table 3.6: Simulation results with 10 binary variables for Ydesign 1

				Propensity Score Specification 1						Propensity Score Specification 2						Propensity Score Specification 3					
	Mahal	Multi	Exact	Pair	Kernel	Ridge	W	W 5	W 20	Pair	Kernel	Ridge	W	W 5	W 20	Pair	Kernel	Ridge	W	W 5	W 20
										σ_ε^2 =100											
n=100																					
Bias	-24.3	-56.5	(16.0)	6.9	-24.2	-6.0	0.1	-42.3	-3.3	-0.4	-31.9	-14.2	-8.9	-48.2	-11.8	-38.8	-78.5	-61.8	-63.0	-85.5	-64.2
MSE	60.7	59.3	(520.0)	51.8	42.1	34.4	69.6	49.3	57.2	50.9	48.7	38.5	66.1	54.2	55.5	72.5	103.5	79.6	89.6	102.9	86.4
MedSE	23.1	34.5	230.9	21.1	17.4	13.8	24.3	27.3	23.5	19.9	19.7	15.3	24.3	28.4	23.6	28.4	63.4	42.3	53.7	72.4	53.5
MAE	59.9	64.5	(181.4)	55.8	51.3	45.6	62.1	57.5	59.0	55.2	54.4	48.1	61.3	60.4	58.6	66.7	85.3	72.9	78.9	88.5	78.1
MedAE	48.0	58.7	152.0	46.0	41.7	37.2	49.3	52.3	48.4	44.6	44.3	39.2	49.3	53.3	48.6	53.3	79.6	65.0	73.3	85.1	73.2
n=400																					
Bias	-7.1	-24.1	66.6	4.2	-4.0	0.6	2.1	-18.2	2.1	-2.5	-12.5	-8.6	-7.5	-24.5	-7.5	-34.2	-59.6	-56.9	-63.0	-67.5	-63.0
MSE	10.1	12.8	69.4	9.4	5.9	5.8	10.0	11.2	10.0	9.6	7.7	6.8	10.1	13.7	10.1	22.4	41.7	38.8	47.1	52.4	47.1
MedSE	4.7	6.6	44.1	4.1	2.4	2.7	4.3	5.6	4.3	4.1	3.6	2.9	4.3	7.1	4.3	13.9	36.0	32.4	40.9	46.3	40.9
MAE	25.3	29.3	71.3	24.5	19.1	19.2	24.9	27.0	24.9	24.4	22.3	20.9	24.9	30.3	24.9	39.6	59.7	57.2	63.3	67.6	63.3
MedAE	21.6	25.6	66.4	20.1	15.5	16.5	20.9	23.7	20.9	20.3	19.0	17.0	20.7	26.7	20.7	37.3	60.0	56.9	64.0	68.0	64.0
n=1600																					
Bias	-5.0	-9.6	37.6	0.9	-0.5	0.1	-0.4	-14.5	-0.4	-2.7	-8.8	-8.0	-9.8	-20.5	-9.8	-21.4	-51.5	-51.5	-64.2	-64.9	-64.2
MSE	5.0	2.6	17.1	4.9	1.5	1.7	2.4	4.0	2.4	4.9	2.3	2.1	3.3	6.1	3.3	9.6	28.4	28.5	43.1	43.8	43.1
MedSE	2.3	1.3	13.0	2.4	0.6	0.5	1.2	2.2	1.2	1.9	1.4	1.2	1.6	4.3	1.6	4.2	27.0	27.4	39.9	40.3	39.9
MAE	17.7	13.1	37.7	17.6	9.7	9.7	12.6	16.6	12.6	17.3	12.5	12.1	14.6	21.3	14.6	24.5	51.5	51.5	64.2	64.9	64.2
MedAE	15.1	11.3	36.1	15.5	7.8	7.3	10.8	15.0	10.8	13.9	11.7	11.2	12.8	20.8	12.8	20.5	52.0	52.3	63.1	63.5	63.1
										σ_ε^2 =400											
n=100																					
Bias	-2.2	-34.4	43.4	10.2	-12.1	-0.4	2.0	-3.6	1.8	8.1	-15.4	-3.4	-2.1	-7.7	-2.3	-8.7	-37.4	-27.1	-28.8	-32.6	-29.0
MSE	39.3	36.3	274.3	56.6	34.9	35.3	34.7	28.8	34.0	56.6	35.7	35.1	34.1	28.4	33.2	55.3	46.1	41.2	39.0	37.0	38.6
MedSE	16.0	17.2	117.1	26.4	15.8	15.2	14.0	12.9	14.0	25.6	16.4	14.5	13.9	13.0	13.9	23.6	23.0	20.0	19.1	18.7	19.1
MAE	49.2	48.7	130.0	60.0	47.5	47.3	46.1	42.7	45.9	59.8	47.8	46.7	45.3	42.4	45.2	58.7	55.0	51.3	50.1	49.1	49.9
MedAE	40.0	41.5	108.2	51.4	39.8	39.0	37.5	35.9	37.4	50.6	40.5	38.1	37.3	36.0	37.3	48.6	47.9	44.7	43.7	43.2	43.7
n=400																					
Bias	-2.3	-27.1	34.1	4.6	-1.7	-0.1	0.5	0.5	0.5	1.7	-5.5	-4.0	-4.0	-4.0	-4.0	-11.0	-26.9	-26.0	-31.4	-31.4	-31.4
MSE	9.3	13.3	35.2	10.6	6.1	6.1	6.1	6.1	6.1	11.0	6.5	6.4	6.2	6.2	6.2	12.4	13.6	13.1	15.7	15.7	15.7
MedSE	4.3	7.9	17.2	4.8	2.6	2.7	2.7	2.7	2.7	4.6	2.9	2.9	2.9	2.9	2.9	5.9	8.0	7.4	10.6	10.6	10.6
MAE	24.5	30.7	47.8	26.0	19.6	19.7	19.7	19.7	19.7	26.1	20.3	20.2	19.9	19.9	19.9	28.2	30.8	30.1	33.9	34.0	33.9
MedAE	20.8	28.1	41.4	22.0	16.2	16.3	16.6	16.5	16.6	21.4	16.9	17.0	17.0	16.9	17.0	24.4	28.2	27.2	32.6	32.6	32.6
n=1600																					
Bias	-1.6	-7.6	20.3	1.9	0.6	0.5	0.3	0.3	0.3	-0.2	-3.5	-3.9	-4.6	-4.6	-4.6	-8.2	-24.7	-25.4	-31.5	-31.5	-31.5
MSE	3.8	2.9	7.0	3.9	1.5	1.5	1.6	1.6	1.6	4.6	2.2	2.1	2.4	2.4	2.4	5.4	8.1	8.4	11.9	11.9	11.9
MedSE	1.9	1.2	4.2	1.9	0.7	0.7	0.8	0.8	0.8	1.9	0.8	0.8	0.8	0.8	0.8	2.3	5.9	6.2	9.8	9.8	9.8
MAE	15.8	13.7	22.4	15.9	9.7	9.9	10.1	10.1	10.1	16.2	10.7	10.7	11.0	11.0	11.0	17.7	25.1	25.6	31.5	31.5	31.5
MedAE	13.7	11.0	20.5	13.8	8.3	8.3	8.9	8.9	8.9	13.9	8.7	8.9	8.7	8.7	8.7	15.1	24.3	25.0	31.3	31.3	31.3

Note: Simulation results for sample sizes n=100, 400, 1600 and σ_ε^2 =100, 400, respectively. Mahal is the Mahalanobis matching estimator. Multi is the multivariate matching estimator. Exact is the matching estimator using only exact matches. Pair, kernel and ridge are the propensity score pair-matching estimator, the propensity score kernel matching estimator and the propensity score ridge matching estimator (both with Epanechnikov kernel). W, W5 and W20 are weighting estimators without and with trimming.

Table 3.7: Simulation results with 10 binary variables for Ydesign 2

				Propensity Score Specification 1						Propensity Score Specification 2						Propensity Score Specification 3					
	Mahal	Multi	Exact	Pair	Kernel	Ridge	W	W 5	W 20	Pair	Kernel	Ridge	W	W 5	W 20	Pair	Kernel	Ridge	W	W 5	W 20
										σ_ε^2 =100											
n=100																					
Bias	25.5	22.6	(64.6)	6.9	1.2	2.5	1.2	-5.5	1.2	12.6	6.0	8.1	8.4	0.8	8.2	51.3	40.6	48.4	55.7	43.4	54.6
MSE	27.8	16.6	(89.1)	34.2	13.3	17.5	21.9	11.4	19.3	33.8	13.0	17.6	21.7	11.4	19.8	51.5	29.2	38.5	52.9	30.2	47.8
MedSE	13.0	7.9	49.4	12.9	5.4	6.5	7.9	4.9	7.8	13.7	5.3	6.6	8.3	5.2	8.1	29.1	15.5	23.1	28.0	19.1	27.8
MAE	42.3	32.9	(78.2)	45.3	28.3	32.2	35.2	26.9	34.2	45.6	27.9	32.2	35.6	26.9	34.8	59.6	44.3	52.2	58.9	46.5	57.7
MedAE	36.1	28.1	70.3	36.0	23.2	25.5	28.1	22.2	27.9	37.1	22.9	25.7	28.8	22.8	28.5	53.9	39.4	48.1	52.9	43.7	52.8
n=400																					
Bias	10.8	18.9	49.7	2.2	-3.9	2.1	-1.3	-4.7	-1.3	8.6	3.8	10.1	7.2	3.7	7.2	39.0	49.3	52.4	56.1	53.5	56.1
MSE	8.3	6.6	32.9	8.7	3.3	3.7	3.5	3.2	3.5	9.0	3.1	4.6	3.8	3.1	3.8	21.4	27.0	30.2	34.1	31.0	34.1
MedSE	3.6	4.0	25.9	4.0	1.6	1.8	1.7	1.6	1.7	4.0	1.5	2.3	1.7	1.4	1.7	15.4	24.1	28.3	31.6	28.6	31.6
MAE	22.7	21.5	51.0	23.5	14.5	15.3	14.8	14.4	14.8	23.8	14.2	17.4	15.5	14.0	15.5	40.4	49.3	52.4	56.1	53.5	56.1
MedAE	18.9	20.0	50.8	20.1	12.7	13.3	13.2	12.6	13.2	19.9	12.1	15.3	13.2	11.9	13.2	39.2	49.1	53.2	56.2	53.5	56.2
n=1600																					
Bias	4.3	13.3	29.2	2.9	0.1	2.2	0.7	-2.5	0.7	7.1	8.1	9.5	8.9	6.2	8.9	23.7	48.0	46.4	56.9	56.5	56.9
MSE	3.6	2.4	9.7	4.0	0.8	0.9	0.8	0.8	0.8	4.2	1.4	1.7	1.6	1.1	1.6	9.2	23.8	22.3	33.0	32.6	33.0
MedSE	1.6	1.8	8.3	2.0	0.3	0.4	0.4	0.4	0.4	1.8	0.8	1.0	1.0	0.6	1.0	5.9	22.9	21.5	32.6	32.2	32.6
MAE	15.2	13.7	29.2	15.9	7.1	7.6	7.3	7.2	7.3	16.3	9.8	10.9	10.5	8.7	10.5	25.8	48.0	46.4	56.9	56.5	56.9
MedAE	12.7	13.4	28.8	14.1	5.9	6.5	6.0	6.2	6.0	13.5	8.8	10.0	9.8	7.9	9.8	24.4	47.9	46.3	57.1	56.8	57.1
										σ_ε^2 =400											
n=100																					
Bias	25.0	24.4	65.0	8.2	5.0	5.1	1.2	0.8	1.2	13.7	7.4	8.1	5.1	4.6	5.1	29.4	23.1	26.2	29.9	27.7	29.7
MSE	21.1	13.0	64.8	21.5	7.9	10.1	11.8	9.3	11.6	23.3	8.4	10.8	12.6	9.9	12.3	28.3	13.9	16.4	20.2	16.7	19.6
MedSE	9.4	6.7	42.9	9.7	3.7	4.2	4.6	4.0	4.6	10.7	3.7	4.6	5.1	4.6	5.1	13.1	6.3	8.0	9.9	9.0	9.9
MAE	36.6	29.5	68.9	36.9	22.3	24.8	26.5	24.3	26.4	38.6	23.1	25.8	27.4	25.2	27.3	42.5	29.7	32.6	35.7	33.3	35.5
MedAE	30.7	26.0	65.5	31.1	19.2	20.4	21.3	20.1	21.3	32.8	19.2	21.3	22.5	21.5	22.5	36.1	25.0	28.3	31.5	30.0	31.5
n=400																					
Bias	9.4	18.6	44.1	3.8	0.4	2.2	-0.3	-0.3	-0.3	6.9	4.1	6.2	4.1	4.1	4.1	23.2	26.6	27.0	29.0	29.0	29.0
MSE	5.8	5.1	23.1	6.2	1.9	2.3	2.1	2.1	2.1	6.6	2.0	2.6	2.2	2.2	2.2	10.8	8.9	9.2	10.2	10.2	10.2
MedSE	2.7	3.5	19.9	2.7	0.8	1.0	0.9	0.9	0.9	3.0	0.9	1.1	0.9	0.9	0.9	6.2	7.1	7.4	8.5	8.5	8.5
MAE	19.2	19.5	44.3	19.9	10.9	12.0	11.3	11.3	11.3	20.5	11.3	12.7	11.6	11.6	11.6	27.1	26.9	27.3	29.2	29.2	29.2
MedAE	16.4	18.6	44.6	16.5	8.8	9.9	9.3	9.3	9.3	17.4	9.4	10.7	9.7	9.6	9.7	24.8	26.7	27.2	29.2	29.1	29.2
n=1600																					
Bias	3.6	11.9	24.0	2.6	0.9	0.6	0.1	0.1	0.1	5.3	4.6	4.7	4.5	4.5	4.5	14.4	26.5	25.1	29.7	29.7	29.7
MSE	2.8	1.8	6.5	2.9	0.4	0.5	0.5	0.5	0.5	3.1	0.7	0.7	0.6	0.6	0.6	4.9	7.5	6.8	9.2	9.2	9.2
MedSE	1.4	1.4	6.1	1.4	0.2	0.3	0.2	0.2	0.2	1.6	0.4	0.4	0.3	0.3	0.3	2.6	7.0	6.3	8.8	8.8	8.8
MAE	13.4	12.1	24.0	13.8	5.3	5.9	5.5	5.5	5.5	14.3	6.6	6.9	6.3	6.3	6.3	18.2	26.5	25.1	29.7	29.7	29.7
MedAE	11.8	11.9	24.6	12.0	4.1	5.0	4.9	4.9	4.9	12.6	5.9	6.4	5.7	5.7	5.7	16.2	26.5	25.2	29.6	29.6	29.6

See note below Table 3.6.

Table 3.8: Simulation results with 10 binary variables for Ydesign 3

				Propensity Score Specification 1						Propensity Score Specification 2						Propensity Score Specification 3					
	Mahal	Multi	Exact	Pair	Kernel	Ridge	W	W 5	W 20	Pair	Kernel	Ridge	W	W 5	W 20	Pair	Kernel	Ridge	W	W 5	W 20
										σ_ε^2 =100											
n=100																					
Bias	-18.1	-29.4	(-46.2)	-0.5	-5.2	-2.3	-0.2	-25.4	-2.2	-1.6	-7.2	-4.5	-2.9	-26.0	-4.6	-11.3	-19.4	-15.7	-16.1	-31.9	-17.1
MSE	20.8	17.9	(123.0)	16.6	9.7	10.7	18.3	15.8	14.6	16.9	10.1	10.7	17.4	16.0	14.4	17.6	13.1	12.9	16.9	18.7	15.6
MedSE	9.1	10.2	37.4	6.9	4.2	4.5	6.4	8.3	6.2	6.7	4.3	4.6	6.0	8.2	5.9	7.9	5.9	5.5	7.8	10.8	7.5
MAE	35.8	35.1	(82.4)	32.0	24.5	25.7	31.6	32.4	29.9	32.2	24.9	25.8	31.0	32.5	29.6	33.2	28.9	28.3	32.0	35.6	31.3
MedAE	30.2	32.0	61.1	26.2	20.4	21.1	25.3	28.7	25.0	25.9	20.7	21.4	24.6	28.7	24.3	28.1	24.3	23.4	27.9	32.9	27.4
n=400																					
Bias	-4.9	-11.0	-11.0	-0.2	-0.3	0.0	0.7	-9.7	0.7	-2.0	-2.4	-2.0	-1.6	-10.4	-1.6	-8.7	-14.0	-12.2	-15.7	-19.1	-15.7
MSE	6.4	3.5	9.3	6.0	2.0	2.1	2.3	2.9	2.3	5.6	1.9	2.0	2.2	3.1	2.2	6.2	3.8	3.4	4.4	5.4	4.4
MedSE	2.8	1.8	4.3	2.8	0.9	0.9	1.0	1.6	1.0	2.8	0.8	0.8	1.0	1.7	1.0	2.8	2.1	1.8	2.6	3.4	2.6
MAE	20.1	15.1	24.1	19.6	11.1	11.5	12.1	13.9	12.1	19.3	11.0	11.1	12.0	14.4	12.0	19.9	16.2	15.1	17.5	19.9	17.5
MedAE	16.9	13.4	20.8	16.8	9.4	9.6	10.2	12.6	10.2	16.8	9.0	9.1	10.2	12.9	10.2	16.6	14.5	13.4	16.0	18.5	16.0
n=1600																					
Bias	-1.5	-7.6	-3.6	0.0	-0.3	-0.3	0.0	-6.3	0.0	-1.0	-2.2	-1.9	-2.3	-6.9	-2.3	-4.8	-12.4	-11.1	-16.1	-16.4	-16.1
MSE	3.3	1.1	1.1	3.3	0.4	0.4	0.5	0.9	0.5	3.3	0.5	0.5	0.6	1.0	0.6	3.5	2.0	1.7	3.1	3.2	3.1
MedSE	1.5	0.7	0.5	1.5	0.2	0.2	0.2	0.5	0.2	1.5	0.2	0.2	0.3	0.6	0.3	1.5	1.6	1.3	2.6	2.7	2.6
MAE	14.5	8.6	8.4	14.4	5.3	5.3	5.8	7.9	5.8	14.5	5.5	5.5	5.9	8.3	5.9	14.8	12.6	11.3	16.2	16.5	16.2
MedAE	12.2	8.1	7.3	12.2	4.4	4.3	4.8	7.1	4.8	12.1	4.7	4.5	5.1	7.5	5.1	12.1	12.6	11.2	16.2	16.5	16.2
										σ_ε^2 =400											
n=100																					
Bias	-15.7	-28.7	-48.5	-2.6	-5.8	-2.0	-0.9	-5.4	-1.2	-2.3	-6.5	-2.9	-1.8	-6.2	-1.8	-6.9	-12.8	-8.9	-9.4	-13.4	-9.5
MSE	15.7	15.1	76.2	14.1	7.7	7.7	8.0	6.9	7.7	14.1	7.8	7.9	7.9	7.0	7.8	15.0	9.1	8.6	9.5	8.8	9.4
MedSE	6.7	8.7	30.5	6.1	3.3	3.2	3.1	3.0	3.1	5.7	3.3	3.3	3.1	3.1	3.1	6.4	3.8	3.9	4.0	3.8	4.0
MAE	31.4	32.3	67.7	29.5	21.8	21.8	21.8	20.9	21.6	29.3	21.8	22.1	21.8	20.9	21.7	30.4	23.9	23.2	24.0	23.4	23.9
MedAE	25.9	29.5	55.3	24.7	18.1	18.0	17.7	17.3	17.7	23.8	18.1	18.2	17.7	17.7	17.6	25.3	19.6	19.7	19.9	19.6	19.9
n=400																					
Bias	-3.5	-15.3	-25.1	-0.6	-1.3	0.2	0.3	0.2	0.3	-1.0	-2.5	-0.9	-1.0	-1.0	-1.0	-4.1	-8.4	-6.1	-8.8	-8.8	-8.8
MSE	4.3	4.2	11.7	4.4	1.5	1.4	1.3	1.3	1.3	4.4	1.5	1.4	1.3	1.3	1.3	4.6	2.2	1.8	2.1	2.1	2.1
MedSE	1.8	2.5	6.5	1.9	0.6	0.7	0.5	0.5	0.5	1.9	0.7	0.6	0.6	0.6	0.6	2.1	1.0	0.9	1.1	1.1	1.1
MAE	16.3	17.2	28.2	16.6	9.5	9.6	9.0	9.0	9.0	16.6	9.8	9.5	9.1	9.1	9.1	17.0	11.9	10.7	11.9	11.9	11.9
MedAE	13.5	15.7	25.4	13.6	7.7	8.1	7.4	7.3	7.4	14.0	8.2	8.1	7.6	7.6	7.6	14.4	10.1	9.3	10.5	10.5	10.5
n=1600																					
Bias	-1.0	-9.6	-10.8	-0.2	-0.7	0.0	0.2	0.2	0.2	-0.3	-1.7	-1.0	-1.1	-1.1	-1.1	-2.5	-7.4	-5.6	-8.9	-8.9	-8.9
MSE	2.8	1.4	1.9	2.9	0.4	0.4	0.3	0.3	0.3	2.9	0.4	0.4	0.4	0.4	0.4	2.9	0.9	0.7	1.1	1.1	1.1
MedSE	1.3	0.9	1.2	1.5	0.2	0.2	0.2	0.2	0.2	1.3	0.2	0.2	0.2	0.2	0.2	1.4	0.5	0.3	0.8	0.8	0.8
MAE	13.4	9.9	11.6	13.6	5.0	5.0	4.7	4.7	4.7	13.7	5.2	5.0	4.8	4.8	4.8	13.8	8.0	6.7	9.3	9.3	9.3
MedAE	11.6	9.4	11.0	12.3	4.2	4.3	4.0	4.0	4.0	11.5	4.4	4.5	4.1	4.1	4.1	11.8	7.3	5.8	9.0	9.0	9.0

See note below Table 3.6.

4 Gender wage gap and propensity score matching

In this section, propensity score matching is applied to examine the gender wage gap in the UK. The gender wage gap has been of long-standing political concern as an indicator of discrimination against women. The fact that women are paid substantially lower wages than men in many countries, may be the result of wage discrimination in the labour market. On the other hand, part of this wage gap may be due to differences in education, experience and other skills, whose distribution differs between men and women. To account for the latter, substantial research has been devoted to simulate the *counterfactual* wages that men would earn if they had the human capital endowment of women.[18]

Most of this research attempts to decompose the *average* gender wage gap, following the parametric approach of Blinder (1973) and Oaxaca (1973). In a Blinder-Oaxaca type decomposition, separate (log) wage regressions are estimated for men ($Y_i = \beta_m X_i + \varepsilon_i$) and for women ($Y_i = \beta_w X_i + \varepsilon_i$), where X are human capital variables characterizing productivity. The average wage gap $\bar{Y}_m - \bar{Y}_w$ can be expressed as the sum of two components: $\beta_m(\bar{X}_m - \bar{X}_w) + (\beta_m - \beta_w)\bar{X}_w$. The first part is attributed to differences in average characteristics between men and women. The second part is due to differences in average returns to the individual characteristics, which may reflect discrimination.[19] $\beta_m \bar{X}_w$ represents the counterfactual wage for women if they were paid as men (i.e. if they received the same returns to their human capital as men do). It also represents the counterfactual wage for men, if they had the average human capital endowment of women.

Several recent studies have demonstrated that the functional form assumptions inherent in the parametric Blinder-Oaxaca decomposition can give misleading results, see e.g. Barsky, Bound, Charles, and Lupton (2001), Mora (2000), Nopo (2002), and they proposed nonparametric methods for the analysis of the gender wage gap. The nonparametric approach differs from the parametric approach in two aspects: First, the regression function is no longer specified as linear but left unspecified. Second, the counterfactual mean wage is simulated only for the common support subpopulation. Nopo (2002) suggested the following decomposition: Let $m_w(x) \equiv E[Y|X = x, D = w]$ denote the mean wage for women with characteristics x, and let

[18] For an overview see Blau and Kahn (1997) and Altonji and Blank (1999).

[19] Usually a constant term is included in X. The component of $\beta_m - \beta_w$ corresponding to the constant captures the pay difference that neither can be attributed to different characteristics nor to different rewards.

$f_w(x)$ denote the distribution of X among women. Define $m_m(x)$ and $f_m(x)$ analogously for men. Let $\mathcal{S}$ denote the common support of f_w and f_m and let $\bar{\mathcal{S}}$ denote its complement. By partitioning the male population into the subpopulation with characteristics belonging to the common support and the subpopulation out of the common support, the average wage of men can be written as:

$$\begin{aligned} E[Y|D=m] &= \underset{\mathcal{S}}{E}[Y|D=m]\cdot P_{\mathcal{S}|m} \;+\; \underset{\bar{\mathcal{S}}}{E}[Y|D=m]\cdot P_{\bar{\mathcal{S}}|m} \\ &= \underset{\mathcal{S}}{E}[Y|D=m]\cdot (1-P_{\bar{\mathcal{S}}|m}) \;+\; \underset{\bar{\mathcal{S}}}{E}[Y|D=m]\cdot P_{\bar{\mathcal{S}}|m} \\ &= \left(\underset{\bar{\mathcal{S}}}{E}[Y|D=m]-\underset{\mathcal{S}}{E}[Y|D=m]\right)\cdot P_{\bar{\mathcal{S}}|m} \;+\; \underset{\mathcal{S}}{E}[Y|D=m], \end{aligned}$$

where $P_{\mathcal{S}|m}=\Pr(X\in\mathcal{S}|D=m)=\int_{\mathcal{S}} f_m(x)dx$ is the measure of the set $\mathcal{S}$ under the distribution f_m. With an analogous expression for women, the total gender wage gap can be written as

$$\begin{aligned} &E[Y|D=m]-E[Y|D=w]= \\ &\left(\underset{\bar{\mathcal{S}}}{E}[Y|D=m]-\underset{\mathcal{S}}{E}[Y|D=m]\right)\cdot P_{\bar{\mathcal{S}}|m}-\left(\underset{\bar{\mathcal{S}}}{E}[Y|D=w]-\underset{\mathcal{S}}{E}[Y|D=w]\right)\cdot P_{\bar{\mathcal{S}}|w} \\ &\qquad +\left\{\underset{\mathcal{S}}{E}[Y|D=m]-\underset{\mathcal{S}}{E}[Y|D=w]\right\}. \end{aligned} \tag{22}$$

The total gender wage gap consists of the wage gap in the common support subpopulation (last term) and the wage gap in the subpopulation with X not in the common support (first two terms). For the latter subpopulation, a nonparametric decomposition into an explained and an unexplained part is not possible, since either $m_m(x)$ or $m_w(x)$ are undefined. These two terms thus take into account the wage difference between men in and out of the common support, and analogously for women.[20]

For the common support subpopulation, the counterfactual wages can be simulated and the gender wage gap can be decomposed into an explained and an unexplained part:

$$\underset{\mathcal{S}}{E}[Y|D=m]-\underset{\mathcal{S}}{E}[Y|D=w]=\int_{\mathcal{S}} m_m(x)\cdot\left(f_m^{\mathcal{S}}(x)-f_w^{\mathcal{S}}(x)\right)dx+\int_{\mathcal{S}}\left(m_m(x)-m_w(x)\right)\cdot f_w^{\mathcal{S}}(x)dx,$$

[20] The second term in (22), for example, represents the expected wage increase for a woman with characteristics out of the common support if her characteristics were changed to characteristics that are in the common support. Nopo (2002) argues that these terms represent (at least partly) some form of pre-market discrimination, e.g. in the access to education or in the set of occupational choices. If females cannot enter certain (well-paid) professions ('barriers to entry') or do not have access to college education, this would be reflected in the first term in (22), which measures the average wage premium men receive for characteristics that women do not possess at all.

where $f_m^{\mathcal{S}}(x) = f_m(x)/P_{\mathcal{S}|m}$ and $f_w^{\mathcal{S}}(x) = f_w(x)/P_{\mathcal{S}|w}$ are the densities of X in the subpopulations of men and women belonging to the common support.[21] As in the Blinder-Oaxaca decomposition, the first term represents the part of the gender wage gap (in the common support) that can be attributed to differences in the distribution of the human capital variables between men and women. The second part is due to differences in the returns to these variables, which may (partly) reflect discrimination. Separating these two components requires an estimate of the counterfactual wage that men would earn if they had the human capital characteristics of women:

$$\int_{\mathcal{S}} m_m(x) \cdot f_w^{\mathcal{S}}(x) dx. \tag{23}$$

This corresponds to $\beta_m \bar{X}_w$ in the Blinder-Oaxaca decomposition.

Decomposing the total gender wage gap thus proceeds in two steps. First, the subpopulations of men and women belonging to the common support are estimated. Second, the adjusted mean wage (23) is estimated for the common support subpopulation.

This approach is used by Nopo (2002) to analyze the gender wage gap in Peru. Since gender occupational segregation is very high in Latin America, the characteristics of men and women have different supports. He proposes to use *exact matching* to estimate simultaneously the common support and the adjusted mean wage (23). This approach, however, gives a very conservative estimate of the common support region and is likely to be sensitive to the discretization of continuous variables.[22] In the application below, an alternative way to estimate the common support is proposed and propensity score matching is used to estimate the adjusted mean wage (23), because exact matching performed very poorly in the Monte Carlo simulations.[23]

Recently, interest has arisen to examine not only the mean gender wage gap but also the

[21] I.e. scaled such that they integrate to one.

[22] In addition, it might easily lead to a non-compact support region. A match might be found for a 36-year old woman and also for a 38-year old, but not for a 37-year old woman. Classifying the latter as out-of-support does not appear reasonable and reduces unnecessarily the size of the subpopulation for which the decomposition of main interest can be made. Indeed, Nopo (2002) classifies 23% of male observations and 30% of female observations as out-of-support.

[23] Nopo (2002, Footnote 6) argues against propensity score matching on the grounds that "the 'ignorability of treatment' assumption ... is not likely to be satisfied" in this context. However, as argued in Theorem 1, the ignorability of treatment assumption is not a pre-condition for the justification of propensity score matching.

whole distribution of wages paid to men and women. The basic idea is to simulate the counterfactual distribution of wages that men would receive if they had the human capital distribution of women. Simulating a counterfactual wage distribution by nonparametric methods has been used prominently in DiNardo, Fortin, and Lemieux (1996), who analyzed the increasing wage inequality in the USA and its relation to changes in labour market institutions. Originating from the observed wage distribution in the year 1988, they simulated the density function of wages that would have prevailed if labour market institutions (e.g. the distribution of workers' characteristics, real value of minimum wage etc.) had remained at their 1979 values, via weighting the kernel density estimator by the ratio of the densities of X in 1979 and in 1988.[24] In the analysis of the gender wage gap, nonparametric methods have been used only since very recently.[25] Barsky, Bound, Charles, and Lupton (2001) analyze to which extent the wealth gap between blacks and whites in the USA can be explained by differences in earnings. They estimate the counterfactual density function of the wealth of whites if they had the earnings distribution of blacks. This counterfactual density is estimated by weighting the kernel density estimate by the conditional propensity of being black given earnings. Their estimation problem, however, is rather simple by being only one-dimensional, as there is only a single characteristic (earnings) that needs to be adjusted for. Nopo (2002) also examines briefly the quantile gender wage gap and seems to be the only study so far that uses nonparametric methods in a *multivariate* context to examine the distribution of the wage gap. Through exact matching, he estimates the counterfactual distribution function for women and derives the unexplained wage gap at different quantiles.

In this paper, the gender wage gap among UK university graduates is analyzed with a

[24] Other studies also simulated counterfactual wage distributions in a similar spirit, but relied on parametric models. Juhn, Murphy, and Pierce (1993) used linear regression to investigate the role of worker characteristics and returns in explaining the increasing wage inequality in the USA. Donald, Green, and Paarsch (2000) compared the wage distributions in Canada and the USA and used a parametric proportional-hazards model to simulate the wage distribution when combining the US wage structure with the distribution of covariates (education, unionization, etc.) in Canada. Gosling, Machin, and Meghir (2000) used linear quantile regression for constructing counterfactual wage distributions to examine the impact of specific variables on wage inequality in the UK. In the treatment evaluation literature, Firpo (2003) estimates quantile treatment effects under the ignorable-treatment assumption (selection on observables) via weighting by the propensity score.

[25] Gardeazabal and Ugidos (2001) use parametric quantile regressions to analyze the gender wage gap at different quantiles and decompose the quantile wage gap into explained and unexplained parts.

particular focus on the importance of the subject of study. Machin and Puhani (2003) pointed out that the substantial wage differences between male and female graduates to a large extent can be explained by the observation that men and women choose to study different subjects. When including subject of degree in the Blinder-Oaxaca decomposition, the fraction of the wage gap explained increases from less than a half to two thirds. Using the same data as Machin and Puhani (2003), I will analyze subsequently the mean wage gap as well as its distribution by propensity score matching.

Table 4.1 shows average hourly earnings, age and type of employment and the subject of degree for men and women. The data are based on the UK labour force survey for 1996, where observations with missing information on earnings or other key explanatory variables were deleted. In the survey, university graduates were asked to report the subject of study of their degree, where they could choose from 124 different subjects, of which 102 subjects are observed in the data. In Table 4.1, for descriptive purposes, these 124 subjects are summarized into eleven broader categories. Clear differences in the patterns of degrees between men and women can be seen. Men are more likely to study physical and mathematical sciences and engineering, while women are more likely to enrol in humanities, creative arts and education.

The average wage gap is 3.10 £ (or 0.208 in log units). Machin and Puhani (2003) conducted various Blinder-Oaxaca decompositions to analyze this wage differential. In their preferred specification,[26] they control for age, age squared, part-time work, public sector employment, 10 regional dummies and 9 industry dummies. The estimated male counterfactual log wage $\beta_m \bar{X}_w$ is 2.37, hence 46% of the wage gap is explained by gender differences in these characteristics. When accounting additionally for the subject of degree, the male counterfactual log wage is 2.33, implying that 66% of the wage gap is explained. (When the same analysis is performed with respect to the wage instead of the log wage, the counterfactual wage is 14.03 £ when not controlling for subject of degree and 13.30 £ when controlling for it. Hence, without controlling for degree virtually nothing can be explained by the characteristics, while 23% is explained with the degree.)

In the following analysis it is investigated, whether and how these results change when lack of common support is taken into account and when the counterfactual is nonparametrically estimated. In addition, the distribution of the wage gap is investigated. Restricting the analysis

[26] Specification 2 in Machin and Puhani (2003).

Table 4.1: Descriptive statistics and Subject of Degree

Variable	Men	Women
Gross hourly wage in £	14.01	10.90
log(hourly wage)	2.47	2.26
Age in years	39.1	36.3
Full-time employed	96.3 %	76.1 %
Employed in private sector	60.2 %	40.5 %
Medicine	4.5 %	8.5 %
Biological Science	4.1 %	6.4 %
Veterinary Science	0.1 %	0.2 %
Agricultural Science	2.5 %	2.2 %
Physical and Mathematical Sciences	16.5 %	7.8 %
Engineering/Technology	18.5 %	1.5 %
Architecture and Planning	3.4 %	0.8 %
Social Sciences and Business	28.7 %	26.1 %
Language Studies	3.6 %	11.4 %
Humanities, Creative Arts, Education	13.4 %	29.1 %
Not Classifiable Combined Studies	4.8 %	6.0 %
Number of observations	2983	2183

Note: UK labour force survey 1996, data coded as in Machin and Puhani (2003).

to the common support population requires estimating where the density $f_m(x)$ or $f_w(x)$ is zero. As outlined in Section 2, a practical approach to estimating the common support is to test where the propensity score $\Pr(D = w|X = x)$ is either zero or one. To this end, the propensity score is estimated by a probit model

$$\Pr(D = w|X = x) = \Phi\left(x'\beta\right)$$

with age and age squared as continuous regressors and dummy variables for full-time employment, public sector employment, 9 regional dummies, 8 industry dummies and 101 subject of degree dummies (in total 122 regressors).[27] The estimation results are shown in Table C.1. With β estimated, the propensity score can be computed for every observation as $\hat{p}_i = \Phi(X_i'\hat{\beta})$. The variance of $\hat{p}_i$ is approximated by the delta method as

$$Var(\hat{p}_i) \approx \phi(X_i'\hat{\beta})^2 \cdot X_i' Var(\hat{\beta}) X_i,$$

where a robust estimator for the covariance matrix $Var(\hat{\beta})$ is used.[28] The corresponding t-statistic for testing whether $p_i = c$ is

$$\frac{\hat{p}_i - c}{\sqrt{\widehat{Var}(\hat{p}_i)}}.$$

With this statistic, for each observation it is tested, whether $p_i = 0$ and whether $p_i = 1$, respectively (using critical values for a one-sided test).

Table 4.2 gives the mean of the estimated propensity scores and shows the number of observations for which the test $p_i = 0$ and $p_i = 1$ rejects. At the significance level $a = 0.05$, the hypothesis $H_0 : p_i = 0$ is rejected for 2515 men, which is 84% of the male observations. For the women, $H_0 : p_i = 0$ is rejected for 2155 women or 99% of the female observations. Hence, at the $a = 0.05$ level, 84 % of the male observations are classified as belonging to the common support population, while the other 16% are classified as out of support. For women, the rejection frequency should be close to 100%, since by definition $p_i > 0$ if the observation i is female.

The lower panel of Table 4.2 provides the corresponding results for the test $H_0 : p_i = 1$. Here, the rejection frequency should be close to 100% for men, since $p_i < 1$ if i is male. Of

[27] Since almost all regressors are dummy variables, this specification is rather flexible. Including additionally a large number of interaction terms between subject of degree and the other variables in the model was not possible because many of the subjects of study were studied by only a very few observations.

[28] (Inverse of expected Hessian) · (Outer product of the gradient) · (Inverse of expected Hessian)

Table 4.2: Testing for common support

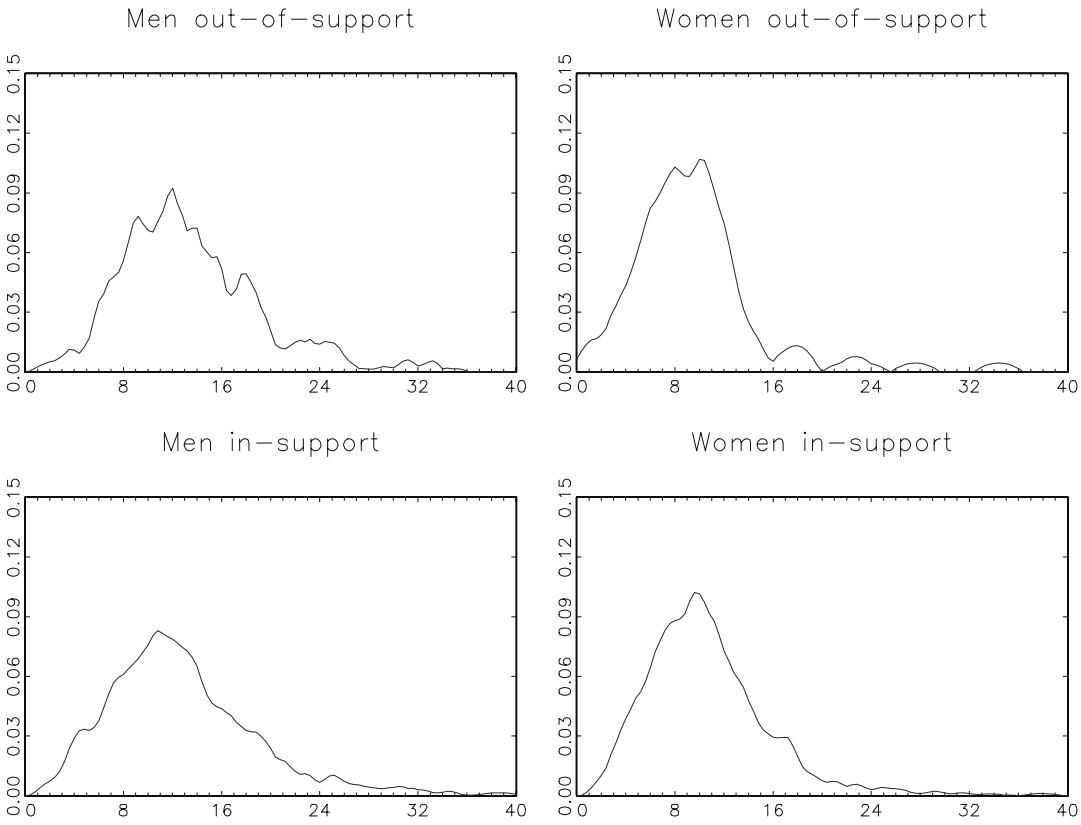
	Men		Women	
Mean of $\hat{p}_i$	0.29		0.60	
$H_0 : p_i = 0$ rejected	obs	%	obs	%
at $\alpha = 10\%$	2743	92.0	2174	99.6
at $\alpha = 5\%$	2515	84.3	2155	98.7
at $\alpha = 1\%$	2222	74.5	2123	97.3
$H_0 : p_i = 1$ rejected				
at $\alpha = 10\%$	2979	99.9	2148	98.4
at $\alpha = 5\%$	2967	99.5	2098	96.1
at $\alpha = 1\%$	2942	98.6	1931	88.5

Note: Upper panel: results for $H_0 : p_i = 0$; lower panel for $H_0 : p_i = 1$. Number of observations (obs) and percentage of observations (%) for which H_0 is rejected at the level α (one sided test).

the female observations, 96% are classified as within-support and 4% as out-of-support, at the $a = 0.05$ level. As seen from the table, these figures clearly depend on the α level, where α represents the probability that an out-of-support observation is classified as within-support. If a larger α is chosen, more observations will be missclassified as within-support. On the other hand, with a very conservative α, the common support population shrinks considerably, such that the subsequent wage gap decomposition is representative only for a much smaller subpopulation. Since the rejection frequency should be about 100% for women when testing $H_0 : p_i = 0$ and for men when testing $H_0 : p_i = 1$, the observed figures can be used to calibrate the α level. From this perspective, $\alpha = 0.05$ seems to be a reasonable choice between retaining a large subpopulation for the subsequent analysis and avoiding nonparametric extrapolation out of support.

From the above analysis of the common support, 15.7 of the male observations but only 3.9% of the female observations are classified as out-of-support. If the same analysis is conducted when neglecting subject of degree in the estimation of the propensity score, (almost) all observations are classified as within support. This indicates that women choose from a smaller set of study subjects. Since the subsequent wage gap decomposition analysis

Figure 4.1: Density of wages for in and out-of-support populations

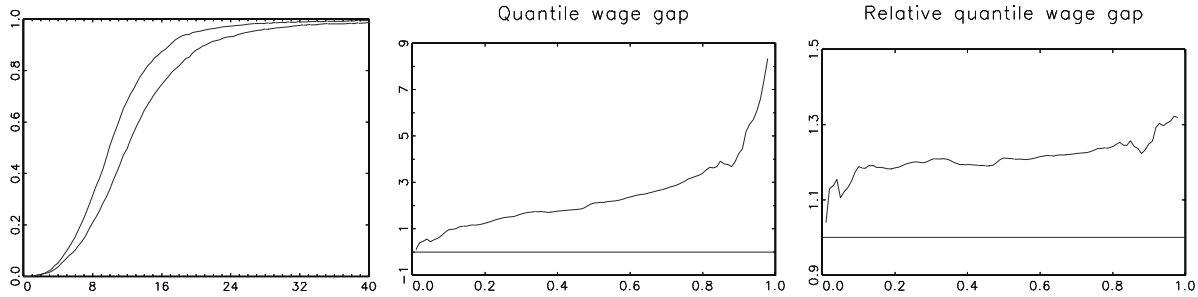


Note: Upper panel: wages for out-of-support subpopulations, lower panel: in-support subpopulations. Results for men on the left, for women on the right.

has to be restricted to the common support population, Figure 4.1 compares the wages between the common support and the out-of-support populations. In the upper half the wage distributions for the out-of-support subpopulations are shown, while the lower half presents the wages for the common support subpopulations. Results for men are on the left and results for women on the right. For both sexes, the out-of-support subpopulation has somewhat less density mass at higher wages than the in-support subpopulation. This is also confirmed by the mean wages, which for men are 13.96 £ for the out-of-support subpopulation and 14.01 £ for the in-support subpopulation; and for women 10.31 £ versus 10.93 £. The differences between out and in-support populations are not very pronounced, though, such that the following analysis, which is restricted to the common support population only, might be roughly informative for the whole population.

The mean gender wage gap in the common support population is 3.09 £ (or 0.193 in log wages) and is thus slightly smaller than for the full population. Figure 4.2 shows the distribution

Figure 4.2: Distribution functions and quantile wage gap for common support population



Left graph: cdf of wages for men and women; middle graph: absolute quantile wage gap ($Q_{\theta|men} - Q_{\theta|women}$); right graph: relative quantile wage gap ($Q_{\theta|men}/Q_{\theta|women}$).

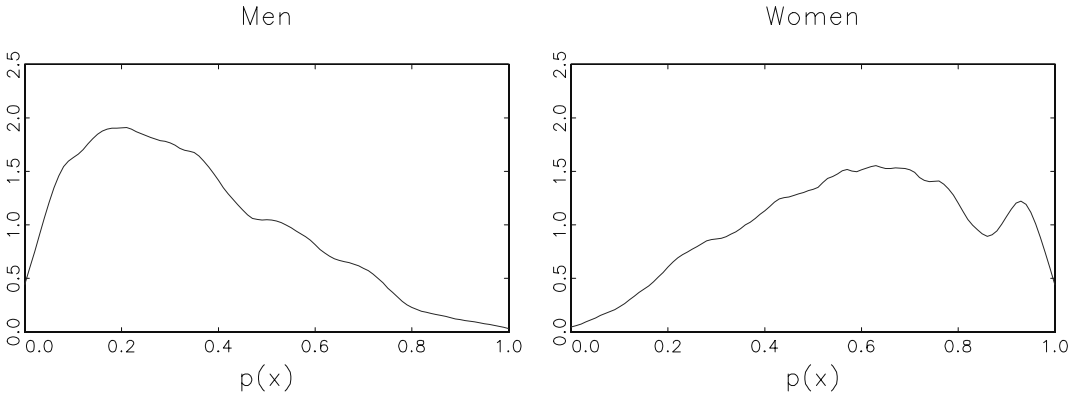
functions for male and female wages and the implied absolute and relative quantile wage gap. The absolute wage gap at, for example, the 20% quantile is the difference between the wage at the 20% quantile in the male distribution and the wage at the 20% quantile in the female distribution.[29] The relative wage gap is the quotient of these two wages. Figure 4.2 thus shows that men at the 20% quantile earn about 1.2 £ (or 18%) more than women at the 20% quantile of their wage distribution. For higher quantiles the wage gap is increasing not only in absolute but also in relative magnitude.

To decompose the *average* wage gap for the common support population, the male counterfactual wage is estimated by adjusting the male wages for the different distribution of the propensity score among men and women. Figure 4.3 shows the distribution of the estimated propensity score in the common support populations. The visible differences between men and women clearly indicate the differences in the observed characteristics between the two sexes. That these differences are largely due to different subjects of study becomes apparent from Figure 4.4, which plots the propensity scores when estimated *without* subject of degree. There the distributions of the propensity score are much more similar, indicating the lower discriminating power of the explanatory characteristics.

The adjusted outcome is estimated by various estimators and the results are shown in Table 4.3. The first rows display the results for the propensity score based matching estimators: pair matching, kernel matching and ridge matching. The bottom rows show the results for the propensity score weighting estimator with various trimming levels. These estimators are implemented exactly as in Section 3. Besides these propensity score based estimators, also the

[29] I.e. it is the *horizontal* distance between the two distribution functions in the leftmost graph in Figure 4.2.

Figure 4.3: Distribution of the propensity score (in common support population)



multivariate kernel matching estimator is examined. (Mahalanobis and exact matching are not included as they performed worse in the Monte Carlo simulations.) Compared to Section 3, the multivariate kernel matching estimator, needs to be adapted to accommodate both continuous and unordered discrete regressors. Therefore, the kernel weights $w(X_j, x)$ are computed as

$$w(X_j, x) = \prod_{k=1}^{1} K\left(\frac{X_{j,k} - x_k}{h}\right) \prod_{k=2}^{5} \lambda_1^{1(X_{j,k} \neq x_k)} \prod_{k=6}^{6} \lambda_2^{1(X_{j,k} \neq x_k)},$$

where X contains the six explanatory variables in the following order: age, part-time employment, sector of employment, region, industry, subject of degree. (K is the Gaussian kernel.) The kernel weights depend on three bandwidths: h, λ_1, λ_2, where h controls the smoothing over the continuous variable *age*, and λ_1 controls the smoothing over the categorical variables: type and sector of employment, region and industry. Finally, λ_2 controls the smoothing over the subject of degree. With $\lambda_2 = 0$, only observations with the same subject of degree would be used in the regression estimator. On the other hand, setting $\lambda_2 = 1$ is equivalent to completely neglecting subject of degree (as in Table 4.4 below). The bandwidths are chosen by cross-validation from a grid of $5 \times 5 \times 5$ values: $(0.01, 0.04, 0.16, 0.64, \infty) \times (0.2, 0.4, .., 1) \times (0.2, 0.4, .., 1)$.

Table 4.3 displays the estimated male counterfactual wage and the implied percentage of the wage gap that is explained by the explanatory characteristics, (with bootstrap standard errors in brackets). The numbers in the left half of the table are estimated with respect to the hourly wage in £, while those in the right half refer to the log wage. The latter estimates are given for comparability with the Blinder-Oaxaca decomposition, which is conventionally conducted with respect to log wages. The results of the three propensity score matching

Table 4.3: Decomposition of the mean gender wage gap (with subject of degree)

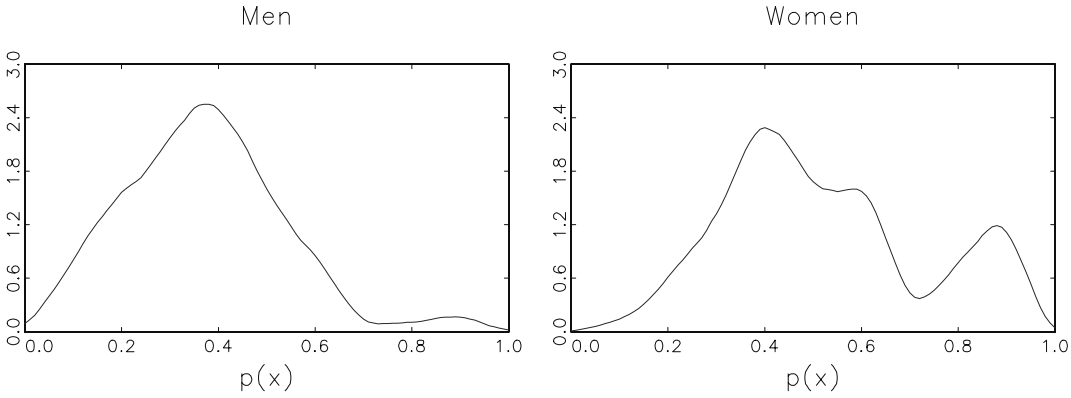
	hourly wages (£)		log wages	
	adjusted mean	% explained	adjusted mean	% explained
Pair Matching	11.60 (0.53)	78 (18.2)	2.305 (0.030)	78 (15.6)
Kernel Matching	12.03 (0.31)	64 (10.8)	2.307 (0.022)	77 (11.6)
Ridge Matching	11.56 (0.35)	79 (12.2)	2.310 (0.024)	75 (12.7)
Multivariate matching	13.41 (0.24)	20 (8.8)	2.411 (0.011)	23 (4.4)
Weighting	13.19 (1.13)	27 (38.8)	2.512 (0.149)	-29 (74.6)
Weighting (c=2)	9.59 (0.26)	143 (12.8)	1.843 (0.034)	317 (28.7)
Weighting (c=5)	10.95 (0.27)	99 (9.5)	2.122 (0.029)	173 (18.8)
Weighting (c=10)	11.55 (0.32)	80 (10.6)	2.250 (0.035)	106 (18.4)
Weighting (c=20)	12.16 (0.49)	60 (16.4)	2.368 (0.050)	46 (24.6)
Weighting (c=50)	13.03 (0.86)	32 (29.2)	2.492 (0.093)	-19 (46.5)

Note: Bandwidths for hourly wages: Kernel (h=0.202), Ridge (h=14.06), Multivariate (h=0.64, λ_1=0.4, λ_2=0.2). For log wages: Kernel (h=0.119), Ridge (h=14.06), Multivariate (h=0.16, λ_1=0.2, λ_2=0.8). Standard errors in brackets (simulated by 100 bootstrap replications).

estimators indicate that about 75 to 80% of the average wage gap can be explained by differences in the explanatory characteristics, which is rather more than the parametric Blinder-Oaxaca decomposition suggested. With respect to log wages, all three matching estimators give the same result. With respect to wages in £, the percentage explained when using kernel matching is somewhat lower. The different results for kernel and ridge matching are partly due to the different bandwidths used. Whereas for kernel matching the cross-validation grid search selected a bandwidth of 0.202 (0.119 for log wages), for ridge matching a very large bandwidth of 14 was chosen. Nevertheless, when using ridge matching with the same bandwidths as selected for kernel matching, the percentage explained with respect to wages in £ reduces to 69.4%, but increases to 79.1% with respect to log wages. Hence, the results are largely robust to bandwidth choice.

These findings are in contradiction with the results of the multivariate matching estimator, which suggests that only about 20% of the average wage gap are explained. However, the results of the multivariate matching estimator do not seem to be very reliable as they are insensitive

Figure 4.4: Distribution of the propensity score without subject of degree



Note: Distribution of propensity score (without subject of degree) in the common support population.

to the inclusion of subject of degree as additional covariate, which is in stark contrast to the findings of Machin and Puhani (2003) and the above analysis. It seems that cross-validation grid search did not succeed in selecting appropriate bandwidths. This is further discussed with Table 4.4 below.

The results of the weighting estimator are given in Table 4.3 for various ceiling levels. The results are highly sensitive to the ceiling level and to whether the analysis is conducted with respect to the wage or the log wage. With respect to the wage and depending on the ceiling level, anything from 27% to 143% of the wage gap is explained.

To assess the explanatory power of the subject of degree, all estimations were repeated *without* the subject of degree variable. When examining the common support without the subject of degree variable, the support analysis yielded that all (but 3) observations can be classified as in-support. (Results of the estimation of the propensity score are given in Table C.2.) For being comparable with the previous analysis, however, the following estimations are performed only with the subpopulation that also belongs to the common support population when subject of degree is included. I.e. Figures 4.3 and 4.4 and Tables 4.3 and 4.4 are based on exactly the same observations (also in the bootstrap). Figure 4.4 gives the distributions of the propensity score when subject of degree is neglected. The differences between men and women are now much less pronounced as when subject of degree was included (as in Figure 4.3).

Table 4.4 gives the estimation results of the average male counterfactual wage without subject of degree. Kernel and ridge matching indicate that about 45% of the wage gap is explained by these variables (age, part-time work, sector of employment, industry and region)

Table 4.4: Decomposition of the mean gender wage gap (without subject of degree)

	hourly wages (£)		log wages	
	adjusted mean	% explained	adjusted mean	% explained
Pair Matching	12.12 (0.82)	61 (25.9)	2.341 (0.028)	59 (14.8)
Kernel Matching	12.66 (0.47)	44 (10.2)	2.336 (0.020)	62 (10.2)
Ridge Matching	12.63 (0.50)	45 (11.3)	2.335 (0.022)	62 (11.1)
Multivariate matching	13.40 (0.24)	20 (8.8)	2.411 (0.011)	23 (4.4)
Weighting	12.79 (0.62)	40 (15.2)	2.383 (0.059)	38 (29.0)
Weighting (c=2)	10.43 (0.53)	116 (17.7)	1.920 (0.040)	277 (25.9)
Weighting (c=5)	11.46 (0.57)	83 (15.8)	2.121 (0.044)	173 (23.7)
Weighting (c=10)	12.28 (0.60)	56 (15.4)	2.278 (0.044)	92 (22.2)
Weighting (c=20)	12.78 (0.62)	40 (15.4)	2.380 (0.054)	39 (26.7)
Weighting (c=50)	12.79 (0.62)	40 (15.2)	2.383 (0.059)	38 (29.0)

Note: Bandwidths for hourly wages: Kernel (h=0.070), Ridge (h=0.343), Multivariate (h=0.64, λ_1=0.4). For log wages: Kernel (h=0.041), Ridge (h=0.070), Multivariate (h=0.16, λ_1=0.2). Standard errors in brackets (simulated by 100 bootstrap replications).

alone. Contrasting these results with Table 4.3, it is seen that subject of degree had contributed substantially to explaining the gender wage gap. The results of the weighting estimator are as sensitive to the ceiling level as before, but also indicate the substantial contribution of subject of degree to the explanation of the gap. Only the results of the multivariate matching estimator seem to indicate that subject of degree is completely irrelevant: The estimated adjusted means are identical in Table 4.3 and 4.4, as well with respect to the wage as to the log wage. This insensitivity also persists when choosing different bandwidth values. Only with a very small bandwidth value λ_2 for the study subject do the results depend on the inclusion of the subject of degree regressor. With $\lambda_2 = 0$, the results of multivariate matching in Table 4.3 change to 12.69 £ (43% explained) and for log wages to 2.373 (43% explained). This indicates that cross-validation did not succeed here in selecting suitable bandwidth values for multivariate matching. Allowing for more bandwidth parameters (e.g. separate bandwidths for region and industry) and a more extensive search grid would be an option, but is beyond the scope of this paper.

The previous analysis had shown that with subject of degree about 75% of the average wage gap can be explained. Without subject of degree, this fraction is much lower. To analyze the gender wage gap in more detail, we would like to compare the gap at different quantiles and examine to which extent it can be explained by observed characteristics. To this end, the distribution function $F_{Y|D=m}(a)$ is adjusted for the different distributions of the covariates among men and women. Noting that $F_{Y|D=m}(a)$ can be written as $E[1(Y \leq a)|D = m]$, the gender gap in the distribution functions $F_{Y|D=m}(a) - F_{Y|D=w}(a)$ can be decomposed analogously to (22). As with the mean wage gap, the decomposition in the common support subpopulation is of particular interest, since the counterfactual simulation is possible only for this subpopulation.

Let $F_{Y|D=m,\mathcal{S}}(a)$ denote the distribution function among men with characteristics belonging to the common support $\mathcal{S}$. The *adjusted* distribution function for men $F^*_{Y|D=m,\mathcal{S}}(a)$ is

$$F^*_{Y|D=m,\mathcal{S}}(a) \;=\; \int_{\mathcal{S}} F_{Y|X,D=m}(a,x) \cdot f_w^{\mathcal{S}}(x)dx,$$

where $F_{Y|X,D=m}$ is the conditional cdf given X. The adjusted cdf $F^*_{Y|D=m,\mathcal{S}}(a)$ represents the wage distribution that would prevail among men if they had the human capital characteristics of women. Writing the adjusted cdf in the following form

$$F^*_{Y|D=m,\mathcal{S}}(a) \;=\; \int_{\mathcal{S}} E\left[1\left(Y \leq a\right) \;|X = x, D = m\right] \cdot f_w^{\mathcal{S}}(x)dx, \tag{24}$$

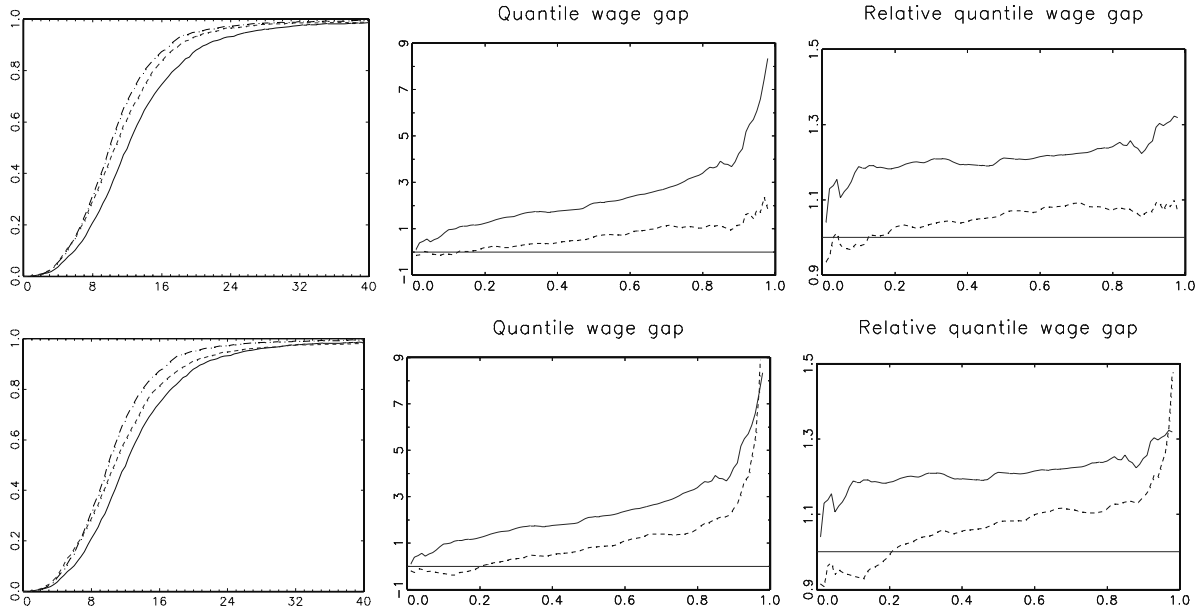
a matching estimator of the adjusted cdf is

$$\hat{F}^*_{Y|D=m,\mathcal{S}}(a) = \frac{1}{n_w^{\mathcal{S}}} \sum_{i:D_i=w,X_i\in\mathcal{S}} \hat{F}_{Y|X,D=m}\left(X_i\right)$$

where $\hat{F}_{Y|X,D=m}\left(x\right)$ is a nonparametric estimate of the conditional cdf $F_{Y|X,D=m}(a,x)$ and $n_w^{\mathcal{S}}$ is the number of female observations with characteristics in the common support. The conditional cdf can be estimated by nonparametric regression of the binary variable $1(Y \leq a)$ on X in the male subpopulation belonging to the common support. As a result of Theorem 1, the adjusted cdf can also be estimated by using the propensity score instead of X.

Figure 4.5 shows the estimated counterfactual cdf with propensity score ridge matching. The top panel shows the results with subject of degree, the bottom panel for without subject of degree. (Bandwidths are as in Table 4.3 and 4.4.) The leftmost graph shows the cdf of male wages (bold), the cdf of female wages (long dashed) and the adjusted male cdf (dashed). The

Figure 4.5: Decomposition of the quantile wage gap with propensity-score ridge matching



Top panel: Adjustment with subject of degree; bottom panel: without subject of degree. Left graph: adjusted cdf; middle graph: absolute quantile wage gap; right graph: relative quantile wage gap. Solid lines are unadjusted and identical to Figure 4.2. Dashed lines are after adjustment.

middle and the rightmost graph display the absolute and relative quantile wage gap derived from the estimated cdf. The solid lines are identical to Figure 4.2 and give the unadjusted wage gap between the observed male cdf and the observed female cdf. The dashed line is the wage gap after adjustment for the covariate distributions, i.e. the difference between the *adjusted* male cdf and the observed female cdf. This line shows which part of the wage gap is *not* explained. These figures clearly show that up to about the 25% quantile, differences in the covariates almost fully explain the gender wage gap. For larger quantiles, however, the unexplained wage differences become more pronounced.

Up to about the 50% quantile, these unexplained differences are of similar size with and without subject of degree. Beyond the 50% quantile, however, subject of degree contributes increasingly to explaining the wage gap, as well with respect to the absolute as to the relative gap. At very high wages (above 90% quantile), the unexplained gap is less than 10% of female wages when subject of degree is accounted for. However, when neglecting subject of degree, the gap after adjustment is (almost) as large as without adjustment. Figures C.1 and C.2 show the analogous results when propensity score pair matching or kernel matching are used. Those results are largely similar and confirm the findings of Figure 4.5.

These results show that the fraction of the wage gap that can be explained by observed covariates depends on the position in the wage distribution. While at low quantiles, the wage gap can be explained almost completely (even without subject of degree), at higher quantiles the power of the regressors to explain the wage gap decreases. It is at these higher quantiles, where the additional explanatory power of the subject of degree variable, whose substantial contribution was already noted in the decomposition of the mean wage gap, becomes most relevant.

5 Conclusions

In this paper, the versatility of propensity score matching has been demonstrated. First, it has been shown that the justification of propensity score matching as an alternative to matching on X does not depend on a conditional independence assumption; it can be used even when selection is on unobservables. Propensity score matching can be used not only to estimate the adjusted mean outcome, but also its density and distribution functions. These results also extend to the case of non-random sampling (stratified sampling, choice-based sampling) under appropriate modifications of a weighted propensity score matching estimator. A small Monte Carlo study showed that propensity score matching may perform superior to matching on X, but can be sensitive to misspecification of the propensity score model.

Propensity score matching was then applied to analyze the gender wage gap among college graduates in the UK. It was found that subject of degree is an important variable to explaining the wage differences between men and women. Its contribution appeared to be even greater than the parametric Blinder-Oaxaca decomposition had suggested. When examining the wage gap at different quantiles, it turned out that the explanatory power of subject of degree is particularly relevant at the higher end of the income distribution. The huge wage differential between high-earning men and high-earning women is thus to a large extent the result of men and women choosing different subjects in university.

A Appendix to Section 2

A.1 Proof of Theorem 1

To prove Theorem 1, it is first shown that $\frac{f_{p|D=t}(\rho)}{f_{p|D=s}(\rho)} = \frac{\rho}{1-\rho}\frac{P_s}{P_t}$ for $\rho < 1$ where $P_t = \Pr(D = t|D \in \{s,t\})$

By Bayes' theorem

$$\Pr\left(D = t|p(X) = \rho, D \in \{s,t\}\right) = \frac{f_{p|D=t}(\rho) \cdot \Pr\left(D = t|D \in \{s,t\}\right)}{f_{p|D\in\{s,t\}}(\rho)}.$$

By iterated expectations and the definition of the propensity score, it follows also that

$$\begin{aligned}\Pr\left(D = t|p(X) = \rho, D \in \{s,t\}\right) &= E\left[\ \Pr\left(D = t|X, D \in \{s,t\}\right)\ |p(X) = \rho, D \in \{s,t\}\right] \\ &= E\left[\ p(X)\ |p(X) = \rho, D \in \{s,t\}\right] = \rho.\end{aligned}$$

Both results together thus imply that

$$\frac{f_{p|D=t}(\rho) \cdot P_t}{f_{p|D\in\{s,t\}}(\rho)} = \rho$$

and

$$\frac{f_{p|D=t}(\rho)}{f_{p|D=s}(\rho)} = \frac{\rho}{1-\rho}\frac{P_s}{P_t} \qquad \text{for } \rho < 1. \tag{25}$$

With this result, Theorem 1 can be proven. The *left-hand side* of (5) can be written as

$$\begin{aligned}&\underset{\mathcal{S}}{E}\left[\xi_{y|x,d}(X,s)\ |D = t\right] \\ &= \int \xi_{y|x,d}(x,s) \cdot f_{x|D=t}(x)\frac{1(x \in \mathcal{S})}{\Pr(X \in \mathcal{S}\ |D = t)}dx \\ &= \int \xi_{y|x,d}(x,s) \cdot f_{x|D=s}(x)\frac{p(x)}{1-p(x)}\frac{P_s}{P_t}\frac{1(x \in \mathcal{S})}{\Pr(X \in \mathcal{S}\ |D = t)}dx\end{aligned}$$

since $\frac{p(x)}{1-p(x)} = \frac{f_{x|D=t}}{f_{x|D=s}}\frac{P_t}{P_s}$ for $x \in \mathcal{S}$ by Bayes' theorem.

$$= E\left[\xi_{y|x,d}(X,s)\frac{p(X)}{1-p(X)}\ |D = s\right]\frac{P_s}{\Pr(X \in \mathcal{S}\ |D = t) \cdot P_t}.$$

Using iterated expectation gives

$$\begin{aligned}&= E\left[E\left[\xi_{y|x,d}(X,s)\frac{p(X)}{1-p(X)}\ |p(X) = P, D = s\right]\ |D = s\right]\frac{P_s}{\Pr(X \in \mathcal{S}\ |D = t) \cdot P_t} \\ &= E\left[\frac{P}{1-P}E\left[\xi_{y|x,d}(X,s)\ |p(X) = P, D = s\right]\ |D = s\right]\frac{P_s}{\Pr(X \in \mathcal{S}\ |D = t) \cdot P_t}.\end{aligned}$$

The *right-hand side* of (5) can be written as

$$\begin{aligned}
& \underset{\mathcal{S}}{E}\left[\xi_{y|p,d}(P,s)\ |D=t\right] \\
= & \int \xi_{y|p,d}(\rho,s)\cdot f_{p|D=t}(\rho)\frac{1(\rho<1)}{\Pr(P<1|D=t)}d\rho \\
= & \int \xi_{y|p,d}(\rho,s)\cdot f_{p|D=s}(\rho)\frac{\rho}{1-\rho}\frac{P_s}{P_t}\frac{1(\rho<1)}{\Pr(P<1|D=t)}d\rho \\
= & \underset{\mathcal{S}}{E}\left[\xi_{y|p,d}(P,s)\cdot\frac{P}{1-P}\ |D=s\right]\frac{P_s}{\Pr(P<1|D=t)\cdot P_t}
\end{aligned}$$

by (25). Noting that $\Pr(X\in\mathcal{S}\ |D=t)=\Pr(P<1|D=t)$ by the definition of the propensity score. Hence, the left-hand and the right-hand side of (5) are identical if condition (4) holds.

A.2 Proof of Theorem 2, Part A

The equality (18)

$$\begin{aligned}
& E^0\left[E^0\left[Y|X,D=s\right]\ |D=t\right] \\
& \qquad = \int E^*\left[W_{yx|t}|p^0(X)=\rho,D=t\right]\cdot\frac{E^*\left[Y\cdot W_{yx|s}|p^0(X)=\rho,D=s\right]}{E^*\left[W_{yx|s}|p^0(X)=\rho,D=s\right]}\cdot f^*_{p^0|t}(\rho)d\rho.
\end{aligned}$$

is proven.

The propensity score under the population distribution is $p^0(x)=f^0_{x|t}P^0_t/f^0_x$. The propensity score under the sampling distribution is $p^*(x)=f^*_{x|t}P^*_t/f^*_x$. Define, similar to Hahn (1998), the equivalence classes $\chi(\rho)$ as the set of all (y,x) values where a deterministic function of (y,x) takes the value ρ, such that the union of the equivalence classes corresponds to the support of (Y,X). In particular, define $\chi^0(\rho)=\{(y,x):p^0(x)=\rho\}$ as the equivalence classes with respect to the propensity score p^0. Let $\chi^*(\rho)=\{(y,x):p^*(x)=\rho\}$ be the equivalence classes with respect to p^*. The measure of the set $\chi^*(\rho)$ under the sampling distribution in the $D=d$ population is

$$f^*_{p^*|d}(\rho) \ = \ \int\limits_{\chi^*(\rho)} f^*_{yx|d}\left(y,x\right)dydx.$$

The measure of $\chi^*(\rho)$ under the population distribution is

$$f^0_{p^*|d}(\rho) \ = \ \int\limits_{\chi^*(\rho)} f^0_{yx|d}\left(y,x\right)dydx = \int\limits_{\chi^*(\rho)} w_{yx|d}\left(y,x\right)f^*_{yx|d}\left(y,x\right)dydx.$$

The corresponding measures for $\chi^0(\rho)$ are

$$f^*_{p^0|d}(\rho) \;=\; \int\limits_{\chi^0(\rho)} f^*_{yx|d}\,(y,x)\,dydx.$$

$$f^0_{p^0|d}(\rho) \;=\; \int\limits_{\chi^0(\rho)} w_{yx|d}\,(y,x)\,f^*_{yx|d}\,(y,x)\,dydx.$$

Now examine the term

$$\begin{aligned} & E^*\left[W_{yx|t}|p^0(X)=\rho, D=t\right] \\ = & \int\limits_{\chi^0(\rho)} w_{yx|t}(y,x)\cdot f^*_{yx|t,p^0=\rho}(y,x)dydx. \end{aligned}$$

By noticing that $f^*_{yx|t,p^0=\rho} \propto f^*_{yx|t}$ for every $(y,x) \in \chi^0(\rho)$, it follows:

$$= \frac{\int\limits_{\chi^0(\rho)} w_{yx|t}(y,x) f^*_{yx|t}(y,x)dydx}{\int\limits_{\chi^0(\rho)} f^*_{yx|t}(y,x)dydx} = \frac{f^0_{p^0|t}(\rho)}{f^*_{p^0|t}(\rho)}. \tag{26}$$

And analogously for the source population:

$$E^*\left[W_{yx|s}|p^0(X)=\rho, D=s\right] = \frac{f^0_{p^0|s}(\rho)}{f^*_{p^0|s}(\rho)}. \tag{27}$$

Now consider the term

$$
\begin{aligned}
& E^*\left[Y \cdot W_{yx|s} | p^0(X) = \rho, D = s\right] \\
= & \int\limits_{\chi^0(\rho)} y \cdot w_{yx|s}(y,x) \cdot f^*_{yx|s,p^0=\rho}(y,x) dy dx \\
= & \frac{\int\limits_{\chi^0(\rho)} y \cdot w_{yx|s}(y,x) \cdot f^*_{yx|s}(y,x) dy dx}{\int\limits_{\chi^0(\rho)} f^*_{yx|s}(y,x) dy dx} \\
= & \int\limits_{\chi^0(\rho)} y \cdot f^0_{yx|s}(y,x)\, dy dx \;/\; f^*_{p^0|s}(\rho) \\
= & \int\limits_{\chi^0(\rho)} y \cdot \frac{f^0_{yx|s}(y,x)}{f^0_{p^0|s}(\rho)} dy dx \cdot \frac{f^0_{p^0|s}(\rho)}{f^*_{p^0|s}(\rho)} \\
= & \int\limits_{\chi^0(\rho)} y \cdot f^0_{yx|s,p^0=\rho}(y,x) dy dx \frac{f^0_{p^0|s}(\rho)}{f^*_{p^0|s}(\rho)} \\
= & E^0\left[Y | p^0(X) = \rho, D = s\right] \frac{f^0_{p^0|s}(\rho)}{f^*_{p^0|s}(\rho)} \qquad (28)
\end{aligned}
$$

Taken (26), (27) and (28) together, gives

$$
\begin{aligned}
& \int E^*\left[W_{yx|t} | p^0(X) = \rho, D = t\right] \cdot \frac{E^*\left[Y \cdot W_{yx|s} | p^0(X) = \rho, D = s\right]}{E^*\left[W_{yx|s} | p^0(X) = \rho, D = s\right]} \cdot f^*_{p^0|t}(\rho) d\rho \\
= & \int \frac{f^0_{p^0|t}(\rho)}{f^*_{p^0|t}(\rho)} \cdot \frac{E^0\left[Y | p^0(X) = \rho, D = s\right] \frac{f^0_{p^0|s}(\rho)}{f^*_{p^0|s}(\rho)}}{f^0_{p^0|s}(\rho) / f^*_{p^0|s}(\rho)} \cdot f^*_{p^0|t}(\rho) d\rho \\
= & \int E^0\left[Y | p^0(X) = \rho, D = s\right] \cdot f^0_{p^0|t}(\rho) d\rho = E^0\left[E^0\left[Y | p^0(X) = P, D = s\right] \;|D = t\right],
\end{aligned}
$$

where all expectation operators are now with respect to the *population distribution*. Hence, Theorem 1 applies:

$$
= E^0\left[E^0\left[Y | X, D = s\right] \;|D = t\right],
$$

which proves the equality.

A.3 Proof of Theorem 2, Part B

The equality (19) of Theorem 2 shows the equivalence of using $p^{0*}(x)$ instead of $p^0(x)$:

$$\int E^*\left[W_{yx|t}|p^0(X)=\rho, D=t\right]\cdot\frac{E^*\left[Y\cdot W_{yx|s}|p^0(X)=\rho, D=s\right]}{E^*\left[W_{yx|s}|p^0(X)=\rho, D=s\right]}\cdot f^*_{p^0|t}(\rho)d\rho$$
$$=\int E^*\left[W_{yx|t}|p^{0*}(X)=\rho, D=t\right]\cdot\frac{E^*\left[Y\cdot W_{yx|s}|p^{0*}(X)=\rho, D=s\right]}{E^*\left[W_{yx|s}|p^{0*}(X)=\rho, D=s\right]}\cdot f^*_{p^{0*}|t}(\rho)d\rho.$$

Since all but the last one of the derivations of the proof of Theorem 2 Part A can also be performed analogously with respect to $p^{0*}(x)$, it remains to be shown that:

$$\int E^0\left[Y|p^{0*}(X)=\rho, D=s\right]\cdot f^0_{p^{0*}|t}(\rho)d\rho=\int E^0\left[Y|p^0(X)=\rho, D=s\right]\cdot f^0_{p^0|t}(\rho)d\rho. \quad (29)$$

By definition of the propensity scores:

$$\frac{p^0(x)}{1-p^0(x)}=\frac{f^0_{x|t}(x)}{f^0_{x|s}(x)}\frac{P^0_t}{P^0_s}$$

and

$$\frac{p^{0*}(x)}{1-p^{0*}(x)}=\frac{f^0_{x|t}(x)}{f^0_{x|s}(x)}\frac{P^*_t}{P^*_s}.$$

Taking the ratio of both expressions gives

$$\frac{p^{0*}(x)}{1-p^{0*}(x)}\frac{1-p^0(x)}{p^0(x)}=\frac{P^*_t}{P^*_s}\frac{P^0_s}{P^0_t}=\alpha+1, \quad (30)$$

with α defined in (13). This gives

$$p^0(x)=\frac{p^{0*}(x)}{\alpha+1-\alpha p^{0*}(x)}$$

Hence, there is a one-to-one correspondence between the propensity score $p^0(x)$ under the population distribution and $p^{0*}(x)$, where choice-based sampling is neglected. Notice that the correspondence is strictly positively monotonous.[30] It follows that a direct mapping between the distribution functions of $p^0(x)$ and $p^{0*}(x)$ exists:

$$F^0_{p^{0*}|t}(\rho)=F^0_{p^0|t}\left(\frac{\rho}{\alpha+1-\alpha\rho}\right)\qquad 0<\rho<1,$$

[30] The first derivative is $\frac{dp^0(x)}{dp^{0*}(x)}=\frac{\alpha+1}{\left(\alpha+1-\alpha p^{0*}(x)\right)^2}$ which is always positive, because $\alpha+1$ is always positive, see (13).

because the sets of (y,x) values where the indicator function is non-zero is identical in $F^0_{p^{0*}|t}(\rho) = \int 1\left(p^{0*}(x) \leq \rho\right) f^0_{yx|t}(y,x)dydx$ and in $F^0_{p^0|t}\left(\frac{\rho}{\alpha+1-\alpha\rho}\right) = \int 1\left(p^0(x) \leq \frac{\rho}{\alpha+1-\alpha\rho}\right) f^0_{yx|t}(y,x)dydx$. With these results, the left-hand side of (29) can be written as

$$\begin{aligned}
& \int E^0\left[Y \mid p^{0*}(X) = \rho, D = s\right] \cdot \frac{dF^0_{p^{0*}|t}(\rho)}{d\rho} d\rho \\
= & \int E^0\left[Y \mid p^0(x) = \frac{\rho}{\alpha+1-\alpha\rho}, D = s\right] \cdot \frac{dF^0_{p^0|t}\left(\frac{\rho}{\alpha+1-\alpha\rho}\right)}{d\rho} d\rho \\
= & \int E^0\left[Y \mid p^0(x) = u, D = s\right] \cdot f^0_{p^0|t}(u)\, du
\end{aligned}$$

where $u = \frac{\rho}{\alpha+1-\alpha\rho}$, which is identical to the right-hand side of of (29). Hence choice-based sampling can be neglected for propensity score matching.

A.4 Proof of Theorem 2, Part C

To proof the equality (21) of Theorem 2, it must be shown that:

$$\int E^0\left[Y|p^*(X) = \rho, D = s\right] \cdot f^0_{p^*|t}(\rho)d\rho = \int E^0\left[Y|p^0(X) = \rho, D = s\right] \cdot f^0_{p^0|t}(\rho)d\rho. \tag{31}$$

With

$$\frac{p^*(x)}{1-p^*(x)} = \frac{f^*_{x|t}(x)}{f^*_{x|s}(x)} \frac{P^*_t}{P^*_s}$$

the odds ratio is

$$\begin{aligned}
\frac{p^*(x)}{1-p^*(x)} \frac{1-p^0(x)}{p^0(x)} &= \frac{f^*_{x|t}(x)}{f^*_{x|s}(x)} \frac{P^*_t}{P^*_s} \frac{f^0_{x|s}(x)}{f^0_{x|t}(x)} \frac{P^0_s}{P^0_t} \\
&= \frac{f^*_{x|t}(x)}{f^0_{x|t}(x)} \frac{f^0_{x|s}(x)}{f^*_{x|s}(x)} (\alpha+1) \\
&= \gamma(\alpha+1),
\end{aligned}$$

by assumption (20). The rest of the proof is identical to the proof of Theorem 2b, starting from (30) with $(\alpha+1)$ replaced by $\gamma(\alpha+1)$.

B Appendix to Section 3

Table B.1: Distribution of X in source and target population, $\sigma_\varepsilon^2 = 30$

	Source population ($D=0$)					Target population ($D=1$)				
	5%	25%	*Median*	75%	95%	5%	25%	*Median*	75%	95%
X_1	-2.71	-1.45	-0.59	0.25	1.45	-1.36	-0.17	0.67	1.52	2.76
X_2	-1.91	-1.01	-0.40	0.21	1.08	-1.02	-0.16	0.45	1.06	1.95
X_3	-1.07	-0.21	0.40	1.01	1.91	-1.94	-1.06	-0.44	0.16	1.02
X_4	-2.58	-1.10	0.45	1.80	2.77	-2.77	-1.82	-0.50	1.05	2.56
X_5	0.00	0.11	0.50	1.46	4.21	0.00	0.09	0.41	1.18	3.41

Note: Percentiles of X in source and target population.

Table B.2: Distribution of X in source and target population, $\sigma_\varepsilon^2 = 100$

	Source population ($D=0$)					Target population ($D=1$)				
	5%	25%	*Median*	75%	95%	5%	25%	*Median*	75%	95%
X_1	-2.64	-1.32	-0.41	0.50	1.80	-1.77	-0.47	0.44	1.35	2.67
X_2	-1.86	-0.92	-0.27	0.37	1.30	-1.28	-0.35	0.29	0.94	1.87
X_3	-1.30	-0.37	0.27	0.92	1.86	-1.87	-0.94	-0.29	0.35	1.28
X_4	-2.63	-1.24	0.31	1.71	2.75	-2.75	-1.72	-0.33	1.23	2.62
X_5	0.00	0.11	0.49	1.42	4.11	0.00	0.09	0.42	1.23	3.54

Table B.3: Means of X in source and target population, $\sigma_\varepsilon^2 = 100$ and 400

$\sigma_\varepsilon^2 = 100$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
$D=0$	-0.13	-0.14	-0.14	-0.15	-0.15	-0.15	-0.15	-0.16	-0.16	-0.16
$D=1$	0.13	0.14	0.14	0.15	0.15	0.15	0.15	0.16	0.16	0.16
$\sigma_\varepsilon^2 = 400$										
$D=0$	-0.07	-0.08	-0.08	-0.08	-0.08	-0.08	-0.08	-0.09	-0.09	-0.09
$D=1$	0.07	0.08	0.08	0.08	0.08	0.08	0.08	0.09	0.09	0.09

C Appendix to Section 4

Figure C.1: Decomposition of the quantile wage gap with propensity score pair matching

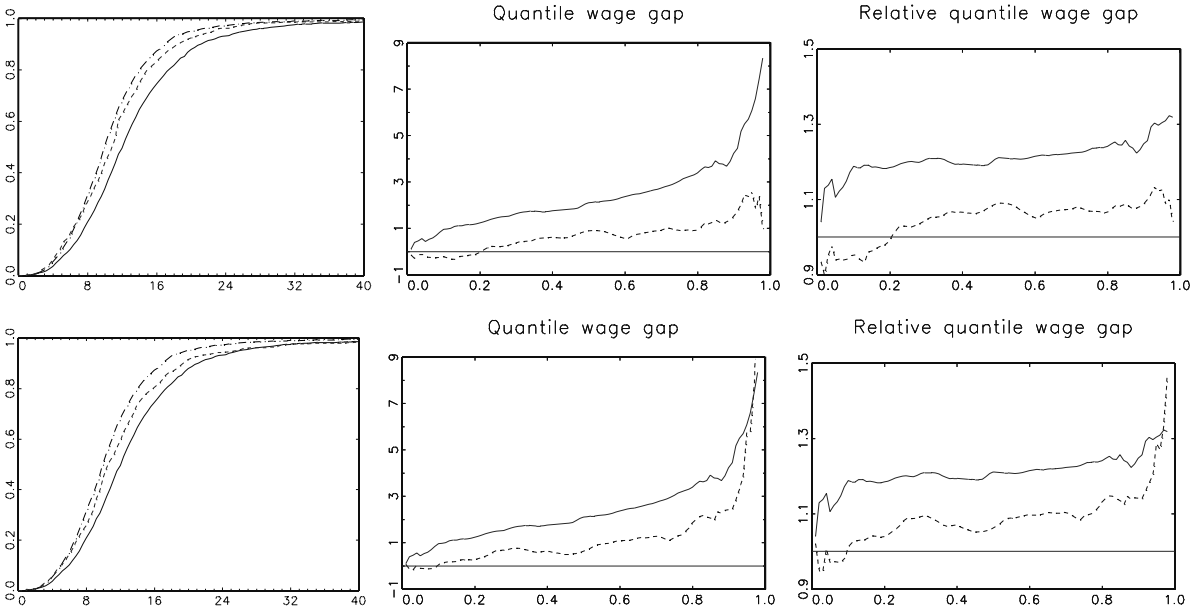


Figure C.2: Decomposition of the quantile wage gap with propensity score kernel matching

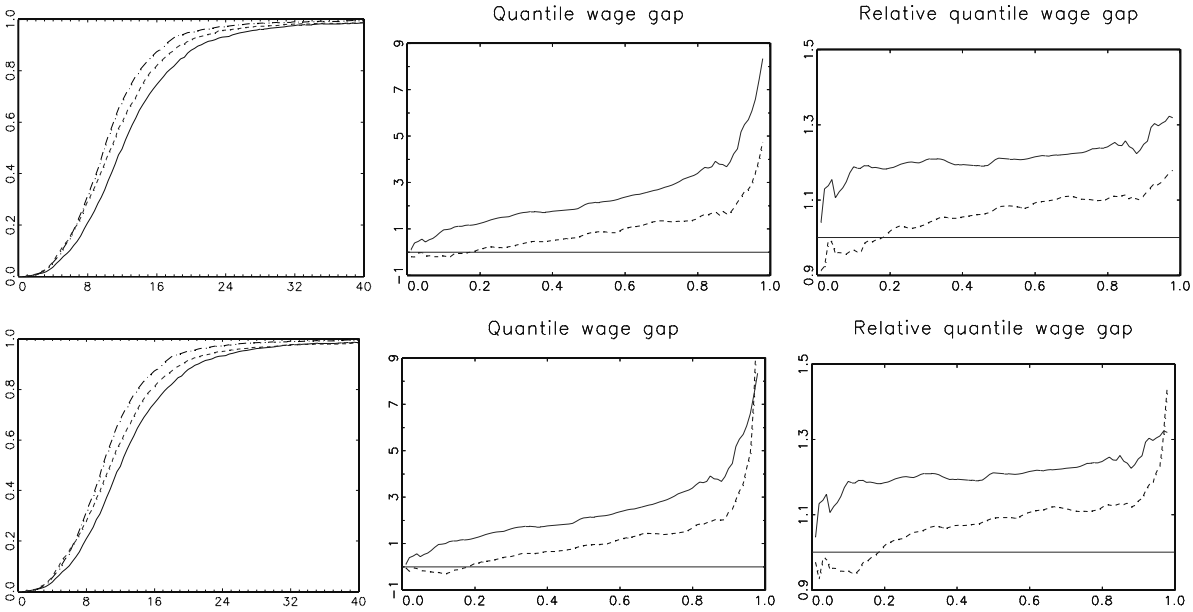


Table C.1: Probit estimation of the propensity score with Subject of Degree

Variable	Estimate	Std.error	t-statistic	Variable	Estimate	Std.error	t-statistic
Const	2.10	0.34	6.13	Subject 46	0.78	0.22	3.57
Age / 10	-0.23	0.16	-1.44	Subject 47	-0.09	0.21	-0.46
Age2 /1000	-0.12	0.20	-0.61	Subject 48	-0.21	0.29	-0.71
Full time work	-1.24	0.07	-17.42	Subject 49	-0.01	0.15	-0.09
Private sector	-0.15	0.06	-2.50	Subject 50	-0.15	0.15	-1.02
Region 1	0.05	0.12	0.44	Subject 51	-0.08	0.17	-0.45
Region 2	0.05	0.09	0.57	Subject 52	-0.20	0.30	-0.67
Region 4	-0.24	0.09	-2.53	Subject 53	-0.39	0.15	-2.54
Region 5	-0.07	0.09	-0.80	Subject 54	0.80	0.30	2.65
Region 6	-0.06	0.13	-0.45	Subject 55	0.20	0.45	0.46
Region 7	-0.03	0.06	-0.43	Subject 56	-0.55	0.39	-1.39
Region 8	-0.08	0.10	-0.87	Subject 57	0.34	0.29	1.19
Region 9	-0.09	0.11	-0.83	Subject 58	1.08	0.57	1.89
Region 10	-0.02	0.09	-0.26	Subject 59	0.70	0.16	4.24
Industry 1	-0.34	0.43	-0.80	Subject 60	0.49	0.60	0.82
Industry 2	-0.30	0.17	-1.79	Subject 61	-0.29	0.31	-0.94
Industry 4	-0.27	0.11	-2.53	Subject 62	0.92	0.24	3.87
Industry 5	-0.13	0.11	-1.20	Subject 63	0.95	0.30	3.12
Industry 6	-0.59	0.17	-3.48	Subject 64	1.41	0.47	2.97
Industry 7	-0.19	0.10	-1.87	Subject 65	0.13	0.16	0.83
Industry 8	-0.23	0.12	-1.91	Subject 66	0.60	0.39	1.52
Industry 9	-0.15	0.07	-2.17	Subject 67	0.30	0.34	0.90
Subject 5	-0.02	0.16	-0.12	Subject 68	-0.64	0.37	-1.71
Subject 6	-0.08	0.50	-0.16	Subject 69	-0.01	0.24	-0.03
Subject 7	0.36	0.29	1.25	Subject 70	-0.51	0.45	-1.15
Subject 8	-0.55	0.38	-1.46	Subject 71	0.55	0.26	2.10
Subject 9	0.42	0.25	1.68	Subject 72	0.44	0.20	2.17
Subject 10	-0.15	0.40	-0.38	Subject 73	0.10	0.24	0.41
Subject 11	1.60	0.33	4.93	Subject 74	0.59	0.45	1.31
Subject 12	0.38	0.20	1.96	Subject 75	0.68	0.14	4.87
Subject 13	0.20	0.32	0.62	Subject 76	0.54	0.15	3.70
Subject 14	-0.54	0.37	-1.45	Subject 77	0.66	0.24	2.77
Subject 16	0.45	0.34	1.33	Subject 78	0.17	0.22	0.76
Subject 17	0.04	0.20	0.19	Subject 79	-0.29	0.36	-0.80
Subject 18	0.46	0.27	1.66	Subject 80	-0.20	0.22	-0.90
Subject 19	0.84	0.43	1.93	Subject 81	-0.22	0.25	-0.87
Subject 20	-0.50	0.33	-1.50	Subject 82	-1.54	0.35	-4.46
Subject 21	0.68	0.26	2.65	Subject 83	0.52	0.17	3.02
Subject 22	-0.47	0.16	-2.94	Subject 84	-0.16	0.17	-0.94
Subject 23	-1.26	0.26	-4.83	Subject 85	0.08	0.77	0.10
Subject 24	0.71	0.55	1.29	Subject 86	0.94	0.20	4.62
Subject 25	-0.24	0.26	-0.90	Subject 87	0.54	0.41	1.31
Subject 26	0.28	0.22	1.28	Subject 88	0.76	0.25	3.06
Subject 27	-0.11	0.16	-0.72	Subject 89	0.60	0.28	2.11
Subject 28	0.29	0.46	0.63	Subject 90	0.59	0.20	3.02
Subject 29	-0.76	0.18	-4.21	Subject 91	0.24	0.32	0.75
Subject 30	-1.03	0.40	-2.54	Subject 92	0.51	0.65	0.79
Subject 31	-1.06	0.24	-4.40	Subject 93	-0.05	0.59	-0.09
Subject 32	-1.55	0.28	-5.49	Subject 94	0.08	0.29	0.28
Subject 33	-0.93	0.24	-3.87	Subject 95	-0.05	0.45	-0.11
Subject 34	-1.68	0.34	-4.91	Subject 96	-0.63	0.52	-1.21
Subject 35	-0.66	0.27	-2.48	Subject 97	-0.84	0.44	-1.91
Subject 36	-0.67	0.68	-0.99	Subject 98	0.13	0.23	0.58
Subject 37	-0.87	0.34	-2.56	Subject 99	-0.23	0.42	-0.54
Subject 38	-0.71	0.26	-2.76	Subject 100	0.01	0.33	0.03
Subject 39	-1.26	0.39	-3.23	Subject 101	1.07	0.35	3.05
Subject 40	-0.24	0.31	-0.78	Subject 102	0.87	0.21	4.07
Subject 41	-0.45	0.19	-2.39	Subject 103	0.34	0.38	0.89
Subject 42	0.74	0.20	3.65	Subject 104	0.54	0.34	1.56
Subject 43	0.59	0.26	2.25	Subject 105	1.29	0.64	2.01
Subject 44	0.61	0.24	2.51	Subject 106	0.55	0.25	2.19
Subject 45	0.05	0.75	0.07				

Note: 5166 observations, 5043 degrees of freedom, robust covariance matrix used, adjusted $R^2 = 0.311$.

Table C.2: Probit estimation of the propensity score without Subject of Degree

	Estimate	Std.error	t-statistic
Const	2.12	0.30	6.96
Age / 10	-0.13	0.15	-0.91
Age squared/1000	-0.20	0.18	-1.11
Full time work	-1.30	0.07	-19.29
Private sector	-0.35	0.05	-6.82
Region 1	0.08	0.11	0.72
Region 2	0.06	0.08	0.77
Region 4	-0.19	0.09	-2.12
Region 5	-0.03	0.09	-0.29
Region 6	-0.08	0.11	-0.73
Region 7	0.01	0.06	0.10
Region 8	-0.07	0.09	-0.84
Region 9	0.00	0.11	-0.01
Region 10	-0.05	0.08	-0.60
Industry 1	-0.50	0.36	-1.39
Industry 2	-0.72	0.14	-5.03
Industry 4	-0.72	0.09	-7.99
Industry 5	-0.06	0.10	-0.59
Industry 6	-1.17	0.14	-8.36
Industry 7	-0.17	0.09	-1.86
Industry 8	-0.46	0.11	-4.12
Industry 9	-0.28	0.06	-4.61

Note: 5166 observations, 5144 degrees of freedom, robust covariance matrix used, adjusted $R^2 = 0.185$.

References

Abadie, A., and G. Imbens (2001): "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," mimeo, Harvard University.

Altonji, J., and R. Blank (1999): "Race and Gender in the Labor Market," in *Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, pp. 3143–3259. North-Holland, New York.

Barnow, B., G. Cain, and A. Goldberger (1981): "Selection on Observables," *Evaluation Studies Review Annual*, 5, 43–59.

Barsky, R., J. Bound, K. Charles, and J. Lupton (2001): "Accounting for the Black-White Wealth Gap: A Nonparametric Approach," NBER 8466.

Blau, F., and L. Kahn (1997): "Swimming Upstream: Trends in the Gender Wage Differential in the 1980s," *Journal of Labor Economics*, 15, 1–42.

Blinder, A. (1973): "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, 8, 436–455.

Brodaty, T., B. Crépon, and D. Fougère (2001): "Using matching estimators to evaluate alternative youth employment programmes: Evidence from France, 1986-1988," in *Econometric Evaluation of Labour Market Policies*, ed. by M. Lechner, and F. Pfeiffer, pp. 85–124. Physica/Springer, Heidelberg.

Dehejia, R., and S. Wahba (1999): "Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programmes," *Journal of American Statistical Association*, 94, 1053–1062.

DiNardo, J., N. Fortin, and T. Lemieux (1996): "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 64, 1001–1044.

Donald, S., D. Green, and H. Paarsch (2000): "Differences in Wage Distributions between Canada and the United States: An Application of a Flexible Estimator of Distribution Functions in the Presence of Covariates," *Review of Economics Studies*, 67, 609–633.

Firpo, S. (2003): "Efficient Semiparametric Estimation of Quantile Treatment Effects," UC Berkeley.

Frölich, M. (2004): "Finite Sample Properties of Propensity-Score Matching and Weighting Estimators," *forthcoming in The Review of Economics and Statistics, February 2004.*

Frölich, M., A. Heshmati, and M. Lechner (2004): "A Microeconometric Evaluation of Rehabilitation of Long-term Sickness in Sweden," *forthcoming in Journal of Applied Econometrics.*

Gardeazabal, J., and A. Ugidos (2001): "Measuring the Gender Gap at Different Quantiles of the Wages Distribution," Universidad del Pais Vasco.

Gerfin, M., and M. Lechner (2002): "Microeconometric Evaluation of the Active Labour Market Policy in Switzerland," *Economic Journal*, 112, 854–893.

Gosling, A., S. Machin, and C. Meghir (2000): "The Changing Distribution of Male Wages in the U.K.," *Review of Economics Studies*, 67, 635–666.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

Heckman, J., H. Ichimura, and P. Todd (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, 64, 605–654.

——— (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

Heckman, J., and R. Robb (1985): "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labour Market Data*, ed. by J. Heckman, and B. Singer. Cambridge University Press, Cambridge.

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

Horowitz, J., and C. Manski (1998): "Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations," *Journal of Econometrics*, 84, 37–58.

Horvitz, D., and D. Thompson (1952): "A Generalization of Sampling without Replacement from a Finite Population," *Journal of American Statistical Association*, 47, 663–685.

Imbens, G. (2000): "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87, 706–710.

——— (2003): "Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review," mimeo, UC Berkeley.

Imbens, G., and T. Lancaster (1996): "Efficient estimation and stratified sampling," *Journal of Econometrics*, 74, 289–318.

Jalan, J., and M. Ravallion (2002): "Estimating the Benefit Incidence of an Antipoverty Program by Propensity Score Matching," *Journal of Business and Economic Statistics*, fortcoming.

Juhn, C., K. Murphy, and B. Pierce (1993): "Wage Inequality and the Rise in Returns to Skill," *Journal of Political Economy*, 101, 410–442.

Larsson, L. (2003): "Evaluation of Swedish Youth Labour Market Programmes," *forthcoming in Journal of Human Resources*.

LECHNER, M. (1999): "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification," *Journal of Business and Economic Statistics*, 17, 74–90.

——— (2001): "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption," in *Econometric Evaluation of Labour Market Policies*, ed. by M. Lechner, and F. Pfeiffer, pp. 43–58. Physica/Springer, Heidelberg.

LECHNER, M., AND R. MIQUEL (2002): "Identification of Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions," *University of St. Gallen Economics Discussion Paper Series*, 2001-07.

LITTLE, R., AND D. RUBIN (1987): *Statistical Analysis with Missing Data*. Wiley, New York.

MACHIN, S., AND P. PUHANI (2003): "Subject of Degree and the Gender Wage Differential: Evidence from the UK and Germany," *Economics Letters*, 79, 393–400.

MANSKI, C., AND S. LERMAN (1977): "The Estimation of Choice Probabilities from Choice-Based Samples," *Econometrica*, 45, 1977–1988.

MORA, R. (2000): "A Nonparametric Decomposition of the Mexican American Average Wage Gap," Universidad Carlos III de Madrid.

NOPO, H. (2002): "Matching as a Tool to Decompose Wage Gaps," Northwestern University, November 2002.

OAXACA, R. (1973): "Male-Female Wage Differences in Urban Labour Markets," *International Economic Review*, 14, 693–709.

PUHANI, P. (1999): *Evaluating Active Labour Market Policies: Empirical Evidence for Poland during Transition*. Physica, Heidelberg.

RACINE, J., AND Q. LI (2000): "Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data," *forthcoming in Journal of Econometrics*.

ROBINS, J., A. ROTNITZKY, AND L. ZHAO (1995): "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of American Statistical Association*, 90, 106–121.

ROSENBAUM, P., AND D. RUBIN (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

——— (1985): "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *The American Statistician*, 39, 33–38.

RUBIN, D., AND N. THOMAS (1992): "Characterizing the Effect of Matching using Linear Propensity Score Methods with Normal Distributions," *Biometrika*, 79, 797–809.

——— (1996): "Matching using Estimated Propensity Scores: Relating Theory to Practice," *Biometrics*, 52, 249–264.

SEIFERT, B., AND T. GASSER (1996): "Finite-Sample Variance of Local Polynomials: Analysis and Solutions," *Journal of American Statistical Association*, 91, 267–275.

——— (2000): "Data Adaptive Ridging in Local Polynomial Regression," *Journal of Computational and Graphical Statistics*, 9, 338–360.

Smith, J., and P. Todd (2000): "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?," mimeo, University of Maryland.

Wooldridge, J. (1999): "Asymptotic Properties of Weighted M-Estimators for Variable Probability Samples," *Econometrica*, 67, 1385–1406.

——— (2002): *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge.