Testing for covariate balance using nonparametric quantile regression and resampling methods

Martin Huber

First draft: Oct 2008, last changes: Feb 2009

Abstract: Consistency of propensity score matching estimators hinges on the propensity score's ability to balance the covariates among treated and non-treated units. Conventional balance tests merely check for differences in covariates' means, but cannot account for differences in higher moments. Specification tests constitute an alternative, but might reject misspecified, but yet balancing propensity score models. This paper proposes balance tests based on (i) nonparametric quantile regression to check for differences in the entire distributions of covariates and (ii) resampling methods to nonparametrically estimate the test statistics' distributions. Simulations suggest that the tests capture imbalance related to higher moments when conventional balance tests fail to do so and correctly accept misspecified, but balancing propensity scores when specification tests reject the null.

Keywords: Balancing property, propensity score matching.
JEL classification: C12, C14, C15, C21

I have benefited from comments by Michael Lechner, Eva Deuchert, and Blaise Melly. Address for correspondence: Martin Huber, SEW, University of St. Gallen, Varnbüelstrasse 14, 9000 St. Gallen, Switzerland, martin.huber@unisg.ch.

1 Introduction

Propensity score matching has become an increasingly popular estimation method in many fields of empirical research. Recent applications include the evaluation of active labor market programmes (Wunsch & Lechner, 2008), the estimation of the health effects of unemployment (Böckerman & Ilmakunnas, 2009), the evaluation of trade gains due to a common currency (Chintrakarn, 2008), and many others. Propensity score matching is attractive because it does not rely on tight functional form assumptions as parametric estimators. Yet, a crucial condition for consistency is the balancing property of the propensity score. I.e., conditional on the propensity score, the covariates' distributions in the pools of treated and non-treated units must be equal.

Balance tests that have been suggested and used in the literature, as the DW test (see Dehejia & Wahba 1999, 2002) or the two-sample t-test for matched samples, merely check for differences in covariates' means. Thus, they cannot account for differences in higher moments and might lack power when imbalance affects distributional features other than the mean. Specification tests as suggested by Shaikh, Simonsen, Vytlacil & Yildiz (2006) constitute an alternative to balance tests. Yet, there might exist propensity score models that balance the covariates despite the fact that they are misspecified. Such models would be unnecessarily rejected by powerful specification tests.

This paper proposes balance tests based on (i) nonparametric quantile estimation, (ii) distribution free Kolmogorov-Smirnov (KS) and Cramer-von-Mises-Smirnov (CMS) test statistics, and (iii) resampling methods to estimate the distribution of the respective KS and CMS statistics required to compute critical values and p-values. Our tests rely on nonparametric quantile regression as implemented by Hayfield & Racine (2008) and resampling methods for test statistics as discussed in Chernozhukov & Fernandez-Val (2005), however, for linear quantile regression processes[1].

The paper contributes to the literature in various dimensions: Firstly, the proposed methods test for differences in the entire quantile functions instead of differences in means alone and

[1] Goh (2007) discusses inference for tests based on quantile regression models consisting of nonparametric additive functions. He provides a generalized likelihood ratio (GLR) statistic that behaves asymptotically under the null like the square of a Kolmogorov-Smirnov statistic. The author shows its test to be asymptotically rate-optimal for nonparametric hypothesis testing.

can be implemented both as full sample and after matching tests. Secondly and similarly to the permutated balance tests investigated by Lee (2006), the test statistics' distributions are not restricted be of any particular parametric form. Thirdly, the paper provides Monte Carlo evidence that the proposed tests capture imbalance related to higher moments when conventional balance tests fail to do so and correctly accept misspecified, but balancing propensity scores when specification tests reject the null. These simulations intuitively highlight the advantages of the methods proposed and the weaknesses of the tests conventionally applied.

The remainder of this paper is organized as follows. Section 1 motivates propensity score matching and more formally discusses the condition to be tested. Section 2 reviews the DW test and introduces full sample balance tests based on resampling and nonparametric quantile regression. Section 3 discusses conventional after matching tests and introduces resampling procedures for matched samples. Section 4 presents simulation results for the proposed KS and CMS resampling tests on full and matched samples as well as for the two sample t-test, the permuted t-test, the DW test of Dehejia & Wahba (1999, 2002), the Ramsey (1969) reset test, and the specification test by Shaikh et al. (2006). Section 5 presents empirical applications of full sample and after matching tests. Section 6 concludes.

2 Propensity score matching and testable conditions

In the treatment evaluation literature, identification strategies based on 'selection on observables' rely on the assumption that all factors jointly affecting the treatment propensity and outcome are observed and thus, can be controlled for. Hence, hypothetical outcomes that would have been realized under alternative treatment states are assumed to be independent of the actual treatment status conditional on the observed covariates. This is known as the conditional independence assumption (CIA), see for instance Lechner (1999) and Imbens (2004) for an in depth discussion. It implies that the effect of the treatment on the outcome of interest is not confounded with other factors. Let Y denote the outcome of interest, D a binary treatment taking either the value 1 (treated) or 0 (nontreated or control)[2], and X a vector of observed covariates with parameter

[2] In contrast, Imbens (2000) and Lechner (2001) discuss effect evaluation in the presence of multiple treatments and Hirano & Imbens (2004) consider continuous treatments and generalized propensity scores. The argumentation in this paper could be easily extended to either framework.

space $\mathcal{X}$. The CIA states that

$$Y^1, Y^0 \perp D | X = x \quad \forall\, x \in \mathcal{X}, \tag{1}$$

where Y^1, Y^0 are the hypothetical outcomes for $D = 0, 1$ and $\perp$ denotes independence.

From a practical perspective, conditioning on a high dimensional X is problematic, as the number of possible combinations of elements in X increases exponentially in the dimension of X such that precise estimation quickly becomes exorbitantly data hungry. In the literature, this problem is referred to as curse of dimensionality. Let $p^*(X) \equiv \Pr(D = 1|X)$ denote the unknown probability of being treated conditional on X, referred to as true propensity score. Rosenbaum & Rubin (1983) have shown that conditioning on the true propensity score is equivalent to conditioning on the covariates directly, as both X and $p^*(X)$ are balancing scores in the sense that they adjust the distribution of covariates in the treatment and in the control group. Thus, if (1) is satisfied, it also holds that the hypothetical outcomes are independent of the treatment conditional on the propensity to be treated:

$$Y^1, Y^0 \perp D | p^*(X). \tag{2}$$

Conditioning on the one dimensional propensity score rather than on the multidimensional vector of covariates circumvents the practical issues related to the curse of dimensionality. For this reason, propensity score matching is a frequently applied in the field of treatment evaluation. If (2) is satisfied, average treatment effects (ATEs) and quantile treatment effects (QTEs) can in principle be consistently estimated, given that there is sufficient common support with respect to $p^*(X)$ among treated and non-treated units. The balancing property of $p^*(X)$ implies that

$$X \perp D | p^*(X). \tag{3}$$

Note that (3) is a mechanical result related to the balancing property and holds even if the CIAs (1) and (2) do not (such that the effect of D on Y is confounded). In reality, the structural form of the true propensity score is unknown to the researcher. In empirical applications it is most commonly modeled parametrically using probit or logit specifications. Let $p(X)$ denote the presupposed specification of the true $p^*(X)$. Whereas the balancing property of $p^*(X)$ follows from the proof in Rosenbaum & Rubin (1983), $p(X)$ might or might not balance X in the pools of treated and controls. However, the balancing property of $p(X)$ is testable by verifying whether

$$F_{X|D=1,p(X)}\left(x|D = 1, p(X)\right) = F_{X|D=0,p(X)}(x|D = 0, p(X)\big) \quad \forall\, x \in \mathcal{X}, \tag{4}$$

where $F_{X|D=d,p(X)}(\cdot|D=d,p(X))$ denotes the conditional cdf of X given $D=d$ and $p(X)$, as this implies that

$$X \perp D | p(X). \tag{5}$$

It is equally valid to check for differences in the conditional quantile functions for $D=1,0$, as the quantile function is simply the inverse of the distribution function. Let Q_A^τ represent the quantile at rank $\tau \in [0,1]$ for some variable A, $Q_A^\tau = \inf\{a : F_A(a) \geq \tau\}$. Then, $F_A(a) = {Q_A^\tau}^{-1}$. For $Q_X^\tau(1,p(x))$ denoting the τth conditional quantile of X given $D=1$ and $p(X)=p(x)$, balancing implies that

$$Q_X^\tau(1,p(X)) = Q_X^\tau(0,p(X)), \quad \forall\, \tau\ \in\ [0,1]. \tag{6}$$

However, coventional balance tests merely capture differences in means by verifying whether

$$E[X|D=1,p(X)] = E[X|D=0,p(X)], \tag{7}$$

which is necessary, but not sufficient for (5). Therefore, these tests do not account for distributional differences related to higher moments and ignore valuable information that might point to the violation of balancing[3]. Furthermore, Lee (2006) provides simulation evidence that conventional balance tests have poor size properties in their original forms where inference is based on asymptotic theory. He suggests to compute p-values using permuted test statistics by randomly shuffling the treatment and control labels 1,0 a large number of times in order to estimate the distribution of the respective test statistic nonparametrically. Even though size properties ameliorate considerably when using permuted versions of the tests, they are as incapable to account for differences in higher moments as the original tests.

Specification tests for the propensity score model constitute an alternative to balance tests. However, the fit of $p(X)$ itself is not an outcome of interest when using propensity score matching. Misspecification is innocuous as long as balancing works and (5) is satisfied, which is sufficient for consistency[4]. Beside the correct, but unknown model, there might exist a misspecified, more parsimonious model that equally satisfies the balancing property and is chosen by the practitioner for the sake of econometric feasibility. This is the case whenever the misspecified model is only a

[3] This point has also been acknowledged by Sekhon (2007*a*) who proposes to include Kolmogorov-Smirnov statistics to test for differences in the covariates densities as one of several balance measures to be optimized.

[4] In contrast, estimators based on inverse probability weighting generally rely on the correctness of the propensity score model such that specification tests are highly relevant for this class of estimators.

monotonic transformation of the true model, such that the order of the individual scores remains unchanged. Indeed, simulation results in Zhao (2008) suggest that ATE estimates based on matching are hardly affected by misspecified, but balancing propensity scores as long as the CIA holds.

For this reason, balance tests appear to be more attractive than specification tests. However, for the reasons discussed it seems more appropriate to use procedures that capture imbalances in the entire distributions than in the means alone. The following sections will propose such test procedures for both full samples and matched samples.

3 Full sample tests

Balance tests can be categorized in methods that check for balance (i) in the full sample (thereafter referred to as full sample tests) or, after having applied the matching algorithm, (ii) in the sample of matched units alone (henceforth after matching tests). An example of the former kind is the DW test which has been applied in Dehejia & Wahba (1999, 2002) and is based on a process originally proposed by Rosenbaum & Rubin (1984) and Rubin (1997). The DW test checks for mean differences of X between strata with the same mean values of $p(X)$ in the pools of treated and control units using two sample t-tests. Lee (2006) describes the DW algorithm used in Dehejia & Wahba (2002)[5] as follows:

1. Start with a parsimonious specification to estimate the score.

2. Split the sample in q equally spaced intervals of the propensity score. For example, using $q = 5$ and dividing observations into strata of equal score range $(0 - 0.2, ..., 0.8 - 1)$. This is usually done over the region of common support.

3. Within each interval, use a t-test to test that the mean $p(X)$ values for treated and comparison units do not differ. If the test fails, spilt the interval in half and test again. The 'optimal' number of intervals is found when the mean $p(X)$ values for treated and comparison units do not differ in all intervals.

4. Within each interval, use a t-test to test that for all covariates, the mean differences treated and comparison units are not significantly different from zero.

5. If covariates are balanced between treated and comparison observations for all intervals, stop. If covariates in any interval are not balanced, modify the logit/probit by using a less parsimonious specification and reevaluate.

[5] This algorithm is available as stata command which was written by Becker & Ichino (2002).

In its original form, the DW test has rather poor size properties, as evidenced by the Monte Carlo results in Lee (2006). The author therefore suggests to estimate the distribution of the test statistic nonparametrically instead assuming it to be t-distributed. This is done by randomly shuffling the treatment and control labels 1,0 in the full sample a large number of rounds in order to compute t-statistics in each round. The permuted distribution of t-values then allows constructing p-values and assessing the significance level of the original t-statistic. Lee (2006) demonstrates that permutation tests have considerably better size properties than the ones relying on parametric distributions of test statistics. However, a second shortcoming is the DW test's incapability to account for differences in higher moments of the covariates. For this reason, we suggest test statistics that are distribution free *and* capture differences in the whole distribution of X.

Let us assume that one would like to test wether the continuously distributed k^{th} covariate in X, denoted as X_k, is balanced conditional on $p(X)$. The null hypothesis is

$$H_0 : Q^{\tau}_{X_k}(1, p(x)) = Q^{\tau}_{X_k}(0, p(x)), \quad \forall\ \tau, p(x) \ \epsilon\ [0, 1]. \tag{8}$$

The proposed testing method is entirely nonparametric. We first estimate the conditional quantiles of X_k by kernel regression, compute Kolmogorov-Smirnov (KS) and the Cramer-Von-Mises-Smirnov (CMS) statistics and finally utilize a resampling methods suggested by Chernozhukov & Fernandez-Val (2005) to compute the p-values of the test statistics.

Let $\hat{p}(X)$ denote the propensity score estimates for specification $p(X)$. In the first step, we estimate $Q^{\tau}_{X_k}(1, \hat{p}(X)), Q^{\tau}_{X_k}(0, \hat{p}(X))$ using local constant kernel regression as discussed in Yu, Lu & Stander (2003) and Li & Racine (2008). For $d = 1, 0$

$$\hat{Q}^{\tau}_{X_k}(d, \hat{p}(x)) = \min_a \sum_{i:D=d}^{n} \rho_\tau(X_{k,i} - a) K\left(\frac{\hat{p}(X_i) - \hat{p}(x)}{h}\right) , \tag{9}$$

where $\rho_\tau(v) = v(\tau - I\{v \leq 0\})$ is the check function suggested by Koenker & Bassett (1978) and $I\{\cdot\}$ denotes the indicator function. As $\hat{p}(X)$ is one dimensional, problems related to the curse of dimensionality as encountered in nonparametric regression on high dimensional covariates do not arise in our framework. The cost in terms of precision when using nonparametric rather than parametric quantile estimation is quite low, while potentially important bias due to misspecification of the quantile process can be avoided. For our Monte Carlo simulations in section 4, we utilize the nonparametric quantile regression procedure in the np package for R, which was written by Hayfield & Racine (2008), and choose the optimal bandwidth h with respect to the data by least squares cross validation.

We would like to infer whether $Q^{\tau}_{X_k}(1, \hat{p}(x)) - Q^{\tau}_{X_k}(0, \hat{p}(x))$, the deviation between the true but unobserved quantiles conditional on $\hat{p}(x)$ for the treated and non-treated, is different from zero. The empirical inference process our hypothesis tests are build upon is the difference of the conditional estimated quantiles,

$$\hat{Q}^{\tau}_{X_k}(1, \hat{p}(x)) - \hat{Q}^{\tau}_{X_k}(0, \hat{p}(x)). \tag{10}$$

The corresponding KS and CMS test statistics of the empirical inference process are defined as

$$\begin{aligned} T_n^{KS} &= \sup_{\tau \in \mathcal{T}, p \in \mathcal{P}} \sqrt{n} ||\hat{Q}^{\tau}_{X_k}(1, \hat{p}(x)) - \hat{Q}^{\tau}_{X_k}(0, \hat{p}(x))||_{\hat{\Lambda}}, \\ T_n^{CMS} &= n \int_{\mathcal{T}} \int_{\mathcal{P}} ||\hat{Q}^{\tau}_{X_k}(1, \hat{p}(x)) - \hat{Q}^{\tau}_{X_k}(0, \hat{p}(x))||^2_{\hat{\Lambda}} d\tau dp, \end{aligned} \tag{11}$$

where $||a||_{\hat{\Lambda}_{\tau}}$ denotes $\sqrt{a'\hat{\Lambda}a}$ and $\hat{\Lambda}$ is a positive weighting matrix satisfying $\hat{\Lambda} = \Lambda + o_p(1)$. Λ is positive definite, continuous and symmetric. $\mathcal{T}, \mathcal{P}$ denote the parameter spaces of τ and $p^*(X)$ and are naturally bounded between 0 and 1. As T_n^{KS}, T_n^{CMS} are distribution-free test statistics, their values are not very telling without knowledge of their asymptotic distribution. Chernozhukov & Fernandez-Val (2005) show for linear quantile regression processes that asymptotically valid critical values for the tests statistics and p-values can be obtained by resampling the recentered test statistics. To this end, we draw J (sub)samples of block size b (the number of observations in each sample) with replacement from the original sample and compute the inference process

$$\hat{Q}^{\tau}_{X_k,b,j}(1, p(x)) - \hat{Q}^{\tau}_{X_k,b,j}(0, p(x)). \tag{12}$$

$\hat{Q}^{\tau}_{X_k,b,j}(d, p(x))$ are the conditional quantile estimates for draw j and block size b, where $1 \leq j \leq J$. The corresponding KS and CMS statistics of the resampled and recentered inference processes are

$$\begin{aligned} T_{n,b,j}^{KS} &= \sup_{\tau \in \mathcal{T}, p \in \mathcal{P}} \sqrt{m} ||\hat{Q}^{\tau}_{X_k,b,j}(1, p(x)) - \hat{Q}^{\tau}_{X_k,b,j}(0, p(x)) - (\hat{Q}^{\tau}_{X_k}(1, \hat{p}(x)) - \hat{Q}^{\tau}_{X_k}(0, \hat{p}(x)))||_{\hat{\Lambda}}, \\ T_{n,b,j}^{CMS} &= m \int_{\mathcal{T}} \int_{\mathcal{P}} ||\hat{Q}^{\tau}_{X_k b,j}(1, p(x)) - \hat{Q}^{\tau}_{X_k b,j}(0, p(x)) - (\hat{Q}^{\tau}_{X_k}(1, \hat{p}(x)) - \hat{Q}^{\tau}_{X_k}(0, \hat{p}(x)))||^2_{\hat{\Lambda}} d\tau dp. \end{aligned} \tag{13}$$

Finally, we compute the distribution free p-values by $1/J \sum_{j=1}^{J} I\{T_{n,b,j} \geq T_n\}$ which is a consistent estimator of $\Pr[T(\hat{Q}^{\tau}_{X_k}(1, \hat{p}(x)) - \hat{Q}^{\tau}_{X_k}(0, \hat{p}(x)) - (Q^{\tau}_{X_k}(1, \hat{p}(x)) - Q^{\tau}_{X_k}(0, \hat{p}(x)))) \geq T_n]$.

Instead of quantiles, one could equivalently use estimates of the conditional densities $f_{X_k}(x|1, p(X))$ or distributions $F_{X_k}(x|1, p(X))$ in the resampling procedure. In fact, the latter methods seem to be logical alternatives to the quantile approach if X_k is discrete. Then, $f_{X_k}(x|d, p(X))$ for $d = 1, 0$ could be evaluated at each mass point of X_k whereas quantile based estimation might suffer a lack of power in particular when the number of mass points is small.

We conclude this section by comparing the advantages of our test statistics vis--vis of specification tests for the propensity score model. It has been argued in the last section that only the balancing properties are decisive for the consistency of treatment effect estimation, as the fit of $p(X)$ itself is not an outcome of interest. Thus, relying on specification tests rather than balance tests might be overly restrictive. A more parsimonious specification might be similarly appropriate with respect to balancing, see Zhao (2008). Yet, it would most likely be rejected by a (powerful) specification test.

A second argument in favor of balance tests is their ease of implementation. Our full sample tests rely on nonparametric regression based on a single regressor, namely the propensity score estimate. Neither the curse of dimensionality, nor the problem of choosing an appropriate parametric regression model are an issue. In contrast, when using parametric specification tests as, for instance, the reset test proposed by Ramsey (1969), one can theoretically choose from an infinite number of alternative specifications to be tested against each other. It is a priori not clear which models should be compared at all. However, parametric specification tests are only powerful if the alternative model is correct, or at least more correct than the model assumed to be true under the null.

Shaikh et al. (2006)provide an alternative method to parametric specification tests. They show that if the propensity score is correctly specified, it holds that

$$f_{p(X)|D=1}(p(x)|D = 1) = \frac{\Pr(D = 1)}{\Pr(D = 0)} \frac{p(x)}{1 - p(x)} f_{p(X)|D=0}(p(x)|D = 0) \quad \forall\ p(x) \in [0, 1], \tag{14}$$

where $f_{p(X)|D=d}(\cdot|D = d)$ denotes the pdf of $p(X)$ conditional on $D = d$. Even though they do not provide a formal testing and inference method in their paper, an asymptotically valid method can easily be obtained by taking the full sample analogues of the densities and probabilities in condition (14) and using the similar resampling methods as for the full sample balance tests. This approach avoids the issue of choosing an alternative model to be tested against the initial propensity score specification. Yet, the first argument in favor of balance tests remains. In section 4, we will present Monte Carlo results for a case where specification tests unnecessarily reject a misspecified, but balancing score, whereas our balance statistics do not.

4 After matching tests

Full sample tests as discussed in the last section check wether the balancing property can be rejected for the population the full sample is randomly drawn from. In contrast, after matching tests use merely the matched sample to test the balance hypothesis in the pools of treated and control units. Unlike full sample tests, they do not condition on the propensity score, as this task is performed by the matching algorithm prior to testing.Lee (2006) argues that matched samples might differ considerably depending on the matching algorithm applied which leads him to conclude that the DW test is only appropriate when the estimation method is stratification on the propensity score which is in fact the estimation method corresponding to the DW test.

Yet, we argue that nonparametric versions of full sample balance tests are useful as check the propensity score's *balancing property* prior to the application of the matching algorithm, whereas after matching tests merely check the *realized balance* in the subsample used for estimation. A rejection by the full sample test indicates that the propensity score fails to balance the distributions of covariates which will most likely lead to imbalance in the matched sample. In contrast, a lack of balance attested by the after matching tests might either be due to the score's failure to balance, or to a lack of common support in the propensity scores induced by the matching algorithm, i.e. the occurrence of bad matches, or both. Thus, after matching tests are sensitive to the common support restrictions imposed in the estimation procedure, while full sample tests are not. The researcher is ultimately interested in the balance of the matched sample, but might apply both tests to trace the reasons for the occurrence of imbalance in order to take the right actions. A rejection by the full sample test suggests the modification of the propensity score specification, whereas a rejection by the after matching test, but not by the full sample test implies a reconsideration of the matching algorithm. Also Lee (2006) acknowledges that matching itself can make balance worse.

We will now review several after matching tests conventionally found in the literature and introduce after matching versions of the KS and CMS resampling procedures. This requires some additional notation. Let $n^m = n^{m1} + n^{m0}$ denote the sample sizes of matches, treated matches, and nontreated matches, respectively. $\bar{X}^{m1}$, $\bar{X}^{m0}$ are the respective sample means $\frac{1}{n^m}\sum_{i=1}^{n^m} D_i X_i$ and $\frac{1}{n^m}\sum_{i=1}^{n^m}(1 - D_i)X_i$. Conventional tests as the hotelling test or the t-test check whether differences in $\bar{X}^{m1}$ and $\bar{X}^{m0}$ or their respective k^{th}elements, denoted as $\bar{X}_k^{m1}$ and $\bar{X}_k^{m0}$, are statistically significant. The two sample t-test for a continuous element X_k is defined as

$$T_m^t = \frac{\bar{X}_k^{m1} - \bar{X}_k^{m0}}{\sqrt{\frac{\hat{V}_k^{m1}}{\frac{n^{m1}}{n^m}} - \frac{\hat{V}_k^{m0}}{1-\frac{n^{m1}}{n^m}}}},$$

where $\hat{V}^{m1}$, $\hat{V}^{m0}$ are the empirical variances $\sum_{i=1}^{n^m} D_i(X_i - \bar{X}^{m1})^2/(n^{m1} - 1)$ and $\sum_{i=0}^{n^m}(1 - D_i)(X_i - \bar{X}^{m1})^2/(n^{m0} - 1)$ and $\hat{V}_k^{m1}$, $\hat{V}_k^{m0}$ denote their k^{th} elements. Conceptually similar to the t-test, the hotelling test checks for joint equality of means in all elements of X. The test of standardized differences constitutes a third alternative and is defined as

$$T_m^{SD} = 100 \cdot \frac{\bar{X}_k^{m1} - \bar{X}_k^{m0}}{\sqrt{\frac{\hat{V}_k^{m1} - \hat{V}_k^{m0}}{2}}}$$

for continuous X_k and

$$T_m^{SD} = 100 \cdot \frac{\Pr(\bar{X}_k^{m1} = x) - \Pr(\bar{X}_k^{m0} = x)}{\sqrt{\frac{\Pr(\bar{X}_k^{m1}=x)(1-\Pr(\bar{X}_k^{m1}=x)) - \Pr(\bar{X}_k^{m0}=x)(1-\Pr(\bar{X}_k^{m0}=x))}{2}}}$$

for discrete X_k, see Baser (2006). According to Rosenbaum & Rubin (1985), a standardized difference larger than 20 has to be considered as "large".

Imai, King & Stuart (2006) argue that standard methods as the t-test have poor properties, as the difference in sample means as measure of balance can be distorted when randomly dropping observations, even if the sample balance does not change. As for the DW test, Lee (2006) suggests to use permutation tests rather than conventional after matching tests to ameliorate the finite sample properties of the test statistics.

Again, we argue that one should not restrict testing to differences in the first moments alone. It is straightforward to implement the KS and CMS resampling procedures as after matching tests to check for imbalances related to the entire distributions of matched treated and control units. We need not condition on the propensity score any more as this has already be done by means of a (hopefully adequate) matching algorithm prior to testing. Let $\hat{Q}_{X_k^m}^{\tau}(d)$ denote the estimate of the τth quantile for all $X_{k,i}$ in the sample of matched units conditional on $D = d$.

The corresponding KS and CMS statistics are

$$\begin{aligned}
T_m^{KS} &= \sup_{x\in\mathcal{X}^m} \sqrt{n^m}||\hat{Q}^\tau_{X_k^m}(1) - \hat{Q}^\tau_{X_k^m}(0)||_{\hat{\Lambda}}, \\
T_m^{CMS} &= n^m \int_{\mathcal{X}^m} ||\hat{Q}^\tau_{X_k^m}(1) - \hat{Q}^\tau_{X_k^m}(0)||^2_{\hat{\Lambda}} dx, \\
T_{m,b,j}^{KS} &= \sup_{x\in\mathcal{X}^m} \sqrt{n^m}||\hat{Q}^\tau_{X_k^m,b,j}(1) - \hat{Q}^\tau_{X_m^k b,j}(0) - (\hat{Q}^\tau_{X_m^k}(1) - \hat{Q}^\tau_{X_m^k}(0))||_{\hat{\Lambda}}, \\
T_{m,b,j}^{CMS} &= n_m \int_{\mathcal{X}_m} ||\hat{Q}^\tau_{X_m^k,b,j}(1) - \hat{Q}^\tau_{X_m^k b,j}(0) - (\hat{Q}^\tau_{X_m^k}(1) - \hat{Q}^\tau_{X_m^k}(0))||^2_{\hat{\Lambda}} dx,
\end{aligned}$$

and p-values are obtained by $1/J\sum_{j=1}^J I\{T_{m,b,j} \geq T_m\}$. Instead of using resampling as suggested by Chernozhukov & Fernandez-Val (2005), computation of $T_{n,b,j}$ can also be based on permutation as advocated by Lee (2006). In this case, treated and control labels are randomly shuffled without replacement to estimate the distribution of the test statistics. Both resampling and permutation are consistent and Monte Carlo results in section 4 suggest that either method works fairly well to assess balance even when sample sizes are rather small. Again, using distributions or densities instead of quantiles is equally valid and most likely more appropriate when considering discrete covariates. Note that the proposed after matching procedures might also be used as *before matching* tests for unconditional balance in the full sample (without conditioning on the propensity score). Differences in before and after matching statistics and p-values indicate the balance gains due to propensity score matching.

5 Monte Carlo results

In this section, we present Monte Carlo evidence on the finite sample properties of KS and CMS test procedures for full sample and after matching balance and run a horse race with several other tests proposed in the literature. Starting with the full sample statistics, we compare the performance of our procedures to the Ramsey (1969) reset test, the specification test suggested by Shaikh et al. (2006), and the DW test[6] (see Dehejia & Wahba 1999, 2002). Of particular interest are test results for scenarios where the propensity score is misspecified, but yet balancing.

[6] We test for equality in mean propensity scores among treated and non-treated units within a stratum at the 10% level of significance.

Therefore, the following data generating process (DGP) is considered:

$$\begin{aligned}
D_i &= I\{\beta_0 + \beta_1 X_{1,i}^3 + \beta_2 X_{2,i} + \varepsilon > 0\}, \\
Y_i &= \gamma_1 X_{1,i}^2 + \gamma_2 X_{2,i} + \gamma_3 D_i + U_i \\
X_1, X_2 &\sim \text{unif}(0,3), \quad \varepsilon \sim N(0,5), \quad U \sim N(0,1) \\
\beta_0 &= -3, \quad \beta_1 = 0.3, \quad \beta_2 = 0.5, \quad \gamma_1 = \gamma_2 = \gamma_2 = 1.
\end{aligned}$$

Treatment effects are homogenous and the ATE is equal to 1. In the scenario considered, the propensity score model is incorrectly specified as

$$p(X) = \Pr(D = 1|X) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2),$$

such that β_1 is estimated with respect to X_1 instead of X_1^3. Thus, it is assumed that the probability to be treated increases linearly in X_1, whereas the true relationship is exponential. Yet, the incorrect model satisfies the balancing property for variable X_1, as the order of the propensity scores is preserved under misspecification. Even though the propensity scores themselves are poorly estimated, the treated units are compared to non-treated units with similar $p^*(X)$ when using propensity score matching. The reason is that the incorrect specification is a monotonous transformation of the true model. To gain some intuition, figure 5.1 displays 1000 simulated values of X_1 along with propensity score estimates (i) using the misspecified probit model (black bubbles) and (ii) based on the correct specification $p^*(X)$ (red bubbles). As the average rank of each observation remains the same in either case such that observations with similar $p^*(X)$ are matched even when using the wrong specification, estimation is consistent[7].

[7] It is, however, less efficient than estimation based on the true propensity score model.

Figure 5.1

Misspecified and balancing scenario ($n = 1000$)

Propensity score estimates under misspecification (black bubbles) and correct specification (red bubbles)

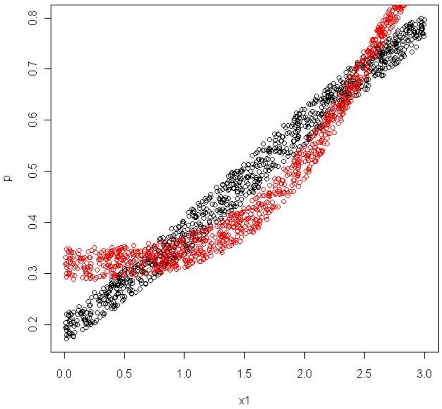


Table (5.1) displays mean p-values and rejection frequencies of the null hypothesis at the 5% and 10% level of significance, i.e. the share of p-values that lie below 0.05 and 0.10, respectively, for 1000 Monte Carlo replications and propensity score estimates obtained from the misspecified probit model. Inference for the KS and CMS balance tests is based on 199 bootstrap draws. For computational ease, the conditional quantiles are evaluated at $\tau \in \mathcal{T}_{[0.25,0.75]} = \{0.25, 0.50, 0.75\}$ and $p(x) \in \mathcal{P}_{[0.25,0.75]} = \{0.25, 0.35, 0.45, 0.55, 0.65, 0.75\}$. More evaluation points in τ and $p(x)$ would most likely ameliorate finite sample properties of the test procedures, but also increase computational time. As already discussed in section 2, Shaikh et al. (2006) showed that condition (14) holds if the propensity score is correctly specified. As for the quantile based procedures, KS and CMS test statistics and p-values can be obtained by resampling methods and again, 199 bootstrap samples are drawn. $p(x)$ is evaluated at the predicted propensity score values for the simulated sample observations. We also investigate the performance of the Ramsey (1969) reset specification test, where the incorrect model is tested against the alternative

$$p(X) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \beta_4 X_2 + \beta_5 X_2^2 + \beta_6 X_2^3).$$

Mean p-values for our full sample balance tests (see ‘KS balance’ and ‘CMS balance’) are high for $n = 1000$ observations, whereas rejection frequencies at the 5 and 10% levels are rather low.

Thus, our procedures correctly accept the balancing property of the misspecified $p(X)$ in most cases. Furthermore, the rejection frequencies approach their theoretical values as the sample size is increased to $n = 4000$. The standard DW test correctly keeps the null most of the time, even though its performance deteriorates in the sample size. Following Lee (2006), whose simulations suggest that the DW test has very poor size properties and rejects the null much too often, we also consider the DW test with an approximation of the Bonferroni adjustment, see 'DW B.a.'. Testing for balance with respect to X_1 alone, the Bonferroni adjustment implies that the significance level (i.e., 5 or 10%) is divided by the number of intervals such that the chance of rejection for each t-test in a particular interval is adjusted downwards to keep the overall probability of incorrect rejection constant as the number of intervals increases. This considerably ameliorates the size properties for $n = 4000$. In contrast, CMS and KS specification tests based on condition (14) reject the misspecified, but balancing model most of the time, see 'CMS spec' and 'KS spec'. The reset test inappropriately rejects the null for any sample size and significance level.

Table 5.1

p-values of balancing and specification tests

1000 Monte Carlo replications, 199 bootstrap draws per replication

			n=1000				
statistic	CMS balance	KS balance	DW	DW B.a.	CMS spec	KS spec	Ramsey reset
mean p-val	0.312	0.393	0.160*	0.160*	0.088	0.183	0.019
share (p-val $\leq$ 0.05)	0.094	0.031	0.074	0.021	0.614	0.302	0.915
share (p-val $\leq$ 0.1)	0.188	0.094	0.082	0.030	0.754	0.470	0.951
			n=4000				
statistic	CMS balance	KS balance	DW	DW B.a.	CMS spec	KS spec	Ramsey reset
mean p-val			0.106*	0.106*	0.000	0.035	0.000
share (p-val $\leq$ 0.05)			0.265	0.047	1.000	0.781	1.000
share (p-val $\leq$ 0.1)			0.301	0.063	1.000	0.934	1.000

* mean minimum p-value of all intervals

Considering the same Monte Carlo set up, we investigate the consistency of 1-nearest neighbor caliper matching on the propensity score. We therefore use the Match command by Sekhon (2007*b*) and set the caliper to 0.25 standard deviations of the propensity score. The ATE estimate $\hat{\Delta} = 1.036$ for $n = 1000$ and the corresponding mean squared error (MSE) is equal to 0.009. For $n = 4000$, $\hat{\Delta} = 1.032$ and MSE= 0.003 such that the estimand gets closer to the true value $\Delta = 1$ as the sample size increases, despite the fact that the propensity score is misspecified. This is, however, not true when using estimators based on inverse probability weighting (IPW), as consistency for this class of estimators relies on the correctness of the propensity score specification. Indeed, using the IPW estimator discussed in Hirano, Imbens & Ridder (2003), the ATE estimate is severely biased ($\hat{\Delta} = 1.293, 1.295$ for 1000 and 4000 observations, respectively) and the MSE is large $(0.096, 0.090)$. Therefore, matching seems to be more robust to propensity score misspecification than IPW.

Secondly, we consider a DGP for which our probit specification is misspecified and non-balancing:

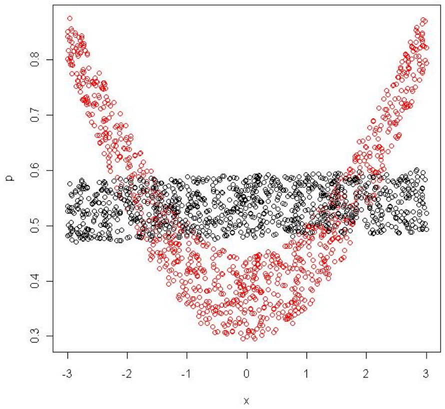
$$\begin{aligned} D_i &= I\{\beta_0 + \beta_1 X_{1,i}^2 + \beta_2 X_{2,i} + \varepsilon > 0\}, \\ Y_i &= \gamma_1 X_{1,i}^2 + \gamma_2 X_{2,i} + \gamma_3 D_i + U_i \\ X_1, X_2 &\sim \text{unif}(-3,3), \quad \varepsilon \sim N(0,5), \quad U \sim N(0,1) \\ \beta_0 &= -3, \quad \beta_1 = 1, \quad \beta_2 = 0.5, \quad \gamma_1 = \gamma_2 = \gamma_3 = 1. \end{aligned}$$

To elucidate the issues of misspecification *and* imbalance, Figure 5.2 displays 1000 simulated realizations of X_1 along with propensity score estimates under misspecification (black bubbles) and under the correct specification (red bubbles).

Figure 5.2

Misspecified and non-balancing scenario ($n = 1000$)

Propensity score estimates under misspecification (black bubbles) and correct specification (red bubbles)



As mentioned before, only treated and non-treated with the same or similar $p^*(X)$ should be compared to each other. It is, however, obvious that matching based on the propensity score estimates under misspsecification fails to do so. The reason is that the specification $p(X)$ cannot handle the V-shaped non-monotonicity in the relation between X_1 and the true propensity score. $p^*(X)$ is zero at the mean of X_1, which is zero, and increases exponentially in either direction. Due to this symmetric relationship, the expected value of the probit estimate for the slope coefficient β_1 is zero, too. Therefore, the expected values of the propensity score estimates are mistakenly independent of X_1. More formally, $E(\hat{\beta}_1) = 0$ implies that $E(X_1|D = d, p(X)) = E(X_1|D = d)$, where ' ˆ ' denotes a parameter estimate, $p(X) = E(\hat{p}(X))$, and d is either one or zero. Therefore, matching is random with respect to the propensity score such that observations with fairly different X_1 are incorrectly compared to each other. Table (5.2) displays the tests' results under the misspecified, non-balancing scenario. The CMS and KS balance tests correctly reject the null most of the times. Imbalance is due to the fact that observations with high absolute values in X_1 are more likely to be treated than those with values close to zero. This is acknowledged by the nonparametric balance tests and table (5.2) shows that their power increases in n.

In contrast, any balance test based on mean differences fails to detect imbalance. Note that the DGP considered, the expected value of X_1 is zero for treated and non-treated. Hence, $E(X|D = d, p(X)) = E(X|D = d)$ and $E(X|D = 1) = E(X|D = 0) = 0$ together imply that conventional balance tests have no power to reject the null. This explains the poor performance of the DW tests (with and without Bonferroni adjustment). Interestingly, the specification tests do no better in the scenario considered. For the reset test, a zero coefficient on X_1 is on average as likely as non-zero coefficients on X_1 and higher order terms and therefore, it has no power to reject the incorrect specification. Also the nonparametric test of Shaikh et al. (2006) performs poorly. Therefore, the only tests that have power in this particular scenario are the CMS and KS balance tests based on quantile regression.

Table 5.2

p-values of balancing and specification tests

1000 Monte Carlo replications, 199 bootstrap draws per replication

n=1000							
statistic	CMS balance	KS balance	DW	DW B.a.	CMS spec	KS spec	Ramsey reset
mean p-val	0.002	0.039	0.182*	0.182*	0.452	0.448	0.418
share (p-val $\leq$ 0.05)	0.984	0.837	0.062	0.009	0.078	0.067	0.138
share (p-val $\leq$ 0.1)	0.997	0.879	0.070	0.013	0.140	0.134	0.202
n=4000							
statistic	CMS balance	KS balance	DW	DW B.a.	CMS spec	KS spec	Ramsey reset
mean p-val	0.000	0.000	0.157*	0.157*	0.442	0.432	0.462
share (p-val $\leq$ 0.05)	1.000	1.000	0.173	0.033	0.097	0.088	0.085
share (p-val $\leq$ 0.1)	1.000	1.000	0.184	0.046	0.164	0.155	0.149

* mean minimum p-value of all intervals

Again, 1-nearest neighbor propensity score matching is used to investigate the impact of imbalance on the accuracy of the estimated effect. For $n = 1000$, the ATE estimate is severely biased ($\hat{\Delta} = 3.025$) and the MSE (4.142) is huge. Things get even worse for $n = 4000$ as $\hat{\Delta} = 3.066$ and MSE= 4.277, which points to inconsistency due to the lacking balancing property[8].

In the remainder of this section, we present simulation results for after matching tests. We compare our KS and CMS tests based on resampling and permutation to permuted and classical (i.e., relying on the asymptotic theory) two sample t-tests. Two random samples of the variables X_1, X_0, each with sample size $n_1 = n_0 = 250$, and examine the tests' performance for various distributions of X_1, X_0. In the resampling and permutation procedures, differences in the quantiles between the two samples are compared to each other. p-values are computed by randomly resampling with replacement in the former and randomly shuffling the labels 1, 0 without replacement[9] in the latter case. The number of bootstrap draws or permutations, respectively, is set to $J = 600$ and the block size of observation in each draw is $b = 500$. The quantile functions are compared on a grid between ranks 0.1 and 0.9 with step size equal to 0.01, i.e., $\tau \in \mathcal{T}_{[0.1,0.9]} = \{0.10, 0.11, ..., 0.89, 0.90\}$. We examine the tests appropriateness considering 4 different scenarios and compute 1000 Monte Carlo replications. In the first case, X_1 and X_0 are standard normal, $X_1, X_0 \sim N(0, 1)$.

[8] The IPW estimator yields $\hat{\Delta} = 3.094, 3.097$ and MSE= $4.414, 4.404$ for $n = 1000, 4000$, respectively.

[9] For permutation with replacement, see Abadie (2002).

Table 5.3 displays the results for the standard normal case. CMS-r, KS-r, denote the KS and CMS statistics obtained by resampling, CMS-p and KS-p stand for the respective permutation statistics. As both variables are drawn from the same distribution, the null hypothesis of equality in distributions (and means) is satisfied. One would therefore expect the p-values to be larger than any conventional level of significance in most Monte Carlo replications. Indeed, both the t-test and the nonparametric alternatives do a good job in evaluating the balance. Average p-values are larger than 0.5 for any test statistic in table 4.1. KS and CMS procedures as well as the standard t-test have empirical rejection frequencies that are lower than the corresponding level of significance, which indicates that tests are conservative, at least for the sample sizes considered. The permuted t-test comes closest to the theoretical rejection frequencies, which confirms its superior small sample properties as evidenced in Lee (2006).

Table 5.3

p-values of balancing tests for $X_1, X_0 \sim N(0,1)$

1000 Monte Carlo replications, 599 bootstrap draws per replication

statistic	CMS-r	KS-r	CMS-p	KS-p	perm.t	t-test
mean p-val	0.534	0.616	0.507	0.507	0.503	0.505
share (p-val $\leq$ 0.05)	0.043	0.018	0.037	0.040	0.048	0.037
share (p-val $\leq$ 0.1)	0.087	0.057	0.086	0.093	0.101	0.082

In the second scenario, the observations are sampled from two identical uniform distributions, $X_1, X_0 \sim U(-1,1)$. Again, all tests yield high average p-values. Even though some statistics are more conservative than others, rejection frequencies are generally speaking fairly close to the theoretical 5% and 10%.

Table 5.4

p-values of balancing tests for $X_1 \sim U(-1,1)$, $X_0 \sim U(-1,1)$

1000 Monte Carlo replications, 599 bootstrap draws per replication

statistic	CMS-r	KS-r	CMS-p	KS-p	perm.t	t-test
mean p-val	0.538	0.613	0.511	0.510	0.510	0.515
share (p-val $\leq$ 0.05)	0.053	0.040	0.048	0.057	0.049	0.046
share (p-val $\leq$ 0.1)	0.101	0.072	0.100	0.110	0.093	0.095

Monte Carlo evidence obtained so far suggests that under the null, tests based on differences in means are as good as procedures based on differences in quantiles. Size properties of any procedure turned out to be quite satisfactory for the distributions considered even in small samples. We will now consider scenarios in which the two samples are drawn from different distributions such that the null is violated. We start out by assuming that X_0 is standard normal, whereas X_1 follows a mixed normal distribution and is defined as

$$\begin{aligned} \nu &= \varepsilon > 0, \quad \varepsilon \sim N(0,1) \\ X_1 &= \nu \cdot \eta_1 + (1-\nu) \cdot \eta_2, \quad \eta_1 \sim N(-3,1), \quad \eta_2 \sim N(3,1) \end{aligned}$$

The distribution of X_1 is bimodal and clearly different from X_0. However, as the means of X_1 and X_0 are both zero, both standard and permuted t-tests fail to detect the distributional imbalance. Results in table 5.5 show that their average p-values remain at 0.5. In contrast, nonparametric procedures yield highly significant mean p-values and reject the null at any conventional level of significance.

Table 5.5

p-values of balancing tests for a bimodal X_1 and $X_0 \sim N(0,1)$

1000 Monte Carlo replications, 599 bootstrap draws per replication

statistic	CMS-r	KS-r	CMS-p	KS-p	perm.t	t-test
mean p-val	0.000	0.000	0.000	0.000	0.500	0.500
share (p-val $\leq$ 0.05)	1.000	1.000	1.000	1.000	0.043	0.047
share (p-val $\leq$ 0.1)	1.000	1.000	1.000	1.000	0.081	0.099

In the fourth scenario, $X_1 \sim U(-1,1)$ and $X_0 \sim U(-1.3,1.3)$. Both variables have a mean equal to zero, but the support of X_0 is somewhat larger. Again, t-tests are not capable to capture this distributional difference. In contrast, the nonparametric procedures are quite powerful and reject equality in distributions in most replications at the 5% and the 10% level. KS statistics seem to have slightly more power than CMS statistics. This is not surprising, as the former are based on maximum difference in quantiles which are theoretically largest at the boundaries of the distributions. The latter are based on integrating differences in quantiles over the whole support. As differences get smaller in the interior and are theoretically zero at the mean, CMS statistics are somewhat less powerful.

Table 5.6

p-values of balancing tests for $X_1 \sim U(-1,1)$, $X_0 \sim U(-1.3,1.3)$

1000 Monte Carlo replications, 599 bootstrap draws per replication

statistic	CMS-r	KS-r	CMS-p	KS-p	perm.t	t-test
mean p-val	0.034	0.013	0.025	0.002	0.510	0.510
share (p-val $\leq$ 0.05)	0.800	0.942	0.862	0.997	0.045	0.047
share (p-val $\leq$ 0.1)	0.925	0.976	0.954	1.00	0.095	0.090

In the next set up, $X_1 \sim N(0.2,1)$ and $X_0 \sim N(0,1)$ such that first moments differ whereas higher moments are the same in both distributions. Table 5.7 reveals that t-tests are more powerful than CMS and KS statistics in the scenario considered. The permuted t-test is most accurate as it yields the lowest mean p-value and the highest rejection rates.

Table 5.7

p-values of balancing tests for $X_1 \sim N(0.2,1)$, $X_0 \sim N(0,1)$

1000 Monte Carlo replications, 599 bootstrap draws per replication

statistic	CMS-r	KS-r	CMS-p	KS-p	perm.t	t-test
mean p-val	0.137	0.208	0.131	0.147	0.054	0.102
share (p-val $\leq$ 0.05)	0.562	0.358	0.559	0.478	0.711	0.601
share (p-val $\leq$ 0.1)	0.669	0.495	0.671	0.627	0.837	0.710

In the final scenario, $X_1 \sim N(0.2,1.3)$ and $X_0 \sim N(0,1)$. The only change to the previous set up is the increase in the variance of X_1 such that the first two moments differ. A comparison of results in tables 5.8 and 5.7 shows that the t-tests do a worse job in rejecting the incorrect null than before. This is due to a loss in precision related to an increased variance in X_1. In contrast, p-values of KS and CMS statistics have decreased considerably, as they capture differences in the entire distribution. Contrarily to the t-tests, the increase in variance comes along with an increase in the statistical power of the KS and CMS statistics, which now outperform the permuted t-ttest with respect to p-values and rejection rates.

Table 5.8

p-values of balancing tests for $X_1 \sim N(0.2,1.3)$, $X_0 \sim N(0,1)$

1000 Monte Carlo replications, 599 bootstrap draws per replication

statistic	CMS-r	KS-r	CMS-p	KS-p	perm.t	t-test
mean p-val	0.045	0.050	0.028	0.017	0.082	0.154
share (p-val $\leq$ 0.05)	0.758	0.751	0.855	0.922	0.607	0.485
share (p-val $\leq$ 0.1)	0.869	0.853	0.933	0.963	0.736	0.600

The simulations results provide us with four insights concerning the properties of the various test statistics. Firstly, permuted t-tests and nonparametric CMS and KS statistics based on quantiles are preferable to conventional t-tests when checking for balance, as at least either of the former is more accurate than the latter in *any* scenario. Secondly, when first moments are equal but higher moments differ, KS and CMS statistics capture the distributional inequality whereas permuted t-tests do not. This shortcoming is inherent to any method merely checking for differences in means. Thirdly, there is a single scenario of distributional difference in which the permuted t-test is strictly more accurate than quantile methods, namely when only first moments differ and higher moments do not. In this particular set up, permuted t-tests are more powerful than KS and CMS statistics, as testing for differences in means is more precise than testing for differences over the whole quantile function. Fourthly, as the extent, to which higher moments are affected by distributional differences, increases, the relative power of the KS and CMS statistics over permuted t-tests increases, too. In practice, we suggest to examine KS, CMS, *and* permuted t-statistics and to be alarmed if either of the three is very low.

6 Empirical application

In this section, we apply full sample and after matching tests to labor market data previously analyzed by Ichino, Mealli & Nannicini (2006) and LaLonde (1986). Ichino et al. (2006) apply propensity score matching to assess the effects of job placements provided by temporary work agencies (TWAs) on the probability to find permanent employment in the two Italian regions Sicily and Tuscany. The data where collected through phone interviews in the beginning of 2001 and the end of 2002. The treatment period (having or not having received a temporary job by a TWA assignment) is the first semester of 2001, whereas the outcome (permanent employment) was measured in November 2002. Pretreatment control variables X include detailed information on demographic characteristics, educational attainments, family background and recent employment history of treated and non-treated units. While Ichino et al. (2006) are interested in the robustness of effect estimates with respect to omitted unobserved factors that would violate the CIA, we use their data to investigate the balancing property of their propensity score specification, which is based on a probit model.

We restrict our attention to the sample drawn in Tuscany, which consists of 281 treated and 628 non-treated individuals. We test wether balance conditional on the predicted propensity scores

$\hat{p}(X_i)$ holds for the variable 'distance from home to the nearest TWA', which is significantly negatively related to the treatment probability. Testing is restricted to the region of common support. Therefore, observations in either treatment group with $\hat{p}(X_i)$ higher than the maximum and lower than the minimum in the other treatment group are discarded from the sample. This leaves us with 255 treated and 519 non-treated individuals.

We test for distributional equality in the distance to the nearest TW at $\tau \in \mathcal{T}_{[0.25,0.75]} = \{0.25, 0.50, 0.75\}$ and $p(x) \in \mathcal{P}_{[0.25,0.75]} = \{0.25, 0.5, 0.75\}$ using 199 bootstrap replications. The results of the full sample CMS and KS tests are presented in table 6.1. Either method rejects the balancing property at any conventional level of significance. This is at odds with Ichino et al. (2006) who use the DW test and conclude that 'distance from home to the nearest TWA' is balanced conditional on the propensity score. Note, however, that the significance level chosen by Ichino et al. (2006) is 0.1 %. Using the DW test algorithm for stata provided by Becker & Ichino (2002) and setting the significance level to 5% would reject the null, but one has to keep in mind that this result comes without Bonferroni adjustment.

Table 6.1

empirical application of full sample tests

'distance from home to the nearest TWA', $n = 774$

statistic	CMS	KS	DW
p-val	0.000	0.028	0.025*

* minimum p-value of all intervals

Secondly, we apply after matching tests to a subsample of experimental data from the National Supported Work Program (NSW) originally analyzed by LaLonde (1986). LaLonde assesses the effectiveness of temporary employment programs to which applicants were randomly assigned in the mid-70s by comparing the earnings of treated and non-treated in 1978. The subsample consists of 185 treated and 260 non-treated individuals and was also investigated by Dehejia & Wahba (1999) and Firpo (2007), among others.

In a first step, we apply CMS and KS tests based on resampling and permutation as well as standard and permuted t-tests to the variable 'age'. When using the former tests, the unconditional quantiles are evaluated at $\tau \in \mathcal{T}_{[0.2,0.8]} = \{0.2, 0.3, ..., 0.8\}$. 999 bootstrap samples are drawn. Due to the randomized set up, one would expect 'age' to be well balanced among treated and non-treated. As the top panel of table 6.2 shows, this seems to be case, as no test rejects the null. In a second step, we discard all non-treated individuals aged 24 to 26, which reduces the pool of non-treated to 211 observations. Even though the mean age among non-treated is hardly affected and increases only slightly from 25.054 to 25.081 years (treated: 25.816), this non-random drop of observations creates a distributional imbalance. The test results for the disturbed data are presented in the bottom panel of table 6.2. The CMS and KS tests reject the null at the 5% level whereas the t-tests fail to do so even at the 10% level due to their reliance on mean differences.

Table 6.2

EMPIRICAL APPLICATION OF AFTER MATCHING TESTS

	'age', $n = 445$					
statistic	CMS-r	KS-r	CMS-p	KS-p	perm.t	t-test
p-val	0.290	0.309	0.171	0.172	0.141	0.266
	'age', $n = 396$, non-treated with age 24-26 discarded					
statistic	CMS-r	KS-r	CMS-p	KS-p	perm.t	t-test
p-val	0.014	0.008	0.019	0.010	0.166	0.329

7 Conclusion

The balancing property of the propensity score is key for the consistency of estimators based on propensity score matching. Thus, the attractiveness of propensity score matching over parametric alternatives with respect to model flexibility is lost when using a propensity score specification that is incapable to balance the covariates' distributions in the groups of treated and non-treated units.

In this paper, we propose balance tests based on nonparametric quantile regression and resampling methods, which do not restrict the Kolmogorov-Smirnov and Cramer-von-Mises-Smirnov test statistics to be of any particular parametric form. These tests check for differences in the entire distributions of covariates. If distributional differences affect higher moments, they are likely to be more powerful than conventional balance tests as the DW test applied in Dehejia & Wahba (1999, 2002) and the two sample t-test for matched samples, which merely check for differences in means. In contrast to specification tests as suggested by Ramsey (1969) and Shaikh et al. (2006), the proposed tests do not reject misspecified, but yet balancing propensity scores. This can be beneficial from a practitioner's point of view who might prefer a misspecified parsimonious model over the true model for the sake of econometric feasibility. As long as the specification balances, propensity score matching is consistent, such that specification tests seem overly restrictive.

The test procedures can either be applied on the full or on the matched sample. Implemented as full sample tests, they check for balancing conditional on the propensity score. Similarly to the DW test, a rejection of the null implies the use of a different, typically more flexible propensity score specification. However, what the researcher is ultimately interested in is the balance in the matched sample, which is used for estimation. As the matching algorithm controls for the propensity score prior to testing, the proposed methods are based on unconditional quantiles when used as after matching tests. If the null cannot be rejected in the full sample, but is rejected in the matched sample, this may point to imbalance introduced by the matching algorithm. Therefore, it makes sense to consider both full sample and after matching tests at the same time. Monte Carlo results suggest that the tests' power and size properties are quite decent in scenarios where conventional balance tests fail to detect imbalance and specification tests incorrectly reject a misspecified, but balancing propensity score model.

References

Abadie, A. (2002), 'Bootstrap tests for distributional treatment effects in instrumental variable models', *Journal of the American Statistical Association* **97**, 284292.

Baser, O. (2006), 'Too much ado about propensity score models? comparing methods of propensity score matching', *Value in Health* **9**(6), 377–385.

Becker, S. & Ichino, A. (2002), 'Estimation of average treatment effects based on propensity scores', *The Stata Journal* **2**(4), 358377.

Böckerman, P. & Ilmakunnas, P. (2009), 'Unemployment and self-assessed health: Evidence from panel data', *Journal of Health Economics* **18**, 161179.

Chernozhukov, V. & Fernandez-Val, I. (2005), 'Subsampling inference on quantile regression processes', *Sankhya: The Indian Journal of Statistics* **67**, 253–276.

Chintrakarn, P. (2008), 'Estimating the euro effects on trade with propensity score matching', *Review of International Economics* **16**(1), 186–198.

Dehejia, R. & Wahba, S. (1999), 'Causal effects in non-experimental studies: Reevaluating the evaluation of training programmes', *Journal of American Statistical Association* **94**, 1053–1062.

Dehejia, R. H. & Wahba, S. (2002), 'Propensity-score-matching methods for nonexperimental causal studies', *The Review of Economics and Statistics* **84**(1), 151161.

Firpo, S. (2007), 'Efficient semiparametric estimation of quantile treatment effects', *Econometrica* **75**, 259–276.

Goh, S. C. (2007), 'Nonparametric inferences on conditional quantile processes', *unpublished manuscript.*

Hayfield, T. & Racine, J. (2008), 'Nonparametric econometrics: The np package', *Journal of Statistical Software.*

Hirano, K. & Imbens, G. (2004), The propensity score with continuous treatments, *in* A. Gelman & X. L. Meng, eds, 'Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives', New York: Wiley, pp. 73–84.

Hirano, K., Imbens, G. & Ridder, G. (2003), 'Efficient estimation of average treatment effects using the estimated propensity score', *Econometrica* **71**, 1161–1189.

Ichino, A., Mealli, F. & Nannicini, T. (2006), 'From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity?', *IZA Discussion Paper No. 2149.*

Imai, K., King, G. & Stuart, E. (2006), 'The balance test fallacy in matching methods for causal inference', *unpublished manuscript.*

Imbens, G. (2000), 'The role of the propensity score in estimating dose-response functions', *Biometrika* **87**, 706–710.

Imbens, G. W. (2004), 'Nonparametric estimation of average treatment effects under exogeneity: a review', *The Review of Economics and Statistics* **86**(1), 429.

Koenker, R. & Bassett, G. (1978), 'Regression quantiles', *Econometrica* **46**(1), 33–50.

LaLonde, R. (1986), 'Evaluating the econometric evaluations of training programs with experimental data', *American Economic Review* **76**(4), 604–620.

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**(1), 74–90.

Lechner, M. (2001), Identification and estimation of causal effects of multiple treatments under the conditional independence assumption, *in* M. Lechner & F. Pfeiffer, eds, 'Econometric Evaluations of Active Labor Market Policies in Europe', Heidelberg: Physica.

Lee, W. (2006), 'Propensity score matching and variations on the balancing test', *unpublished manuscript.*

Li, Q. & Racine, J. (2008), 'Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data', *Journal of Business and Economic Statistics* **26**(4), 423–434.

Ramsey, J. B. (1969), 'Tests for specification errors in classical linear least squares regression analysis', *Journal of the Royal Statistical Society, series B* **31**, 350371.

Rosenbaum, P. & Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**(1), 41–55.

Rosenbaum, P. R. & Rubin, D. B. (1984), 'Reducing bias in observational studies using subclassification on the propensity score', *Journal of the American Statistical Association* **79**, 516–524.

Rosenbaum, P. R. & Rubin, D. B. (1985), 'Constructing a control group using multivariate matched sampling methods that incorporate the propensity score', *American Statistician* **3**, 33–38.

Rubin, D. B. (1997), 'Estimating causal effects from large data sets using propensity scores', *Annals of Internal Medicine* **127**(5), 757–763.

Sekhon, J. S. (2007*a*), 'Alternative balance metrics for bias reduction in matching methods for causal inference', *unpublished manuscript.*

Sekhon, J. S. (2007*b*), 'Multivariate and propensity score matching software with automated balance optimization: The matching package for r', *forthcoming in the Journal of Statistical Software.*

Shaikh, A. M., Simonsen, M., Vytlacil, E. J. & Yildiz, N. (2006), 'On the identification of misspecified propensity scores', *unpublished manuscript.*

Wunsch, C. & Lechner, M. (2008), 'What did all the money do? on the general ineffectiveness of recent west german labour market programmes', *Kyklos* **61**(1), 134–174.

Yu, K., Lu, Z. & Stander, J. (2003), 'Quantile regression: applications and current research areas', *The Statistician* **52**, 331350. Part 3.

Zhao, Z. (2008), 'Sensitivity of propensity score methods to the specifications', *Economics Letters* **98**(3), 309–319.