




## Testing for covariate balance using quantile regression and resampling methods


Martin Huber


To cite this article: Martin Huber (2011) Testing for covariate balance using quantile regression and resampling methods, Journal of Applied Statistics, 38:12, 2881-2899, DOI: [10.1080/02664763.2011.570323](https://doi.org/10.1080/02664763.2011.570323)

To link to this article: <http://dx.doi.org/10.1080/02664763.2011.570323>

 Published online: 21 Apr 2011.

 Submit your article to this journal [↗](#)

 Article views: 287

 View related articles [↗](#)

 Citing articles: 2 View citing articles [↗](#)

# Testing for covariate balance using quantile regression and resampling methods

Martin Huber\*

*Department of Economics, SEW, University of St. Gallen, Varnbuelstrasse 14, 9000 St. Gallen, Switzerland*

*(Received 30 November 2010; final version received 20 February 2011)*

Consistency of propensity score matching estimators hinges on the propensity score's ability to balance the distributions of covariates in the pools of treated and non-treated units. Conventional balance tests merely check for differences in covariates' means, but cannot account for differences in higher moments. For this reason, this paper proposes balance tests which test for differences in the entire distributions of continuous covariates based on quantile regression (to derive Kolmogorov–Smirnov and Cramer–von-Mises–Smirnov-type test statistics) and resampling methods (for inference). Simulations suggest that these methods are very powerful and capture imbalances related to higher moments when conventional balance tests fail to do so.

**Keywords:** balancing property; balance test; propensity score matching

*JEL classification:* C12, C15, C21

## 1. Introduction

Propensity score matching (PSM) [33,35] has become an increasingly popular estimation method in many fields of empirical research concerned with the evaluation of treatment effects in a conditional independence or selection on observables framework (see [21]). Applications include the evaluation of active labor market policies [14,15], the estimation of the health effects of unemployment [7], the evaluation of trade gains due to a common currency [31] and many others.

PSM is attractive because it does not rely on tight functional form assumptions as parametric estimators, nor is it prone to the curse of dimensionality issue inherent in matching on a high dimensional covariate vector directly. However, one condition for consistency of PSM is the balancing property of the presumed propensity score model. It states that conditional on the

---

\*Email: martin.huber@unisg.ch

I have benefited from comments by Eva Deuchert, Bernd Fitzenberger, Michael Lechner, Enno Mammen, Blaise Melly, Conny Wunsch, conference/seminar participants in Linz (Annual Meeting of the Austrian Economic Association), London (cemmap conference on quantile regression), Geneva (Annual Meeting of the Swiss Society of Economics and Statistics), and Freiburg i.B. (seminar in empirical economics), and two anonymous referees.

propensity score, the distributions of the covariates in the pools of treated and non-treated units must be equal, i.e., balanced.

All balance tests conventionally used in the literature, as the DW test (see [10,11]), the regression test of Smith and Todd [40], or the two-sample  $t$ -test for matched samples, merely check for differences in the means of covariates. Thus, they might lack power when imbalances affect distributional features other than the mean. An alternative is the specification test proposed by Shaikh *et al.* [39] which tests the entire propensity score model, not just the balancing property. However, the caveat of this method is that it may reject misspecified propensity score models which are nevertheless balancing, e.g., when the misspecified propensity score is only a monotonic transformation of the true one. Therefore, the availability of suitable balance tests for PSM appears to be unsatisfactory.

This paper aims at filling this gap by suggesting test procedures for continuous covariates<sup>1</sup> which account for differences in the entire distribution. In contrast to commonly applied mean difference tests, the proposed methods also capture distributional imbalances related to higher moments. The procedures are based on (i) quantile regression (see, for instance [5,8,23,24]), (ii) the computation of Kolmogorov–Smirnov (KS) and Cramer–von-Mises–Smirnov (CMS) test statistics on the empirical inference process, and (iii) resampling in order to estimate the distributions and  $p$ -values of the KS and CMS statistics (see [9]). We discuss the implementation of these methods as full sample tests (based on the entire sample) and after-matching tests (based on the sample of matched units alone) and point to differences in the interpretation of the results. Furthermore, we provide simulation evidence on the performance of the tests relative to existing balance tests. Therefore, this paper complements the analysis of the finite sample properties of balance tests by Lee [27] and extends the range of tests investigated.

The remainder of this paper is organized as follows. Section 2 motivates PSM and more formally discusses the condition to be tested. Section 3 briefly reviews the literature on balance tests and introduces our full sample and after-matching tests for continuous covariates based on resampling and quantile regression. Section 4 presents simulation results about the finite sample properties of our methods and the tests applied in the literature. Section 5 presents an empirical application of full sample and after-matching tests to Italian labor market data. 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 probability and the 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, Imbens [21] for an in-depth discussion. It implies that the effect of the treatment on the outcome is conditionally unconfounded. Let  $Y$  denote the outcome variable,  $D$  a binary treatment taking either the value 1 (treated) or 0 (non-treated),<sup>2</sup> and  $X$  a vector of observed covariates with support  $\mathcal{X}$ . The CIA states that

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

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

From a practitioner’s perspective, conditioning on a high dimensional  $X$  may be problematic, as the number of possible combinations of elements in  $X$  increases exponentially in the dimension of  $X$  such that a precise estimation quickly becomes exorbitantly data hungry. This problem is known as curse of dimensionality. Let  $p^*(X) \equiv \Pr(D = 1 | X)$  denote the unknown probability of being treated conditional on  $X$ , henceforth referred to as true propensity score. Rosenbaum and

Rubin [33] 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 distributions of covariates in the treatment and in the control (or non-treated) group. Thus, if Equation (1) is satisfied, it also holds that the hypothetical outcomes are independent of the treatment conditional on the propensity score:

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

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, e.g., the occurrence of empty cells for particular combinations of covariates. For this reason, PSM is frequently used in empirical applications. If Equation (2) is satisfied, average treatment effects (ATEs) and quantile treatment effects (QTEs) can 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 Equation (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 the real world, the structural form of the true propensity score is usually unknown to the researcher. In empirical applications it is most commonly modeled parametrically using probit or logit specifications. Let  $p(X)$  denote the presumed specification of the true  $p^*(X)$ . Whereas the balancing property of  $p^*(X)$  follows from the proof in Rosenbaum and Rubin [33], it is a priori not clear whether  $p(X)$  balances  $X$  in the pools of treated and non-treated units. However, the balancing property of  $p(X)$  is testable by verifying whether

$$F(x|D = 1, p(X) = \rho) = F(x|D = 0, p(X) = \rho) \quad \forall x \in \mathcal{X}, \quad \forall \rho \in (0, 1), \tag{4}$$

where  $F_{X|D=d,p(X)}(\cdot|D = d, p(X))$  denotes the conditional cdf of  $X$  given  $D = d$  and  $p(X)$ . If Equation (4) is satisfied, it holds that

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

Instead of building tests for equality of the conditional distribution functions<sup>3</sup> given  $p(X)$  it is equally valid to test for differences in the conditional quantile functions for  $D = 1, 0$ , as the quantile function is simply the inverse of the distribution function. Let  $Q_A^\tau$  represent the  $\tau$ th quantile ( $\tau \in (0, 1)$ ) for some variable  $A$ ,  $Q_A^\tau = \inf\{a : F_A(a) \geq \tau\}$ . Then,  $F_A(a) = Q_A^{\tau^{-1}}$ . For  $Q_X^\tau(d, \rho)$  denoting the  $\tau$ th conditional quantile of  $X$  given  $D = d$  and  $p(X) = \rho$ , the balancing property implies that

$$Q_X^\tau(1, \rho) = Q_X^\tau(0, \rho), \quad \forall \tau, \rho \in (0, 1), \quad \forall x \in \mathcal{X}. \tag{6}$$

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

$$E[X|D = 1, p(X) = \rho] = E[X|D = 0, p(X) = \rho], \quad \forall x \in \mathcal{X}, \quad \forall \rho \in (0, 1), \tag{7}$$

which is necessary, but not sufficient for Equations (4) and (5). That is, these tests do not account for distributional differences related to higher moments and ignore valuable information that might point to the violation of covariate balance, see also the discussion in Sekhon [37]. Thus, it appears more appropriate to use procedures that capture imbalances in the entire distributions rather than in the means alone. For this reason, we propose quantile-based tests for continuously distributed covariates in Section 3, which may be applied to both full and matched samples.

### 3. Testing

#### 3.1 Full sample tests

Balance tests can be categorized into methods testing the balancing property (i) in the entire 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). Two commonly applied tests of the former kind are the DW test used in Dehejia and Wahba [10,11], which is based on a process originally proposed by Rosenbaum and Rubin [34] and Rubin [36], and the regression test of Smith and Todd [40].

Smith and Todd [40] suggest regressing the covariate of interest on a quartic polynomial of the estimated propensity score, the treatment state, and its interaction terms with the polynomial. Using a Wald-statistic one tests whether the coefficients on the treatment dummy and the interaction terms are jointly equal the zero. If the latter holds true, the conditional mean of the covariate is independent of the treatment, which is necessary, albeit not sufficient for balancing. However, one may extend this approach to higher moments. In the simulations and the application further below, we therefore also consider a joint regression test for both the mean and the variance in a GMM framework. The DW test is based on testing for mean differences in the covariates across treated and non-treated units within strata defined upon the estimated propensity scores, see Dehejia and Wahba [11] for further details. Lee [27] argues that the standard DW test has poor size properties and suggests to estimate the distribution of the test statistic based on permutation (see [32]) instead of asymptotic approximation. However, this permuted version is incapable to account for differences in higher moments as the original test.

Due to this shortcoming of conventional methods, we propose a test procedure that captures differences in the entire distribution of a continuous covariate. Our method is based on the results of Chernozhukov and Fernandez-Val [9] who developed tests based on KS and CMS statistics derived from quantile regression processes. As these statistics are non-pivotal, the authors propose resampling methods to compute the critical values and  $p$ -values. Our balance test can be divided into four steps. Prior to testing, we predict the propensity scores for the units in the sample based on the presumed model  $p(X)$ . Secondly, we estimate the covariate's conditional quantiles given the estimated propensity score. In the third step, KS and CMS statistics are computed based on the differences in the conditional quantiles across treatment states. Finally, we use bootstrapping to estimate the distributions of the test statistics required for inference.

Our approach differs from Chernozhukov and Fernandez-Val [9] with respect to one important feature, namely that the regressors are known in their framework, whereas we need to estimate the propensity score (which serves as the regressor in our test procedure). To the best of our knowledge no analytical results for resampling methods of statistics on quantile regression processes exist when the regressor is estimated. However, as the propensity scores are re-estimated in each resampling step, the bootstrap procedure takes account of the uncertainty coming from this estimation. Simulation results in Section 4 suggest that the test procedures perform well at least when the propensity score is estimated parametrically.

As a further remark, note that our methods are not necessarily restricted to continuous covariates. They may also be used for count variables, given that they are artificially smoothed. E.g., we can add uniformly distributed noise to the discrete covariate of interest in order to create a continuous variable (with all the desired properties of smoothness). Machado and Silva [30] show that under certain conditions, we may replace the original covariate by the new variable when conducting inference based on quantile regression, because there exists a one-to-one relationship between the conditional quantiles of the latter and the former variable. This allows circumventing the problems related to the non-smoothness of the objective function in quantile regression when data are discrete. In the subsequent exposition we will, however, focus on the continuous case.

To formally discuss the test procedure, we denote the (continuously distributed) covariate of interest as  $X_k$ , indicating that it is the  $k$ th element in the covariate vector  $X$ . The null hypothesis is

$$H_0 : Q_{X_k}^\tau(1, \rho) = Q_{X_k}^\tau(0, \rho), \quad \forall \tau, \rho \in (0, 1), \tag{8}$$

i.e. that the conditional quantiles of  $X_k$  given  $p(X)$  are equal across treatment states  $D = 1, 0$  at all ranks and for all values of the propensity score. This would imply that Equation (5) holds.

We estimate  $Q_{X_k}^\tau(1, \rho)$ ,  $Q_{X_k}^\tau(0, \rho)$  by a quantile regression of  $X$  on a constant and a polynomial of the propensity score estimate, e.g., on the score itself, its square and its cubic. Let  $\hat{p}(X_i)$  denote the propensity score estimate for unit  $i$  and specification  $p(X)$ . For treatment state  $d = 1, 0$ , the quantile coefficients  $\beta_d^\tau$  are estimated by solving the following minimization problem:

$$\hat{\beta}_d^\tau = \min_{\beta} \frac{1}{n_d} \sum_{i:D=d} \eta_\tau \left( X_{k,i} - \sum_{l=0}^L (\hat{p}^l(X_i)) \beta \right), \tag{9}$$

where  $n_d$  is the number of observations with  $D = d$ .  $\eta_\tau(v) = v(\tau - I\{v \leq 0\})$  is the check function, an asymmetric loss function, suggested by Koenker and Bassett [24] in their seminal paper on quantile regression. By setting  $L = 3$ , we regress  $X_k$  on a constant and the third-order polynomial of the propensity score estimate. We suspect this specification to be sufficiently flexible for a univariate regression, but also try lower orders in our simulations presented in Section 4. The conditional quantile of  $X_k$  given  $D = d$  at  $p(X) = \rho$  is predicted by

$$\hat{Q}_{X_k}^\tau(d, \rho) = \sum_{l=0}^L (\rho^l) \hat{\beta}_d^\tau. \tag{10}$$

We would like to infer whether the process  $Q_{X_k}^\tau(1, \rho) - Q_{X_k}^\tau(0, \rho)$ , which is not observed, is different from zero. However, we only observe the empirical inference process

$$\hat{Q}_{X_k}^\tau(1, \rho) - \hat{Q}_{X_k}^\tau(0, \rho), \tag{11}$$

i.e. the difference between the conditional quantile estimates. We use these differences to compute KS and CMS test statistics, denoted as  $T_n$ , which account for differences in the conditional quantile estimates across different ranks of the covariate distribution and across propensity scores ( $\rho$ ). Let  $n, n_1$  and  $n_0$  denote the total sample size, the number of treated and the number of non-treated observations, respectively. The KS statistic is based on the supremum of the difference across ranks and scores, the CMS statistic on the integration over the squared differences:

$$\begin{aligned} T_n^{\text{KS}} &= \sup_{\tau \in \mathcal{T}, \rho \in \mathcal{P}} \sqrt{\frac{n_1 \cdot n_0}{n}} \|\hat{Q}_{X_k}^\tau(1, \rho) - \hat{Q}_{X_k}^\tau(0, \rho)\|_{\hat{\Lambda}}, \\ T_n^{\text{CMS}} &= \frac{n_1 \cdot n_0}{n} \int_{\mathcal{T}} \int_{\mathcal{P}} \|\hat{Q}_{X_k}^\tau(1, \rho) - \hat{Q}_{X_k}^\tau(0, \rho)\|_{\hat{\Lambda}}^2 d\tau d\rho. \end{aligned} \tag{12}$$

$\mathcal{T}$  and  $\mathcal{P}$  denote the support of  $\tau$  and  $p(X)$  and are naturally bounded between 0 and 1.  $\|a\|_{\hat{\Lambda}}$  denotes  $\sqrt{a' \hat{\Lambda} a}$  and  $\hat{\Lambda}$  is a positive weighting matrix satisfying  $\hat{\Lambda} = \Lambda + o_p(1)$ .  $\Lambda$  is positive definite, continuous and symmetric.

$T_n^{\text{KS}}$  and  $T_n^{\text{CMS}}$  are non-pivotal such that their distributions do not converge to any known distribution. For linear quantile regression processes as considered in this paper, Chernozhukov and Fernandez-Val [9] show in Theorem 1 that the distributions of  $T_n^{\text{KS}}$  and  $T_n^{\text{CMS}}$  can be consistently estimated by resampling the recentered test statistics under their Assumptions A.1–A.3.

These assumptions state that the data are stationary and strongly mixing (which is satisfied in i.i.d. samples) and that the uniformly consistent parameters entering the null hypothesis, in our case the quantile coefficient estimates, are asymptotically Gaussian under local and global alternatives. Following their approach, we draw  $J$  samples of size  $n$  with replacement from the original sample. For each bootstrap sample, we estimate the propensity scores and the conditional quantiles to compute the bootstrapped inference process

$$\hat{Q}_{X_k,j}^\tau(1, \rho) - \hat{Q}_{X_k,j}^\tau(0, \rho). \quad (13)$$

$\hat{Q}_{X_k,j}^\tau(1, \rho)$  and  $\hat{Q}_{X_k,j}^\tau(0, \rho)$  denote the conditional quantile estimates for sample draw  $j$ , where  $(1 \leq j \leq J)$ . The corresponding KS and CMS statistics of the bootstrapped and recentered inference processes are

$$T_{n,j}^{\text{KS}} = \sup_{\tau \in \mathcal{T}, \rho \in \mathcal{P}} \sqrt{\frac{n_1 \cdot n_0}{n}} \|\hat{Q}_{X_k,j}^\tau(1, \rho) - \hat{Q}_{X_k,j}^\tau(0, \rho) - (\hat{Q}_{X_k}^\tau(1, \rho) - \hat{Q}_{X_k}^\tau(0, \rho))\|_{\hat{\Lambda}},$$

$$T_{n,j}^{\text{CMS}} = \frac{n_1 \cdot n_0}{n} \int_{\mathcal{T}} \int_{\mathcal{P}} \|\hat{Q}_{X_k,j}^\tau(1, \rho(x)) - \hat{Q}_{X_k,j}^\tau(0, \rho) - (\hat{Q}_{X_k}^\tau(1, \rho) - \hat{Q}_{X_k}^\tau(0, \rho))\|_{\hat{\Lambda}}^2 d\tau d\rho. \quad (14)$$

Note that these statistics differ slightly to Chernozhukov and Fernandez-Val [9] in that  $n_1 \cdot n_0/n$  is used instead of  $n$  as we consider a two samples testing problem. Finally, we compute the  $p$ -values by  $J^{-1} \sum_{j=1}^J I\{T_{n,j} > T_n\}$  which is a consistent estimator of  $\Pr[T(\hat{Q}_{X_k}^\tau(1, \rho) - \hat{Q}_{X_k}^\tau(0, \rho) - (Q_{X_k}^\tau(1, \rho) - Q_{X_k}^\tau(0, \rho))) > T_n]$ .

Having outlined our procedure we would like to point out that this is not the only possibility to build balance tests based on quantile regression processes. E.g., equivalent to Smith and Todd [40], one might regress the covariate on a function of the propensity score, the treatment state, and the interaction terms, however, not at the mean, but at different quantiles. Based on this regression and the goodness-of-fit measure for quantile regression introduced by Koenker and Machado [25] the balancing property may be tested in the following way: Firstly, we estimate the full model and a restricted model without the treatment state and the interaction terms. Secondly, we construct LR statistics based on the differences in the model fits at each quantile as outlined in Koenker and Machado [25]. Finally, we take the supremum of the LR statistics across quantiles and use the critical values provided in Andrews [2] for such suprema to test the balancing property. We include this approach in the simulations and the application presented further below.

### 3.2 After-matching tests

The most popular after-matching test among practitioners appears to be the two sample  $t$ -test for mean differences in covariates across treated and non-treated matches. As for the DW test, Lee [27] suggests us to use permuted  $t$ -tests to improve the finite sample properties. A further issue is the test's sensitivity to the sample size, see the discussion in Imai *et al.* [19]. I.e., the test statistic can be distorted by randomly dropping observations, even though the balance is unaffected. In contrast, the test of standardized differences suggested by Rosenbaum and Rubin [35] is robust to variations in the sample size. It is based on normalizing the mean difference across treated and non-treated matches by the square root of the variances in the full sample (but not by the sample size). According to Rosenbaum and Rubin a standardized difference greater than 20 is 'large', i.e., pointing to imbalance. Among all after-matching procedures, the only test also accounting for imbalances in higher moments appears to be the permuted KS test for equality in distributions advocated by Diamond and Sekhon [12].

We will now propose an alternative to this approach based on quantiles. In contrast to the full sample tests we need not condition on the propensity score as this task is performed by a (hopefully accurate) matching algorithm prior to testing. Therefore, the after-matching KS and CMS resampling procedures consist of three steps: The estimation of the unconditional quantiles in the pools of treated and non-treated matched units, the computation of the test statistics, and the resampling procedure to compute the  $p$ -values. Let  $\hat{Q}_{X_k^m}^\tau(d)$  denote the  $\tau$ th unconditional quantile in the sample of matched units with  $D = d$ , where the superscript  $m$  indicates ‘matched’. The KS and CMS statistics for the empirical inference process  $\hat{Q}_{X_k^m}^\tau(1) - \hat{Q}_{X_k^m}^\tau(0)$  are

$$\begin{aligned}
 T_{n^m}^{\text{KS}} &= \sup_{x \in \mathcal{X}^m} \sqrt{\frac{n_1^m \cdot n_0^m}{n^m}} \|\hat{Q}_{X_k^m}^\tau(1) - \hat{Q}_{X_k^m}^\tau(0)\|_{\hat{\Lambda}}, \\
 T_{n^m}^{\text{CMS}} &= \frac{n_1^m \cdot n_0^m}{n^m} \int_{\mathcal{X}^m} \|\hat{Q}_{X_k^m}^\tau(1) - \hat{Q}_{X_k^m}^\tau(0)\|_{\hat{\Lambda}}^2 dx.
 \end{aligned}
 \tag{15}$$

We draw  $J$  bootstrap samples from the matched sample, estimate the quantiles  $\hat{Q}_{X_k^m,j}^\tau(1)$  and  $\hat{Q}_{X_k^m,j}^\tau(0)$  and compute the statistics on the bootstrapped and recentered inference processes:

$$\begin{aligned}
 T_{n^m,j}^{\text{KS}} &= \sup_{x \in \mathcal{X}^m} \sqrt{\frac{n_1^m \cdot n_0^m}{n^m}} \|\hat{Q}_{X_k^m,j}^\tau(1) - \hat{Q}_{X_k^m,j}^\tau(0) - (\hat{Q}_{X_k^m}^\tau(1) - \hat{Q}_{X_k^m}^\tau(0))\|_{\hat{\Lambda}}, \\
 T_{n^m,j}^{\text{CMS}} &= \frac{n_1^m \cdot n_0^m}{n^m} \int_{\mathcal{X}^m} \|\hat{Q}_{X_k^m,j}^\tau(1) - \hat{Q}_{X_k^m,j}^\tau(0) - (\hat{Q}_{X_k^m}^\tau(1) - \hat{Q}_{X_k^m}^\tau(0))\|_{\hat{\Lambda}}^2 dx.
 \end{aligned}
 \tag{16}$$

Finally, the  $p$ -values are obtained by  $J^{-1} \sum_{j=1}^J I\{T_{n^m,j} > T_{n^m}\}$ . Note that these  $p$ -values do not bear the same interpretation as in classical hypothesis tests (e.g., when testing the balancing property using the full sample tests). Firstly, they are asymptotically not valid for testing the balancing property for the population, because the matched sample is a non-random draw that depends on the matching algorithm. Therefore, judgments about balance strictly refer to the matched sample. Secondly and as argued by Imai *et al.* [19] and Sekhon [37], the  $p$ -values are not to be used as stopping rules for covariate balancing in matched samples, where the researcher seeks to maximize balance without limit. I.e., one strives for identical covariate distributions in the pools of treated and non-treated matches, implying a  $p$ -value of 1.

As an alternative to bootstrapping, one may estimate the distribution of the test statistics by permutation, i.e., by randomly shuffling treatment and control labels among matched observations without replacement. Permutation tests are valid when shuffling the labels does not affect the results under the null hypothesis, see Good [13]. As this is satisfied in balance tests, where the covariate distribution is independent of the treatment label under null, Diamond and Sekhon [12] use a permuted KS distribution test to assess covariate balance. This test was proposed by Abadie [1] in a different context, namely to test for distributional treatment effects in an IV framework. Abadie shows that the procedure has correct asymptotic size under the weak condition that the variable (in our case  $X_k^m$ ) has a non-degenerate distribution with bounded support. In the simulations and the application both resampling- and permutation-based versions of the tests are considered.

As a final remark, note that the proposed methods cannot be easily applied to matching algorithms that do not create an explicit matched sample. E.g., kernel matching as discussed in Heckman *et al.* [14,15] merely provides weights which balance the propensity scores of treated and non-treated units and allow predicting the counterfactual outcomes. These weights do not reveal the value of the counterfactual covariate, as there is no one-to-one correspondence between the propensity score and the covariate. The same value of the propensity score can in principle be



obtained by many combinations of the covariates. For mean difference tests, it suffices that the weights allow estimating the conditional mean of the counterfactual covariate given the propensity score, as the tests average over the covariates in the matched sample. This is neither the case for the proposed CMS and KS procedures, nor for the KS distribution test, which require the knowledge of the distributions of the covariates in the matched sample. With this respect, full sample tests appear to be more generally applicable than the after-matching tests considered in this section.

#### 4. Monte Carlo simulations

In this section, we present Monte Carlo evidence on the finite sample properties of KS and CMS full sample and after-matching tests and run a horse race with other tests proposed in the literature. Concerning the propensity score model, we consider three different scenarios: Correct specification of the propensity score, misspecification of the propensity score but satisfaction of the balancing property, and misspecification and violation of the balancing property. Accurate balance tests should keep the null in the first and second scenario, but reject it in the third. It is the aim of the scenarios to give an intuition about the strengths and weaknesses of alternative (classes of) tests, but of course, they do not claim completeness, as many more data generating processes could be considered.

##### 4.1 Full sample tests

Starting with the full sample tests, we compare the performance of our procedures to another quantile regression-based test using the results of Koenker and Machado [25] (see the bottom of Section 3.1), the DW test<sup>4</sup> (see [10,11]), the regression test of Smith and Todd [40] both in its original version (for the mean only, henceforth denoted as ST1) and for the first and the second moment jointly (henceforth ST2), and the specification test proposed by Shaikh *et al.* [39].<sup>5</sup> The first data generating process (DGP) considered is

$$\begin{aligned} D_i &= I\{\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \varepsilon > 0\}, \\ Y_i &= \gamma_1 X_{1,i}^2 + \gamma_2 X_{2,i} + \gamma_3 D_i + U_i \\ X_1, X_2 &\sim \text{unif}(0, 3), \quad \varepsilon \sim N(0, 2), \quad U \sim N(0, 1) \\ \beta_0 &= -1.5, \quad \beta_1 = \beta_2 = 0.5, \quad \gamma_1 = \gamma_2 = \gamma_3 = 1. \end{aligned}$$

Treatment effects are homogenous w.r.t.  $X$  and equal to 1. The constant in the treatment equation ( $\beta_0$ ) is chosen such that the unconditional probability to receive the treatment is roughly 50%, and the same applies to the other scenarios considered further below. In the first scenario, the propensity score is correctly specified and characterized by the following probit model:

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

where  $\Phi(\cdot)$  denotes the normal cdf. We test whether the continuous covariate  $X_1$  is balanced conditional on the propensity score for two sample sizes,  $n = 1000, 4000$ . The latter are comparable to many recent empirical studies using PSM estimators, e.g., Berger and Hill [4], Blundell *et al.* [6], Jalan and Ravallion [22] and Loecker [29].

Table 1 reports the rejection frequencies of the null hypothesis at the 5% and 10% significance levels, i.e., the share of  $p$ -values that lie at or below 0.05 and 0.10, respectively, for 1000 Monte Carlo replications. Inference for the KS and CMS balance tests is based on 499 bootstrap draws. The conditional quantiles are evaluated at  $\tau \in \mathcal{T}_{[0.25, 0.75]} = \{0.25, 0.30, 0.35, \dots, 0.75\}$ .

Table 1. Full sample tests: rejection frequencies under correct specification.

Rejection rates at	$n = 1000$		$n = 4000$	
	5%	10%	5%	10%
CMS (var) order 1*	0.000	0.000	0.000	0.000
CMS (var) order 2*	0.010	0.037	0.021	0.084
CMS (var) order 3*	0.000	0.001	0.000	0.000
CMS (dens) order 1*	0.000	0.002	0.000	0.000
CMS (dens) order 2*	0.009	0.032	0.020	0.078
CMS (dens) order 3*	0.000	0.000	0.000	0.000
KS (var) order 1*	0.011	0.034	0.033	0.061
KS (var) order 2*	0.033	0.076	0.044	0.109
KS (var) order 3*	0.009	0.025	0.013	0.036
KS (dens) order 1*	0.008	0.021	0.006	0.025
KS (dens) order 2*	0.037	0.090	0.064	0.132
KS (dens) order 3*	0.004	0.019	0.012	0.031
Koenker and Machado	0.000	0.000	0.000	0.000
DW	0.044	0.047	0.134	0.155
DW Bonferroni adj.	0.007	0.010	0.018	0.032
Smith and Todd	0.015	0.045	0.027	0.067
Smith and Todd mean + var.	0.008	0.017	0.135	0.235
Shaikh <i>et al.</i> spec. test**	0.006	0.009	0.006	0.011

Notes: 1000 Monte Carlo replications. \*, 499 bootstrap draws per replication; \*\*, bandwidth for kernel density estimation according to ML cross validation.

The propensity score  $p(X)$  is evaluated on an equidistant grid consisting of 10 values between the 0.25th and 0.75th quantile of the estimated propensity score, which ensures that boundary regions with sparse data are not used in the test procedures. We consider different combinations of smoothing and weighting schemes  $\Lambda$  for the KS and CMS balance tests: We weight differences in conditional quantiles (i) by the inverse of their respective variance (CMS (var), KS (var)), which gives more weight to differences that are precisely estimated, and (ii) by the densities of the predicted propensity scores (CMS (dens), KS (dens)), which gives more weight to differences in areas with large densities of the propensity score. Furthermore, smoothing is varied by using only the propensity score or 2nd and 3rd order polynomials of the propensity score in the quantile regressions, respectively.

In the Koenker and Machado [25] test, the regression functions are evaluated at the same conditional quantiles as in the CMS and KS procedures. Concerning the DW test, we present the results for both the standard version and a modified DW test with an approximation of the Bonferroni adjustment (DW Bonferroni adj.). The motivation for the latter is the simulation evidence in Lee [27], which suggests that the standard DW test has very poor size properties and rejects the null much too often. Testing for balance with respect to  $X_1$ , 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.

As expected, all tests correctly keep the null in most Monte Carlo replications. The CMS test is very conservative and rejects the balancing hypothesis substantially less frequently than the theoretical rates of 5% and 10% and even more so when using propensity score density weighting. However, when using a second-order polynomial of the propensity score, the empirical size of the test improves as the sample size increases. The rejection frequencies of the KS test are generally closer to the theoretical size, again in particular when using a second-order polynomial. Note that the rejection rates of either test are non-monotone in the order of the propensity score. The Koenker and Machado [25] test is very conservative under any sample size. The Shaikh *et al.* specification

test, the DW test with Bonferroni adjustment and the ST1 test are conservative for both sample sizes whereas the standard DW and the ST2 tests reject the null too often for  $n = 4000$ . With the exception of the DW test without Bonferroni adjustment, the empirical size of which deteriorates in the sample size, no class of tests seems to do strikingly better or worse than any other.

To check the accuracy of propensity score methods under the correct specification we apply two nearest neighbors caliper matching and inverse probability weighting (IPW) estimators to the simulated data. For matching we use the Match command by Sekhon [38] and set the caliper to 0.1 standard deviations of the propensity score. The ATE estimate is  $\hat{\Delta} = 1.004$  for  $n = 1000$  and the mean squared error (MSE) is 0.008. For  $n = 4000$ ,  $\hat{\Delta} = 1.002$  and MSE= 0.002. The IPW estimator, see, for instance, Horvitz and Thompson [17] and Hirano *et al.* [16], performs similarly well.  $\hat{\Delta} = 0.998, 1.002$  and MSE= 0.007, 0.002 for  $n = 1000, 4000$ .

We now turn to a more interesting scenario where the propensity score is misspecified, but yet balancing. We investigate the performance of the tests when data are drawn from the following DGP:

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

We incorrectly use the same propensity score model as before,  $p(X) = \Pr(D = 1|X) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$ , such that  $\beta_1$  is estimated with respect to  $X_1$  instead of  $X_1^3$ . Thus, it is assumed that the index model that underlies the treatment probability is linear in  $X_1$ , whereas the true relationship is cubic. Yet, the incorrect model satisfies the balancing property for variable  $X_1$ , as it is only a monotonous transformation of the true model such that the order of the propensity scores is preserved under misspecification. Even though the propensity scores themselves are poorly estimated, the treated are matched to non-treated units with similar  $p^*(X)$  when using PSM.

To gain some intuition, Figure 1 displays 1000 simulated values of  $X_1$  along with propensity score estimates (i) using the misspecified probit model (dark bubbles) and (ii) based on the correct specification  $p^*(X)$  (light bubbles). As the rank of each observation on average remains the same in either case such that observations with similar  $p^*(X)$  are matched even when using the wrong specification, estimation is consistent.<sup>6</sup>

Table 2 reports the rejection frequencies under the misspecified, but balancing scenario where the propensity score is estimated based on the misspecified probit model. All versions of the CMS test are either on the conservative side or have rejection frequencies that are not too far from the theoretical sizes. Note that there seems to be no clear relationship between the empirical size and the order or the weighting scheme. Also the results for the KS test are quite satisfactory, with the exception of the test versions using a third order polynomial under the larger sample size which rejects the null too often. As in the first scenario, the Koenker and Machado [25] is very conservative and, therefore, appears to have less favorable size properties than the CMS and KS procedures.

The standard DW test is quite accurate for  $n = 1000$ , but its performance deteriorates in the sample size. The Bonferroni adjustment considerably improves the size properties of the DW test for  $n = 4000$ . The rejection frequencies of the ST1 test are already too high for  $n = 1000$  and severely increase in the sample size. This is somewhat surprising, as the ST1 procedure should theoretically test for covariate balance, not for misspecification. Still, it seems to have power in the wrong direction. In contrast, the empirical size of the ST2 test is decent under both sample sizes. As expected, the rejection rates of the Shaikh *et al.* specification test increase in the sample

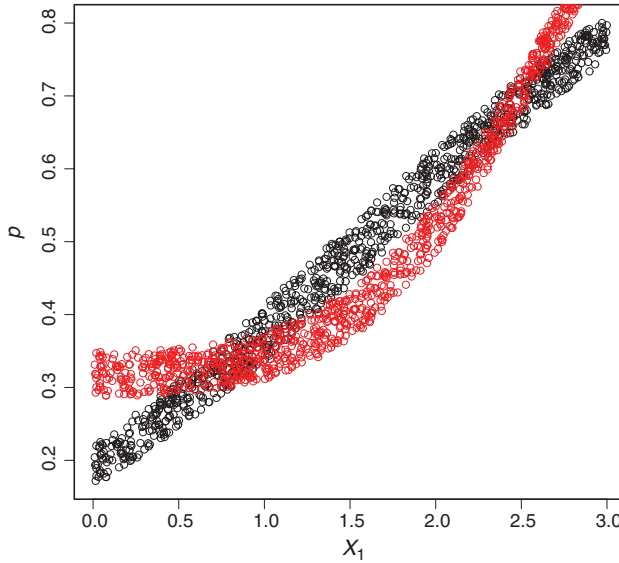


Figure 1. Misspecified and balancing scenario. Propensity scores under misspecification (dark bubbles) and correct specification (light).

Table 2. Full sample tests: rejection frequencies under misspecification and balance.

Rejection rates at	$n = 1000$		$n = 4000$	
	5%	10%	5%	10%
CMS (var) order 1*	0.007	0.013	0.001	0.004
CMS (var) order 2*	0.056	0.079	0.005	0.010
CMS (var) order 3*	0.009	0.038	0.039	0.105
CMS (dens) order 1*	0.008	0.015	0.000	0.003
CMS (dens) order 2*	0.049	0.080	0.005	0.008
CMS (dens) order 3*	0.011	0.030	0.037	0.107
KS (var) order 1*	0.019	0.037	0.003	0.011
KS (var) order 2*	0.073	0.123	0.063	0.103
KS (var) order 3*	0.047	0.115	0.121	0.227
KS (dens) order 1*	0.035	0.070	0.043	0.077
KS (dens) order 2*	0.047	0.081	0.026	0.062
KS (dens) order 3*	0.051	0.113	0.176	0.296
Koenker and Machado	0.000	0.000	0.000	0.000
DW	0.074	0.082	0.265	0.301
DW Bonferroni adj.	0.021	0.030	0.047	0.063
Smith and Todd	0.182	0.274	0.747	0.850
Smith and Todd mean + var.	0.029	0.044	0.091	0.137
Shaikh <i>et al.</i> spec.test**	0.508	0.588	1.000	1.000

Notes: 1000 Monte Carlo replications. \*, 499 bootstrap draws per replication; \*\*, bandwidth for kernel density estimation according to ML cross validation.

size. It rejects the misspecified, but balancing model in all replications for  $n = 4000$ . We conclude that only the quantile-based procedures as well as the ST2 test and the DW test with Bonferroni adjustment yield satisfactory results under the misspecified, but balancing scenario.

Again, we investigate the finite sample properties of two nearest neighbors caliper matching on the propensity score.  $\hat{\Delta} = 1.033$  for  $n = 1000$  and the MSE is equal to 0.008. For  $n = 4000$ ,  $\hat{\Delta} = 1.031$  and MSE= 0.003. Similar to the results in Zhao [41], the misspecification of the

propensity score does not much affect PSM. This is, however, not true for IPW estimators, as consistency of this class of estimators is contingent on the correctness of the propensity score specification. Indeed, the IPW estimates are substantially biased ( $\hat{\Delta} = 1.293, 1.295$ ) and the MSEs are large (0.096, 0.090) for 1000 and 4000 observations, respectively. Therefore, PSM seems to be more robust to propensity score misspecification.

Thirdly, we consider a DGP under which the probit specification is misspecified and not 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 clarify the issues of misspecification *and* imbalance, Figure 2 displays 1000 simulated realizations of  $X_1$  along with propensity score estimates under misspecification (dark bubbles) and under the correct specification (light bubbles).

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. Only treated and non-treated with the same or similar  $p^*(X)$  should be compared to each other. It is obvious that matching on estimates of  $p(X)$  fails to do. The reason is that the incorrect model  $p(X)$  cannot handle the U-shaped non-monotonicity in the relation between  $X_1$  and the true propensity score.  $p^*(X)$  is minimized at the mean of  $X_1$ , which is zero, and increases in either direction. Due to this symmetric relationship, the expected value of the slope coefficient estimate  $\beta_1$  is zero. Therefore, the expected values of the propensity score estimates are independent of  $X_1$ , implying that  $E(X_1|D = d, p(X)) = E(X_1|D = d)$ . Hence, matching is random with respect to the true propensity score such that observations with fairly different  $X_1$  are incorrectly compared to each other.

Table 3 reports the results under the misspecified, non-balancing scenario. Already for  $n = 1000$ , the CMS and KS tests are quite powerful and even more so when using inverse variance

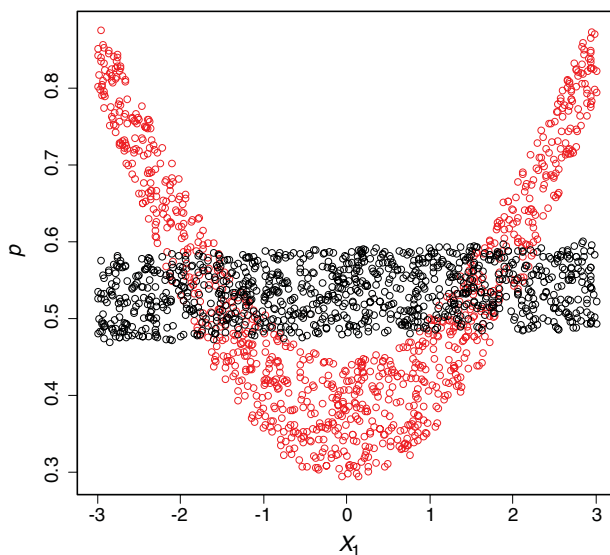


Figure 2. Misspecified and non-balancing scenario. Propensity scores under misspecification (dark bubbles) and correct specification (light).

Table 3. Full sample tests: rejection frequencies under misspecification and imbalance.

Rejection rates at	$n = 1000$		$n = 4000$	
	5%	10%	5%	10%
CMS (var) order 1*	0.930	0.953	1.000	1.000
CMS (var) order 2*	0.975	0.991	1.000	1.000
CMS (var) order 3*	0.997	1.000	1.000	1.000
CMS (dens) order 1*	0.904	0.949	1.000	1.000
CMS (dens) order 2*	0.896	0.970	1.000	1.000
CMS (dens) order 3*	0.955	0.992	1.000	1.000
KS (var) order 1*	0.996	0.997	1.000	1.000
KS (var) order 2*	0.999	0.999	1.000	1.000
KS (var) order 3*	1.000	1.000	1.000	1.000
KS (dens) order 1*	0.982	0.995	1.000	1.000
KS (dens) order 2*	0.997	0.998	1.000	1.000
KS (dens) order 3*	0.997	1.000	1.000	1.000
Koenker and Machado	0.990	1.000	1.000	1.000
DW	0.062	0.070	0.173	0.184
DW Bonferroni adj.	0.009	0.013	0.033	0.046
Smith and Todd	0.037	0.088	0.137	0.219
Smith and Todd mean + var.	0.049	0.110	0.501	0.630
Shaikh <i>et al.</i> spec. test**	0.010	0.013	0.001	0.001

Notes: 1000 Monte Carlo replications. \*, 499 bootstrap draws per replication; \*\*, bandwidth for kernel density estimation according to ML cross validation.

weighting. In the latter case, the null is always rejected in more than 90% of the simulations. For  $n = 4000$ , the rejection rates amount to 100% for any test version, independent of the order and the weighting scheme. The Koenker and Machado [25] is similarly powerful as the CMS and KS procedures. In contrast, the power of balance tests based on mean differences is low. Note that for the DGP considered, the expected value of  $X_1$  is zero for the treated and for the 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 test (with and without Bonferroni adjustment) and the ST1 test. Interestingly, also the Shaikh *et al.* test keeps the null most of the time. Under the larger sample size, the ST2 test does better than the mean tests, as it also accounts for imbalances w.r.t. the second moment. Nevertheless, it is considerably less powerful than the quantile-based procedures.

How is the PSM estimator affected by the imbalance? For  $n = 1000$ , the ATE estimate is severely biased ( $\hat{\Delta} = 3.038$ ) and the MSE (4.191) is huge. For  $n = 4000$ ,  $\hat{\Delta} = 3.071$  and MSE = 4.297. The IPW estimator yields  $\hat{\Delta} = 3.094, 3.097$  and MSE = 4.414, 4.404 for  $n = 1000, 4000$ , respectively. Thus, the imbalance is not innocuous and entails severe biases and inconsistency. In summary, the quantile-based procedures, i.e., the KS/CMS tests and the method based on Koenker and Machado [25], appear to be superior to the balance tests conventionally used by practitioners. While their rejection frequencies are low when the balancing property holds, they are very powerful when it is violated, at least in the scenarios considered.

## 4.2 After-matching tests

This section presents simulations on the finite sample properties of after-matching tests and considers the same DGPs as for the full sample tests. We compare our CMS and KS tests based on resampling (in our case bootstrapping) and permutation to the permuted KS distribution test [12], the permuted and classical (i.e., relying on asymptotic theory) two sample  $t$ -tests, and the test of standardized differences of Rosenbaum and Rubin [35].

For the CMS and KS resampling tests, we again consider two different weighting schemes  $\Lambda$ : We weight differences in quantiles (i) by the inverse of their respective variances (CMS resampling (var), KS resampling (var)) and (ii) by the densities of the predicted propensity scores (CMS resampling (dens), CMS resampling (dens)). To be precise, we weight the differences in quantiles by the product of the densities at the respective quantiles in the samples of treated and non-treated matches. The quantiles are evaluated at  $\tau \in \mathcal{T}_{[0.1,0.9]} = \{0.10, 0.11, 0.12, \dots, 0.90\}$  and inference relies on 499 bootstrap draws or permutations, respectively.

Table 4 displays the results for the correctly specified (and balancing) scenario. Even though the balancing property holds, the rejection frequencies of the CMS and KS tests, including the KS distribution test, are much higher than the theoretical sizes and increase with the sample size. The tests seem to detect the slightest imbalances not eliminated by the matching algorithm. This is unsatisfactory, as the caliper matching procedure yields estimates which are close to the true value even without perfect balance. Note that the empirical sizes of the CMS and KS resampling tests are more accurate when weighting by the propensity score densities, but are still far from being acceptable. The KS distribution test used by Diamond and Sekhon [12] performs even worse. In contrast, the rejection frequencies of permuted and standard  $t$ -tests are not too far from the theoretical sizes, whereas the test of standardized differences is very conservative.

In the misspecified but balancing scenario (see Table 5), the CMS and KS resampling tests with propensity score density weighting have accurate sizes for  $n = 1000$ , but reject the null much too often for  $n = 4000$ . Again, they perform better than the CMS and KS tests based on inverse variance weighting. Also the KS distribution test rejects the null much too often whereas the  $t$ -tests and the test of standardized differences are overly conservative for both sample sizes.

Under misspecification and imbalance all CMS and KS procedures are very powerful and reject the null all the time (see Table 6). In contrast, mean difference tests fail to detect the imbalance related to higher moments. The rejection frequencies of the  $t$ -tests are fairly low and the test of standardized differences has no power at all. Summing up, simulation evidence on after-matching tests is ambiguous about the relative performance of the proposed tests. Even though the CMS and KS tests are very powerful under imbalance, they reject the null much too often when the balancing property holds. This suggests that we should have more confidence in the CMS and KS full sample tests than in the after-matching versions. Using the density of the propensity score estimates as weights in the after-matching tests partly alleviates the problem of over-rejection. Therefore, more research is required with regard to the optimal choice of the weighting matrix in balance tests.

Table 4. After-matching tests: rejection frequencies under correct specification.

Rejection rates at	$n = 1000$		$n = 4000$	
	5%	10%	5%	10%
CMS resampling (var)*	0.239	0.371	0.624	0.783
CMS resampling (dens)*	0.150	0.274	0.560	0.721
KS resampling (var)*	0.366	0.490	0.788	0.871
KS resampling (dens)*	0.168	0.271	0.616	0.752
CMS permutation*	0.218	0.355	0.622	0.772
KS permutation*	0.385	0.531	0.807	0.885
KS distribution*	0.695	0.828	0.989	0.998
permuted $t$ -test*	0.015	0.042	0.068	0.119
standard $t$ -test	0.010	0.024	0.066	0.118
test of standardized differences**	0.000		0.000	

Notes: 1000 Monte Carlo replications. \*, 499 bootstrap draws/permutations per replication; \*\*, rejection if absolute standardized difference  $> 20$ .

Table 5. After-matching tests: rejection frequencies under misspecification and balance.

Rejection rates at	$n = 1000$		$n = 4000$	
	5%	10%	5%	10%
CMS resampling (var)*	0.082	0.160	0.410	0.573
CMS resampling (dens)*	0.047	0.097	0.335	0.481
KS resampling (var)*	0.160	0.234	0.603	0.722
KS resampling (dens)*	0.054	0.111	0.446	0.588
CMS permutation*	0.093	0.158	0.411	0.579
KS permutation*	0.184	0.251	0.653	0.749
KS distribution*	0.418	0.561	0.940	0.973
permuted $t$ -test	0.000	0.000	0.000	0.001
standard $t$ -test	0.000	0.000	0.000	0.000
test of standardized differences**	0.000		0.000	

Notes: 1000 Monte Carlo replications. \*, 499 bootstrap draws/permutations per replication; \*\*, rejection if absolute standardized difference >20.

Table 6. After-matching tests: rejection frequencies under misspec. and imbalance.

Rejection rates at	$n = 1000$		$n = 4000$	
	5%	10%	5%	10%
CMS resampling (var)*	1.000	1.000	1.000	1.000
CMS resampling (dens)*	1.000	1.000	1.000	1.000
KS resampling (var)*	1.000	1.000	1.000	1.000
KS resampling (dens)*	1.000	1.000	1.000	1.000
CMS permutation*	1.000	1.000	1.000	1.000
KS permutation*	1.000	1.000	1.000	1.000
KS distribution*	1.000	1.000	1.000	1.000
permuted $t$ -test	0.050	0.110	0.115	0.167
standard $t$ -test	0.064	0.106	0.144	0.213
test of standardized differences**	0.000		0.000	

Notes: 1000 Monte Carlo replications. \*, 499 bootstrap draws/permutations per replication; \*\*, rejection if absolute standardized difference >20.

## 5. Empirical application

In this section, we apply full sample and after-matching tests to labor market data previously analyzed by Ichino *et al.* [18].

### 5.1 Full sample tests

Ichino *et al.* [18] use PSM to evaluate the effects of job placements by temporary work agencies (TWAs) on the probability to find permanent employment later on in the two Italian regions of Sicily and Tuscany. The data were collected by phone interviews. The treatment period (having or not having a temporary job by TWA assignment) covers the first semester of 2001, the outcome (permanent employment) was measured in November 2002. Pre-treatment covariates  $X$  include detailed information about demographic characteristics, educational attainment, family background and the recent employment history of treated and non-treated individuals. While Ichino *et al.* [18] are interested in the robustness of the estimated effects 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 the balancing property of the propensity score specification



Table 7. Application of full sample tests.

	Fraction unemployed ( $p$ -value)
CMS (var) order 1*	0.615
CMS (var) order 2*	0.724
CMS (var) order 3*	0.562
CMS (dens) order 1*	0.590
CMS (dens) order 2*	0.638
CMS (dens) order 3*	0.382
KS (var) order 1*	0.624
KS (var) order 2*	0.768
KS (var) order 3*	0.644
KS (dens) order 1*	0.676
KS (dens) order 2*	0.630
KS (dens) order 3*	0.704
Koenker and Machado	>0.100
DW**	0.009
Smith and Todd	0.003
Smith and Todd mean + var.	0.260
Shaikh <i>et al.</i> spec. test <sup>†</sup>	0.189

Notes: \*: 999 bootstrap draws. \*\*, minimum  $p$ -value of all intervals; <sup>†</sup>, bandwidth for kernel density estimation according to ML cross validation.

used in Ichino *et al.* [18] for the variable ‘fraction of the school-to-work period that the worker spent as unemployed’ (in %), which characterizes the relative time spent in unemployment after finishing education. Before matching, the fraction is 37.9% for the treated and 47.7% for the non-treated individuals in the sample. We apply the CMS and KS full sample tests to the region of common support in the predicted propensity scores  $\hat{p}(X_i)$ . Therefore, observations in any 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 the null hypothesis at  $\tau \in \mathcal{T}_{[0.25, 0.75]} = \{0.25, 0.25, 0.30, \dots, 0.75\}$  and  $p(x) \in \mathcal{P}_{[0.20, 0.80]} = \{0.20, 0.25, 0.30, \dots, 0.80\}$  using 999 bootstrap replications.

Table 7 presents the test results. All CMS and KS balance tests keep the null at the 5% level, irrespective of the order of the propensity score and the weighting scheme. Also the test based on Koenker and Machado [25] does not reject the test at the 10% level of significance.<sup>7</sup> Ichino *et al.* [18] use the DW test algorithm for Stata provided by Becker and Ichino [3] and do not reject the balancing property either. Note, however, that the significance level chosen by the authors is 0.1%. Setting the significance level to just 1% would reject the null, but one has to bear in mind that this result comes without the Bonferroni adjustment. This example highlights the arbitrariness of the standard DW test with respect to the significance level to be chosen when there are many propensity score intervals.

The ST1 test, which uses a quartic of the propensity score in the regression, rejects the null at the 1% level. However, the test is very sensitive to the choice of the order. Versions based on squared and cubic expansions of the propensity score yield  $p$ -values larger than 5%, which is in line with the insignificant  $p$ -value of the ST2 test. Whereas the tests based on quantile regression unanimously keep the null under various propensity score polynomials, the conclusions drawn from the ST1 and DW tests depend on the choice of the functional form and the level of significance that is considered to be appropriate in the light of stratification, respectively.

## 5.2 After-matching tests

We apply the CMS and KS after-matching tests based on resampling and permutation, the permuted KS distribution test, the standard and permuted  $t$ -tests, and the test of standardized

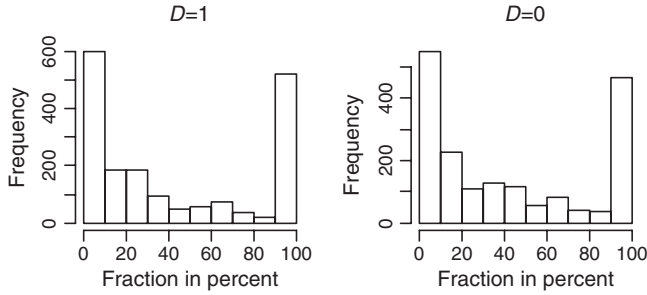


Figure 3. Fraction in unemployment (in %) of treated (left) and non-treated (right) matches.

Table 8. Application of after-matching tests.

	Fraction unemployed ( <i>p</i> -value)
CMS resampling (var)*	0.270
CMS resampling (dens)*	0.303
KS resampling (var)*	0.348
KS resampling (dens)*	0.553
CMS permutation*	0.075
KS permutation*	0.046
KS distribution*	0.000
permuted <i>t</i> -test*	0.664
standard <i>t</i> -test	0.675
test of standardized differences**	-1.434

Notes: \*, 999 bootstrap draws/permutations per replication; \*\*, rejection if absolute standardized difference >20.

differences to the same variable after the application the two nearest neighbors caliper matching algorithm.<sup>8</sup> Figure 3 presents the distributions of the variables ‘fraction of the school-to-work period that the worker spent as unemployed’ for treated and non-treated matches. The distributions appear to be similar and also the sample means are quite close, namely 43.072% for the treated and 43.626% for the non-treated individuals.

Table 8 reports the results of the CMS and KS tests, which evaluate the quantiles at  $\tau \in \mathcal{T}_{[0.1,0.9]} = \{0.1, 0.11, 0.12, \dots, 0.9\}$  and are based on 999 bootstrap samples. Most of the CMS and KS tests yield *p*-values larger than 5% for balance of the variable ‘fraction of period in unemployment’, which is in line with the CMS and KS full sample tests. Only the KS distribution test is highly significant, whereas the *t*-tests and the test of standardized difference suggest that the variable is well balanced. Summing up, neither the full sample nor the after-matching tests based on quantile regressions suggest that the balancing property fails for the variable considered.

### 6. Conclusion

The balancing property of the propensity score is key to the consistency of PSM estimators. Thus, the attractiveness of this class of estimators over parametric alternatives in terms of model flexibility is lost when using a propensity score specification that is incapable to balance the distributions of the covariates in the groups of treated and non-treated units. In this paper, we propose a new class of balance tests for continuous covariates based on quantile regression and bootstrapping Kolmogorov–Smirnov and Cramer–von-Mises–Smirnov-type test statistics. As these tests account for differences in the entire distributions of the covariates, they are most likely more

powerful than conventional balance tests like the DW test used in Dehejia and Wahba [10,11], the regression test by Smith and Todd [40], and the two sample  $t$ -test for matched samples, which merely check for differences in means.

The proposed tests may either be applied in full or in matched samples. Implemented as full sample tests, they test covariate balance conditional on the propensity score. Similar to the DW test, a rejection of the null implies the use of a different, typically more flexible propensity score specification. Monte Carlo results suggest that the power and size properties are satisfactory in scenarios where conventional balance tests fail to detect imbalances and specification tests incorrectly reject a misspecified, but balancing propensity score model. Implemented as after-matching tests, they apply to the unconditional quantiles in the pools of treated and non-treated units, as the matching algorithm (hopefully) eliminates differences in the common support of the propensity score prior to testing. The suggested methods are very powerful when the matched sample is not balanced, but reject the null too often when the balancing property holds. This suggests that we should have more confidence in the CMS and KS full sample tests than in the after-matching versions.

## Notes

1. Note that these tests may also be applied to count data if they are artificially smoothed as outlined in Machado and Silva [30].
2. In contrast, Imbens [20] and Lechner [26] discuss effect evaluation for multiple treatments. The discussion in this paper could be easily extended to their framework.
3. Testing for equality of conditional distributions is discussed in Li *et al.* [28], although for discrete conditioning variables, whereas we need to condition on a continuous  $p(X)$ .
4. We test for equality in mean propensity scores among treated and non-treated units within a stratum at the 10% level of significance.
5. Shaikh *et al.* [39] show that  $f_{p(X)|D=1}(\rho|D=1) = \Pr(D=1)/\Pr(D=0)\rho/1 - \rho f_{p(X)|D=0}(\rho|D=0) \quad \forall \rho \in (0, 1)$ , with  $f_{p(X)|D=d}(\cdot|D=d)$  being the pdf of  $p(X)$  conditional on  $D=d$ , is a testable implication of a correctly specified propensity score and propose a specification test based on kernel density estimation.
6. It is, however, less efficient than estimation based on the true propensity score model.
7. Note that we do not know the exact  $p$ -value of the Koenker and Machado [25] test statistic because Andrews [2] only provides us with the critical values up to the 10% level, but not across the entire distribution.
8. The caliper is set to 0.1 standard deviations of the propensity score and 59 observations (6.5%) are dropped due to a lack of common support.

## References

- [1] A. Abadie, *Bootstrap tests for distributional treatment effects in instrumental variable models*, J. Amer. Statist. Assoc. 97 (2002), pp. 284–292.
- [2] D.W.K. Andrews, *Tests for parameter instability and structural change with unknown change point*, Econometrica 61 (1993), pp. 821–856.
- [3] S. Becker and A. Ichino, *Estimation of average treatment effects based on propensity scores*, Stata J. 2 (2002), pp. 358–377.
- [4] L.M. Berger and J. Hill, *Maternity leave, early maternal employment and child health and development in the US*, Econom. J. 115 (2005), pp. F29–F47.
- [5] P.K. Bhattacharya, *On an analog of regression analysis*, Ann. Math. Statist. 34 (1963), pp. 1459–1473.
- [6] R. Blundell, M.C. Dias, C. Meghirs, and J.V. Reenen, *Evaluating the employment impact of a mandatory job search program*, J. Eur. Econom. Assoc. 2 (2004), pp. 569–606.
- [7] P. Böckerman and P. Ilmakunnas, *Unemployment and self-assessed health: Evidence from panel data*, J. Health Econom. 18 (2009), pp. 161–179.
- [8] M. Buchinsky, *Recent advances in quantile regression models: A practical guideline for empirical research*, J. Hum. Resour. 33 (1998), pp. 88–126.
- [9] V. Chernozhukov and I. Fernández-Val, *Subsampling inference on quantile regression processes*, Sankhya: Indian J. Statist. 67 (2005), pp. 253–276.

- [10] R.H. Dehejia and S. Wahba, *Causal effects in non-experimental studies: Reevaluating the evaluation of training programmes*, J. Amer. Statist. Assoc. 94 (1999), pp. 1053–1062.
- [11] R.H. Dehejia and S. Wahba, *Propensity-score-matching methods for nonexperimental causal studies*, Rev. Econom. Statist. 84 (2002), pp. 151–161.
- [12] A. Diamond and J.S. Sekhon, *Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies*, Institute of Governmental Studies Working Paper, 2006.
- [13] P. Good, *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, Springer, New York, 2001.
- [14] J.J. Heckman, H. Ichimura, and P. Todd, *Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme*, Rev. Econom. Stud. 64 (1997), 605–654.
- [15] J.J. Heckman, H. Ichimura, and P. Todd, *Matching as an econometric evaluation estimator*, Rev. Econom. Stud. 65 (1998), pp. 261–294.
- [16] K. Hirano, G.W. Imbens, and G. Ridder, *Efficient estimation of average treatment effects using the estimated propensity score*, Econometrica 71 (2003), pp. 1161–1189.
- [17] D.G. Horvitz and D.J. Thompson, *A generalization of sampling without replacement from a finite universe*, J. Amer. Statist. Assoc. 47 (1952), pp. 663–685.
- [18] A. Ichino, F. Mealli, and T. Nannicini, *From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity?*, J. Appl. Econometrics 23 (2008), pp. 305–327.
- [19] K. Imai, G. King, and E. Stuart, *The balance test fallacy in matching methods for causal inference*, unpublished manuscript, 2006.
- [20] G.W. Imbens, *The role of the propensity score in estimating dose–response functions*, Biometrika 87 (2000), pp. 706–710.
- [21] G.W. Imbens, *Nonparametric estimation of average treatment effects under exogeneity: A review*, Rev. Econom. Statist. 86 (2004), pp. 4–29.
- [22] J. Jalan and M. Ravallion, *Estimating the benefit incidence of an antipoverty program by propensity-score matching*, J. Business Econom. Statist. 21 (2003), pp. 19–30.
- [23] R. Koenker, *Quantile Regression*, Cambridge University Press, Cambridge, 2005.
- [24] R. Koenker and G. Bassett, *Regression quantiles*, Econometrica 46 (1978), pp. 33–50.
- [25] R. Koenker and J.A.F. Machado, *Goodness of fit and related inference processes for quantile regression*, J. Amer. Statist. Assoc. 94 (1999), pp. 1296–1310.
- [26] M. Lechner, *Identification and estimation of causal effects of multiple treatments under the conditional independence assumption*, in *Econometric Evaluations of Active Labor Market Policies in Europe*, M. Lechner and F. Pfeiffer, eds., Physica, Heidelberg, 2001.
- [27] W. Lee, *Propensity score matching and variations on the balancing test*, unpublished manuscript, 2006.
- [28] Q. Li, E. Maasoumi, and J. Racine, *A nonparametric test for equality of distributions with mixed categorical and continuous data*, J. Econometrics 148 (2009), pp. 186–200.
- [29] J.D. Loecker, *Do exports generate higher productivity? Evidence from Slovenia*, J. Int. Econom. 73 (2007), pp. 69–98.
- [30] J.A.F. Machado and J.M.C.S. Silva, *Quantiles for counts*, J. Amer. Statist. Assoc. (100) (2005), pp. 1226–1237.
- [31] T. Persson, *Currency unions and trade: How large is the treatment effect?*, Econom. Policy 16 (2001), pp. 435–448.
- [32] E.J.G. Pitman, *Significance tests which may be applied to samples from any populations*, Suppl. J. Roy. Statist. Soc. 4 (1937), pp. 119–130.
- [33] P.R. Rosenbaum and D.B. Rubin, *The central role of the propensity score in observational studies for causal effects*, Biometrika 70 (1983), pp. 41–55.
- [34] P.R. Rosenbaum and D.B. Rubin, *Reducing bias in observational studies using subclassification on the propensity score*, J. Amer. Statist. Assoc. 79 (1984), pp. 516–524.
- [35] P.R. Rosenbaum and D.B. Rubin, *Constructing a control group using multivariate matched sampling methods that incorporate the propensity score*, Amer. Statist. 39 (1985), pp. 33–38.
- [36] D.B. Rubin, *Estimating causal effects from large data sets using propensity scores*, Ann. Int. Med. 127 (1997), pp. 757–763.
- [37] J.S. Sekhon, *Alternative balance metrics for bias reduction in matching methods for causal inference*, unpublished manuscript, 2007.
- [38] J.S. Sekhon, *Multivariate and propensity score matching software with automated balance optimization: The matching package for R*, J. Statist. Softw., 2007, forthcoming.
- [39] A.M. Shaikh, M. Simonsen, E.J. Vytlačil, and N. Yildiz, *A specification test for the propensity score using its distribution conditional on participation*, J. Econometrics 151 (2009), pp. 33–46.
- [40] J. Smith and P. Todd, *Rejoinder*, J. Econometrics 125 (2005), pp. 365–375.
- [41] Z. Zhao, *Sensitivity of propensity score methods to the specifications*, Econom. Lett. 98 (2008), pp. 309–319.