On Assessing the Specification of Propensity Score Models

Wang-Sheng Lee*

Melbourne Institute of Applied Economic and Social Research
The University of Melbourne

May 18, 2007

Abstract

This paper discusses a graphical method and a closely related regression test for assessing the specification of the propensity score, an area which the literature currently offers little guidance. Based on a Monte Carlo study, it is found that the proposed regression test has power to detect a misspecified propensity score in the case when the propensity score model is under-specified (i.e., needs higher order terms or interactions) as well as in the case when a relevant covariate is inadvertently excluded. In light of findings in the recent literature that suggest that there appears to be little penalty to over-specifying the propensity score, a possible strategy for applied researchers would be to use the proposed tests in this paper as a guard against under-specification of the propensity score.

JEL Classifications: C21, C52
Key words: Propensity score, Specification, Monte Carlo

*Research Fellow, Melbourne Institute of Applied Economic and Social Research, The University of Melbourne, Level 7, 161 Barry Street, Carlton, Victoria 3010, Australia. E-mail: wang.lee@unimelb.edu.au. Thanks to Jeff Borland, Michael Lechner, Jim Powell and Chris Skeels for comments on an earlier version of this paper. All errors are my own.

1. Introduction

Despite being an important first step in implementing propensity score methods, the effect of misspecification of the propensity score model has not been subject to much scrutiny by researchers. In general, if the propensity score model is misspecified, the estimator of the propensity score will be inconsistent for the true propensity score and the resulting propensity score matching estimator may be inconsistent for estimating the average treatment effect. Recent papers that discuss the issue of the misspecification of propensity scores include Millimet and Tchernis (2007), whose Monte Carlo findings suggest that over-specifying the propensity score model might be beneficial, and Shaikh et al. (2006), who propose a specification test based on a certain restriction between the estimated densities of the treatment and comparison groups. In current practice, a heuristic approach used by some applied researchers to check the specification of the propensity score is the propensity score specification test first suggested by Rosenbaum and Rubin (1984) and elaborated in more detail by Dehejia and Wahba (1999, 2002) (henceforth referred to as the DW algorithm). However, with the possible exception of Lee (2006), the properties of the DW algorithm have to date not been subject to much investigation and its use as a specification test is by no means universally accepted.[1]

This paper discusses a graphical method and a closely related regression test for assessing the specification of the propensity score, an area which the literature currently offers little guidance. The graphical method is an original proposal that integrates a suggestion of Rubin (1984) with the ideas of regression graphics for binary responses in Cook (1996), Cook (1998) and Cook and Weisburg (1999). It is based on using two-dimensional scatter plots and is useful for visualising balance in continuous covariates.

[1] For example, many papers in the published literature that use propensity score methods do not use the DW algorithm as a specification test.

In addition, a more general regression interpretation of the graphical method is made. Based on a Monte Carlo study, it is found that the regression test has power to detect a misspecified propensity score in the case when the propensity score model is under-specified (i.e., needs higher order terms or interactions) as well as in the case when a relevant covariate is inadvertently excluded. However, the test has low power to detect a misspecified propensity score in the case when an irrelevant covariate is included in the propensity score model. Despite this shortcoming of the regression test, as findings in the recent literature by Millimet and Tchernis (2007) suggest that there appears to be little penalty to over-specifying the propensity score, a possible strategy for applied researchers would be to use the proposed tests in this paper as a guard against under-specification of the propensity score.

The layout of the paper is as follows. In section 2, some background information is provided on diagnostics for propensity score methods to set the context for understanding where the graphical diagnostics fit in. In section 3, we introduce the use of two-dimensional scatter plots to guide in the specification of the propensity score. In section 4, we use simulated data to illustrate the graphical test. Section 5 discusses a regression interpretation of the graphical test for dimension reduction and performs a Monte Carlo study. Finally, section 6 concludes.

2. Propensity Score Diagnostics

One advantage of matching over regression based methods for estimating average treatment effects is that diagnostics are available for the former that do not involve use of the outcome variable. Only information on the observable covariates are needed for such a diagnostic so there is no way that performing these diagnostics can systematically influence estimates of the average causal effect. In the case of propensity

score matching, balancing tests that do not involve the outcome variable are often performed to check if the comparison group can be considered to be an adequate counterfactual representation of the treatment group. The idea behind balancing tests is to verify if X has the same distribution for the treatment and comparison groups conditional on the propensity score:

$$D \perp X \mid p(X)$$

where D is the binary treatment group indicator, X is a set of covariates that are chosen to fulfil the conditional independence assumption (CIA) and $p(X)$ is the propensity score. The basic intuition is that after conditioning on $p(X)$, additional conditioning on X should not provide new information on D.

2.1 *Precision in Estimating the Propensity Score*

In general, the literature has placed much less attention on the specification of the propensity score (both the choice of variables and functional form), compared with the issues regarding how the propensity score is to be used (choice of matching algorithm and matching structure).[2] In part, this could be because the results of Drake (1993) suggest that misspecification of the propensity score model does not lead to large biases relative to misspecification of the response model. In addition, Rubin and Thomas (1992a, 1996) and Gu and Rosenbaum (1993) show that much of the balance resulting from matching on $p(X)$ alone can be achieved with relatively coarse matching. For example, when stratification (or blocking) on the propensity score is done, the intuition is that there is less reliance on correct specification of the model used to estimate the propensity score since the probabilities are used only to partially order the sample.

[2] See, for example, Augurzky and Kluve (2007), Frölich (2004), and Zhao (2004) who all emphasize the latter.

However, precision in the estimated propensity score can be important in some cases. For example, in the case when the propensity score is used as a Horvitz-Thompson type estimator based on weighting by the inverse of the estimated propensity scores (e.g., Hirano and Imbens 2001, Hirano, Imbens and Ridder 2003), the propensity score is used directly in the process for estimating treatment effects. Rubin (2004) notes that "[i]n such cases, the estimated probabilities can be very influential on the estimated effects of treatment versus control, and so the probabilities themselves must be very well-estimated." In this case, specification tests that can check the precision of the estimated propensity score would clearly be important. The graphical method discussed in this paper attempts to measure precision in the estimate of the propensity score by checking if there are regions of the propensity score where *X* is unbalanced.

2.2 *Model Checking Diagnostics versus Propensity Score Model Diagnostics*

Little advice is currently available regarding which model or functional form (e.g., logistic models, probit models, semi-parametric models etc.) is best for estimating the propensity score, although slightly more advice is available regarding the inclusion (or exclusion) of covariates in the propensity score model (see Caliendo and Kopeinig 2006, section 3). Regardless of the method used to estimate the propensity score, as propensity scores serve only as devices to balance the observed distribution of covariates between the treated and comparison groups, many model checking diagnostics that pertain to specific model classes are arguably irrelevant. For example, in the case of logistic models, these would be tests like the Pearson χ^2 goodness-of-fit test and the ROC curve. Instead, the success of propensity score estimation should be assessed by the resultant balance rather than by the fit of the models used to create the estimated propensity scores. Rubin (2004) distinguishes between: (a) diagnostics for the

successful prediction of probabilities and parameter estimates underlying those probabilities, possibly estimated using logistic regression; and (b) diagnostics for the successful design of observational studies based on estimated propensity scores, possibly estimated using logistic regression. He states that "there is no doubt in my mind that (b) is a critically important activity in most observational studies, whereas I am doubtful about the importance of (a) in these." Rubin further notes that many applications of propensity score methods have not appreciated this subtle point and continue to use (a) for propensity score diagnostics. The graphical methods discussed in this paper add to the toolbox available for task (b).

2.3 *Before-matching versus After-matching Balancing Tests*

In order to understand the context in which the graphical methods discussed in this paper are relevant, it is important to make a distinction between "before-matching" and "after-matching" balancing tests. Although Caliendo and Kopeinig (2006) classify the DW algorithm as part of their step (iv) (see their Figure 1), we prefer to classify it as part of their step (i) and would only classify the other balancing tests they mention in their step (iv). This is because the DW algorithm is implemented for the full sample that is used to estimate the propensity score model *before* any matching is done (and hence is more like a specification test of the propensity score model) while other balancing tests like two sample *t*-tests and the Hotelling test are done *after* matching and use a smaller subsample of the original data.[3]

[3] For example, when matching is only done over the region of common support, units outside the common support are discarded. Similarly, in one-to-one nearest neighbor matching, all non-matched units are dropped. Lee (2006) provides a more detailed discussion regarding the difference between before-matching and after-matching diagnostics.

2.4 *Choosing the Relevant Diagnostic*

Combining the three concepts discussed above – the relative importance of precision in the estimated propensity score, the difference between model checking diagnostics and propensity score model diagnostics, and the difference between before-matching and after-matching tests – this suggests that diagnostic tests for propensity score methods likely work best when closely aligned with how the estimated propensity scores are to be used.

Before-matching balance diagnostics like the DW algorithm are more relevant if the full sample that is used to estimate the propensity score is to be used for estimating average causal effects. In the case of stratification on the propensity score, as the probabilities are used only to partially order the sample, less precise estimates of the propensity score might be sufficient. However, in the case of weighting on the propensity score, since small changes in estimated probabilities can have large effects on the estimated average causal effects, more precision would be needed. Here, tests that can check for the precision of the estimated propensity score would be very useful.

On the other hand, after-matching balance diagnostics such as the test of standardized differences (Rosenbaum and Rubin 1985) or two sample *t*-tests would appear to be more relevant if a smaller matched sample is used for estimating average causal effects. In this case, the estimated average causal effects could be less sensitive to the estimated probabilities if relative coarse matching is done (e.g., one-to-one nearest neighbor matching). However, precision of the estimated propensity score could once again be more important if the estimated probabilities are used to create weights. For example, in kernel matching, the weight that is given to a comparison group unit is proportionate to the closeness of its characteristics to a treated unit (i.e., the further away the comparison unit is from the treated unit in terms of the estimated propensity

score, the lower the weight). It would therefore appear that tests that can check for the precision of the estimated propensity score would again be helpful in this context.

In the remainder of this paper, two before-matching diagnostics are discussed: a graphical approach (sections 3 and 4) and a closely related regression test (section 5).

3. Dimension Reduction and Regression Graphics

Propensity score matching is a dimension reduction technique that attempts to avoid the problem of the curse of dimensionality when the number of covariates is large. There simultaneously exists a large literature on dimension reduction in statistics. But thus far, there has been little or no connection made between the two parallel literatures of dimension reduction. As we show below, the idea of a dimension reduction subspace for the regression of Y on X (Li 1991), which represents a sufficient reduction in the dimension of the set of covariates X, is closely related to the balancing hypothesis of propensity scores given in Theorem 2 of Rosenbaum and Rubin (1983).

3.1 *Overview of Dimension Reduction Regression*

In the general regression problem, letting n be the number of observations, we have a $k \times n$ response Y (usually k =1) and a $p \times n$ predictor X, and the goal is to learn about how the conditional distribution $F(Y \mid X)$ varies as X varies through its sample space. In parametric regression, we specify a functional form for the conditional distribution. In non-parametric regression, no functional form assumptions are made about $F(Y \mid X)$, but progress is really only possible if the dimensions p and k are small. Dimension reduction regression is one intermediate possibility between the parametric and non-parametric extremes. Dimension reduction without loss of information is a dominant theme of regression graphics. An attempt is made to reduce the dimension of

X without losing information on $Y \mid X$. Although the idea of dimension reduction regression was originally introduced for continuous response variables (Cook 1994), extensions have been made for the case of binary response variables (Cook 1996, Cook and Lee 1999). These extensions are important as it is through these extensions that the ideas from regression graphics can then be related to the propensity score framework in which there is a binary treatment indicator.

In dimension reduction regression (see, for example, Cook 1998 and Cook and Weisburg 1999), it is assumed without loss of information that the conditional distribution $F(Y \mid X)$ can be indexed by d linear combinations of X, or for some unknown $p \times d$ matrix B, with $d \leq p$

$$F(Y \mid X) = F(Y \mid B'X) \tag{1}$$

This statement is equivalent to saying that the distribution of $Y \mid X$ is the same as that of $Y \mid B'X$ for all values of X. It implies that the $p \times n$ matrix of covariates X can be replaced by the $d \times n$ predictor matrix $B'X$ that contains d linear combinations without loss of regression information. In other words, using the conditional independence notation introduced by Dawid (1979), we can write:

$$Y \perp X \mid B'X \tag{2}$$

This represents a potentially useful reduction in the row dimension of the matrix of covariates X from p to d. Such a B always exists because (2) is trivially true when $B = I_P$, the $p \times p$ identity matrix. The relation in (2) can be viewed as a statement about a dimension reduction subspace for the regression of Y on X (see, for example, Cook 1996 and Cook and Lee 1999). The d linear combinations $B'X = (b_1'X, b_2'X, ..., b_d'X)$ are referred to as *sufficient predictors* in the regression graphics literature because together they contain all the regression information that X has about Y.

In the case when $d = 1$, replacing Y with D in (2) (where D is a binary indicator denoting treatment group membership) and the $1 \times n$ predictor vector $B'X$ with the $1 \times n$ vector $p(X)$, it is immediately obvious that (1) is the balancing hypothesis of propensity scores. Having $B = I_P$ corresponds to the propensity score being equal to the "most trivial balancing score", which is when the balancing score is X (see Rosenbaum and Rubin 1983, p. 42).

In the case when $d > 1$, then $p(X)$ is only one of the sufficient predictors. In other words, conditioning only on $p(X)$ will lead to missing relevant information and complete dimension reduction cannot be achieved. Matching on this value of $p(X)$ would not be expected to lead to unbiased average causal effects. One way forward in this situation would be to add higher-order or interaction terms in X to remove the necessity for the $d > 1$ structure. This echoes the ad hoc procedure used in the DW algorithm of fine-tuning the specification of $p(X)$.

3.2 *Binary Response Plots*

For the purposes of this paper, we are interested in using a particular interpretation of an auxiliary plot used in regression graphics for binary variables.[4] In particular, these auxiliary plots are the *binary response plots* in Cook (1996, Figure 1), Cook (1998, Figure 5.3) and Cook and Weisburg (1999, Figure 22.6). Here, we integrate the interpretation of binary response plots with a suggested graphical diagnostic of Rubin (1984) for logistic regression models in order to create a diagnostic for the specification of the propensity score. By viewing Rubin's diagnostic using Cook's insights, a graphical method can be used as a diagnostic to check balance for any continuous X variable, *regardless of the method used to estimate p(X)*. Rather than

[4] The regression graphics literature focuses on estimating d and the subspace $B'X$. An overview of regression graphics is given in the appendix.

trying to estimate $B'X$ and the dimension d in a data set, the graphical diagnostic proposed in this section assumes that the structure of the data can be represented by a 1D model (i.e., $d = 1$) like a logistic model and that $B'X$ is exogenously given. By focusing on the $d = 1$ case, we are attempting to verify if the p-dimensional vector X can be replaced by an exogenously given 1-dimensional vector $p(X)$ without loss of regression information. In other words, the dimension reduction hypothesis is that when the regression vector X is augmented by the propensity score, the $(p+1)$-dimensional vector $(X, p(X))$ can be reduced to $p(X)$ for the conditioning set. Assuming that propensity score matching is feasible and that complete dimension reduction is feasible using a single scalar, the focus is on using binary response plots to check if $Y \perp X \mid B'X$ holds for each X in the data set.[5]

The exposition of how binary response plots can be used to assess balance in the context of propensity score matching is best developed using an example that appears in the literature. Although the context of the original discussion of the example in Landwehr, Pregibon, and Shoemaker (1984) is the specification of the fit of a logistic model or what Rubin (2004) terms "diagnostics for the successful prediction of probabilities and parameter estimates," we illustrate how a slight change in interpreting a graph that arose from the discussion of the paper allows for an assessment of balance to be made or what Rubin (2004) refers to as "diagnostics for the successful design of observational studies based on estimated propensity scores."

3.3 *Example from Landwehr, Pregibon, and Shoemaker (1984)*

This section continues a discussion between Landwehr, Pregibon, and Shoemaker (1984) and Rubin (1984). In Example 4 of Landwehr, Pregibon, and

[5] If $d > 1$ and $p(X)$ is only one of the sufficient predictors, then conditional independence would not be obtained as complete dimension reduction is not possible. The failure to find conditional independence in this case could be because $d > 1$ or that $d = 1$ but that $p(X)$ is specified incorrectly.

Shoemaker (1984, p. 67), one hundred observations are generated according to the model: $\text{logit}(D) = -1 + X_5 + X_6 + 2X_6^2$ (with X_5 distributed -1(0.02)1 and X_6 distributed U(-1, 1)). Presuming that the particular functional form or nature of the dependence on X_5 and X_6 is unknown, they start by fitting a model linear in X_5 and X_6 and calculate their suggested *partial residual plots* (which is discussed in more detail in their paper but not relevant to the present discussion) for X_6. They find that dependence on X_6 is non-linear and that an additional term such as X_6^2 or $|X_6|$ should be included in the model.

In commenting on Landwehr, Pregibon, and Shoemaker (1984), Rubin (1984) suggested that one alternative way of assessing the fit of the logistic regression model is to plot *X* versus the predicted values of the dependent variable (i.e., the propensity score) using different symbols for the *D* = 1 and *D* = 0 points (see Rubin 1984, example 1).[6] But in his written comments, Rubin (1984) did not further discuss how the plot should be interpreted. He also did not make the distinction between diagnostics for the successful prediction of probabilities and parameter estimates underlying those probabilities, and diagnostics for the successful design of observational studies based on estimated propensity scores, which he does more recently in Rubin (2004).

The discussion continued with Landwehr, Pregibon, and Shoemaker (1984a) responding to Rubin (1984). They used the generated data from the model described above, created the plot Rubin suggested based on estimating the misspecified model that is linear in X_5 and X_6 but found that "[this] plot does not reveal to us anything about

[6] Incidentally, another suggestion of Rubin for assessing the fit of a logistic regression model (Rubin 1984, example 3), where he suggested categorising the fitted probabilities and comparing the distributions of *X* for the *D* = 1 and *D* = 0 units within categories of the fitted probabilities, is a precursor and an informal version of the specification test described in more detail in Dehejia and Wahba (1999, 2002).

the inadequate fit of the model nor the need to transform X_6." Moreover, they state that the standard test for comparing two regression lines (the conditional distribution of X_6 given the fitted probabilities for $D = 1$ and $D = 0$) is not significant whereas their partial residual plot clearly demonstrates the basic model inadequacy.

Figure 1 below reproduces Figure 1 from Landwehr, Pregibon, and Shoemaker (1984a). Their focus is on the two regression lines which they highlight are not significantly different.

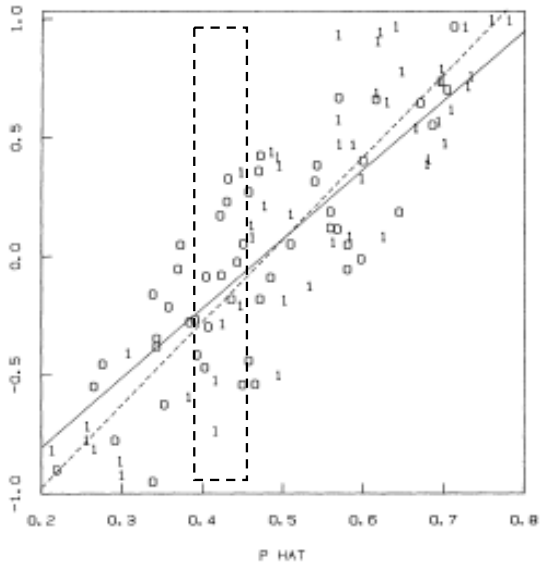

Figure 1: From Landwehr, Pregibon, and Shoemaker (1984a, Figure 1). The plot is a scatter plot of X_6 against the fitted probabilities. Plotted characters show the values of *D*. The solid diagonal line is the least squares regression of X_6 on the fitted probabilities for $D = 0$, the dashed diagonal line for $D = 1$. The meaning of the rectangular window (or vertical slice) is discussed in the text.

However, Rubin's suggestion of plotting *X* versus the fitted probabilities using different symbols for the $D = 1$ and $D = 0$ points has a subtle interpretation that was not noted by Landwehr, Pregibon, and Shoemaker (1984a). This interpretation is similar to how binary response plots are interpreted in Cook (1996, Figure 1), Cook (1998, Figure 5.3) and Cook and Weisburg (1999, Figure 22.6).

Instead of focusing on the two regression lines, imagine viewing Figure 1 as a small rectangular window is moved horizontally across the plot (like the one over the value of PHAT = 0.4). To assess whether $D \perp X_6$ given the fitted probabilities (i.e., whether there is balance in X_6), Cook suggests looking for non-uniform intraslice densities within slices of the fitted probabilities, which effectively holds the fitted probabilities constant. Slices with only one type of symbol are consistent with the conditional independence statement, and so are any slices that have a roughly constant fraction of each type of symbol as one moves from the bottom to the top of the slice. Symbol density does not need to be similar from slice to slice, as long as it is constant within each slice. However, conditional independence (i.e., balance) is violated if in any slice, there is non-random variation in symbol density as one moves from the bottom to the top of the slice.

More formally, following Cook (1996), let the small rectangular window Δ for the *j*th predictor be:

$$\Delta = \{X \mid a_j \leq X_j \leq b_j,\ j = 1, ..., p\}$$

where a_j and b_j are the left and right boundaries of the window. The ratio of the counts in each window can be expressed in terms of the fraction of one symbol in Δ:

$$\frac{\Pr(X \in \Delta, D = 1)}{\Pr(X \in \Delta)} = E(D \mid B'X \in \Delta_B)$$

where $\Delta_B = \{B'X \mid X \in \Delta\}$. This expression states that if $D \perp X \mid B'X$ then the fraction of each symbol should be roughly constant within each Δ. Although one could technically count the ratio of one symbol relative to the total, Cook suggests that "[w]ith a little practice, it seems possible to assess relative symbol density visually without actually counting" (Cook 1996, p. 985).

The rectangular window depicted in Figure 1 (over the value of PHAT = 0.4) shows that within that slice, it is more likely for group '1' to have lower values of X_6 than group '0'. In other words, if we were assessing whether $D \perp X_6 \mid p(X)$ for this chosen model specification, it would be possible to conclude that the conditional independence relation is not supported by the data. Contrary to the conclusion reached by Landwehr, Pregibon, and Shoemaker (1984a), by focusing on rectangular windows instead of the two regression lines in Figure 1, it would therefore be possible to conclude that the model specification that is linear in X_5 and X_6 is inappropriate.

For the remainder of this paper, we refer to this plot as the *Rubin-Cook plot*, as Rubin had first suggested it as a possible diagnostic for fitting logistic regression models while Cook's interpretation of such binary response plots is what makes it useful as a propensity score model diagnostic. Recall that the intuitive idea behind propensity score matching is that there are many miniature experiments at each value of the propensity score (see Zhao 2004). In other words, holding the propensity score constant, there should be no relationship between each covariate and the treatment indicator. The graphical diagnostic described in this section essentially puts into practice this intuitive idea. It is useful as a visual method for checking for balance and precision in the estimated propensity score and can be viewed as a substitute or complement to the DW algorithm.[7]

Although graphical tests like the Rubin-Cook plot that depend on eyeballing the data might be considered subjective and not have the rigorous flavor that statistical tests that involve the use of non-subjective and easy to interpret p-values might have,

[7] From the practical perspective, Rubin-Cook plots can be created by any standard statistical software packages that can create scatter plots. Specialized software like *Arc* that are necessary for regression graphics are not needed in order for the plots to be created.

complements to statistical tests can be useful if they help researchers see different aspects of their data more clearly.

4. Simulated Data Example

In this section, we generate 200 observations according to the model: $\text{logit}(D) = -1 + X_1 + X_2 + 3X_1X_2$, where X_1 and X_2 are drawn from N(0, 1) distributions.[8] As we draw error terms such that they are independent of *X*, the generated data set is balanced by construction.[9] We estimate the propensity score using two specifications. In the first specification (S1), we do not correctly model the interaction term in a logistic model and instead specify a model linear in X_1 and X_2 while in the second specification (S2), we correctly include X_1, X_2 and the interaction term X_1X_2 in our logistic model. The question examined in this section is how well diagnostics for assessing the specification of *p(X)* work under S1 and S2. Will a researcher correctly choose S2 over S1 based on diagnostics performed on the estimated propensity score? As the data generated under S1 are unbalanced by construction while the data under S2 are balanced by construction, we can evaluate the ability of "diagnostics for the successful design of observational studies based on estimated

[8] We draw values from N(0, 1) instead of U(-1, 1) distributions like in Landwehr, Pregibon, and Shoemaker (1984) because matching theory is more developed under the assumption of X having an ellipsoidal distribution (e.g., distributions such as the normal or *t*). For example, Rubin and Thomas (1992b) show that affinely invariant matching methods, such as Mahalanobis metric matching and propensity score matching (if the propensity score is estimated by logistic regression), are equal percent bias reducing if all of the covariates used have ellipsoidal distributions. The seed number used in the simulations in *Stata /SE 9.2* is 55623.

[9] Generating a balanced data set under the null in order to perform the simulations was done as follows. In the binary choice selection equation, because we assume that the error term in the selection equation is independent of the *X*s, when we use the error term, arbitrary values of β and *X* to generate *D*, it is true that:

$$D \perp X \mid X\beta$$

As only monotonic transformations are performed, it therefore follows that

$$D \perp X \mid \text{logit}(X\beta) \quad \text{or} \quad D \perp X \mid p(X)$$

Therefore by construction, these data sets satisfy the balancing property of propensity scores: $D \perp X \mid p(X)$.

propensity scores" to uncover the truth. The issue of which variables to include in *X* is not examined here as it is assumed that the researcher knows that X_1 and X_2 both need to be included in the model in order to fulfil the conditional independence assumption underlying matching estimators, but does not know the functional form they take.

4.1 *Rubin-Cook Plots for Simulated Data*

The Rubin-Cook plots for both covariates X_1 and X_2 are constructed in this section. Figure 2 plots the Rubin-Cook plots under the incorrect specification of the propensity score while Figure 4 does the same using the correct specification of the propensity score.

In the top panel of Figure 3, when assessing balance for X_1, checking for constant symbol density within vertical slices shows that in at least three intervals of the propensity score, there is non-constant symbol density. The findings are similar when checking for balance in X_2 in the bottom panel of Figure 2, where four occasions of non-constant symbol density are highlighted using rectangular windows. As the Rubin-Cook plots in Figure 2 depict that it is not the case that $X_1 \perp D \mid p(X)$ and $X_2 \perp D \mid p(X)$, this suggests that S1 is not a useful balancing score. A parallel and equivalent interpretation is that using S1 as the estimated propensity score does not lead to sufficient dimension reduction so it cannot be concluded that information in *p(X)* can be used to replace information in *X*. It is also worth noting that if the data set contains many more observations (e.g., *n* = 10,000), the Rubin-Cook plot can still be used. The only adjustment that needs to be made is to zoom in on smaller regions of the propensity score. For example, each graph in Figure 3 could be replaced by 10 separate graphs plotting the intervals *p(x)* = 0 to 0.1, 0.1-0.2 etc.

Figure 3 repeats the same exercise for the correct specification of the propensity score. Relative to Figure 2, one immediate obvious finding from the figures is that there is more separation between the $D = 1$ and $D = 0$ observations, with more of the former clustered around high values of $p(X)$ and more of the latter clustered around low values of $p(X)$. Recall that each slice where there are only $D = 1$ or $D = 0$ observations can be ignored when assessing balance. Assessing balance in Figure 3 therefore involves focusing on rectangular windows that start from approximately $p(X) = 0.2$ and end with $p(X) = 0.7$ where there are both $D = 1$ or $D = 0$ observations. Compared to Figure 2, it is clear that the symbols are distributed much more randomly within each vertical slice, reflecting the intuition of miniaturized experiments at each value of $p(X)$. Compare, for example, the highlighted rectangular windows in Figure 3 with the windows in Figure 2. Based on the Rubin-Cook plot for S2, a researcher would be much more likely to conclude that S2 serves as a useful balancing score.

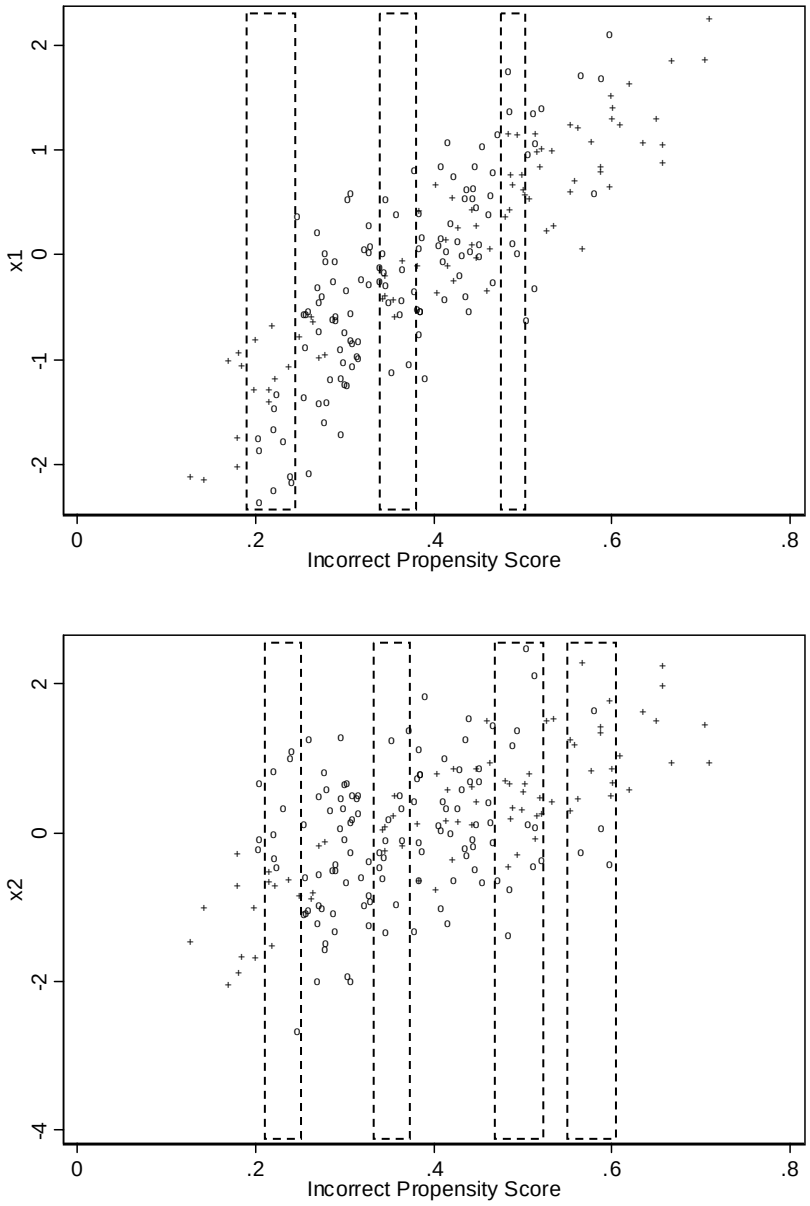

Figure 2: Rubin-Cook Plots for X_1 versus the Incorrect Propensity Score (top panel) and X_2 versus the Incorrect Propensity Score (bottom panel)

Notes: The '+' symbol denotes treatment group members while the 'o' symbol denotes comparison group members.

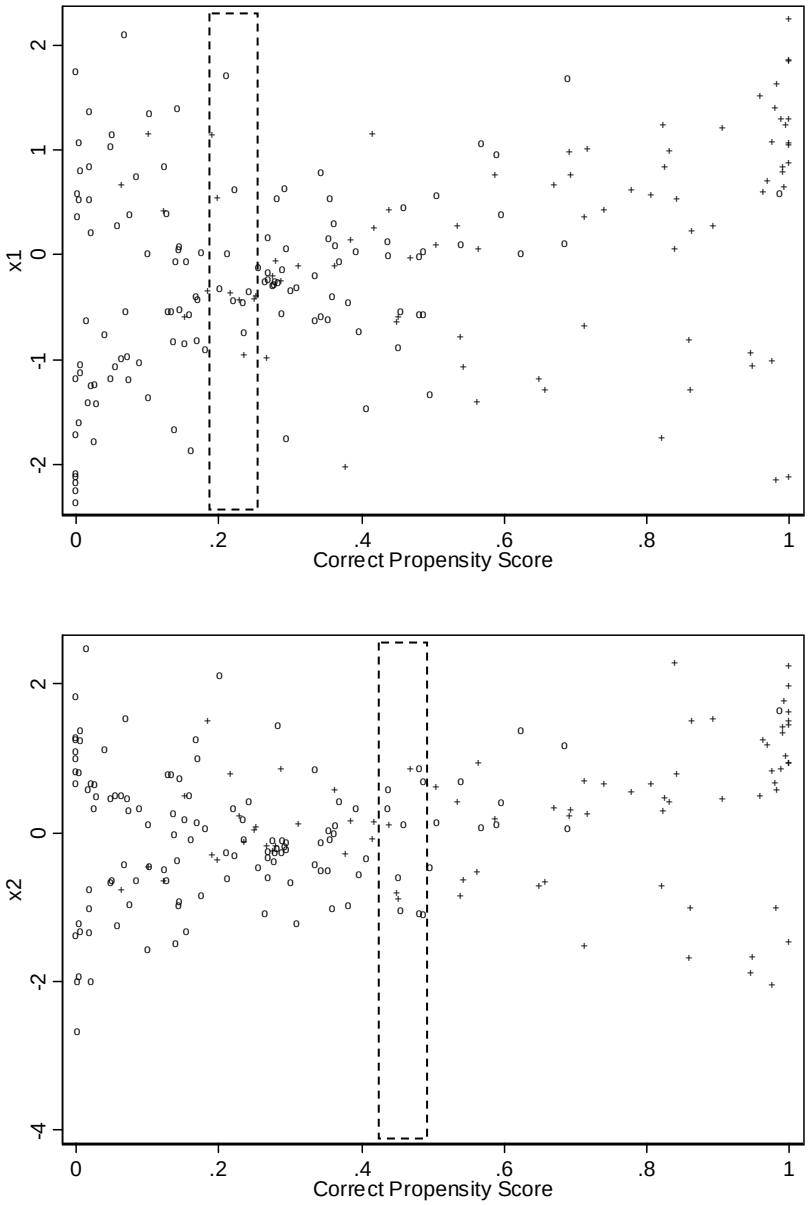

Figure 3: Rubin-Cook Plots for X_1 versus the Correct Propensity Score (top panel) and X_2 versus the Correct Propensity Score (bottom panel)

Notes: The '+' symbol denotes treatment group members while the 'o' symbol denotes comparison group members.

5. A Regression Test Interpretation of the Rubin-Cook Plot

Although Rubin-Cook plots can be informative, their use is rather limited when the number of covariates in *X* is large (as there is one plot per regressor and it is not clear what the implications are if conflicting results for different regressors show up) or when there are categorical variables. For example, in the simulated example in section 4, if X_1 were binary instead of continuous, then a Rubin-Cook plot of X_1 versus *p(X)* would have values of X_1 lined up along $X_1 = 1$ and $X_1 = 0$, making any kind of graphical interpretation impossible, even if the points are jittered. One possible way around this would be to have different Rubin-Cook plots for each value of the categorical variable. For example, in a model with X_1 binary and X_2 continuous, instead of plotting the two graphs of X_1 versus *p(X)* and X_2 versus *p(X)*, one could plot two graphs of X_2 versus *p(X)*, one each for $X_1 = 1$ and $X_1 = 0$. However, if there are many categorical variables in the data set, it is not difficult to see that such an approach can quickly become impractical.

In addition, it is difficult to generalize the results in section 4 from a single simulated data set. Although one might consider performing a Monte Carlo study of Rubin-Cook plots, that would require a method of summarising the information in the plots obtained from the different replications. Unfortunately, it does not appear that there is a simple automated method of summarising information from the plots. We therefore turn to a regression test interpretation of the Rubin-Cook plot that is more amenable to a Monte Carlo study.

Recall the dimension reduction hypothesis which contends that when the regression vector *X* is augmented by the propensity score, the *(p+1)*-dimensional vector (*X*, *p(X)*) can be reduced to *p(X)* for the conditioning set. This hypothesis suggests a

natural regression test. After *p(X)* is estimated, one can estimate the following regression:

$$D = \alpha + \sum X\beta + p(X) \tag{3}$$

An *F*-test can then be conducted to test that all the β coefficients are jointly equal to zero. In other words, under the null, if *p(X)* is sufficient for dimension reduction, after it is conditioned on, information from the vector *X* would no longer provide useful information on *D*. Note that although bearing some similarities in spirit, this regression test is different from the one described in Smith and Todd (2005) who suggest estimating regressions for each variable included in estimating *p(X)* (i.e., if there are 8 covariates there will be 8 regressions estimated), with each *X* in turn as the dependent variable and *D*, *p(X)* and interactions between *D* and *p(X)* as the right-hand-side variables (and possibly higher orders of *p(X)* and interactions between *D* and higher orders of *p(X)*). The regression test described in this paper is a more literal translation of the dimension reduction hypothesis.[10] A Monte Carlo simulation is conducted in the next section in order to better understand its statistical properties.

5.1 *Monte Carlo Simulation*

The setup in the Monte Carlo simulation in this section builds on the setup used in section 3.2.2 of Millimet and Tchernis (2007). They assume that the data generating process for the outcome involves the same set of *X* variables used to generate the propensity score, and their emphasis in their Monte Carlo simulations is on estimating the bias and mean squared error of the estimated average treatment effect. In contrast, the emphasis of the simulations done in this paper is on the performance of a specification test for the propensity score. Therefore, information on outcomes is not

[10] It is also possible to add higher order terms of *p(X)* to the regression, but doing so would make it less closely related to the Rubin-Cook plot.

required and not used in our simulations. The focus is solely on how often the regression test is able to correctly identify the true specification of *p(X)* (size of the test) and how often it correctly rejects incorrect specifications of *p(X)* (power of the test).

Several different setups are considered in the simulations that differ in terms of the number of covariates (two or three) in the true propensity score and the functional form of the true propensity score (linear, quadratic, cubic, pairwise interactions). For each of the setups, we simulate 1000 data sets. Each data set contains four variables for each of the *n* observations (*n* = 1000 or 20000): *D* and the covariates X_1, X_2 and X_3. Following the data generating process in Millimet and Tchernis (2007), the three covariates are independently drawn from a U[0,1] distribution. To simulate treatment assignment, we first simulate the true propensity score *P** for each observation and draw a random number *u* from U[0,1] distribution. We then assign *D* = 0 if *P** < *u* and *D* = 1 otherwise. The propensity scores are simulated from two logit specifications, where:

$$P = \frac{\exp(A)}{1+\exp(A)}$$

and *A* takes the following forms:[11]

I. True specification of *p(X)* includes three variables (X_1, X_2 and X_3)

(a) Linear: $A = 0.33[(-3+6X_1)+(-3+6X_2)+(-3+6X_3)]$

(b) Quadratic:

$$A = 0.33[\ \{2.5-[2.5(1-1.5X_1)]^2\}+\{2.5-[2.5(1-1.5X_2)]^2\}+\{2.5-[2.5(1-1.5X_3)]^2\}\ ]$$

(c) Cubic:

[11] Specifications I(a), I(b), II(a) and II(b) are exactly the same specifications used in Millimet and Tchernis (2007), who perform 1000 Monte Carlo simulations on simulated data sets with 1000 observations. See their Table 3, panels III, IV, VII and VIII, and their Table 4, panels III and IV.

$$A = 0.33[\ \{10-[2.5(1-1.5X_1)]^3\}+\{10-[2.5(1-1.5X_2)]^3\}+\{10-[2.5(1-1.5X_3)]^3\}\ ]$$

(d) Linear and Pairwise interactions:

$$A = 0.33[(-3+6X_1)+(-3+6X_2)+(-3+6X_3)]+8X_1X_2+8X_1X_3+8X_2X_3$$

II. True specification of *p(X)* includes two variables (X_1 and X_2)

(a) Linear: $A = 0.33[(-3+6X_1)+(-3+6X_2)]$

(b) Quadratic:

$$A = 0.33[\ \{2.5-[2.5(1-1.5X_1)]^2\}+\{2.5-[2.5(1-1.5X_2)]^2\}\ ]$$

(c) Cubic:

$$A = 0.33[\ \{2.5-[2.5(1-1.5X_1)]^3\}+\{2.5-[2.5(1-1.5X_2)]^3\}\ ]$$

(d) Linear and Pairwise interactions:

$$A = 0.33[(-3+6X_1)+(-3+6X_2)]+8X_1X_2$$

Based on the above parameters, the treatment-control group ratios in each simulated data set is approximately between 10-90 and 20-80.

Under the true logit specifications I(a) to I(d), the true propensity score model includes the variables X_1, X_2 and X_3 in various alternative specifications. In the simulations, the propensity score is estimated using a logit model with 6 different specifications involving the variables X_1, X_2 and X_3. These 6 specifications constitute a researcher's *guess* as to what the true functional form of *p(X)* might be. Specification 1 adds linear terms for X_1, X_2 and X_3 (first-order linear approximation). Specification 2 adds quadratic terms for each *X* to specification 1. Specification 3 adds interaction terms between each of the *X*s to specification 1. Specification 4 adds both quadratic terms for each *X* and interaction terms between each of the *X*s to specification 1 (second-order

linear approximation). Specification 5 adds interactions between first and second order terms to specification 4. Specification 6 adds cubic terms for each X to specification 5 (third-order linear approximation). When the true propensity score is linear, specification 1 is the correct model. When the true propensity score has linear terms and quadratic terms, specification 2 is the correct model. When the true propensity score has linear terms and pairwise interaction terms but no higher order terms, specification 3 is the correct model. For the case when the true propensity score has linear terms and cubic terms, none of the 6 specifications is the correct model. This particular setup sheds light on the scenario where the researcher correctly guesses the set of variables to include in the propensity score (X_1, X_2 and X_3) but might not correctly guess their functional form.

In the estimated models where a relevant regressor is excluded (to determine properties of the test when a relevant variable is incorrectly excluded in the estimated propensity score model), all terms involving X_3 are excluded in specifications 1 to 6.

The true logit specifications II(a) to II(d) are similar to specifications I(a) to I(d) except that X_3 is excluded from the true propensity score model. The purpose of specifications II(a) to II(d) is to determine what happens when an additional variable is incorrectly included in the estimated propensity score model.

5.2 *Monte Carlo Simulation Results*

The results of the Monte Carlo simulations are given in Tables 1 to 4. Tables 1 and 2 are based on n = 1000 and 1000 replications (as in Millimet and Tchernis 2007) while Tables 3 and 4 are based on n = 20000 and 1000 replications. The latter is an attempt to obtain some large sample properties of the regression test.

In the top panel of Table 1, based on n = 1000, results are shown for the case when both the true model and the estimated model include the variables X_1, X_2 and X_3. Four alternative specifications (a) to (d) for the true model are given in the first four rows, while 6 alternative specifications for the estimated model are given in columns 1 to 6. Shaded cells denote an exact match between the true model and the estimated model. Based on this small sample, it does not appear that the regression test has much power to detect misspecification. A possible exception is when the true model is I(d). It is clear, however, that in general the test does not distinguish between the correct specification and over-specification.

In the bottom panel of Table 1, based on n = 1000, results are shown for the case when the true model includes the variables X_1, X_2 and X_3 but the estimated model only includes the variables X_1 and X_2. In other words, this is the case of having an omitted variable. The regression test in this case has some power to detect misspecification if the researcher guesses that the true model is linear (specification 1) when it is not, and when the researcher guesses that the true model is linear with pairwise interaction terms (specification 3) when it is not. However, it does not appear that the test has much power to detect that the estimated *p(X)* is misspecified in the other scenarios considered. In other words, the regression test does not appear to be powerful when it comes to detecting omitted variables in small samples.

In Table 2, again based on n = 1000, results are shown for the case when the true model includes only the variables X_1 and X_2 but the estimated model includes the variables X_1, X_2 and X_3. With p-values of the F-test > 0.10 in almost all simulations, the regression test appears to have little power to detect misspecification when we include an irrelevant variable in estimating *p(X)*.

We now turn to a similar set of simulations based on n = 20000 instead of n = 1000. In the top panel of Table 3, based on n = 20000, the results for this larger sample are more pronounced and suggest that the test is more powerful in larger samples for detecting misspecification. In the bottom panel of Table 3, the companion results (based on n = 20000) to the bottom panel of Table 1 (based on n = 1000) are shown, suggesting that the test appears to have more power in detecting a misspecification of $p(X)$ due to an omitted variable in larger samples. Finally, in Table 4 which is again based on n = 20000, relative to the test based on n = 1000 in Table 2, the larger sample increases the power of the test to detect misspecification in two scenarios (when the true model is II(d) and when the researcher estimates specification 1 or 2). In general, however, the regression test still does not do a very good job of detecting the fact that an irrelevant variable has been included.

Summarising the results of the Monte Carlo study in Tables 1 to 4:

- *When the correct set of Xs are chosen to estimate p(X), the regression test has some power to identify under-specification but cannot distinguish between the correct specification and over-specification.*
- *When we exclude a relevant X variable in estimating p(X), except for the case when the true p(X) model is linear, the regression test has some power to detect that the estimated p(X) is misspecified.*
- *When we include an irrelevant X variable in estimating p(X), the test has little power to detect misspecification.*

Table 1: True specification of *p(X)* includes 3 variables X_1, X_2 and X_3 (*n* = 1000, 1000 replications)

	Specification of estimated *p(X)*					
	1	2	3	4	5	6
Include correctly x1, x2 and x3 in estimated p(X)						
Linear true *p(X)*	0%	0%	0.1%	0%	0%	0%
Quadratic true *p(X)*	0%	0%	6.4%	0%	0%	0%
Cubic true *p(X)*	0%	0%	0%	0%	0%	0.4%
True *p(X)* includes pairwise interaction terms but no higher order terms	27.3%	15.0%	0%	0%	0%	0%
Exclude incorrectly x3 in estimated p(X)						
Linear true *p(X)*	0.5%	0.1%	0.6%	0.5%	0.2%	0.1%
Quadratic true *p(X)*	58.9%	0%	100%	0%	0%	0%
Cubic true *p(X)*	69.4%	0%	100%	0%	0%	0%
True *p(X)* includes pairwise interaction terms but no higher order terms	0.1%	0%	5.5%	0.6%	0%	0%

Table 2: True specification of *p(X)* includes 2 variables X_1 and X_2 (*n* = 1000, 1000 replications)

	Specification of estimated *p(X)*					
	1	2	3	4	5	6
Include incorrectly irrelevant variable x3 in estimated p(X)						
Linear true *p(X)*	0%	0%	0%	0%	0%	0%
Quadratic true *p(X)*	0%	0%	0%	0%	0%	0%
Cubic true *p(X)*	0%	0%	0%	0%	0%	0%
True *p(X)* includes pairwise interaction terms but no higher order terms	0.2%	0%	0%	0%	0%	0%

Table 3: True specification of *p(X)* includes 3 variables X_1, X_2 and X_3 (*n* = 20000, 1000 replications)

	Specification of estimated *p(X)*					
	1	2	3	4	5	6
Include correctly x1, x2 and x3 in estimated p(X)						
Linear true *p(X)*	0%	0%	0%	0%	0%	0%
Quadratic true *p(X)*	21.9%	0%	100%	0%	0%	0%
Cubic true *p(X)*	73.4%	0%	100%	0%	0%	0%
True *p(X)* includes pairwise interaction terms but no higher order terms	100%	100%	0%	0%	0%	0%
Exclude incorrectly x3 in estimated p(X)						
Linear true *p(X)*	18.9%	2.3%	8.2%	0%	0.1%	0%
Quadratic true *p(X)*	100%	7.2%	100%	92.5%	67.6%	0%
Cubic true *p(X)*	100%	100%	100%	100%	100%	100%
True *p(X)* includes pairwise interaction terms but no higher order terms	19.7%	99.9%	100%	98.7%	33.7%	25.4%

Table 4: True specification of *p(X)* includes 2 variables X_1 and X_2 (*n* = 20000, 1000 replications)

	Specification of estimated *p(X)*					
	1	2	3	4	5	6
Include incorrectly irrelevant variable x3 in estimated p(X)						
Linear true *p(X)*	0%	0%	0%	0%	0%	0%
Quadratic true *p(X)*	0%	0%	5.6%	0%	0%	0%
Cubic true *p(X)*	0%	0%	0%	0%	0%	0%
True *p(X)* includes pairwise interaction terms but no higher order terms	100%	100%	0.1%	0%	0%	0%

Notes: Tables 1 to 4 are simulation results of *F*-tests conducted on equation (3) that all estimated coefficients for *X* are jointly zero after conditioning on *p(X)*. Figures reported in the tables are the percentage of times the p-value of the F-test is < 0.10. A high percentage of p-values < 0.10 implies frequent rejection of the null hypothesis (i.e., estimated *p(X)* is insufficient for dimension reduction). A low percentage of p-values < 0.10 implies non-rejection of the null hypothesis (i.e., estimated *p(X)* is sufficient for dimension reduction, information on *X* is not needed). Shaded areas depict the corresponding correct specification and give an indication of the size of this test. Non-shaded areas depict incorrect specifications and give an indication of the power of this test. The treatment-control ratio is between 10-90 and 20-80 in the simulations. All specifications satisfy the common support condition.

Specifications 1 to 6 increase in the degree of overfitting *p(X)*:
1 = 1 in Millimet and Tchernis (2007). Linear model.
2 is not in Millimet and Tchernis (2007). Includes specification 1 and quadratic terms.
3 = 2 in Millimet and Tchernis (2007). Includes specification 1 and interaction terms.
4 = 3 in Millimet and Tchernis (2007). Includes specification 1 and quadratic terms and interaction terms.
5 = 4 in Millimet and Tchernis (2007). Includes specification 4 and pairwise interactions between first and second order terms.
6 = 5 in Millimet and Tchernis (2007). Includes specification 5 and cubic terms.

6. Conclusion

There is an unambiguous need for more options for assessing the specification of the propensity score, in particular if the propensity score is to be used for weighting. Rubin (2004) highlights the importance of distinguishing between regression model diagnostics like goodness-of-fit tests and tests for the design of observational studies. According to a recent survey by Caliendo and Kopeinig (2006), this is an area which the literature currently offers little guidance. The main contributions of this paper are that it proposes a new graphical approach for propensity score model diagnostics based on two-dimensional scatter plots which we term Rubin-Cook plots, and a closely related and more general regression test interpretation of it.

The DW algorithm when used as a before-matching specification test relies on dividing the data into 'optimal' intervals of *p(X)* and are heavily dependent on how the interval is chosen. In contrast, the flexibility the researcher has in being able to choose meaningful intervals of the propensity score is an advantage of the Rubin-Cook plot. Rubin-Cook plots can help the researcher visualize clearly where regions of thin and thick support for each continuous covariate are and make it easy to assess the quality of randomisation at each small moving window of *p(X)*. Although Rubin-Cook plots can be helpful for making assessments regarding whether the estimated propensity score is a useful balancing score and is precisely estimated, they are not complete tests as they do not allow checks on balance to be made for categorical covariates. In such instances, statistical tests like the chi-square test would still have to be relied on to check for balance in covariates conditional on the propensity score.[12] Here, the problem of how best to condition on the propensity score (e.g., what interval widths to use) remains an open problem.[13]

For applied researchers, a possible simple to implement strategy involves routinely applying a more general regression based test that attempts to literally translate the dimension reduction hypothesis underlying Rubin-Cook plots, and using Rubin-Cook plots as a complementary diagnostic. Monte Carlo simulations suggest that the regression test has power to detect under-specification of the propensity score as well as misspecifications due to an omitted variable. The regression test, however, has little power to detect over-specification and the inclusion of irrelevant variables. As a companion diagnostic, Rubin-Cook plots can be used for verifying the results of the regression test for the continuous *X* variables.

[12] In practice, the *t*-test is often used to test for group mean differences for categorical covariates (as in the DW algorithm) even though it is not specifically designed to do so.
[13] See, for example, a related discussion in Dahiya and Gurland (1973).

Biases in the estimation of average treatment effects using propensity score methods could arise from several avenues. They could arise, for example, from a poor specification of the propensity score, not fulfilling the conditional independence assumption, little overlap in the common support region, imbalance in covariates, or the use of an inappropriate matching algorithm. The emphasis in this paper has been on finding a diagnostic for identifying the true specification of the propensity score, the first stage in the propensity score method. So far, no discussion of the effect of bias in outcomes has been made. A useful set of complementary findings to the simulation results in this paper is provided by Millimet and Tchernis (2007) who focus on how misspecification in the propensity score model can bias estimated average treatment effects. They find little penalty in terms of bias to estimated average treatment effects to over-specification of the propensity score and the inclusion of irrelevant variables in the propensity score model. However, they do not discuss how in practice one decides if a chosen model specification is under-specified or over-specified.

When the insights from this paper are combined with the results from Millimet and Tchernis (2007), a natural implication for applied researchers arises.[14] In order to determine if the estimated propensity score model is under-specified or excludes relevant variables, the regression test described in this paper can first be conducted. After fitting a model (that might include higher order terms and/or interactions) that passes this regression test, Rubin-Cook plots can be used (for continuous X variables) as a complementary diagnostic to further assess relative symbol density at small rectangular windows of the propensity score and verify the results of the regression test.

[14] Recall that the setup of the Monte Carlo simulations and data generating processes in this paper build on the setup used in Millimet and Tchernis (2007). Their results on bias (which assume that the same set of X variables are related to both the probability of treatment and the outcome) can therefore be viewed as results that could have been performed by this author and included in this paper in order to derive implications regarding bias for estimated average treatment effects. While it would be interesting to determine if the same implication arises in scenarios where there are separate sets of variables that are related to the probability of treatment and the outcome (e.g., Brookhart et al. 2006), such work is beyond the scope of the current paper.

Once both these tests are passed, suggesting that the propensity score model is not under-specified, the estimated propensity score can be used as an input for estimating average treatment effects. The results from Millimet and Tchernis (2007) then can be used to argue that even if the researcher guessed incorrectly and over-specified the propensity score model or included irrelevant variables in the propensity score model (both which the regression test has low power to detect), there should be little penalty in terms of bias in estimates of the average treatment effects. This strategy of erring on the side of over-specification also finds support in Rubin and Thomas (1996) who argue that a variable should only be excluded from analysis if there is consensus that the variable is either unrelated to the outcome or not a proper covariate. If there are doubts about these two points, they explicitly advise to include the relevant variables in the propensity score estimation.

References

Augurzky, B. and J. Kluve. (2007). "Assessing the Performance of Matching Algorithms when Selection into Treatment is Strong." Forthcoming in the *Journal of Applied Econometrics*.

Brookhart, M. S. Schneeweiss, K. Rothman, R. Glynn, J. Avorn and T. Stürmer. (2006). "Variable Selection for Propensity Score Models," *American Journal Of Epidemiology*, 163, pp. 1149-1156.

Caliendo, M. and S. Kopeinig. 2006. "Some Practical Guidance for the Implementation of Propensity Score Matching." Forthcoming, *Journal of Economic Surveys*.

Cook, D. (1994). "On the Interpretation of Regression Plots." *Journal of the American Statistical Association*, 89, pp. 177-189.

Cook, D. (1996). "Graphics for Regression with a Binary Response." *Journal of the American Statistical Association*, 91, pp. 983-992.

Cook, D. (1998). *Graphical Regression*. New York: Wiley.

Cook, D. and H. Lee. (1996). "Dimension Reduction in Binary Response Regression." *Journal of the American Statistical Association*, 94, pp. 1187-1200.

Cook, D. and S. Weisburg. (1999). *Applied Regression Including Computing and Graphics.* New York: Wiley.

Dahiya, R. and J. Gurland. (1973). "How Many Classes in the Pearson Chi-Square Test?" *Journal of the American Statistical Association*, 68, pp. 707-712.

Dawid, A. (1979). "Conditional Independence in Statistical Theory." *Journal of the Royal Statistical Society*, Series B, 41, pp. 1-15 (with discussion).

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

Dehejia, R. and S. Wahba. (2002). "Propensity Score Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics*, 84(1), 151-161.

Frölich, M. (2004). "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators." *Review of Economics and Statistics*, 86, pp. 77-90.

Gu, X. and P. Rosenbaum. (1993). "Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms." *Journal of Computational and Graphical Statistics*, 2, pp. 405-420.

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

Landwehr, J., D. Pregibon, and A. Shoemaker. (1984). "Graphical Methods for Assessing Logistic Regression Models." *Journal of the American Statistical Association*, 79, pp. 61-71.

Landwehr, J., D. Pregibon, and A. Shoemaker. (1984a). "Rejoinder." *Journal of the American Statistical Association*, 79, pp. 81-83.

Lee, W. (2006). "Propensity Score Matching and Variations on the Balancing Test." Manuscript. (Available at http://ssrn.com/abstract=936782).

Li, K.C. (1991). "Sliced Inverse Regression for Dimension Reduction (with discussion)." *Journal of the American Statistical Association*, 86, pp. 314-342.

Millimet, D. and R. Tchernis. (2007). "On the Specification of Propensity Scores: with Applications to the Analysis of Trade Policies." Manuscript. (Available at http://faculty.smu.edu/millimet/pdf/mt.pdf).

Rosenbaum, P. and D. Rubin. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, pp. 41-55.

Rosenbaum, P. and D. Rubin. (1984). "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, pp. 516-524.

Rubin, D. (1984). "Comment: Assessing the Fit of Logistic Regressions Using the Implied Discriminant Analysis." Comment on Landwehr, J., D. Pregibon, and A. Shoemaker (1984). *Journal of the American Statistical Association*, 79, pp. 79-80.

Rubin, D. (2004). "On Principles for Modeling Propensity Scores in Medical Research." *Pharmacoepidemiology and Drug Safety*, 13, pp. 855-857.

Rubin, D. and N. Thomas. (1992a). "Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions." *Biometrika*, 79, pp. 797-809.

Rubin, D. and N. Thomas. (1992b). "Affinely Invariant Matching Methods with Ellipsoidal Distributions." *Annals of Statistics*, 20, pp. 1079-1093.

Rubin, D. and N. Thomas. (1996). "Matching Using Estimated Propensity Scores: Relating Theory to Practice." *Biometrics*, 52, pp. 249-264.

Shaikh, A., M. Simonsen, E. Vytlacil and N. Yildiz. (2006) "On the Identification of Misspecified Propensity Scores." Stanford University. Manuscript. (Available at http://www.econ.au.dk/vip_htm/msimonsen/matching.pdf)

Smith, J. and P. Todd. (2005). "Rejoinder." *Journal of Econometrics*, 125, pp. 365-375.

Zhao, Z. (2004). "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence." *Review of Economics and Statistics*, 86, pp. 91-107.

Appendix: Structural Dimension in Graphical Regression

The minimal number *d* of sufficient predictors is called the *structural dimension* of the regression. A model where (2) holds for dimension *d* is referred to as having *d*D structure. If $d < p$, then a sufficient reduction in the regression is achieved which in turn leads to *sufficient summary plots* of *Y* versus $B'X$ as graphical displays of all the necessary modelling information for the regression of *Y* on *X*. As discussed by Cook (1996) and Cook and Lee (1999), having binary response variables instead of continuous response variables cause no conceptual complications, but construction and interpretation of summary plots must recognize the nature of the response.

Suppose that *X* comprises of two variables. Dimension reduction to $d = 1$ would mean reducing two covariates to the single linear combination $B'X = b_1'X$ without any evidence in the data that this reduction would result in loss of information on $Y \mid X$. An example of a 1D model is the single index model:

$$\begin{aligned} Y &= m(\alpha + B'X) + \varepsilon \\ &= m(\alpha + b_1'X) + \varepsilon \end{aligned}$$

where $\varepsilon \perp X$ and *m* is a link function (e.g., the link function is the identity function in the case of multiple linear regression). Alternatively, a regression has 2D structure if two linear combinations $B'X = (b_1'X, b_2'X)$ are needed to characterize the regression, so that *Y* is independent of *X* given $b_1'X$ and $b_2'X$. An example of a 2D model is:

$$Y = m(b_1'X, b_2'X) + \varepsilon$$

where b_1 and b_2 are not collinear. More generally, in a *d*D model, $B'X = (b_1'X, b_2'X, ..., b_d'X)$ and all the regression information is contained in the *d* linear combinations $(b_1'X, b_2'X, ..., b_d'X)$.

Related to the idea that there exist many balancing scores for propensity score matching (and that controlling for any balancing score is sufficient for the theory of propensity score matching to be valid), sufficient predictors are not unique. If $B'X$ is a vector of d sufficient predictors and A is any $d \times d$ full rank matrix then $AB'X$ is another set of sufficient predictors. In practice, however, this non-uniqueness of sufficient predictors is not an important issue in regression graphics as the distribution of $Y \mid B'X$ and $Y \mid AB'X$ contain the same statistical information so sufficient summary plots of Y versus $B'X$ and Y versus $AB'X$ would be identical.

An emphasis of the literature on regression graphics has been to find the B of lowest possible dimension d for which the representation in (2) holds and to use sufficient summary plots as a guide to formulate appropriate models for $F(Y \mid X)$.[15] For example, if d = 2 and B is known, then a three-dimensional plot of Y versus $(b_1'X, b_2'X)$ can be used as a sufficient summary plot for the regression. In general, both d and B are usually unknown and need to be estimated. In general, if non-linearities are present and not represented by the predictor variables, then the dimension of the regression cannot be 1D. Cook (1996) discusses in more detail in the context of a logistic model how a specially written *Xlisp-Stat* program *Arc* for regression graphics can be used to check the structural dimension of the regression and help assess candidate models.[16] This graphical assessment involves rotating graphical displays based on lower dimensional projections of the data in *Arc* to get the "best" visual fits. See, for example, Figure 2 in Cook (1996) where it is illustrated how the goal of the visual fit is to stop rotation at a point where the relative intraslice density is constant in any slice parallel to the stationary vertical axis. The horizontal axis in such a plot

[15] Techniques for estimating the dimension of a regression include sliced inverse regression, sliced average variance estimation and principal Hessian directions. See Cook (1998) and Cook and Weisburg (1999) for more discussion.

[16] This software is available at no cost at: http://www.stat.umn.edu/arc.

corresponds to a particular combination of the predictors $B'X$ that *Arc* can provide estimates of.

An estimate of the subspace spanned by B is obtained once the dimension and directions have been estimated. Given the estimate of the subspace, the next phase involves recovering information about F and identifying the best model for the reduced data, where both parametric and non-parametric models can be used. Cook (1998) and Cook and Weisburg (1999) provide more details on how graphical regression can be implemented in practice.