Using Entropic Distance Metric to Rank Control Distributions in Evaluation Studies.

Adeola Oyenubi

PhD Student
University of Cape Town (UCT)
South Africa
OYNADE001@myuct.ac.za

Working Paper

(Please do not cite)

Summary of Hypothesis and Results

In causal evaluation failure to compare like with like can result in substantial bias in the treatment effect estimate. This is often referred to as the balancing condition or Common Support Assumption (CSA). Under experimental situations randomization, if successful, produces a control group that meets the requirement of being a credible counterfactual because its relevant covariate distribution is identical to the covariate distribution of the treatment group. For quasi-experiments one can often define plausible control groups based on theory or some proximity factor. However, these control groups may not have the property of having a covariate distribution that is identical to that of the treatment group. This paper introduces a novel way to assess how close the covariate distribution of a given control group is to the treatment group whose counterfactual it is supposed to represent and the implication this has for bias and robustness.

The proposed approach quantifies balance by measuring the entropic distance between summarized characteristic distributions of conditioning variables across treatment status. Based on the property of randomized control groups we posit that a theoretically plausible control that is "closest" to the treatment group by our proposed proximity metric will perform better than control groups that are further away. This performance is measured in terms of lower bias and/or robustness of treatment effect estimate.

To investigate the proposition we rank plausible control groups in the popular National Supported Work Demonstration (NWS) programme in terms of their proximity to the treatment group. We show that the mean balancing approach masks differences in characteristic distribution at other parts of the distribution and this may result in substantial bias and less robust treatment effect estimates. Furthermore, we show that using the entropic distance approach allow for comparison of balance across conditioning variables since this metric is a function of the distribution and not the levels of the conditioning variables.

This paper is written under the supervision of Patrizio Piraino (UCT) and Martin Wittenberg (UCT). However all errors therein are mine as this paper has not been fully scrutinized by my supervisors.

1. Introduction

The last chapter (first substantive chapter of thesis) introduced the entropic distance measure as a way of assessing violation of Common Support. It was argued that causal inference required identical characteristic distribution in the treatment and control groups. The entropic measure assess balance by summarizing the information from all the moments of the characteristic distributions being compared as against a few moments (typically mean and variance). The efficacy of the entropic measure was demonstrated by a simulation study. It was shown that control samples whose propensity score distribution are closer to the treatment sample of interest yield better results in terms of bias and robustness of treatment effect across econometric techniques.

In this chapter (or paper) we investigate the performance of this measure on real data. This serves as a more realistic setting than the simulation study of chapter 2. Using data for which we have both experimental control group and theoretically plausible control groups from survey data sets we show that: (i) Entropic measure is correlated with the size of bias when different theoretically plausible control samples are used to estimate Average treatment Effect (ATE); i.e. results from control samples with higher entropic distance from the treatment sample are more biased compared to the experimental benchmark than control samples with lower entropic distance in the case of simple average treatment effect; (ii) For a control sample whose entropic distance is "close" to the treatment sample, different econometric techniques gives similar result for ATE i.e. ATE is more robust when treatment and control samples are balanced as measured by the entropic distance metric. In general we show that the entropic measure is a more reliable way of assessing balance.

We assess the performance of the entropic measure on real data using the National Supported Work Demonstration (NWS) programme data. This data contains observations from a Randomized Control Trial (RCT). In our analysis, ATE estimate from the RCT (i.e. using the randomized control group) is used as the benchmark. Following Lalonde (1986), plausible control samples from the Panel Study of Income Dynamics (PSID) and Westat's matched Current Population Survey (CPS) were used to calculate ATE in a non-experimental setting. Lalonde (1986) found that treatment effects vary by control group. He also found that most of the ATE calculated using different control groups and under different econometric techniques deviate substantially from the experimental benchmark with no way of detecting which of the estimates is closer to the benchmark. Using the entropic distance measure we show that the size of bias and variability of ATE under different samples can be explained by balance/lack thereof as measured by the entropic distance metric. In other words, when CSA assumption holds, treatment effect estimates tend to be robust across econometric methods. This is similar to what randomization achieves under RCTs and suggests that the choice of econometric technique and/or identification assumption becomes less important when balancing condition holds.

2. Literature Review

The link between bias, robustness and balance is well highlighted by the papers that have previously analysed the NWS programme. Lalonde (1986) use the NWS data set to evaluate

the performance of non-experimental estimators using experimental estimates as a benchmark. His results suggest that it is unlikely for an econometrician to recover a treatment effect estimate that is comparable to the one that would have been obtained under a RCT as such estimates are often biased and sensitive to econometric approach and/ or model specification. He showed this for simple ATE calculated as difference in mean outcome and using standard non-experimental estimators[1] to adjust for bias. The variability (or lack of robustness) in estimates under various econometric techniques observed by Lalonde (1986) is not necessarily surprising given that the result from a non-experimental approach depends on the assumption(s) that validate the approach. When there are no selection problems, at most one set of assumption will be satisfied (Smith & Todd, 2004).

Dehejia and Wahba (1999, 2002) use a subsample (henceforth referred to as the DW Sample) of the NWS data used by Lalonde (1986) (henceforth referred to as the Lalonde Sample). They showed that Propensity Score Matching (PSM) can be used to obtain estimates that are more comparable with the experimental benchmark estimates under the assumption that selection is on observables. They attributed the success of PSM to its ability to flexibly control for observable differences by selecting a subset of the PSID and CPS control sample that are more comparable with the treated units in the NWS programme. Their result suggests that adopting the right econometric method reduces bias.

In contrast, Smith and Todd (2004) found that estimates of the impact of NWS based on PSM are highly sensitive to both the set of variables included in the propensity score equation and the particular analysis sample used. In their words PSM does not constitute a "magic bullet" that solves the selection problem in every context. Instead, the goal should be to develop a mapping function from characteristics of data and institutions available in a particular evaluation context to the optimal non-experimental estimator (Smith and Todd, 2004). They showed that a difference-in-difference matching estimator or Conditional difference-in-difference (CDiD) exhibits better performance than cross-sectional estimators (such as PSM) for the NWS example because it controls for time invariant unobservables like differences in survey instruments and labour market conditions.

Literature therefore suggests that correcting for imbalance through the right econometric approach reduces bias in non-experimental situations. Nevertheless, this does not solve the problem of lack of robustness i.e. sensitivity to econometric method and model specification. Those we take a different approach and consider the characteristics of a control group that reduces bias and increase robustness across econometric methods.

Treatment effect estimates vary under non-experimental conditions because the assumption(s) that potential outcomes are independent of treatment status[2] is valid to different degrees in different samples. This assumption requires that a plausible control has identical distribution with the treatment sample i.e.

[1] Lalonde (1986) considered Regression, Difference-in-difference and latent variable selection approach.
[2] This assumption is the same as the Conditional Independence Assumption (CIA) and the Common Support Assumption (CSA).

$$f(W|D=1) \equiv f(W|D=0) \ldots \ldots (1)$$

Where W represent relevant characteristics and $f(W|D=i)\ i=0\ or\ 1$ is the conditional distribution of observed and unobserved characteristic(s) given the treatment status (1 for treated and 0 for untreated). This means that randomization when done properly guarantees (asymptotically) that treatment status is independent of observable and unobservable covariates.

RCTs work, not because randomization removes the selection bias but because it balances the bias in the treatment and control groups (Heckman & Smith, 1995). The implication is that under RCTs, balance between the treatment and control group is the way Conditional Independence Assumption (CIA) is mitigated (i.e. by balancing the bias due to unobserved variables in both groups). Therefore under an RCT the two groups will have empirical distributions that are comparable in terms of shape and support as indicated in equation 1. Under equation 1, the difference in outcome between the two groups will have unbiased causal interpretation because controlling for covariates will not be necessary when this condition holds (Iacus *et. al.*, 2011).

This highlights the central role of the balancing property in treatment effect estimation. One can then conclude that for causal inference the "best" counterfactual is the experimental control group. Therefore, under RCT, impact estimates are robust irrespective of econometric method i.e. treatment effect estimates are less model dependent. An Impact estimate computed as simple mean difference often does not depart significantly from the one calculated under any other econometric method. This is why RCTs are often referred to as gold standard when it comes to causal evaluations. Estimates from randomized control groups are often robust across econometric methods.

For quasi-experimental methods the "holy grail" is an analysis that recovers a treatment effect estimate that is comparable to the one that would have been obtained under an RCT (Shadish, 2013). Achieving this involves finding a control group that satisfies a unique condition obtainable under RCT (at least in terms of observable characteristics). If one can somehow find a control group that has similar shape and support as the treatment group i.e. verify equation 1, the results from such quasi-experiments are more likely to be close to what would have been obtained under an RCT. Although it must be noted that even when this is achieved it is not clear if one can infer that unobserved attributes are balanced like under RCT since in this case balance is not by randomization. However, one can safely assume that the more balanced observable attributes there are in a particular setting the greater comfort one would tend to have about how balanced the unobservables are (here we mean relevant unobserved variables). This would be the case when the observables are correlated with the unobservables (Imai *et. al* 2008).

To estimate an unbiased treatment effect various econometric techniques are used to correct for possible violation of equation 1. These methods could be considered a way of obtaining estimated treatment effect that is close to what would have been obtained under an RCT. While these econometric approaches offer a way to estimate an unbiased

treatment effect under some assumptions, having a balanced sample is often preferred for two reasons. First, when the assumption(s) that validate a particular method is not testable, justifying the result from such method becomes more difficult. Second, it is often desirable to check the robustness of results but when covariates are not balanced robustness of result across methods or/and model specification[3] should not be expected. This is because at most only one of the various econometric approaches available will yield valid results (Smith and Todd, 2004).

The literature highlights the importance of balance for causal inference under experiments or quasi-experiments (see Heckman, Ichimura & Todd (1997); Hainmueller (2011); Imai *et. al* (2008); King & Zeng (2006); Stuart *et. al.* (2013) and Lechner (2001)). However, the literature is thin on how to verify that balancing condition actually holds in data. Our approach introduces a way to quantify balance in a way that is stronger than comparing the first moment across treatment status and in a way that different level of balance can be assessed. Being able to differentiate levels of balance in different samples can help rank plausible control group in terms of their ability to replicate experimental results.

3. Data

To show the relationship between bias and robustness of a treatment effect estimate as stated in (i) and (ii) (section 1) we use the NWS dataset used by Lalonde (1986), Dehejia and Wahba (1999, 2002) and Smith and Todd (2004) among others[4]. The experimental sample includes male respondents in the NWS's ex-addict, ex-offender and high school dropout target groups. The original Lalonde experimental sample includes 297 treatment and 425 control observations. Dehejia and Wahba (1999, 2002) used a subsample of Lalonde sample. Their sample has 185 treatment and 260 control observations. This sample was selected to include two years of pre-program earnings which eliminated about 40% of observations in the original Lalonde sample. It is this ability to control for pre-programme earning that is the major difference between the two samples. The non-experimental control groups CPS and PSID samples contain 15,992 and 2,490 potential control observations respectively. Lalonde (1986) defines plausible control groups that are subsamples of the PSID and CPS data.

TABLE 1:Sample Description based on Lalonde (1986)

PSID-1	All male hh heads continuously from 1975 through 1978 who are less than 55 years old and did not classify themselves as retired in 1975
PSID-2	Selects from PSID-1 all men who were not working when surveyed in the Spring of 1976.
PSID-3	Selects from PSID-1 all men who were not working when surveyed in either 1975 or 1976.
CPS-1	All males based on Westat's criteria, except those over 55 years old
CPS-2	Selects from CPS-1 unemployed males in 1976.
CPS-3	Selects from CPS-1 unemployed males in 1976 whose income in 1975 was

[3] Hainmueller (2011) showed using the NWS treatment group with PSID control that when covariates distribution are balanced treatment effect is invariant to model specification.

[4] The data is available online from "http://users.nber.org/~rdehejia/data/nswdata2.html".

below the poverty level.

Table 1 above shows how these subsamples were selected from the original PSID and CPS dataset. The subsamples can be thought of as crude matched samples of the original control data sets or matching on a few variables (gender, age, employment status in 1975 and 1976). This represents plausible control groups that a researcher might want to use to estimate the impact of the NWS programme. Even though each of the sample represent a plausible control group one would expect sample 3 to be a better counterfactual for the NWS treatment group than sample 2 and sample 2 in turn to be better than sample 1 because of Ashenfelter's dip (Ashenfelter, 1978). Literature suggests that it is important to look at several years of pre-intevention earnings in determining the effect of job training programmes (Ashenfelter 1978; Dehejia & Wahba 1999). Theory and literature would suggest that based on the way the subsamples are selected, samples that match on pre-programme employment status should be better counterfactuals for the NWS treatment group. We show that the proposed entropic metric provide evidence that supports this theory. For more details on the data see Lalonde (1986) and Dehejia and Wahba (1999 & 2002).

3.1 Method

Our proposed approach of assessing balance among various control groups involves two steps. In the first step we summarize the multivariate distribution of conditioning variables with the univariate propensity score distribution. This approach is often used under PSM (Rosenbaum and Rubin, 1983). It has also been used to identify region of Common Support (screen data) before using other methods like regression on the reduced sample (see Crump *et. al.* 2006; Angrist & Pischke 2008). In the second step the S_ρ metric is used to quantify the entropic distance between the distribution of propensity scores in the treatment and control groups. The measure is given by

$$\mathrm{S}_\rho = \frac{1}{2}\int_{-\infty}^{\infty}\left(f_1^{1/2} - f_2^{1/2}\right)^2 \mathrm{dx}$$

Where f_1 and f_2 represent the density of the two distributions being compared (treatment and control in our case). Crump *et. al.* (2006) and Angrist and Pischke (2008) use this density comparison approach to screen their sample (identify region of Common Support) before using regression to estimate treatment effect on the screened sample. In our case we use the propensity score distribution to approximate the "difference" between joint distributions of conditioning variables across treatment status.

This idea is based on the result of Rosenbaum and Rubin (1983) which suggests that conditioning on propensity scores is equivalent to conditioning on the set of covariate (W) under PSM. The assumption here is that the entropic distance between estimated propensity score distributions will provide credible signal of how balanced the sample is. So that relevant differences in the multivariate distribution of characteristics is captured in part by the difference in their corresponding propensity score distributions.

After quantifying the entropic distance between the treatment group and various theoretically plausible control groups, the control groups are then ranked by their entropic distance from the treatment. Our hypothesis is that treatment effect (calculated as difference in mean outcome after treatment) estimated from control groups that are further away from the treatment group are more biased than the ones calculated from control groups that are closer to the treatment group. Furthermore treatment effects from control groups that are close to the treatment group are more robust across econometric methods when compared with the ones are not. We define

$$\gamma_i = \| f(W|D=1) - f_i(W|D=0) \| \dots\dots (2)$$

$i = 1 \dots \dots . k$. Here i is an index of the space of plausible (theoretically sound) control distributions and k is the number of such control distributions in this space. γ_i is the entropic distance between the treatment group and the i^{th} control group. We seek a control group that minimize the entropic distance in equation 2 among a finite set of plausible control groups.

$$\gamma_{min} = \min_{f_i(W|D=0)} \{ \gamma_i \} \dots\dots\dots (3)$$

Note that this comparison only makes sense if the same of W are available in both groups. The comparison would be consistent if there are different set of controls in the control groups being compared[5].

For a control group obtained by randomization one would expect its propensity score distribution to be very close to that of the treatment group so that we can write $\gamma_{RAND} \cong 0$ (where γ_{RAND} is the entropic distance associated with the Randomized control group, we assume here that randomization has been successful). However, under a non-experimental setting it is unlikely that a control distribution with $\gamma_i = 0$ will be found. Therefore, what an econometrician should seek is a control distribution that satisfies equation (3) in the universe of plausible control groups.

For both the PSID and CPS samples, γ_{min} represents the minimum distance between plausible control groups from a particular dataset (e.g. PSID 1, 2 or 3) and the treatment group. Treatment effect calculated from the control group associated with γ_{min} should result in lower bias (using simple difference in mean outcome) and is expected to yield more robust treatment effect estimate across econometric techniques. For our analysis we can write

$$\gamma_{PSIDi} = \| f(W|D=1) - f_{PSIDi}(W|D=0) \|$$

$i = 1 \dots 3$. Where PSID1 are male respondents selected from the PSID sample that are less than 55years old (exact binary matching on gender and age), PSID2 is a subsample of PSID1 selected by unemployment duration of one year prior to the programme (exact matching on gender, age and employment status one year prior to the programme), and PSID3 is a

[5] Note that this point has slightly bigger reach, because even if both datasets have nominally the same variables (e.g. income) but measure them in different ways this might mean that the Ws are not the same

subsample of PSID1 selected using unemployment duration of two years (exact matching on gender, age and employment status two years prior to the programme). See table 1 for details of how these subsamples were selected. We can also write

$$\gamma_{CPSi} = \| f(W|D=1) - f_{CPSi}(W|D=0) \|$$

$i = 1 \ldots 3$. Where CPS 2 and 3 are subsamples of CPS1 similar to the PSID subsamples.

To assess the robustness of treatment effect estimates under various econometric techniques we also estimate the treatment effect under various techniques. We considered (along with simple unadjusted average mean difference) unadjusted difference in difference, difference in difference controlling for covariates, regression and various weighting schemes under PSM. These include Nearest Neighbour, Radius, Kernel, Stratification and Conditional difference in difference which combines PSM with difference in difference method.

4. Results

4.1 Univariate Case

We start the analyses by exploring the possibility of using the entropic distance measure to assess how different each conditioning variable is across treatment and control group in the various NWS samples. We restrict the analyses to continuous variables. Tables 2 A & B show the univariate comparison of control variables analysed in this section. The table compare the NWS treatment and various control groups in terms of the balance between covariate distributions. The control groups are NWS experimental control (Randomized control), DW control samples (PSID and CPS 1, 2 and 3) and Lalonde control samples (PSID and CPS 1, 2 and 3). In these tables we compare the conventional mean difference and the entropic distance between the selected covariates i.e. re75 (income in 1975), re74 (income in 1974 applicable to DW sample only), Education and Age. Mean_re75 and Dist_re75 for example, represent the mean difference and the entropic distance between distributions of income of respondents in 1975 respectively.

Tables 2A and 2B show the result for DW and Lalonde samples respectively. The calculation in both tables is for the screened sample[6]. This is done to mitigate bias due to difference in support. The implementation involves using a propensity score specification that balances covariates in terms of means in the treatment and control groups as suggested by Dehejia and Wahba (2002) to screen the samples. Then mean difference and distributional difference is estimated for each of the variables considered. A table (table 2C) showing the propensity score specification for each treatment/control combination is included in Appendix. In estimating the propensity scores all variables that are theoretically relevant are included in the propensity score specification. Then following the advice in literature (Dehejia & Wahba, 2002) we included interaction terms of variables that are not balanced.

[6] i.e. sample is restricted to the region of Common Support based on propensity scores. The general pattern is not very different for the unscreened sample (result not presented here).

Table 2A: Mean and Entropy test DW Sample (Screened)

	mean_re75	Dist_re75	mean_re74	Dist_re74	mean_Edu	Dist_Edu	mean_Age	Dist_Age
Trt vs PSID1	-7753.53	0.717648	-8916.47	**1.003847**	-0.79324	0.051801	-4.30126	0.044807
Trt vs PSID2	-3056.74	0.192647	-5410.69	**0.589722**	**-0.18347**	**0.02045**	-4.63757	0.042118
Trt vs PSID3	**-961.349**	**0.08566**	-2449.57	**0.30232**	**-0.21466**	**0.02606**	-4.18378	**0.038868**
Trt vs CPS1	-3308.19	0.122598	-3813.54	**0.248029**	-0.57655	0.034827	**0.198932**	0.021592
Trt vs CPS2	-1961.13	**0.076152**	-2406.61	**0.210576**	**-0.28359**	**0.006958**	2.515886	0.041067
Trt vs CPS3	-587.009	**0.035401**	-1740.95	**0.22045**	**0.095946**	**0.012615**	**0.460953**	0.020783
NWS	**245.3906**	**0.215515**	**228.6073**	**0.215191**	**0.245544**	**0.019661**	**0.543124**	**0.006992**

Table 2B: Mean and Entropy test Lalonde Sample (Screened)

	mean_re75	Dist_re75	mean_re74	Dist_re74	mean_Edu	Dist_Edu	mean_Age	Dist_Age
Trt vs PSID1	-6965.83	0.185794			-0.82962	0.084095	-6.52262	**0.082274**
Trt vs PSID2	-3540.18	0.106071			**-0.26882**	0.06782	-8.71971	**0.120516**
Trt vs PSID3	**-22.6198**	**0.026746**			**-0.18275**	0.05052	-7.845	**0.108823**
Trt vs CPS1	-8254.87	0.172296			-1.25108	0.277864	-4.39955	**0.049912**
Trt vs CPS2	-3140.11	0.052697			-0.73015	0.041081	**-0.85278**	**0.03399**
Trt vs CPS3	**514.5977**	**0.011305**			**0.060674**	0.02698	-1.86105	**0.041457**
NWS	**26.99831**	**0.011859**			**0.202746**	**0.00932**	**0.161808**	**0.007333**

*Bold text not significant at 5%

The “pscore” routine (Becker, Ichino *et. al.,* 2002) in Stata is used to estimate the propensity scores.

To verify the mean balancing condition, the routine divides the propensity score distribution into blocks such that mean propensity score within each block is not significantly different across groups using t-test. Furthermore the routine checks for mean balance in each variable for each block. The text in ***Bold*** in tables 2A & B represent cases where (in the full population) the mean difference or the entropic distance is not significant at the 5% level[7].

In contrasting the entropic distance measure with the mean balancing approach it is important to note one key advantage of the distance measure (which is also our main result in this section). The entropy measure is a function of the distribution of the variables being compared and not a function of the actual values of the variable like the mean. Since the difference in mean is a function of the data points the magnitude of the test statistic cannot be compared across variables. The entropy measure on the other hand can be used across variables, therefore the entropic distance tell us how bad balance is in one variable compared to another.

For example, if we consider only the magnitudes (ignoring the hypothesis test), it is fair to say that the distribution of income in 1974 (Dist_74) shows more imbalance than the distribution of income in 1975, education and age (Dist_75, Dist_edu & Dist_age) across groups defined by treatment status. Specifically for table 2A, even though the imbalance in each variable tends to reduce as we move from sample 1 to 3 (as predicted by Ashenfelter’s dip), the biggest shape difference in each instance is associated with income in 1974. Under the assumption that pre-treatment labour condition matter this observation suggests that this variable is a potential source of bias in the DW sample. This agrees with the theory that variables related to employment history are important in evaluation of job training programs (Ashenfelter, 1978). Therefore the entropic distance measure in this case confirms what theory suggests that could not be confirmed using the conventional mean difference approach.

This kind of information may help a researcher to select which variable to do crude matching on. As noted by Smith and Todd (2004) this is the reason why the Dehejia and Wahba (2002) got better results than Lalonde (1986). By accounting for 2 years of pre-treatment income DW sample was able to better solve the evaluation problem. Therefore lower bias in the DW sample may have more to do with the control sample used rather than the estimation method.

For the Lalonde sample 1974 income (Dist_74) is not part of the conditioning variables. This omission ordinarily falls under violation of Conditional Independence Assumption (CIA). We can however also interpret it as violation of Common Support Assumption (CSA) in an unmeasured variable when compared with the DW sample.

[7] However we lay little emphasis on the result of the significant test because of the balance test fallacy see Imai *et. al* (2008), Ho *et. al.* (2007) and Senn (1994)

This result underscores the importance of theory as mentioned earlier, without the DW sample one may erroneously have false confidence in the level of balance in the Lalonde sample. This is because the variable that showed the highest imbalance also happen to be unobserved. Balance measures cannot fix this kind of problem but theory will suggest that the result from Lalonde sample may be more biased. We provide more discussion on this situation and other situations under which our proposed method may not capture differences in propensity score distribution that is correlated with bias in section 5.

4.2. Multivariate Case

We analyse the relationship between bias, robustness and balance in the samples considered in a multivariate seting. Tables 3A & 4A show absolute value of bias when treatment effect estimate from the various control samples are compared with the experimental benchmark. In these tables the samples were screened before treatment effect was estimated. We estimate treatment effect under various econometric methods. Tables 3B & 4B contain similar result for the unscreened samples. The result shown represent the bias measured as the absolute value of the difference between the treatment effects calculated in each case and the benchmark estimate.

In the first two columns (1 & 2) we have the unadjusted treatment effect i.e. simple mean difference in outcome and unadjusted difference in difference (Unadj-Diff). The next three columns (3, 4 and 5) control for covariates using Regression (Regression)[8], propensity score weighting (pscore weighting)[9] and Difference in difference with controls (Diff-in-Diff). The next five columns (6 to 10) use various weighting methods under PSM i.e. Nearest Neighbour (NN), Radius weighting (Radius), Kernel weighting (Kernel), Stratification (Strat) and Conditional Difference in difference[10] (CDiD). The next three columns (11, 12 and 13) contain the entropic distance between the propensity score distribution of the treatment and control groups (S_ρ pscore), difference in mean of propensity scores across groups (pscore Mean diff) and variance of the treatment effect estimate across econometric methods (Var of estimates). Variance is used to assess the robustness of results under each sample.

The specification of the parametric methods (in column 3, 4 and 5) is based on the specification of the propensity score that balance covariates within strata. It is important to note that while the pscore routine judges the samples as "balanced" based on its approach, the values in column 11 (pscore Mean diff) are all significant at the 5% level. There are two reasons for this. First, the routine only checks for balance within blocks which means when there are large number of blocks the test may have low power because of small within block

[8] Regression is treated as a cross-sectional estimator so pre-programme income were not included in its estimate i.e. income in 1974 was not included for the DW sample and income in 1975 was not included for the

[9] Following Stuart *et. al*. (2013) treated observations get a weight of 1 while control observations get a weight equal to propensity score over one minus propensity score; this serves to weight both groups to resemble the treatment group.

[10] Here difference in difference method is used after matching using Nearest Neighbour Matching

3:DW Screened and Unscreened sample (bias estimates)													
	Unadjusted	**Unadj-Diff**	**Regression**	**pscore weighting**	**Diff-in-Diff**	**NN**	**Radius**	**Kernel**	**Strat**	**CDiD**	S_ρ **pscore**	**pscore Mean diff**	**Var of estimates**
3A:DW Screened													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
PSID1	8643	**502**	1465	1602	**597**	**501**	7434	**400**	**124**	**337**	*0.62*	*0.61*	*3335*
PSID2	2881	2758	1272	**730**	**29**	**499**	**1317**	**829**	**484**	**318**	*0.24*	*0.42*	*1394*
PSID3	**904**	1774	**130**	**67**	**270**	**990**	**923**	**1096**	**889**	**482**	*0.18*	*0.27*	*800*
CPS1	4197	**155**	**627**	**693**	**401**	**252**	4217	**914**	**330**	**278**	*1.13*	*0.36*	*1626*
CPS2	3010	**375**	**1096**	1350	**498**	**392**	2885	**519**	**45**	**233**	*0.50*	*0.47*	*1087*
CPS3	1463	**507**	**516**	**427**	**411**	**1120**	**1319**	**344**	**480**	**480**	*0.25*	*0.42*	*485*
NWS*	*1854*	*1625*	*1664*	*1664*	*1795*	*2156*	*1905*	*1907*	*1788*	*1786*	*0.03*	*0.03*	*155*
3B:DW Unscreened													
PSID1	16999	**323**	**784**	1455	**592**	**501**	15658	**1010**	**124**	**337**	*5.02*	*0.65*	*6792*
PSID2	5441	3479	1581	**803**	**25**	**603**	2464	**841**	**484**	**318**	*0.42*	*0.60*	*2247*
PSID3	**724**	2735	**721**	**162**	**350**	**1213**	**1035**	**1125**	**889**	**482**	*0.37*	*0.51*	*1151*
CPS1	10292	1618	**872**	**839**	**411**	**286**	10321	2888	**330**	**278**	*28.33*	*0.38*	*4292*
CPS2	5616	**1005**	1341	1604	**505**	**392**	5305	**870**	**45**	**233**	*1.09*	*0.50*	*2227*
CPS3	2429	**1083**	**539**	**155**	**427**	**1120**	1965	**412**	**480**	**480**	*0.34*	*0.48*	*945*
NWS*	*1794*	*1806*	*1672*	*1672*	*1799*	*2156*	*1899*	*1897*	*1788*	*1786*	*0.04*	*0.04*	*138*

Table shows size of bias for treatment effect estimate measured against the experimental estimate for each econometric method (except for NWS* row)

NWS* Experimental estimate or benchmark estimate

Estimates in bold are within the 95% confidence interval of the true treatment effect estimate

Columns (1) unadjusted treatment effect estimate (2) unadjusted difference in difference estimate (3) Regression estimate with controls used for propensity score estimation (with the exception of income in 1974) (4) propensity score weighting with cross-sectional controls (5) like 3 but with 1974 income (6) Nearest Neighbour matching (7) Radius matching radius=0.1 (8) kernel matching (Gaussian kernel was used) (9) Stratification matching (10) Conditional Difference-in-difference, stratification matching technique was used (11) Entropic Distance S_ρ between propensity score kernel densities of the treatment and control group. Gaussian kernel was used for the kernel density estimation. (12) t statistic of test of propensity score means (13) variance of treatment effect estimates across methods.

In each case the propensity score specification that that satisfy the mean balancing condition as in Becker et. al. (2002) i.e. mean propensity score and conditioning variables are balanced within each strata.

4:Lalonde Screened and Unscreened sample (bias estimates)													
	Unadjusted	**Unadj-Diff**	**Regression**	**pscore weighting**	**Diff-in-Diff**	**NN**	**Radius**	**Kernel**	**Strat**	**CDiD**	**S_ρ pscore**	**pscore Mean diff**	**Var of estimates**
4A:Lalonde Screened													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
PSID1	8793	1800	8831	2990	1900	3407	7515	2688	2409	1779	*0.47*	*0.55*	*2936*
PSID2	4740	1173	2726	1610	1329	1521	2276	1923	1955	**1050**	*0.31*	*0.48*	*1072*
PSID3	1187	1137	**156**	**223**	1136	1592	1574	1433	1362	1904	*0.28*	*0.36*	*543*
CPS1	8303	**21**	6096	1710	1477	**919**	8104	2602	1276	1302	*2.88*	*0.33*	*3083*
CPS2	4579	1412	2696	1371	1098	1699	4173	1400	1182	**786**	*0.50*	*0.45*	*1313*
CPS3	1931	2418	**17**	**376**	**585**	**983**	1145	**709**	**388**	**483**	*0.29*	*0.46*	*726*
NWS*	*902*	*875*	*809*	*809*	*825*	*818*	*921*	*916*	*788*	*903*	*0.06*	*0.01*	*52*
4B:Lalonde Unscreened													
PSID1	16464	**427**	6750	2523	1895	3404	14280	3391	2409	1779	*1.19*	*0.59*	*5540*
PSID2	4906	**363**	3078	2059	1323	1418	2307	1908	1955	**1050**	*0.35*	*0.53*	*1246*
PSID3	**189**	**605**	**376**	**449**	1141	1587	1533	1417	1362	1904	*0.38*	*0.51*	*586*
CPS1	9757	**868**	4403	1689	1470	**915**	9574	3013	1276	1302	*8.76*	*0.33*	*3618*
CPS2	5081	**711**	2696	1905	1091	1696	4647	1458	1182	**786**	*0.58*	*0.46*	*1541*
CPS3	1894	2454	**252**	**580**	**577**	**981**	1147	**693**	**388**	**483**	*0.31*	*0.48*	*698*
NWS*	*886*	*847*	*802*	*802*	*818*	*815*	*912*	*904*	*788*	*903*	*0.05*	*0.01*	*49*

Table shows size of bias for treatment effect estimate measured against the experimental estimate for each econometric method (except for NWS* row)

NWS* Experimental estimate or benchmark estimate

Estimates in bold are within the 95% confidence interval of the true treatment effect estimate

Columns (1) unadjusted treatment effect estimate (2) unadjusted difference in difference estimate (3) Regression estimate with controls used for propensity score estimation (with the exception of income in 1974) (4) propensity score weighting with cross-sectional controls (5) like 3 but with 1974 income (6) Nearest Neighbour matching (7) Radius matching radius=0.1 (8) kernel matching (Gaussian kernel was used) (9) Stratification matching (10) Conditional Difference-in-difference, stratification matching technique was used (11) Entropic Distance S_ρ between propensity score kernel densities of the treatment and control group. Gaussian kernel was used for the kernel density estimation. (12) t statistic of test of propensity score means (13) variance of treatment effect estimates across methods.

In each case the propensity score specification that that satisfy the mean balancing condition as in Becker et. al. (2002) i.e. mean propensity score and conditioning variables are balanced within each strat

sample size. Second, the 1% level of significant used by default in the pscore routine loads the die in favour of the null hypothesis of no difference.

We also note that there is a difference between the screened and unscreened samples. The relationship between bias and balance in the screened sample captures bias that is due to shape differences alone. On the other hand for the unscreened sample the relationship captures both shape and support differences. Therefore one should expect bias estimate to be bigger for the unscreened samples relative to the screened ones.

In terms of (i) and (ii) the result show that the benchmark estimate from the experimental control group (*NWS** row in tables 3A & B and 4A & B) has the minimum entropic distance between itself and the treatment group. Under our conjecture this represents the best and most robust estimate of treatment effect. The unbiasedness follows from the strength of RCT which is confirmed by the entropic distance measure (column 11). The robustness is reflected in that across econometric methods the treatment effect using this sample has the minimum variance (column 13). This result also justifies the use of estimates from randomize control as the benchmark.

4.2.1 Bias in Non-experimental samples

In this section we discuss result under point (i) for non-experimental controls. To reiterate, our hypothesis is that control samples that are closer to the treatment sample in terms of the entropic distance result in lower bias for unadjusted average treatment effect. We expect unadjusted mean to yield unbiased result only when all covariates (both observed and unobserved) are balanced in similar manner to the experimental control group. However, various control samples are not likely to display the kind of balance obtainable under an RCT. Ashenfelter's dip predicts that the level of imbalance should decrese as we move from sample 1 to 3 for the PSID and CPS controls (Ashenfelter, 1978). This is exactly what the entropic measure (S_ρ outcome in column 11) shows in all the tables.

The treatment effect estimates in column 1 show that the size of bias for the unadjusted mean agrees with the ranking suggested by the entropic distance measure and the theory based on Ashenfelter's dip. For example, in table 3A the entropic distance for the control groups PSID 1, 2 and 3 are approximately 0.62, 0.24 and 0.18 respectively (refer to the first three entries in column 11 of table 3A) while the size of bias for the unadjusted mean estimate are 8 643, 2 881 and 904 respectively (first three entries in column 1 of table 3A). We note that the estimate for PSID sample 3 is not only less biased it lies within one standard deviation of the benchmark treatment effect estimate (Bold text represent situations when the estimate is within one standard deviation of the treatment effect). Although the entropic distance does not track the size of bias perfectly, it suggests a ranking of control samples that reflect the size of bias in each control group.

The CPS samples in the same table reflect similar ranking i.e. entropic distance 1.13, 0.50, 0.25 and size of bias 4 197, 3 010 and 1 463 respectively. Therefore for table 3A (DW screened sample) one could write

$$\gamma_{PSID3} = \min_{f_{PSID}(W|D=0)} \{\gamma_{PSID1}, \gamma_{PSID2}, \gamma_{PSID3}\} \ldots\ldots.4$$

and

$$\gamma_{CPS3} = \min_{f_{CPS}(W|D=0)} \{\gamma_{CPS1}, \gamma_{CPS2}, \gamma_{CPS3}\} \ldots\ldots.5$$

The important thing to note about this result is that the mean balancing test as implemented by the "pscore" routine (which implements the algorithm put forward in Dehejia & Wahba (2000)) suggests that in terms of mean balancing all samples are as balanced as they could be. The mean balancing condition is a binary decision rule that is based on a t-test such that it is either the sample is balanced or it is not. As noted earlier, t-tests could just be picking up differences in sample size (especially when it is applied within blocks as is the case here). In addition to this, our result suggests that there can be varying degree of bias in samples that satisfy the mean balancing condition as implemented by the "pscore" routine. Furthermore, this bias is tracked by the entropic distance metric which is a continuous decision rule that allows the comparison of level of balance in different samples.

As one may suspect, the mean difference in propensity scores is also correlated with the size of bias, however its correlation is weaker. While it performs well for the PSID samples, its ranking fails to capture the size of bias for the CPS sample. For example, in table 3A, "pscore mean diff" are 0.36, 0.47 and 0.42 while the size of bias is 4,197, 3,010 and 1,463 for CPS 1, 2 and 3 respectively. To get a better idea of the relationship between bias and the two balancing measures we calculate the rank correlation between balance and bias for the two measures. To improve on the sample size we calculate placebo treatment effects by using the randomized control group as the treatment group and then calculating treatment effect between this "treatment" group and the PSID and CPS controls (the results are shown in the appendix). The rank correlation between the bias in column 1 and entropic distance for the DW screened and unscreened samples are 0.83 and 0.86 respectively. While the corresponding rank correlation between bias and mean measure are 0.58 and 0.16 respectively. The same pattern is observed in the Lalonde sample suggesting that the entropic measure is more correlated with bias than the mean balancing measure.

Finally, we note that for the unscreened samples (tables 3B and 4B) the entropic distances are higher than their corresponding screened values. This is attributable to the fact that both shape and support components of the bias decomposition (Heckman, Ichimura and Todd, 1997) are captured in the unscreened sample while only the shape component is captured in the screened sample. As expected the bias associated with unadjusted mean difference (column 1) in the unscreened sample is also higher compared to the corresponding value for the screened sample.

4.2.2. Robustness in Non-experimental samples

Variance of treated effect estimate across econometric methods is used to assess the relationship between robustness of treatment effect estimate and balance as measured by the entropic distance metric (point ii). Column 13 shows the variance of the treatment

effect estimates across the econometric methods considered. Estimates from samples with higher entropic distances vary more across econometric techniques (i.e. are less robust) compared to those with low entropic distances. Among the non-experimental samples, sample 3 yields treatment effect estimates with lower variability across methods than sample 2, and sample 2 in turn yield results with lower variability than sample 1. The result suggests that while one can often use various econometric techniques to estimate an unbiased treatment effect (giving the assumption(s) of the method), the entropic metric can help identify samples that yield more robust result across econometric techniques. Such samples should be preferred because their result does not rely heavily on model assumption(s).

4.3. General discussion of the Results

In general the results suggest that bias and robustness of ATE across econometric methods can be signalled by entropic distance between the distribution of propensity scores. This supports the notion that the entropic distance is a better way of assessing balance both in the univariate and the multivariate case as our results have shown. As noted earlier for the screened sample, the bias that is captured by the entropic metric is attributable to differences in shape of characteristic distribution alone. This provides some evidence that shape differences alone can result in bias and that restricting estimation to region of Common Support does not take care of this component of bias. Furthermore entropic distance tends to be larger for the unscreened sample than the screened sample suggesting that the proposed measure captures both shape and support differences in propensity score distribution across treatment status.

The observed relationship between bias and robustness tends to be distorted when various econometric approaches are used to estimate ATE. The rank correlation between the entropic distance and bias under various econometric techniques tends to be much lower and less meaningful (sometimes negative) except for Radius matching. This should be expected since various econometric approaches correct for the imbalance based on the specific assumption(s) that validate them. These approaches do fare better for the screened sample (shape differences only) as bias tends to be lower for the screened sample under different methods that adjust for covariate differences. Bias due to the difference in the shape of the distributions over the Common Support is also referred to as bias due to wrong-weighting (Heckman, Ichimura & Todd, 1997) or interpolation bias (King & Zeng, 2006). This bias can be taken care of by adopting the “appropriate” weighting function to correct for the difference in shape over the region of Common Support. Literature suggests that various econometric methods corresponds to different weighting function that control for imbalance. For example, Angrist and Pischke (2008) compared PSM and regression, they showed analytically that regression is equivalent to using a different weighting scheme than the one used under PSM. PSM weights covariate-specific estimate into an estimate of the treatment effect on the treated using the distribution among the treated units while regression produces a variance weighted average of these effects (Angrist and Pischke, 2008 pg 54; Forbes and Klein, 2015 pg 213). For the unscreened samples bias tends to be larger,

suggesting that the correction made by various techniques is less successful when there are support differences.

PSM based methods outperform other methods in the DW sample. More of the PSM based estimates lie within one standard deviation of the benchmark treatment effect. Across the DW and Lalonde samples Stratification and Conditional Difference-indifference (note that Nearest Neighbour matching scheme is used for the conditional diff-in-diff) are robust. This means that these methods result in the same treatment effect irrespective of whether the sample is screened or unscreened. This can be attributed to the fact that the weighting under these methods will automatically screen the samples before estimating treatment effect. Both methods tend to automatically keep estimation sample within the Common Support.

Lastly, the DW sample appears to be a better control sample as a whole as more of the estimates from this sample are within one standard deviation of the benchmark compared with the Lalonde sample. It may be tempting to assume that this is reflected in the entropic distance measure (the entropic distance tends to be smaller for the DW sample than the Lalonde sample); however, we would advise caution when it comes to comparing the entropic measure across samples. We deal with this issue and other related issues in the next section.

5. Possible Complications

In assessing balance in different control samples, the proposed measure should be used with caution. The entropic measure can only measure observable differences in characteristic distribution which means that comparison across samples may be problematic. For example, the DW sample controls for 1974 income while the Lalonde sample does not. It is not clear how this variable will affect the entropic distance in the two samples. One may expect that since the Lalonde sample does not have to control for a variable that shows the highest level of imbalance it should have the lower entropic distance compared to corresponding DW sample. However, our result does not always portray this. More research will be needed to investigate the effect of unobservables on the entropic measure.

In general, using the entropic distance metric to rank control distributions that are not identical in terms of conditioning variables or survey instrument might yield dubious results. When comparing across samples, it is not guaranteed that the set of conditioning variables will be identical. Even when they are, differences in survey instrument may mean that samples are different in terms of unobservables. When the set of W is not the same in the control samples being compared, the entropic measure may not be consistent. This is because in such cases it is not comparing identical things.

Another relevant issue is how to interpret the differences between the entropic distances associated with different samples. Our analysis based on this dataset (and the simulation study in the previous chapter) suggests that a control group whose propensity score distribution has an entropic distance of the order of magnitude -2 (i.e. $x * 10^{-2}$ where $x <$

9) from the propensity score distribution of the treatment group performs better. Such control samples will almost always recover treatment effect estimate that are comparable to the one obtainable under an RCT. When the distance is more than this one may expect varying degree of bias that is correlated with the distance. We however cannot reliably pin down based on this analysis, how different entropic distances must be for there to be a significant difference between the level of bias in different samples. Our approach relies on the result of Rosenbaum and Rubin (1983) which suggests that conditioning on propensity scores is equivalent to conditioning on the set of covariate (W) under PSM. Since propensity score is a coarse balancing score[11] which we posit contains information about the suitability of control distributions, it is possible that some differences in the joint distribution of covariates will not be captured by the entropic distance that is based on propensity score distributions. For extreme cases of large and low differences between these joint distributions, we expect that the entropic distance will reflect this. However, in-between these extremes more research will be needed to uncover how different entropic distances between two samples must be for it to help separate close control groups from extreme ones.

6. Conclusion

We examined the plausibility of using the entropic distance measure to rank non-experimental control distributions. Our results show that in non-experimental situations where one can define plausible control groups with identical set of control variables, the entropic distance measure can help identify control groups that will result in estimates that are comparable to the one obtainable under an RCT.

[11] the covariate themselves are the finest balancing score Rosenbaum and Rubin (1983)

Bibliography

Angrist, Joshua D, and Jorn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press, 2008.

Ashenfelter, Orley C, and David Card. "Using the longitudinal structure of earnings to estimate the effect of training programs." *Using the longitudinal structure of earnings to estimate the effect of training programs.* National Bureau of Economic Research Cambridge, Mass., USA, 1984.

Ashenfelter, Orley. "Estimating the effect of training programs on earnings." *The Review of Economics and Statistics* (JSTOR), 1978: 47-57.

Becker, Sascha O, Andrea Ichino, and others. "Estimation of average treatment effects based on propensity scores." *The stata journal* 2, no. 4 (2002): 358-377.

Caliendo, Marco, and Sabine Kopeinig. "Some practical guidance for the implementation of propensity score matching." *Journal of economic surveys* (Wiley Online Library) 22, no. 1 (2008): 31-72.

Cochran, William G. "Analysis of covariance: its nature and uses." *Biometrics* (JSTOR) 13, no. 3 (1957): 261-281.

Cochran, William G, and Donald B Rubin. "Controlling bias in observational studies: A review." *Sankhya: The Indian Journal of Statistics, Series A* (JSTOR), 1973: 417-446.

Cook, Thomas D, Donald Thomas Campbell, and Arles Day. *Quasi-experimentation: Design & analysis issues for field settings.* Vol. 351. Houghton Mifflin Boston, 1979.

Crump, Richard, V Joseph Hotz, Guido Imbens, and Oscar Mitnik. "Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand." *Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand.* National Bureau of Economic Research Cambridge, Mass., USA, 2006.

Dehejia, Rajeev H, and Sadek Wahba. "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs." *Journal of the American statistical Association* (Taylor & Francis Group) 94, no. 448 (1999): 1053-1062.

Dehejia, Rajeev H, and Sadek Wahba. "Propensity score-matching methods for nonexperimental causal studies." *Review of Economics and statistics* (MIT Press) 84, no. 1 (2002): 151-161.

Granger, CW, Esfandiar Maasoumi, and Jeffrey Racine. "A dependence metric for possibly nonlinear processes." *Journal of Time Series Analysis* (Wiley Online Library) 25, no. 5 (2004): 649-669.

Hainmueller, Jens. "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies." *Political Analysis* (SPM-PMSAPSA), 2011: mpr025.

Hansen, Ben B. "The prognostic analogue of the propensity score." *Biometrika* (Biometrika Trust) 95, no. 2 (2008): 481-488.

Heckman, J, H Ichimura, J Smith, and P Todd. "Characterizing selection bias using experimental data." *Econometrica* (Wiley-Blackwell) 66, no. 5 (1998): 1017-1098.

Heckman, James J, and Jeffrey A Smith. "Assessing the case for social experiments." *The Journal of Economic Perspectives* (JSTOR), 1995: 85-110.

Heckman, James J, and V Joseph Hotz. "Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training." *Journal of the American statistical Association* (Taylor & Francis Group) 84, no. 408 (1989): 862-874.

Heckman, James J, Hidehiko Ichimura, and Petra E Todd. "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *The Review of economic studies* (Oxford University Press) 64, no. 4 (1997): 605-654.

Ho, Daniel E, Kosuke Imai, Gary King, and Elizabeth A Stuart. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference." *Political analysis* (SPM-PMSAPSA) 15, no. 3 (2007): 199-236.

Iacus, Stefano M, Gary King, and Giuseppe Porro. "Causal inference without balance checking: Coarsened exact matching." *Political analysis* (SPM-PMSAPSA), 2011: mpr013.

Imai, Kosuke, Gary King, and Elizabeth A Stuart. "Misunderstandings between experimentalists and observationalists about causal inference." *Journal of the royal statistical society: series A (statistics in society)* (Wiley Online Library) 171, no. 2 (2008): 481-502.

Imbens, Guido M, and Jeffrey M Wooldridge. "Recent developments in the econometrics of program evaluation." Tech. rep., National Bureau of Economic Research, 2008.

Imbens, Guido W, and D.[VNV] Rubin. *Causal inference in statistics, and in the social and biomedical sciences.* Cambridge University Press New York, 2009.

King, Gary, and Langche Zeng. "The dangers of extreme counterfactuals." *Political Analysis* (SPM-PMSAPSA) 14, no. 2 (2006): 131-159.

LaLonde, Robert J. "Evaluating the econometric evaluations of training programs with experimental data." *The American Economic Review* (JSTOR), 1986: 604-620.

Lechner, Michael, and others. *The estimation of causal effects by difference-in-difference methods.* the essense of knowledge, 2011.

Lee, Wang-Sheng. "Propensity score matching and variations on the balancing test." *Empirical economics* (Springer) 44, no. 1 (2013): 47-80.

Maasoumi, Esfandiar, and Jeffrey S Racine. "A robust entropy-based test of asymmetry for discrete and continuous processes." *Econometric Reviews* (Taylor & Francis) 28, no. 1-3 (2008): 246-261.

Maasoumi, Esfandiar, and Le Wang. "The Gender Earnings Gap: Measurement and Analysis." Tech. rep., Emory University, 2013.

Rosenbaum, Paul R, and Donald B Rubin. "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome." *Journal of the Royal Statistical Society. Series B (Methodological)* (JSTOR), 1983: 212-218.

Rubin, Donald B. "Using propensity scores to help design observational studies: application to the tobacco litigation." *Health Services and Outcomes Research Methodology* (Springer) 2, no. 3-4 (2001): 169-188.

Rubin, Donald B, and Neal Thomas. "Combining propensity score matching with additional adjustments for prognostic covariates." *Journal of the American Statistical Association* (Taylor & Francis Group) 95, no. 450 (2000): 573-585.

Senn, Stephen. "Testing for baseline balance in clinical trials." *Statistics in medicine* (Wiley Online Library) 13, no. 17 (1994): 1715-1726.

Shadish, William R. "Propensity score analysis: promise, reality and irrational exuberance." *Journal of experimental criminology* (Springer) 9, no. 2 (2013): 129-144.

Smith, Jeffrey. "A critical survey of empirical methods for evaluating active labor market policies." Tech. rep., Research Report, Department of Economics, University of Western Ontario, 2000.

Smith, Jeffrey A, and Petra E Todd. "Does matching overcome LaLonde's critique of nonexperimental estimators?" *Journal of econometrics* (Elsevier) 125, no. 1 (2005): 305-353.

Stuart, Elizabeth A. "Matching methods for causal inference: A review and a look forward." *Statistical science: a review journal of the Institute of Mathematical Statistics* (NIH Public Access) 25, no. 1 (2010): 1.

Stuart, Elizabeth A, Brian K Lee, and Finbarr P Leacy. "Prognostic score--based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research." *Journal of clinical epidemiology* (Elsevier) 66, no. 8 (2013): S84--S90.

APPENDIX

DW SAMPLE (SCREENED SAMPLE)													
	Unadjusted	**Unadj-Diff**	**Regression**	**Pscore weighting**	**Diff-in-Diff**	**NN**	**Radius**	**Kernel**	**Strat**	**CDiD**	**S_ρ pscore**	**pscore Mean diff**	**Var of estimates**
PSID1	-6788.7	**2127.77**	199.4958	62.58652	**2392.029**	**1654.566**	-5528.48	**1507.102**	**1911.89**	**2122.985**	*0.616698*	*0.61268*	*3335.057*
PSID2	-1026.89	4383.805	392.0281	**934.0052**	**1823.728**	**1656.124**	**588.7399**	**1077.822**	**1304.313**	**2104.06**	*0.242193*	*0.417552*	*1393.631*
PSID3	**950.1351**	3399.704	**1534.186**	**1731.428**	**1525.094**	**1165.918**	**982.1399**	**811.3037**	**898.5208**	**2267.823**	*0.178692*	*0.274721*	*800.3414*
CPS1	-2342.97	**1470.566**	**1037.358**	**971.7719**	**1393.792**	**2407.145**	-2311.07	**993.4656**	**1457.419**	**1507.67**	*1.134834*	*0.358532*	*1626.116*
CPS2	-1156.36	**1250.253**	**568.1828**	314.2746	**1297.256**	**1763.394**	-979.161	**1388.517**	**1832.825**	**1553.479**	*0.50334*	*0.466277*	*1087.227*
CPS3	391.1736	**2132.121**	**1148.201**	**1237.323**	**1384.021**	**1035.418**	**586.662**	**1563.159**	**1307.868**	**1306.527**	*0.254329*	*0.421196*	*485.143*
PSID1_C	-7931.67	**1207.295**	-2477.27	-1849.29	**431.3745**	**535.3451**	-6642.19	**35.58736**	**-135.706**	**165.6904**	*0.661455*	*0.704406*	*3176.404*
PSID2_C	-3221.74	3462.489	-1533.27	**-384.05**	**460.8504**	**952.3425**	**-1218.08**	**-451.598**	**-79.3301**	**261.7725**	*0.386561*	*0.5659*	*1746.177*
PSID3_C	**-862.285**	1536.636	**-665.299**	**-22.8348**	**-661.664**	**-449.423**	**-1091.52**	**-925.118**	**-984.483**	**343.0954**	*0.36189*	*0.473184*	*809.0368*
CPS1_C	-5510.34	**327.6729**	**-435.777**	**-579.069**	**-281.284**	**817.5674**	-5426.91	**-691.173**	**-187.675**	**-538.403**	*1.987298*	*0.475425*	*2269.597*
CPS2_C	-3900.91	**-100.909**	**-694.856**	**-734.899**	**-255.683**	**-381.498**	-3604.03	**42.35908**	**95.83374**	**-196.269**	*0.606713*	*0.557722*	*1491.853*
CPS3_C	-1744	**618.4929**	**-306.929**	**-194.298**	**-230.015**	**-119.626**	**-826.157**	**137.5337**	**-98.9192**	**-198.92**	*0.32918*	*0.502103*	*623.2711*
NWS	*1854.101*	*1625.494*	*1664.472*	*1664.472*	*1794.982*	*2155.553*	*1905.494*	*1907.145*	*1787.857*	*1786.14*	*0.034409*	*0.029072*	*155.4148*
Corr_rho	0.839161	-0.65035	0.286713	0.538462	0.286713	-0.23776	0.811189	-0.25175	-0.37063	-0.2028			
P_value	0.000643	0.022034	0.366251	0.070894	0.366251	0.456801	0.001363	0.429919	0.235621	0.527302			
Corr_pscore_t	0.587413	-0.03497	0.608392	0.426573	0.328671	-0.03497	0.412587	-0.74126	-0.57343	-0.51049			
P_value	0.044609	0.914093	0.035806	0.1667	0.296904	0.914093	0.182564	0.005801	0.051266	0.089914			

Bold text represents treatment effects that are 95% confidence interval of the benchmark treatment effect estimate (NWS)

All the entropic distance values are significant at 5%.

Corr_rho, Corr_rho_y and Corr_pscore_t represents the rank correlation coefficient between bias and S_ρ pscore, S_ρ outcome and pscore Mean diff respectively

The rows control_C represent the placebo treatment effect calculated by using the randomized control group as the treatment group. The treatment effect in this case is zero

DW SAMPLE (UNSCREENED SAMPLE)

	Unadjusted	Unadj-Diff	Regression	Pscore weighting	Diff-in-Diff	NN	Radius	Kernel	Strat	CDiD	S_ρ pscore	pscore Mean diff	Var of estimates
PSID1	-15204.8	**2128.395**	**888.0269**	216.846	**2391.057**	**1654.566**	-13759	**887.6175**	**1911.89**	**2122.985**	*5.021759*	*0.647435*	*6791.894*
PSID2	-3646.81	5284.923	91.24919	**868.7802**	**1774.346**	**1552.283**	-564.127	**1056.717**	**1304.313**	**2104.06**	*0.420392*	*0.595847*	*2247.45*
PSID3	**1069.85**	4541.15	**951.4557**	**1509.847**	**1449.428**	**942.2618**	**864.0239**	**772.475**	**898.5208**	**2267.823**	*0.370667*	*0.509874*	*1151.145*
CPS1	-8497.52	3423.711	**800.4338**	**833.5229**	**1388.002**	**2441.301**	-8422.02	-990.734	**1457.419**	**1507.67**	*28.33365*	*0.377369*	*4291.989*
CPS2	-3821.97	**2810.415**	331.1468	68.03685	**1294.03**	**1763.394**	-3405.65	**1027.809**	**1832.825**	**1553.479**	*1.091239*	*0.499785*	*2227.079*
CPS3	-635.026	**2888.637**	**1133.567**	**1516.979**	**1371.957**	**1035.418**	-65.9151	**1485.756**	**1307.868**	**1306.527**	*0.342643*	*0.482023*	*945.4416*
PSID1_C	-16999.1	**322.5994**	**-946.386**	**-993.189**	**824.7642**	**531.3666**	-15028.3	**-254.148**	**-135.706**	**165.6904**	*9.130998*	*0.743232*	*6767.35*
PSID2_C	-5441.15	3479.128	-1831.46	**-818.348**	**461.1278**	**1029.071**	-1808.52	**-483.361**	**-79.3301**	**261.7725**	*0.478017*	*0.665571*	*2305.761*
PSID3_C	**-724.493**	2735.355	**-1196.41**	**-490.863**	**-658.185**	**-432.64**	**-1076.34**	**-917.567**	**-984.483**	**343.0954**	*0.439144*	*0.583965*	*1165.686*
CPS1_C	-10291.9	1617.915	**-807.434**	**-658.197**	**-281.929**	**830.219**	-10221.3	-1753.58	**-187.675**	**-538.403**	*84.36048*	*0.491782*	*4327.875*
CPS2_C	-5616.31	**1004.62**	**-1198.23**	**-1158.67**	**-256.327**	**-369.403**	-5240.31	**-134.704**	**95.83374**	**-196.269**	*1.15745*	*0.579455*	*2260.222*
CPS3_C	-2429.37	**1082.841**	**-211.387**	**-44.6179**	**-231.619**	**-119.626**	**-1011.93**	**120.5001**	**-98.9192**	**-198.92**	*0.384565*	*0.540206*	*899.9392*
NWS	1794.342	1805.795	1672.203	1672.203	1799.282	2155.553	1899.383	1897.366	1787.857	1786.14	*0.038095*	*0.03699*	*138.0612*
Corr_rho	0.862239	-0.35338	0.237762	0.636364	0.237762	-0.32168	0.811189	0.377622	-0.30769	-0.18881			
P_value	0.000309	0.259828	0.456801	0.026097	0.456801	0.30791	0.001363	0.226206	0.330589	0.556737			
Corr_pscore_t	0.169621	-0.05654	0.426573	0.27972	0.335664	0.090909	0.076923	-0.3986	-0.18182	-0.41958			
P_value	0.598178	0.861449	0.1667	0.378569	0.286123	0.778725	0.812183	0.199335	0.571701	0.174519			

Bold text represents treatment effects that are 95% confidence interval of the benchmark treatment effect estimate (NWS)

All the entropic distance values are significant at 5%.

Corr_rho, Corr_rho_y and Corr_pscore_t represents the rank correlation coefficient between bias and S_ρ pscore, S_ρ outcome and pscore Mean diff respectively

The rows control_C represent the placebo treatment effect calculated by using the randomized control group as the treatment group. The treatment effect in this case is zero

LALONDE (SCREENED SAMPLE)

	Unadjusted	Unadj-Diff	Regression	Pscore weighting	Diff-in-Diff	NN	Radius	Kernel	Strat	CDiD	S_ρ pscore	pscore Mean diff	Var of estimates
PSID1	-7891.19	-925.361	-8022.62	-2181.04	-1075.15	-2589.29	-6594.42	-1772.28	-1621.13	-875.114	*0.46556*	*0.545216*	*2936.061*
PSID2	-3837.9	-297.727	-1917.61	-801.547	-504.304	-702.619	-1355.33	-1006.62	-1166.4	**-146.983**	*0.30716*	*0.475633*	*1071.982*
PSID3	-285.132	-262.512	**652.5788**	**585.8592**	-311.439	-773.505	-652.716	-516.775	-573.685	-1000.76	*0.280456*	*0.356407*	*542.9932*
CPS1	-7400.65	**854.2245**	-5287.64	-901.793	-651.769	**-100.663**	-7183.15	-1686.28	-487.429	-398.791	*2.883506*	*0.328763*	*3082.501*
CPS2	-3676.81	-536.692	-1886.93	-562.704	-273.233	-881.314	-3251.67	-483.762	-393.816	**117.2828**	*0.497523*	*0.447481*	*1313.081*
CPS3	-1028.99	-1543.58	**791.7031**	**432.7568**	**240.3608**	**-164.995**	-224.006	**206.9935**	**399.8197**	**420.6184**	*0.290747*	*0.462064*	*726.354*
PSID1_C	-11453.8	-1578.76	-8854.81	-3392.21	-1596.3	-1549.22	-9181.88	-2431.4	-1967.45	-1218.36	*0.536594*	*0.623906*	*3905.644*
PSID2_C	-5452.1	**-1076.84**	-2964.01	-2327.17	-1193.36	-1228.29	-2101.15	-2016.95	-2371.3	**-993.439**	*0.302958*	*0.449236*	*1329.823*
PSID3_C	-942.149	-1269.57	**-586.508**	-1123.53	-1602.87	**34.40152**	-1731.85	-1499.25	-1874.06	-1953.11	*0.405968*	*0.471481*	*623.9183*
CPS1_C	-7548.98	**-195.969**	-6131.98	-1697.75	-1411.19	**-924.553**	-7248.71	-1956.17	-1443.33	-1412.39	*1.697526*	*0.42349*	*2808.27*
CPS2_C	-4582.71	-1280.98	-2207.07	-1268.11	-957.586	**-828.049**	-4053.25	-1036.23	**-869.712**	**-861.984**	*0.518013*	*0.526209*	*1395.281*
CPS3_C	-1742.36	-2260.56	**286.4577**	**-347.306**	**-477.494**	**-617.793**	**-681.047**	**-429.824**	**-308.194**	**-604.795**	*0.328817*	*0.50971*	*738.4816*
NWS	901.8913	874.893	808.6908	808.6908	825.0522	818.1426	920.9375	915.9641	788.1665	903.4402	*0.055794*	*0.011393*	*51.63043*
Corr_rho	0.601399	-0.26573	0.671329	0.545455	0.405594	-0.16783	0.811189	0.426573	0.034965	0.153846			
P_value	0.038588	0.403833	0.016831	0.066612	0.190836	0.602099	0.001363	0.1667	0.914093	0.633091			
Corr_pscore_t	0.230769	0.671329	0.223776	0.258741	0.125874	0.041958	0.104895	0	0.230769	-0.16084			
P_value	0.470532	0.016831	0.484452	0.416775	0.696683	0.896986	0.745609	1	0.470532	0.617523			

Bold text represents treatment effects that are 95% confidence interval of the benchmark treatment effect estimate (NWS)

All the entropic distance values are significant at 5%.

Corr_rho, Corr_rho_y and Corr_pscore_t represents the rank correlation coefficient between bias and S_ρ pscore, S_ρ outcome and pscore Mean diff respectively

The rows control_C represent the placebo treatment effect calculated by using the randomized control group as the treatment group. The treatment effect in this case is zero

LALONDE (UNSCREENED SAMPLE)

	Unadjusted	Unadj-Diff	Regression	Pscore weighting	Diff-in-Diff	NN	Radius	Kernel	Strat	CDiD	S_p pscore	pscore Mean diff	Var of estimates
PSID1	-15577.6	**419.6706**	-5947.6	-1721.41	-1076.83	-2589.29	-13367.4	-2486.91	-1621.13	-875.114	*1.190808*	*0.591845*	*5539.845*
PSID2	-4019.6	**483.5259**	-2275.53	-1256.65	-505.381	-603.109	-1395.24	-1003.85	-1166.4	**-146.983**	*0.353266*	*0.525899*	*1245.815*
PSID3	**697.0584**	**241.6561**	**425.9468**	**353.0403**	-323.405	-772.51	-620.641	-513.383	-573.685	-1000.76	*0.384856*	*0.508375*	*586.485*
CPS1	-8870.31	**1714.398**	-3601.09	-887.109	-651.786	**-100.663**	-8662.27	-2108.79	-487.429	-398.791	*8.759456*	*0.333748*	*3618.334*
CPS2	-4194.76	**136.3773**	-1893.8	-1103.19	-273.222	-881.314	-3734.97	-554.278	-393.816	**117.2828**	*0.58459*	*0.456905*	*1541.057*
CPS3	-1007.82	-1607.43	**550.3258**	**222.0463**	**240.7971**	**-166.622**	-234.41	**210.5863**	**399.8197**	**420.6184**	*0.311607*	*0.481013*	*697.5367*
PSID1_C	-16463.9	**-427.218**	-6928.09	-3176.14	-1597.47	-1514.45	-13736.5	-2736.89	-1967.45	-1218.36	*1.050478*	*0.655978*	*5657.26*
PSID2_C	-4905.9	**-363.362**	-3572.35	-2828.32	-1203.21	-1228.29	-2144.34	-1995.17	-2371.3	**-993.439**	*0.398758*	*0.568357*	*1350.956*
PSID3_C	**-189.245**	**-605.232**	**-856.626**	-1361.94	-1602.47	**34.40152**	-1714.28	-1495.69	-1874.06	-1953.11	*0.462571*	*0.560214*	*712.126*
CPS1_C	-9756.61	**867.5094**	-4228.15	-1697.06	-1411.35	**-918.037**	-9490.79	-2417.91	-1443.33	-1412.39	*14.66341*	*0.435161*	*3616.668*
CPS2_C	-5081.06	**-710.511**	-2396.14	-1808.5	-957.63	**-820.003**	-4531.45	-1080.7	**-869.712**	**-861.984**	*0.608773*	*0.53885*	*1618.518*
CPS3_C	-1894.12	-2454.32	**110.8061**	**-506.379**	**-477.631**	**-589.838**	**-729.921**	**-432.089**	**-308.194**	**-604.795**	*0.356515*	*0.538079*	*775.7939*
NWS	886.3037	846.8883	802.0486	802.0486	818.0273	814.6747	912.2283	904.0836	788.1665	903.4402	*0.050713*	*0.010949*	*49.01733*
Corr_rho	0.777429	0.014135	0.741259	0.34965	0.594406	0.055944	0.839161	0.713287	0.307692	0.454545			
P_value	0.00292	0.965224	0.005801	0.265239	0.041521	0.862898	0.000643	0.009202	0.330589	0.137658			
Corr_pscore_t	0.113081	-0.62194	0.223776	0.524476	0.328671	0.132867	0.118881	0.167832	0.566434	0.146853			
P_value	0.726406	0.030827	0.484452	0.080019	0.296904	0.680598	0.712884	0.602099	0.054842	0.648796			

Bold text represents treatment effects that are 95% confidence interval of the benchmark treatment effect estimate (NWS)

All the entropic distance values are significant at 5%.

Corr_rho, Corr_rho_y and Corr_pscore_t represents the rank correlation coefficient between bias and S_p pscore, S_p outcome and pscore Mean diff respectively

The rows control_C represent the placebo treatment effect calculated by using the randomized control group as the treatment group. The treatment effect in this case is zero

TABLE 2C: PROPENSITY SCORE SPECIFICATION THAT BALANCE MEANS	
TREATTMENT VS CONTROL	
LALONDE SAMPLE	
Lalonde PSID1	age education married nodegree black hispanic re75 married_re75 hispanic_re75 age2
Lalonde PSID2	age education married nodegree black hispanic re75 married_re75 hispanic_re75 age2
lalonde PSID3	age education married nodegree black hispanic re75 age2 edu2
Lalonde CPS1	age education married nodegree black hispanic re75 age2 edu2 re752
LalondeCPS2	age education married nodegree black hispanic re75 age2 edu2 re752 nodegree_re75
LalondeCPS3	age education married nodegree black hispanic re75 age2 edu2 re752 black_married hispanic_z75
DW SAMPLE	
DW PSID1	age education married nodegree black hispanic re74 re75 age2 edu2 re742 re752 black_z74 (DW,2000)
DW PSID2	age education married nodegree black hispanic re74 re75 age2 edu2 re742 re752 nodegree_re75 edu_re74 married_re75
DW PSID3	age education married nodegree black hispanic re75 re74 age2 edu2 nodegree_re74
DW CPS1	age education married nodegree black hispanic re74 re75 age2 edu2 re742 re752 black_z74
DWCPS2	age education married nodegree black hispanic re74 re75 age2 edu2 re742 re752 black_z74
DWCPS3	age education married nodegree black hispanic re74 re75 age2 edu2 re752