

Case study on program evaluation in the presence of selection bias: an application of propensity score matching to the national supported work demonstration data

Boyang Xin

University of Waterloo, Waterloo, Canada

b3xin@uwaterloo.ca

Abstract. Randomised controlled trials are often impractical for evaluating labour market support programmes, since random assignment carries high costs, long timelines and considerable implementation difficulty. How to obtain an accurate and reliable estimate of a programme's treatment effect from observational data subject to selection bias has therefore been debated in labour economics and causal inference for four decades. This study takes the National Supported Work Demonstration, a randomised experiment whose true effect is known, deletes the experimental control group, substitutes a non-experimental comparison group drawn from the Current Population Survey, and asks how close matching comes to an answer that is already established. Against an experimental benchmark of \$1,794, the naive comparison returns -\$8,498, with the wrong sign, regression adjustment yields \$699, and propensity score matching yields \$1,984, with 181 of 185 treated units on common support. The practical implication is that a matching estimate is credible only to the extent that the diagnostics printed beside it are, since balance, overlap and sensitivity to the matching rules are what an evaluator can inspect when no benchmark exists.

Keywords: propensity score matching, selection bias, programme evaluation, causal inference, National Supported Work Demonstration

1. Introduction

Whether a training programme raises the earnings of its participants is a long-standing question in labour economics. The difficulty lies not in measurement but in selection: participants differ from nonparticipants before the programme begins, and those differences persist in post-programme earnings. LaLonde showed the consequences by applying the non-experimental estimators then in common use to a case whose true answer was already known, and recording how far they missed it, sometimes by enough to reverse the sign [1]. Four decades of work on estimators that might survive without it followed. Propensity score matching became the method most applied researchers reached for, and Dehejia and Wahba reported that it closed most of the distance to the experimental answer [2-4]. Their claim drew an immediate challenge [5]. The disagreement it opened has still not been resolved [6]. This study addresses not whether the method exists but how seldom its diagnostics are shown. A published matching estimate frequently arrives with no evidence that the covariates

balanced, no count of treated units outside the overlap region, and no indication of how far the number would move if the neighbour count or caliper were set differently. The estimate and its diagnostics are therefore treated here as equally important outputs. The design is as follows. Covariate selection follows the backdoor criterion, propensity scores are estimated by logit, and treated units are matched to nearest neighbours on the score. Balance, the overlap region and the estimate's sensitivity to the neighbour count and caliper are then reported alongside it. Because the experimental target is observable, every choice can be scored against it rather than defended on theoretical grounds alone.

2. Literature review

The modern framework begins with the potential outcomes model, which defines a causal effect as the contrast between the outcomes the same unit would record under treatment and under non-treatment, only one of which is ever observed [7]. Identification, therefore, requires an assumption about the untreated outcomes of the treated. The most common assumption is unconfoundedness: conditional on observed covariates, treatment assignment is independent of potential outcomes.

Conditioning on many covariates simultaneously is infeasible in finite samples. Rosenbaum and Rubin resolved this by proving that the propensity score, the conditional probability of treatment given the covariates, is a balancing score [2]. Conditioning on this single index therefore suffices to remove the bias that conditioning on the full covariate vector would remove. This result makes matching practical.

LaLonde supplied the field's most influential test. Using the National Supported Work Demonstration, he compared experimental estimates against non-experimental estimators applied to survey comparison groups and found the latter unreliable [1]. Dehejia and Wahba revisited the same data with propensity score methods and reported estimates far closer to the benchmark, a result that established matching's credibility in applied economics [3, 4].

Smith and Todd disputed that conclusion, showing that the performance depended heavily on which analysis subsample was used and how the covariates were specified, and that difference-in-differences matching, which exploits panel structure to remove time-invariant unobservables, performed more consistently in their tests [5]. The theory behind such estimators had been developed earlier [8, 9], but estimators of that family require repeated observations of the same individuals, which a cross-sectional design like this one cannot provide.

Later work clarified the conditions under which matching is trustworthy. A survey of the assumptions behind treatment effect estimation under unconfoundedness singled out overlap as a separate requirement, frequently violated and rarely reported [10], and credibility has since been argued to rest on design-stage reasoning and validation against known benchmarks rather than on the choice of estimator [11]. Working directly with LaLonde's data, Imbens and Xu find that modern methods can reproduce the experimental benchmark when overlap is adequate and covariates are rich, though they warn that agreement is never guaranteed [6].

Two points of consensus emerge from this literature. The propensity score is now accepted as a valid means of reducing dimensionality when unconfoundedness holds, and the choice of comparison group, together with the richness of the pre-programme covariates, matters more to the result than the choice among matching estimators. What remains unresolved is the second-order question: given that matching sometimes reproduces an experimental benchmark and sometimes does not, which conditions distinguish the two cases, and which reported diagnostics would enable a reader to distinguish them in a study where no benchmark exists. The contribution of this paper lies there. Rather than adding another estimate to the record, it reports the

full diagnostic apparatus of a single matching exercise, balance, overlap and specification sensitivity, at the same length as the estimate itself, so that the conditions under which the estimate can be believed are visible on the page.

3. Data, identification strategy and methodology

3.1. Data and variables

Two datasets are combined here, both carrying the same eleven variables. The first is the Dehejia-Wahba extract of the National Supported Work Demonstration, a mid-1970s American programme that placed disadvantaged workers, among them former offenders and long-term welfare recipients, in subsidised temporary jobs [3]. Random assignment of 185 individuals to treatment and 260 to control makes the earnings gap between the two arms an unbiased estimate of the programme effect, which has made this dataset the standard benchmark for non-experimental estimators [1, 6].

The second dataset is a non-experimental comparison group of 15,992 Current Population Survey observations, and substituting it for the experimental controls reproduces the situation an evaluator ordinarily faces: a treated group, a large survey of untreated individuals, and no randomisation linking the two [1].

Both files record age, years of education, indicators for Black and Hispanic ethnicity, marital status, an indicator for lacking a high-school degree, and real annual earnings for 1974, 1975 and 1978. The 1978 figure is the outcome, though most of the identifying work is done by the two earlier earnings years, which capture a labour market history that age and schooling alone would miss. Table 1 shows why the naive comparison must fail. The CPS sample is older, better educated and far more often married than the treated group, seven years older and two years better educated on average, and 84.3% of participants are Black against 7.4% of CPS respondents. The largest disparity between the two groups stems from pre-programme income. CPS respondents earned roughly \$14,000 in 1974 and 1975 against about \$2,000 for the treated group, and 60% of participants recorded zero earnings in 1975 compared with 10.9% of the CPS sample. The standardised mean difference for several variables before matching is accordingly far beyond the empirical threshold of 0.10, reaching 2.43 for the Black indicator and -1.75 for 1975 earnings, so a direct comparison would be severely biased by selection [12].

Table 1. Descriptive statistics, mean (standard deviation)

Variable	Treated (185)	Experimental control (260)	CPS-1 (15,992)
Age	25.8 (7.2)	25.1 (7.1)	33.2 (11.0)
Years of education	10.3 (2.0)	10.1 (1.6)	12.0 (2.9)
Black	0.843 (0.365)	0.827 (0.379)	0.074 (0.261)
Hispanic	0.059 (0.237)	0.108 (0.311)	0.072 (0.259)
Married	0.189 (0.393)	0.154 (0.361)	0.712 (0.453)
No high-school degree	0.708 (0.456)	0.835 (0.372)	0.296 (0.456)
Earnings 1974	2,096 (4,887)	2,107 (5,688)	14,017 (9,570)
Earnings 1975	1,532 (3,219)	1,267 (3,103)	13,651 (9,270)
Earnings 1978	6,349 (7,867)	4,555 (5,484)	14,847 (9,647)
Zero earnings 1975	0.600 (0.491)	0.685 (0.466)	0.109 (0.312)

3.2. Identification strategy and propensity score matching

Identification rests on the potential outcomes framework [7]. Let D indicate programme participation, and let $Y(1)$ and $Y(0)$ denote the 1978 earnings a given individual would record under participation and under non-participation, respectively. The estimand throughout is the average treatment effect on the treated, as given in Equation (1).

$$ATT = E[Y(1) - Y(0)|D = 1] \quad (1)$$

Because $Y(1)$ is observed only for participants and $Y(0)$ only for non-participants, the second term is counterfactual for every treated individual and must be imputed from the comparison group.

Which variables belong in X follows from representing the problem as a causal graph, since age, education, ethnicity, marital status and pre-programme earnings each plausibly influence both enrolment and later earnings, opening backdoor paths between treatment and outcome. Pearl's backdoor criterion identifies the effect once a set blocking all such paths is conditioned on, while warning against conditioning on variables that would open new ones [13]. Post-treatment variables are the clearest such case: causal identification principles forbid controlling for a variable realised after the intervention, and 1978 earnings therefore enter the analysis only as the outcome.

Two conditions license the imputation. Unconfoundedness, Equation (2), requires that assignment carry no information about potential outcomes once X is held fixed:

$$(Y(0), Y(1)) \perp D | X \quad (2)$$

while overlap, Equation (3), requires that any covariate profile occurring among participants could also have occurred among non-participants:

$$0 < Pr(D = 1|X) < 1 \quad (3)$$

The first cannot be tested, since it concerns outcomes never observed; its plausibility rests on whether X is rich enough, which here depends on the two years of pre-programme earnings. Overlap is an empirical property of the sample, and Section 5.2 examines it directly [10].

Matching on the full vector X is infeasible in finite samples. Rosenbaum and Rubin resolve this by showing that the propensity score in Equation (4) [2], is a balancing score: treatment status is independent of X conditional on $e(X)$. Conditioning on this scalar therefore removes the same bias that conditioning on the entire vector would. The score is estimated by logit with age, education, their squares, marital status, the no-degree indicator, ethnicity, 1974 and 1975 earnings, and indicators for zero earnings in each of those years. This list is reproduced from Dehejia and Wahba and left unchanged so that the results here stay comparable with theirs [4]. Squaring age and schooling makes it possible for the enrolment probability to rise across part of their range and fall across the rest, rather than to move in one direction only. The zero-earnings dummies exist because someone with no recorded earnings in a year is in a different state from someone with low earnings, and a specification placing both on one continuous scale would lose that distinction.

$$e(X) = Pr(D = 1|X) \quad (4)$$

The procedure then involves five steps: fit the score, examine balance before any matching is performed, restrict the sample to the overlap region, match on the fitted score, and check whether balance improved. Write $J(i)$ for the set of control units that minimise $|\hat{e}(X_i) - \hat{e}(X_j)|$ for treated unit i , and N_1 for the number of treated units on common support. The estimator is then given by Equation (5):

$$ATT = \frac{1}{N_1} \sum_i [Y_i - \frac{1}{|J(i)|} \sum_{j \in J(i)} Y_j] \quad (5)$$

Averaging over all members of $J(i)$ rather than selecting one member keeps the estimator well defined when several controls tie at the minimum distance, which happens for 29 of the 181 matched treated units. Matching is with replacement, reducing bias at the cost of increased variance [14]. Because the score is estimated rather than known, conventional standard errors understate uncertainty; the appropriate correction is derived by Abadie and Imbens and applied by the Stata implementation used here [15, 16].

4. Empirical results

4.1. The experimental benchmark

Because the assignment was randomised, the difference in mean 1978 earnings between the 185 treated individuals and the 260 experimental controls estimates the programme's effect without further assumptions. Treated participants earned \$6,349 on average against \$4,555 for controls. The average treatment effect on the treated is therefore \$1,794, with a standard error of 671, statistically distinguishable from zero and equal to roughly 39% of control-group earnings. Every subsequent estimate is judged against this figure.

Table 1 also permits a check that randomisation was successful, and on most covariates the two experimental arms are close: mean age 25.8 against 25.1, education 10.3 years against 10.1, and 84.3% against 82.7% Black. One gap is wider than chance comfortably explains, since 70.8% of the treated group lacked a high-school degree against 83.5% of controls, but the earnings histories that drive selection are nearly identical at \$2,096 and \$2,107 in 1974. This is the standard the non-experimental comparison group must meet.

4.2. The naive non-experimental comparison

Discarding the experimental controls and comparing the treated group directly against the 15,992 CPS respondents produces an estimate of -\$8,498 with a standard error of 583, which, if read literally, would mean the programme cost its participants more than eight thousand dollars a year. Since the estimate is at once precisely determined and wrong in sign, it exhibits the most dangerous combination a policy evaluation can produce, and the total bias of roughly \$10,300 has an origin that Table 1 makes plain: because the comparison group was already earning far more before the programme began, the raw difference is measuring pre-existing inequality between two populations rather than anything the programme did. LaLonde's original warning reappears intact [1].

A larger sample would reduce the standard error and leave a systematic selection bias of this kind unchanged. The standard error of 583 is in fact smaller than the one attached to the experimental estimate, so a researcher with only these data would report a result that is precise, highly significant, and completely wrong, with nothing in the output to signal the problem. Confidence intervals reflect sampling variation alone and do not correct for bias arising from an inappropriate comparison group, which is why doubling the CPS sample would halve the standard error while leaving the \$10,300 bias unchanged.

4.3. Regression adjustment

The conventional first response is to control for the covariates in a linear regression of 1978 earnings on a treatment indicator, age, education, ethnicity, marital status, the no-degree indicator, and 1974 and 1975 earnings. This yields \$699, with a standard error of 548.

Correcting the sign is a substantial improvement, though the resulting estimate is statistically indistinguishable from zero and understates the true effect by roughly 61%. Extrapolation causes the trouble, since a single functional form is fitted across the whole covariate space and will produce a predicted

counterfactual for a treated individual even when no CPS respondent resembles that person on any dimension that matters. Given how badly these two groups overlap, extrapolation is carrying most of the estimate, and whether it carries it accurately depends on a linearity assumption for which the data offer no support whatever.

A second difficulty underlies the first and is easier to miss. Regression weights every observation by how much it contributes to the variance of the treatment indicator, so that the comparison individuals least like participants exert the greatest influence on the estimate. Set 15,992 CPS observations against 185 treated ones and the fitted earnings function belongs, in any practical sense, to a population the programme never enrolled.

4.4. Propensity score matching

Matching inverts that ordering. The model is for participation, not the outcome, and each treated individual is then matched with untreated individuals carrying a similar estimated probability of having enrolled. Fitting the logit of Section 3.2 and pairing every treated unit on common support with its closest neighbour returns \$1,984, standard error 766, across 181 matched treated units.

That leaves a gap of \$190 to the experimental benchmark, well within a single standard error. Sign, rough magnitude and policy reading all come back, none of which survived the naive comparison and all of which regression adjustment had badly understated. The standard error of 766 nevertheless leaves a null effect inside the interval, which the point estimate on its own would not suggest.

How ties are handled turns out to matter, because it changes the number. Identical covariate profiles are common among CPS respondents, and identical profiles produce identical propensity scores. In about one in six cases—29 of the 181—no single control is closest. If any one of the tied controls is selected without averaging, the estimate shifts by more than \$150, which is most of the remaining distance to the benchmark. Everything reported here averages over ties, which makes the figures deterministic and exactly reproducible.

The matches are tight, with a mean absolute difference of 0.0018 in propensity score between the treated and control units and a range between about 0.001 and 0.890. Proximity on the index is necessary rather than sufficient, since two individuals may share a score while differing on the individual covariates that compose it, and Section 5.1 therefore checks balance rather than assuming it.

4.5. Comparison across methods

As shown in Table 2, four estimates of the same programme effect span more than \$10,000, and the choice of comparison group rather than the outcome model accounts for almost all of that spread.

Table 2. Estimated average treatment effect on the treated (1978 earnings, USD)

Method	Comparison group	Estimate	SE	Bias vs benchmark
Experimental benchmark	NSW experimental controls	1,794	671	n/a
Naive difference	CPS-1	-8,498	583	-10,292
OLS with covariates	CPS-1	699	548	-1,095
Propensity score matching	CPS-1	1,984	766	+190

Matching estimates carry a matched-pair standard error, which should be read as a lower bound. Intervals widen under the corresponding Abadie-Imbens correction, since the propensity score is estimated rather than known [15, 16]. Point estimates are unaffected, and were checked against an independent implementation before reporting.

Moving from the raw comparison to regression adjustment eliminates about 89% of the bias, and moving from regression adjustment to matching removes most of what remains. This reproduces the qualitative

finding of Dehejia and Wahba on these same data: with rich pre-programme earnings histories and a comparison group confined to the region of overlap, matching can approximate an experimental benchmark [3, 4].

Both methods employ the same covariates, and the difference lies in how each treats the data before the effect is modelled. Matching asks which treated subjects have genuinely similar counterparts and discards the rest, so the modelling step operates only where the data can support it, whereas regression does not ask the question and extrapolates silently into regions with no support. That the two approaches differ by almost \$1,300 on identical variables argues for locating credibility in the design rather than in the choice of estimator [11].

5. Diagnostics and robustness

5.1. Covariate balance

Balance is measured by the standardised mean difference of Equation (6), which expresses the gap in group means for a covariate as a fraction of its pooled standard deviation [17], with absolute values under 0.10 conventionally taken as acceptable [12]. Because the measure carries no units, years of schooling, dollars of earnings and 0/1 indicators can all be read off the same scale.

$$SMD = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{s_t^2 + s_c^2}{2}}} \quad (6)$$

Table 3. Covariate balance, standardised mean differences

Variable	Before matching	After matching	Reduction
Age	-0.796	0.029	96%
Years of education	-0.679	0.070	90%
Black	2.428	0.061	97%
Hispanic	-0.051	-0.002	97%
Married	-1.233	-0.174	86%
No high-school degree	0.904	0.025	97%
Earnings 1974	-1.569	0.109	93%
Earnings 1975	-1.746	0.109	94%

As Table 3 shows, balance improves for every covariate and in most cases substantially: the Black indicator drops from 2.43 to 0.06, while both pre-programme earnings variables fall from below -1.5 to approximately 0.11, leaving six of the eight below the 0.10 threshold with the two earnings measures missing it only narrowly.

Marital status does not improve, staying at -0.174 because marriage remains appreciably more common in the matched comparison group than among the treated. Since the labour literature generally treats marital status as an independent predictor of earnings, some bias evidently survives the procedure, and this residual imbalance is one plausible contributor to the \$190 gap between the matched estimate and the benchmark.

Whether that residue reflects an inadequate logit specification or a variable that a scalar propensity score simply cannot balance in this sample can be settled empirically, and the answer is the former. Marital status enters the Dehejia-Wahba specification linearly. Re-estimating the score with marriage interacted with age,

schooling, ethnicity, the no-degree indicator and both pre-programme earnings terms reduces the post-matching standardised mean difference on marital status from -0.174 to -0.037, comfortably inside the threshold, indicating that the comparison pool is rich enough to support balance on this covariate. The pool is in fact already working in that direction: marriage stands at 71.2% across the full survey but at 41.7% among the 4,458 CPS respondents falling inside the overlap region, against 18.9% among the treated.

The augmented specification is nonetheless not adopted here, for two reasons. Balance on marital status is achieved at the expense of balance elsewhere, since the same score moves the no-degree indicator from 0.025 to 0.126 and so outside the threshold, and the estimate moves from \$1,984 to \$1,700. More importantly, the reported specification is held to the published Dehejia-Wahba covariate list so that the results remain comparable with theirs [4]. Re-specifying a propensity score until a chosen covariate balances, while a different one drifts out, is the specification search that Section 5.3 is designed to expose rather than to perform.

Reading Table 3 as evidence that the procedure succeeded requires two qualifications. The standardised mean difference compares means alone, and two distributions can share a mean while differing widely in spread, so balance of this kind is necessary without being sufficient. Matching here is also carried out with replacement, so that a small number of CPS respondents serve as counterfactuals for several treated units apiece: match quality improves, since each treated unit draws its best available comparator instead of being forced onto a worse one, but the effective number of distinct comparison observations is smaller than the count of matched pairs implies, and the true variance is correspondingly larger than a naive calculation would give [14].

5.2. Common support

The estimated propensity scores range from 0.001 to 0.890 in the overlap region. Of 185 treated individuals, 181 fall within it and 4 do not. Those four are so unlike any CPS respondent that no credible counterfactual exists for them, and including them would mean extrapolating exactly as regression does.

What is estimated is therefore an effect for the 181 rather than for the 185, a small change of estimand but a real one that applied papers frequently leave unstated. Overlap in this dataset is in any case unusually good, since losing 4 units out of 185 amounts to a mild trim, whereas the same exercise conducted with a comparison group with weaker overlap would discard far more and yield an estimate correspondingly less representative of the treated population [10, 11]. Nor are the four excluded individuals a random subset: carrying the highest estimated participation probabilities in the sample, they are the people whose characteristics most strongly predict enrolment and least resemble anyone in a general population survey. Excluding them is methodologically correct, since no comparable untreated individual exists, but it removes precisely the units the programme most clearly targeted, and in a sample with poorer overlap that trimming could remove the policy-relevant population altogether.

5.3. Sensitivity to matching choices

Table 4 substantially qualifies that agreement. Opening the comparison to two neighbours or more settles the estimate somewhere between roughly \$1,430 and \$1,550, precision bought at the cost of bias as steadily more distant controls are admitted. Loose calipers hold the estimate close to \$1,980 while discarding almost nothing. The 0.001 caliper behaves differently, driving the estimate to \$1,292 while leaving only 112 of the 181 treated units in the sample, and a drop of that size may reflect who remains in the sample rather than how good the matches are.

Table 4. Sensitivity of the ATT to specification (USD)

Specification	Estimate	SE	Matched treated
1 nearest neighbour, no caliper	1,984	766	181
2 nearest neighbours	1,492	705	181
3 nearest neighbours	1,546	669	181
5 nearest neighbours	1,429	641	181
1 neighbour, caliper 0.01	1,991	783	177
1 neighbour, caliper 0.005	1,970	825	166
1 neighbour, caliper 0.001	1,292	903	112

Every estimate in the table reflects the same trade-off. Each neighbour added lowers the standard error, from 766 at one neighbour to 641 at five, because averaging over more controls reduces noise in the imputed counterfactual. The point estimates fall in the same direction and for an evident reason, since each control admitted resembles the treated unit slightly less than the previous one. Tightening the caliper reverses this, improving match quality at the cost of a standard error rising from 766 to 903 and a matched sample 38% smaller. Neither direction is correct in the abstract, and nothing in the data settles which one to prefer.

Estimates between roughly \$1,290 and \$1,990 are therefore all defensible on the same data and the same covariates, with the experimental figure of \$1,794 falling comfortably inside that range. This is exactly the sensitivity Smith and Todd identified, and the reason an estimate reported without a specification table should never be taken on its own [5].

6. Conclusion

On these data, propensity score matching recovers the experimental answer more closely than either of the alternatives considered. The naive comparison is wrong in sign, regression adjustment corrects the sign but understates the effect badly, and matching yields an estimate that agrees with the experimental benchmark in sign, in rough magnitude and in the policy reading it supports.

The agreement is produced by the design stage rather than the estimator. Confining the comparison to the region where treated and untreated individuals genuinely resemble one another, and conditioning on a covariate set that includes several years of pre-programmed earnings, allows a survey comparison group to stand in for an experimental one. A comparison group lacking such an earnings history would not support the same conclusion.

Three qualifications belong with the result. Identification rests on unconfoundedness, which no test can reach. Balance improves greatly without becoming complete, since marital status remains above the conventional threshold. The estimate also varies with the matching rules, across a range wide enough that the agreement with the benchmark is partly a property of the specification chosen. A specification table printed beside the headline figure is therefore not an appendix but part of what is being reported.

Several extensions follow. The panel structure left unused in these data would support difference-in-differences matching, which relaxes the reliance on unconfoundedness by differencing away any factors that are unobserved but fixed over time. Kernel and radius matching deserve comparison against the nearest-neighbour estimator on the same sample. Most importantly, the whole procedure should be repeated where overlap is poor, since the near-complete common support available here is a favourable case rather than a typical one.

References

- [1] LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4), 604-620.
- [2] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.
- [3] Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448), 1053-1062.
- [4] Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1), 151-161.
- [5] Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1-2), 305-353.
- [6] Imbens, G. W., & Xu, Y. (2025). Comparing experimental and nonexperimental methods: What lessons have we learned four decades after LaLonde (1986)? arXiv:2406.00827.
- [7] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688-701.
- [8] Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4), 605-654.
- [9] Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2), 261-294.
- [10] Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1), 4-29.
- [11] Imbens, G. W. (2024). Causal inference in the social sciences. *Annual Review of Statistics and Its Application*, 11, 123-152.
- [12] Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3), 399-424.
- [13] Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press, pp. 78-86.
- [14] Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1), 31-72.
- [15] Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1), 235-267.
- [16] Abadie, A., & Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84(2), 781-807.
- [17] Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33-38.