Model-based matching

Kellie Ottoboni

May 12, 2016

1 Introduction

Observational studies in a range of fields including social sciences, epidemiology, and ecology are used to make inferences about cause and effect. An ideal situation is the randomized control trial, in which treatment is assigned at random. Randomization ensures that treatment is statistically independent of all variables. This enables us to disentangle the effect of treatment from the effect of other variables. This is rarely the case in observational studies. There, the treatment effect is obscured by "confounders," variables that affect both the assignment of treatment and the outcome.

In principle, one could solve the problem of confounding by grouping individuals with identical values of each of their pretreatment covariates, then estimating treatment effects within groups and averaging to obtain an overall estimate. In practice, high dimensionality makes this impossible. Studies typically measure a large number of covariates, so groups of individuals matched exactly on all pretreatment covariates can be too small to provide adequate statistical power to detect a significant treatment effect. Post hoc methods to construct a group of controls whose covariates balance the covariates in the treatment group include exact matching on covariates, matching and stratifying by propensity score (Rosenbaum and Rubin [1983], Austin [2014], Lunceford and Davidian [2004]), genetic matching (Diamond and Sekhon [2013]), propensity score weighting (Hirano et al. [2003]), and maximum-entropy balancing (Hainmueller [2012]). These balanced control groups are then used to estimate average treatment effects using standard methods such as unadjusted differences of means or linear regression controlling for other covariates. Estimates of the average treatment effect using matching or weighting can be biased if the propensity score model is misspecified (Drake [1993]) or if the outcome model is misspecified (Freedman and Berk [2008]).

We propose two steps to improve precision and reduce bias, both of which rely on a prediction of

the outcome in the control regime. First, we replace the observed outcomes with their residuals after subtracting off the predicted control outcome. Rosenbaum [2002] used this type of residualization to stabilize the outcomes, since much of their variance may be attributed to observed covariates. Second, we stratify observations on their predicted control outcomes. If strata are chosen appropriately, then treatment assignment is "as if" random within strata. This enables us to estimate treatment effects unbiasedly and to carry out randomization inference within strata.

We develop the theory for this procedure, which we call "model-based matching," both for estimating treatment effects and carrying out hypothesis tests. Throughout, we assume that there are only two levels of treatment and that outcomes are continuous. In the first section, we develop the notation and theory to estimate the average treatment effect. In the next section, we introduce randomization inference and extend it to observational studies in which we want to test whether there is any treatment effect whatsoever. Finally, we discuss benefits and shortcomings of the proposed framework.

2 Estimation

We begin this section with a review of matching and stratification estimators. Next, we present the proposed estimator and show under what conditions it is unbiased for the average treatment effect. Finally, we show that a special case of the proposed estimator has smaller variance than the usual difference in means estimator in a completely randomized experiment.

2.1 Matching and Stratified Estimators

We use the potential outcomes (Neyman [1923], but see Splawa-Neyman et al. [1990]) notation throughout the paper. Suppose that we randomly sample N individuals from a super population.

Assume that every individual, indexed by i, has two fixed potential outcomes: we observe outcome $Y_i(0)$ if they are assigned to control or $Y_i(1)$ if they are assigned to treatment. We may never observe both. The observed outcome can be written $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$, where T_i is an indicator for treatment. For individual i, the effect of the treatment is thus $Y_i(1) - Y_i(0)$. The main estimand of interest is the average treatment effect:

$$ATE = \mathbb{E}\left[Y(1) - Y(0)\right]$$

In randomized control trials, potential outcomes are balanced between treatment and control groups by construction. Mathematically, $(Y(1), Y(0)) \perp\!\!\!\perp T$. This enables us to estimate the ATE unbiasedly, because individuals in different treatment groups are, on average, alike in every way but the treatment. In observational studies, this is no longer guaranteed.

One way to address this issue is to achieve a weaker form of balance by conditioning on other observed covariates, X. Rosenbaum and Rubin [1983] achieve this weaker form of balance by conditioning on the propensity score, $p(x)$, the probability of receiving treatment conditional on having covariates $X = x$. They show that the propensity score is a balancing score, such that

$$X \perp\!\!\!\perp T \mid p(X)$$

Intuitively, matching or stratifying individuals by their propensity score approximates a randomized experiment within groups. Then, one can obtain unbiased estimates of ATE within groups and of the overall ATE by averaging over group-wise treatment effects.

One issue is that in observational studies, the propensity score is unknown and must be estimated. It is common to estimate the propensity score using logistic or probit regression with a set of pretreatment covariates. This may be problematic if the propensity model is misspecified, either in its functional form or due to omitted variables. When propensity score estimates are incorrect, then the estimated ATE may be biased (Drake [1993]). Even if one uses a correctly-estimated propensity score to weight regressions instead of match, the model of the outcome must be correctly specified, otherwise estimates may be biased and standard errors may be under- or overestimated (Freedman and Berk [2008]).

Matching and stratification may balance the treatment and control groups even if the propensity score estimates are wrong, as long as the right units are matched. However, there is no "optimal" procedure to group observations. There are many variants on matching such as matching with or without replacement; one-to-one, one-to-many, or many-to-many; and within a fixed distance or nearest neighbor (Austin [2014]). Which method achieves the best balance on covariates depends on the particular dataset. Similarly, there is no optimal way to choose a stratification rule. The literature generally suggests using five strata based on quintiles of the estimated propensity scores, though there is little justification for this (Austin et al. [2007]).

2.2 Stratified Estimation

Instead of relying on the propensity score to achieve balance, we'd like to try to condition on something else. We propose using $\hat{Y}$, a model-based prediction of Y, absent knowledge of the treatment.

Hansen [2008] defines a prognostic score as a function $\Psi(X)$ such that $Y(0) \perp\!\!\!\perp X \mid \Psi(X)$ for all X. Conditioning on a prognostic score induces balance in the covariate distributions of individuals with contrasting potential outcomes. Hansen [2008] shows the following useful result:

Lemma 2.1. *Suppose that there is no hidden bias, so* $(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$. *Then conditioning on a prognostic score deconfounds potential response under the control regime from treatment assignment:*

$$Y(0) \perp\!\!\!\perp T \mid \Psi(X)$$

Furthermore, if there is no effect modification, then

$$Y(1) \perp\!\!\!\perp T \mid \Psi(X)$$

We omit the details of effect modification and the proof.

Let $\hat{Y} = f(X)$ be a prediction of the response under the control regime. If there is no hidden bias, then $\hat{Y}$ is a prognostic score. Matching individuals exactly on $\hat{Y}$ presents the same issues as matching on the propensity score: incorrect estimates of $\hat{Y}$ may lead to bad matches that worsen balance between groups. Instead, we propose stratifying on $\hat{Y}$. This allows us some error in estimating $\hat{Y}$, as long as our estimates are "close enough" to correct.

Let $\Psi_s(X)$ be a function that classifies individuals into one of S strata based on their predicted control potential outcome. That is, $\Psi_s(X) = \Pi(\hat{Y}) = \Pi(f(X))$ for some $\Pi : \mathbb{R} \to \{1, \ldots, S\}$. The stratification rule Ψ_s and estimation procedure for $\hat{Y}$ should be defined independently of the observed sample to avoid overfitting. Define $S_i = \Psi_s(X_i)$ to be the stratum assignment of unit i. The sth stratum contains N_{st} treated individuals and N_{sc} control individuals for a total of $N_s = N_{st} + N_{sc}$ individuals, so $N = N_1 + \cdots + N_S$. We estimate the ATE by

$$\hat{\tau}^{strat} = \sum_{s=1}^{S} \frac{N_s}{N} \left(\frac{1}{N_{st}} \sum_{i:T_i=1, S_i=s} (Y_i - \hat{Y}_i) - \frac{1}{N_{sc}} \sum_{i:T_i=0, S_i=s} (Y_i - \hat{Y}_i) \right) \tag{1}$$

$\hat{\tau}^{strat}$ is the weighted average of within-stratum estimated treatment effects with weights proportional to the stratum sizes. For this estimator to be unbiased, we need the stratum-specific

estimators to be unbiased. The following assumption is a sufficient condition for unbiasedness within strata.

Assumption 1. *Conditional independence between potential outcomes and treatment given stratum assignment:* $(Y(1), Y(0)) \perp\!\!\!\perp T \mid S$.

For Assumption 1 to hold, the stratification must be so fine that it variations in $\hat{Y}$ within strata are unrelated to treatment assignment. Intuitively, the results within each strata must act as though they had come from a randomized experiment. This condition implies that $\mathbb{P}(T_i = 1 \mid S_i = s, Y_i(1), Y_i(0)) = \mathbb{P}(T_i = 1 \mid S_i = s)$, so the probability of any particular unit receiving treatment is constant within strata. In an actual randomized experiment, the analogous condition is that the treatment is assigned at random within strata.

Theorem 2.2. *If* $(Y(1), Y(0)) \perp\!\!\!\perp T \mid S$ *and* $0 < N_{st} < N_s$ *for* $s = 1, \ldots, S$*, then* $\hat{\tau}^{strat}$ *is unbiased for the ATE.*

The proof appears in the appendix.[1] The result generalizes to Bernoulli treatment assignment as well, so long as treatment probabilities are constant within strata. See Lemma 5.1 for details.

2.2.1 Variance in a randomized experiment

Assume that we sample N units from a super population and do a complete randomization with N_t units assigned to treatment. Instead of the stratified estimator used above, consider the special case of a single stratum:

$$\hat{\tau}^{adj} = \frac{1}{N_t} \sum_{i:T_i=1} (Y_i - \hat{Y}_i) - \frac{1}{N_c} \sum_{i:T_i=0} (Y_i - \hat{Y}_i)$$

Randomization ensures that this is an unbiased estimate of the ATE.

Define $\sigma_1^2 = \text{Var}_{SP}(Y(1))$, $\sigma_0^2 = \text{Var}_{SP}(Y(0))$, and $\sigma_{01}^2 = \text{Var}_{SP}(Y(1) - Y(0)) = \text{Cov}_{SP}(Y(1), Y(0))$. In addition, define the variance of the predicted outcomes in the population as $\nu^2 = \text{Var}_{SP}(\hat{Y}) = \mathbb{E}_{SP}\left[(\hat{f}(X) - \mathbb{E}_{SP}\left[\hat{f}(X)\right])^2\right]$. Finally, define the correlation coefficient between the predicted outcomes $\hat{Y}$ and true outcomes: $\rho_1 = \frac{\text{Cov}_{SP}(Y(1), \hat{Y})}{\sigma_1 \nu}$, and $\rho_0 = \frac{\text{Cov}_{SP}(Y(0), \hat{Y})}{\sigma_0 \nu}$.

[1] Throughout the following results, we follow Miratrix et al. [2013] and condition on the event that $N_{st} > 0$ and $N_{sc} > 0$ for all strata $s = 1, \ldots, S$. This is the set of possible treatment assignments for which the stratified estimator is defined. We omit this notation in the mathematics for legibility.

Theorem 2.3.

$$Var(\hat{\tau}^{adj}) = \frac{(\sigma_1 - \nu)^2 - 2(1-\rho_1)\sigma_1\nu}{N_t} + \frac{(\sigma_0 - \nu)^2 - 2(1-\rho_0)\sigma_0\nu}{N_c} \tag{2}$$

The proof appears in the appendix.

Compare this to $\hat{\tau}^{diff}$, the unadjusted difference in means estimator:

$$\hat{\tau}^{diff} = \frac{1}{N_t}\sum_{i:T_i=1} Y_i - \frac{1}{N_c}\sum_{i:T_i=0} Y_i$$

Using the same proof as above but substituting Y for $Y - \hat{Y}$, one can show that the variance of $\hat{\tau}^{diff}$ is

$$\text{Var}\left(\hat{\tau}^{diff}\right) = \frac{\sigma_1^2}{N_t} + \frac{\sigma_0^2}{N_c} \tag{3}$$

To guarantee $\text{Var}(\hat{\tau}^{adj}) < \text{Var}(\hat{\tau}^{diff})$, it is sufficient to have $\sigma_1 \geq \frac{\nu}{2(2-\rho_1)}$ and $\sigma_0 \geq \frac{\nu}{2(2-\rho_0)}$. Indeed, this always holds true when the variance in predictions is no more than four times the variance in outcomes. This will hold for any reasonable predictor $\hat{Y}$. Therefore, we conclude that residualizing improves precision.

Does stratification improve precision beyond this? Analytically, the variance of $\hat{\tau}^{strat}$ from Equation 1 depends on the number of strata and how they partition the covariate space. Let $\hat{\tau}_s$ be the difference in mean outcomes for the s-th stratum. Then, it is possible for stratification to improve variance.

Theorem 2.4. *Suppose we do complete randomization of N individuals sampled from a super population. If $Var(\hat{\tau}_s) \leq Var(\hat{\tau}^{adj})$ for all s, then $Var(\hat{\tau}^{strat}) < Var(\hat{\tau}^{adj})$.*

The proof appears in the appendix. We confirm this empirically in the next section.

2.3 Empirical Results

2.3.1 Constant Additive Treatment Effects

We generate a population of $N = 100$ individuals. For each unit, we observe two independent standard normal covariates X_1 and X_2. We assign treatment in two ways:

- Random treatment assignment: 50 individuals are assigned to treatment and the remaining 50 are assigned to control.
- Correlated with X_1: Among the 50 individuals with lowest X_1 values, 10 are randomly assigned treatment. Among the remaining 50 individuals with largest X_1 values, 40 are randomly assigned treatment.

The outcome is given by

$$Y_i = 1 + 2X_{1i} + 4X_{2i} + \tau T_i + \varepsilon_i$$

In this section, we fix the treatment effect at $\tau = 1$. We generate the errors in two ways:

- Homoskedastic: the ε_i are IID standard normal
- Heteroskedastic: the ε_i are independent and normal with $\mathbb{E}(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_i) = X_{1i}^2 + X_{2i}^2$

We run all four combinations of treatment assignment and error distributions. We compare seven methods of estimating the treatment effect: the raw difference in means, OLS controlling for X_1 and X_2, model-based matching with fine stratification, model-based matching with coarse stratification, residualized difference in means without stratification, one-to-one propensity score matching, and entropy balancing (Hainmueller [2012]). We run all of the methods under "best case" circumstances, assuming that we have measured the relevant variables and know the correct functional forms. To estimate the predictor $\hat{f}$ for residualization, we split the sample into a training set and test set to avoid overfitting the predictive model. This overfitting would produce biased estimates of subgroup treatment effects: in strata with low predicted control outcomes, estimates would be biased upward, whereas strata with high predicted control outcomes would have downward biased estimates (Abadie et al. [2013]). We use a random 30% of the dataset to estimate $\hat{Y}$, estimating a correctly-specified linear model using the control observations. Then using the remaining 70%, we predict their outcomes under control and use these $\hat{Y}$s to stratify finely, with 20 strata, and coarsely, with 5 strata. We estimate the propensity score with a correctly-specified probit regression and match each treated individual to a control, doing one-to-one matching with replacement.

Figure 1 shows the distribution of these estimates over the 1000 simulations. Table 1 compares the RMSE of the seven methods and Table 2 compares their variances. In all cases, OLS performs the best. The Gauss-Markov theorem predicts this, as we specified the model correctly. Entropy balancing has the next smallest RMSE and variance. Model-based matching performs nearly as

well, with slightly higher RMSE and variance. Propensity score matching and the raw difference in means are highly variable, with estimates as high as 6 in the worst case. The unadjusted difference in means is the only estimator that appears to be biased.

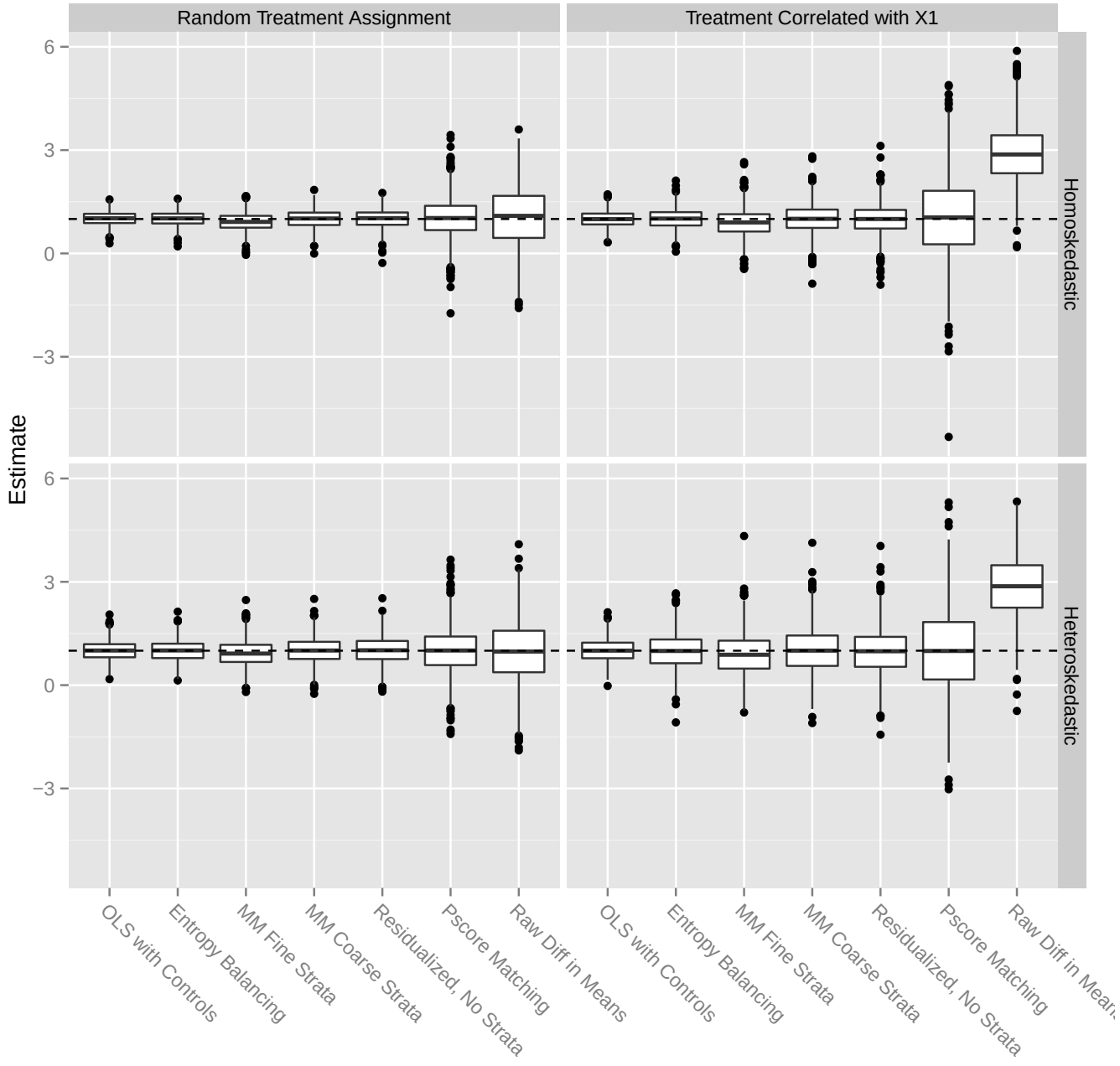

Figure 1: Estimates of the average treatment effect over 1000 simulations (constant treatment effect design)

	Random treatment assignment		Treatment Correlated with X_1	
	Homoskedastic	Heteroskedastic	Homoskedastic	Heteroskedastic
OLS with Controls	0.20	0.29	0.23	0.34
Entropy Balancing	0.20	0.30	0.28	0.51
MM Fine Strata	0.28	0.39	0.41	0.63
MM Coarse Strata	0.26	0.38	0.42	0.67
Residualized, No Strata	0.27	0.39	0.44	0.70
Propensity Score Matching	0.61	0.68	1.19	1.19
Raw Diff in Means	0.89	0.93	2.06	2.07

Table 1: RMSE over 1000 simulations (constant treatment effect design)

	Random treatment assignment		Treatment Correlated with X_1	
	Homoskedastic	Heteroskedastic	Homoskedastic	Heteroskedastic
OLS with Controls	0.04	0.08	0.05	0.11
Entropy Balancing	0.04	0.09	0.08	0.26
MM Fine Strata	0.07	0.15	0.15	0.39
MM Coarse Strata	0.07	0.15	0.18	0.45
Residualized, No Strata	0.07	0.15	0.20	0.48
Propensity Score Matching	0.37	0.46	1.42	1.43
Raw Diff in Means	0.79	0.87	0.76	0.79

Table 2: Variance over 1000 simulations (constant treatment effect design)

2.3.2 Heterogeneous Treatment Effects

We use the same set-up as in Section 2.3.1, but now the outcome is

$$Y_i = 1 + 2X_{1i} + 4X_{2i} + \tau(1 + X_1 + X_2)T_i + \varepsilon_i$$

Again, we set $\tau = 1$. Then $\mathbb{E}(Y(1) - Y(0) \mid X_1, X_2) = 1 + X_1 + X_2$, but unconditionally, the ATE is 1. We keep the estimators the same except for OLS, to which we add interactions between the covariates and treatment.

Figure 2 shows the distribution of estimates for the seven estimators. Table 3 compares the RMSE of the seven methods and Table 4 compares their variances. OLS still dominates the other methods. However, model-based matching performs nearly as well as entropy balancing when treatment assignment is random and performs better when treatment assignment is correlated with X_1. Figure 2 shows that when the treatment is correlated with X_1, there is bias in all of the estimators besides OLS. Model-based matching reduces bias, with the smallest bias coming from the finest stratification.

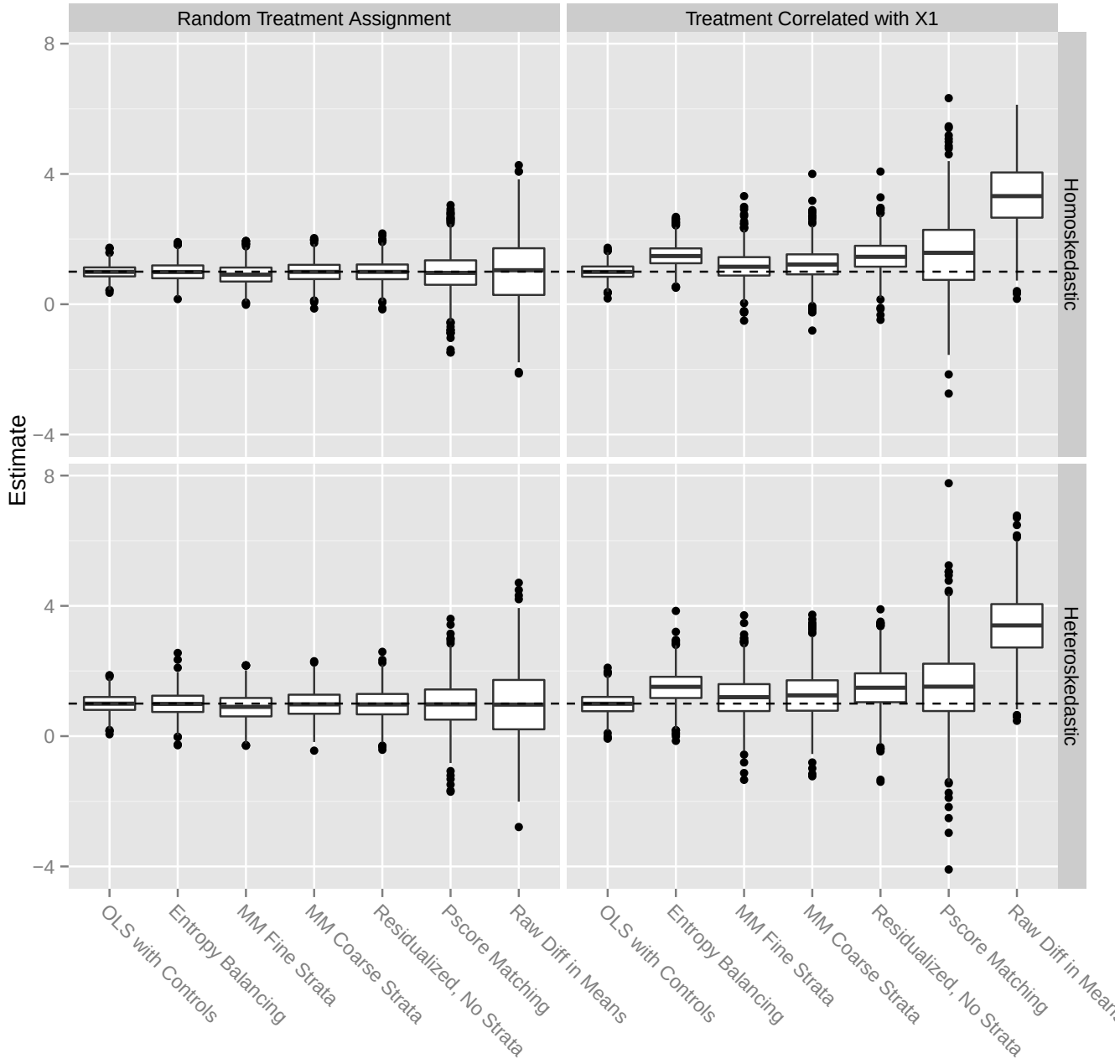

Figure 2: Estimates of the average treatment effect over 1000 simulations (heterogeneous design)

2.3.3 Conclusions

The three residualized estimators are neither best nor worst, as measured by both RMSE and variance. Of the three, the bias and RMSE are smallest when we use fine strata. Coarse strata perform nearly as well, while the unstratified estimator is the worst of the three. One reason that the residualized estimators have a greater variance than OLS and entropy balancing because the

	Random treatment assignment		Treatment Correlated with X_1	
	Homoskedastic	Heteroskedastic	Homoskedastic	Heteroskedastic
OLS with Controls	0.20	0.29	0.24	0.32
Entropy Balancing	0.29	0.37	0.60	0.72
MM Fine Strata	0.33	0.43	0.47	0.68
MM Coarse Strata	0.33	0.43	0.54	0.77
Residualized, No Strata	0.36	0.46	0.69	0.85
Propensity Score Matching	0.63	0.71	1.31	1.29
Raw Diff in Means	1.05	1.08	2.55	2.59

Table 3: RMSE over 1000 simulations (heterogeneous design)

	Random treatment assignment		Treatment Correlated with X_1	
	Homoskedastic	Heteroskedastic	Homoskedastic	Heteroskedastic
OLS with Controls	0.04	0.08	0.06	0.11
Entropy Balancing	0.08	0.14	0.12	0.26
MM Fine Strata	0.10	0.18	0.20	0.42
MM Coarse Strata	0.11	0.19	0.24	0.52
Residualized, No Strata	0.13	0.21	0.25	0.48
Propensity Score Matching	0.40	0.50	1.42	1.43
Raw Diff in Means	1.10	1.17	0.95	0.95

Table 4: Variance over 1000 simulations (heterogeneous design)

sample size used to estimate the effect is cut when we split the sample into test and training sets. One work-around would be the repeated sample splitting proposed by Abadie et al. [2013]: one would carry out the sample splitting procedure and estimate the treatment effect on a held-out test set M times, then average over the M estimates to obtain a grand estimate.

In the constant treatment effects example, OLS and entropy balancing appear to work well. Compared to the raw difference in means, the residualized estimators improve RMSE and reduce variance by introducing a small amount of bias. In the heterogeneous treatment effects example, OLS dominates. Model-based matching is next best because stratification reduces bias compared to the other methods. Without knowing a priori whether treatment effects are constant or heterogeneous, it is unclear what the model-based matching estimators will do to the bias-variance trade-off. Model-based matching is most beneficial, compared to other methods, when outcomes are highly variable and the treatment effect varies according to observed covariates.

3 Hypothesis Testing

Fisher [1935] developed and popularized the use of permutation inference to analyze randomized

experiments. Combined with the Neyman-Rubin causal model, they offer a powerful framework for assessing the effect of a treatment. In this framework, one assumes that individuals' potential outcomes are fixed and what is random is the treatment assignment. Permutation inference tests the strong null hypothesis that there is no effect whatsoever, individual by individual. Under the null, one knows both the mechanism of random treatment assignment and both potential outcomes – namely, they're equal. This information is sufficient to compute the null distribution of any test statistic. Such a test is guaranteed to have the correct significance level. Furthermore, one can choose the test statistic to maximize power against any alternative. For instance, if one is comparing two experimental groups and wants power against the alternative hypothesis that treatment shifts the response by a constant, then a popular test statistic is the mean of the treatment group outcomes. The Wilcoxon rank sum test is a particular case of a permutation test comparing two groups, obtained by replacing outcomes by their ranks (Wilcoxon [1945]).

Classical statistical tests are used more prominently than permutation tests. The most commonly used tests are based on likelihood ratios of parametric distributions (Neyman and Pearson [1933]). Such tests assume that the data come from some parametric distribution and test the null hypothesis that some parameter(s) of the distribution take a certain value. Likelihood ratio tests have been shown to be uniformly most powerful in certain situations, such as testing a simple null hypothesis against a simple alternative hypothesis. Such tests are usually calibrated using asymptotic normal theory to approximate the true null distribution. In the example of comparing two experimental groups, one might test the null hypothesis of no average treatment effect by conducting a t-test for the coefficient of the treatment indicator in a linear regression. This parametric null hypothesis is a weaker null: it amounts to hypothesizing that there is no effect *on average*. While the strong null implies the weak null, the converse is not true. In this framework, the outcomes are random variables; randomness comes from sampling from one or more distributions, not from treatment assignment per se. While some parametric tests are asymptotically equivalent to the permutation inference (Samii and Aronow [2012]), the methods may give significantly different results in finite samples.

We propose several modifications to the standard permutation test to render it amenable to observational studies. These modifications follow the same idea as the previous section: we rely on the predicted control outcome, $\hat{Y}$. We use the predictions to residualize the outcomes. This improves precision by removing extraneous variance due to covariates X (Rosenbaum [2002]). In addition, we stratify on the predicted outcomes. The reason for stratification is twofold. First, it

allows us some room for error in the predicted outcomes, as long as units within strata are "close enough." Second, it allows for the possibility that the outcome depends on the treatment in a non-constant way. We only look for associations between treatment and outcome within small ranges of the predicted control outcome. In effect, we relax the typical assumption of a constant treatment effect to one of locally constant effects.

In the next section, we develop the notation for permutation inference in randomized experiments. Next, we extend the idea to observational studies, where treatment assignment is "as if" random among subsets of individuals. We discuss the assumptions necessary for the two-step residualization and stratification test to work. Finally, we present simulations comparing the proposed method with common hypothesis tests for a treatment effect.

3.1 Hypothesis testing in randomized experiments

Suppose we run a completely randomized experiment with N individuals, N_t of whom receive treatment. We use the same notation as above: T_i is an indicator for whether individual i received treatment, $(Y_i(0), Y_i(1))$ are individual i's potential outcomes under the control and treatment regimes, respectively, and S_i is unit i's stratum assignment.

Let $\mathcal{W}$ denote the set of all possible treatment assignment vectors $\mathbf{T}$. Under complete randomization, there are $\binom{N}{N_t}$ equally likely elements in $\mathcal{W}$. Suppose we'd like to test the strong null hypothesis of no treatment effect against the alternative hypothesis that there is some effect:

$$H_0 : Y_i(1) = Y_i(0) \text{ for all } i = 1, \ldots, N \tag{4}$$

$$H_1 : Y_i(1) \neq Y_i(0) \text{ for some } i \tag{5}$$

Under the strong null, we know both potential outcomes for every individual because they are identical. We test the null by computing some test statistic $\tau(\mathbf{Y}, \mathbf{T})$, a function of treatment assignment $\mathbf{T} = (T_1, \ldots, T_N)$ and observed responses $\mathbf{Y} = \mathbf{Y}(\mathbf{T}) = (Y_1(T_1), \ldots, Y_N(T_N))$. Note that under the strong null, $\mathbf{Y}$ does not actually depend on $\mathbf{T}$, and so for any permutation $\mathbf{T}^*$ of treatment assignments, $\mathbf{Y}(\mathbf{T}) = \mathbf{Y}(\mathbf{T}^*)$.

The null distribution of τ is given by the distribution of $\tau(\mathbf{Y}, \mathbf{T}^*)$ for all $\mathbf{T}^* \in \mathcal{W}$. In principle, we could compute all $\binom{N}{N_t}$ possible values of $\tau(\mathbf{Y}, \mathbf{T}^*)$ to find the exact distribution of τ. In practice, we estimate the null distribution by permuting the treatment assignments across individ-

uals (essentially sampling from $\mathcal{W}$ with equal probability) for some large number of times B and calculating $\tau(\mathbf{Y}, \mathbf{T}_1^*), \ldots, \tau(\mathbf{Y}, \mathbf{T}_B^*)$. The p-value is approximated by comparing our observed test statistic to the estimated distribution. Extreme values of $\tau(\mathbf{Y}, \mathbf{T})$ give evidence for rejecting the null hypothesis.

We would like to stabilize the variance of Y by removing some extra variance coming from known covariates X. Rosenbaum [2002] devises more powerful randomization tests by replacing outcomes by residuals after predicting the outcome with covariates. Given a function $\hat{Y}$ which predicts $Y(0)$ using X, the residuals $r_i = Y_i - \hat{Y}_i$ are fixed values computed from the data, not stochastic. Even if the randomization had been different, the residuals would be the same. Under the null hypothesis, $Y_i(0) = Y_i(1)$ for all i, so we may think of $\hat{Y}$ as predicting the outcome using no information on treatment assignment. Then the vector of observed residuals $\mathbf{r} = \mathbf{r}(\mathbf{T})$ does not actually vary with the treatment, so $\mathbf{r}(\mathbf{T}) = \mathbf{r}(\mathbf{T}^*)$ for all $\mathbf{T}, \mathbf{T}^* \in \mathcal{W}$.

Further, one may conduct a conditional test by conditioning on an ancillary statistic and thereby restricting the set of possible treatment assignments. For instance, we may want to incorporate information on the S strata in order to preserve aspects of the covariate distributions of the treatment and control groups. Stratifying allows us to detect subgroup treatment effects that may be obscured if we only take averages over all individuals. Then instead of permuting treatment assignments of all individuals, we would only permute treatment assignment within strata, independently across strata. This corresponds to restricting the set of all possible treatment assignments to some $\mathcal{W}_S \subset \mathcal{W}$. We show in the next section that this conditional test is also valid.

3.2 Hypothesis testing in observational studies

Suppose that instead of a randomized experiment, we observe N individuals, N_t of whom received the active treatment. We would like to test hypothesis (4) against (5). However, now we don't know the treatment assignment mechanism. In particular, the elements of $\mathcal{W}$ of may not be equally likely treatment assignments. We'd like to identify an ancillary statistic such that conditioning on it yields a set $\mathcal{W}_S$ where the elements are all equally likely. Assumption 1 gives us one, namely strata of $\hat{Y}$.

Suppose we've set up a stratification rule Ψ_s such that $S_i = \Psi_s(X_i)$ is the stratum assignment of unit i. Once we've observed $(\mathbf{Y}, \mathbf{X})$, we can compute $\mathbf{S}$. Then the number of individuals in each stratum $N_1, \ldots, N_S$ are fixed. The number of treated and control units in each stratum partition

the set of possible randomizations $\mathcal{W}$. For each set of positive integers $\{N_{1t}, \ldots, N_{St}, N_{1c}, \ldots, N_{Sc}\}$ satisfying

1. $N_{jt} + N_{jc} = N_j$ for all j
2. $\sum_{j=1}^{S} N_{jt} = N_t$
3. $\sum_{j=1}^{S} N_{jc} = N_c$

we define the subset

$$\mathcal{W}_s = \left\{ \mathbf{T} \in \mathcal{W} \mid N_{jt} = \sum_{i:S_i=j} T_i, \quad N_{jc} = \sum_{i:S_i=j} (1 - T_i), \quad N_{jt} + N_{jc} = N_j \ \ \forall j \right\}$$

These subsets form a partition of $\mathcal{W}$: that is, $\cap_s \mathcal{W}_s = \emptyset$ and $\cup_s \mathcal{W}_s = \mathcal{W}$. Conditioning on the set $\mathcal{W}_s$ of treatment assignments corresponding to the realized values of $\{N_{1t}, \ldots, N_{St}, N_{1c}, \ldots, N_{Sc}\}$, i.e., permuting treatment assignments within strata and independently across strata, yields a valid test.

For the stratified permutation test, we use the absolute value of the difference in means within strata, averaged across strata. As above, we replace outcomes by their residuals, $r_i = Y_i - \hat{Y}_i$. The test statistic is

$$\tau(\mathbf{r}, \mathbf{T}) = \sum_{s=1}^{S} \frac{N_s}{N} \left| \frac{1}{N_{st}} \sum_{\substack{i:S_i=s \\ T_i=1}} r_i - \frac{1}{N_{sc}} \sum_{\substack{i:S_i=s \\ T_i=0}} r_i \right|$$

In order for the stratified permutation test based on residuals, rather than outcomes, to be valid, we need exchangeability within strata. In addition to Assumption 1, we need the following stronger assumption:

Assumption 2. *$\hat{Y}$ is constant within strata. That is, for all units i and j with $S_i = S_j = s$, $\hat{Y}_i = \hat{Y}_j$.*

We can obtain constant predictions within strata one of two ways: either by stratifying extremely finely if we have sufficient data, or by using a prediction rule that gives piecewise constant predictions, such as a tree-based method.[2] Together, Assumptions 1 and 2 make the residuals exchangeable within strata. They imply the following:

[2] Wager and Athey [2015] make this assumption about the leaves of their causal trees. In effect, their estimates are constant within leaves.

Assumption 3. *Conditional independence between residuals and treatment given stratum assignment:* $(r(1), r(0)) \perp\!\!\!\perp T \mid S$.

If Assumption 1 or 2 is violated, Assumption 3 is sufficient to create a valid permutation test.

Theorem 3.1. *Under Assumptions 1 and 2 or Assumption 3, the stratified permutation test of residuals controls the type one error rate at level* α.

Proof.

We follow the notation of Hennessy et al. [2015]. Suppose we observe treatment assignment $\mathbf{T}$ and test statistic $\tau_{obs} = \tau(\mathbf{r}, \mathbf{T})$. Let $p_s = \mathbb{P}\left(|\tau(\mathbf{r}, \mathbf{T}^*)| \geq |\tau_{obs}| \mid \mathcal{W}_s\right)$ be the p-value of the conditional test given $\{N_{1t}, \ldots, N_{St}, N_{1c}, \ldots, N_{Sc}\}$. Let U be a random variable with the same distribution as $|\tau(\mathbf{r}, \mathbf{T}^*)|$ conditional on $\mathbf{T}^* \in \mathcal{W}_s$ and let F_U be its CDF.

$$\begin{aligned} p_s &= \mathbb{P}\left(|\tau(\mathbf{r}, \mathbf{T}^*)| \geq |\tau_{obs}| \mid \mathbf{T}^* \in \mathcal{W}_s\right) \\ &= 1 - F_U(|\tau_{obs}|) \\ &= 1 - F_U(|\tau(\mathbf{r}, \mathbf{T})|) && \text{by definition of } \tau_{obs} \end{aligned}$$

p_s has a distribution over all possible randomizations $\mathbf{T} \in \mathcal{W}_S$. Since all elements $\mathbf{T}^*$ are equally likely under the assumption(s),

$$\begin{aligned} p_s &\stackrel{d}{=} 1 - F_U(|\tau(\mathbf{r}, \mathbf{T}^*)|) \\ &\stackrel{d}{=} 1 - F_U(U) \end{aligned}$$

Since $F_U(U)$ has a uniform distribution on $[0, 1]$, so does p_s. Therefore, $\mathbb{P}(p_s \leq \alpha \mid H_0) \leq \alpha$. It remains to show that the unconditional test has level no greater than α. We use the law of total

probability.

$$\begin{aligned}
\mathbb{P}(\text{reject } H_0 \mid\mid H_0) &= \sum_s \mathbb{P}(p_s \leq \alpha \mid \mathbf{T} \in \mathcal{W}_s \mid\mid H_0)\mathbb{P}(\mathbf{T} \in \mathcal{W}_s) \\
&\leq \sum_s \alpha \mathbb{P}(\mathbf{T} \in \mathcal{W}_s) \\
&= \alpha \sum_s \mathbb{P}(\mathbf{T} \in \mathcal{W}_s) \\
&= \alpha
\end{aligned}$$

□

Assumption 1 is crucial for the conditional test to have level α. The assumption that treatment assignments in $\mathcal{W}_s$ are equally likely would give the same result.[3] Furthermore, combining Assumption 2 with Assumption 1 is important when we take residuals: $\hat{Y}$ should be constant within strata so the exchangeability of $(Y(1), Y(0))$ translates to exchangeability of the residuals. It is unclear to what extent Assumption 2 may be violated before the test becomes invalid; this will be data-dependent. The stratified permutation test is valid for fully randomized experiments, because randomization guarantees independence between treatment and potential outcomes and that every treatment assignment is equally likely. If some individuals within a stratum are more likely than others to receive treatment (i.e., they have different propensity scores), then permuting treatments with equal probability will not estimate the correct null distribution.

3.3 Empirical Results

3.3.1 Constant Additive Treatment Effects

The first set of simulations follows the set up in Section 2.3.1. We draw 100 random X values. Then, for each draw, we randomly generate ϵ and treatment assignments according to the procedures described in Section 2.3.1. We vary the treatment effect from $\tau = 0$ (null hypothesis, no treatment effect), to 0.25 and 0.5. We compare five tests of the strong null: the usual t-test of the OLS coefficient in a regression with controls, the Wilcoxon rank sum test, model-based matching with fine and coarse strata, and Rosenbaum's permutation test of residualized outcomes with no stratification.

[3] Rosenbaum [2002] proposes a similar method for observational studies, grouping observations on their estimated propensity scores rather than predicted outcomes. It follows from the theorems in Rosenbaum and Rubin [1983] that the potential outcomes will be independent of treatment within groups.

Figure 3 shows the power curves: they plot the nominal significance level against the probability of rejecting the null hypothesis. The top row shows results for a treatment effect of size 0; these curves show the level of the test. If the test has the correct level, the curves should overlap with the dashed identity line. The tests appear to have the correct level when treatment assignment is random, but only OLS has the right level when treatment is correlated with X_1. The Wilcoxon rank sum test does not control for X_1, so it incorrectly rejects the null far more often than the nominal level. Aside from the Wilcoxon rank sum test, which always has high power in the correlated case, OLS tends to have the highest power. The residualized, unstratified test has next highest power, while the test with finest stratification has the lowest power. This is likely because the number of possible treatment assignments decreases as the strata become smaller, so the effect must be more extreme to detect it.

Table 5 shows the actual level of the test when we reject the null hypothesis at level 0.05. This is a snapshot of the first row of curves in Figure 3. The results are worrisome: only OLS appears to have the correct level when treatment is correlated with X_1. These poor results suggest that Assumption 1 is not met by the stratification chosen. The finest stratification improves the test level compared to the other residualized tests, but it is still higher than nominal level. Perhaps an even finer stratification would grant Assumption 1.

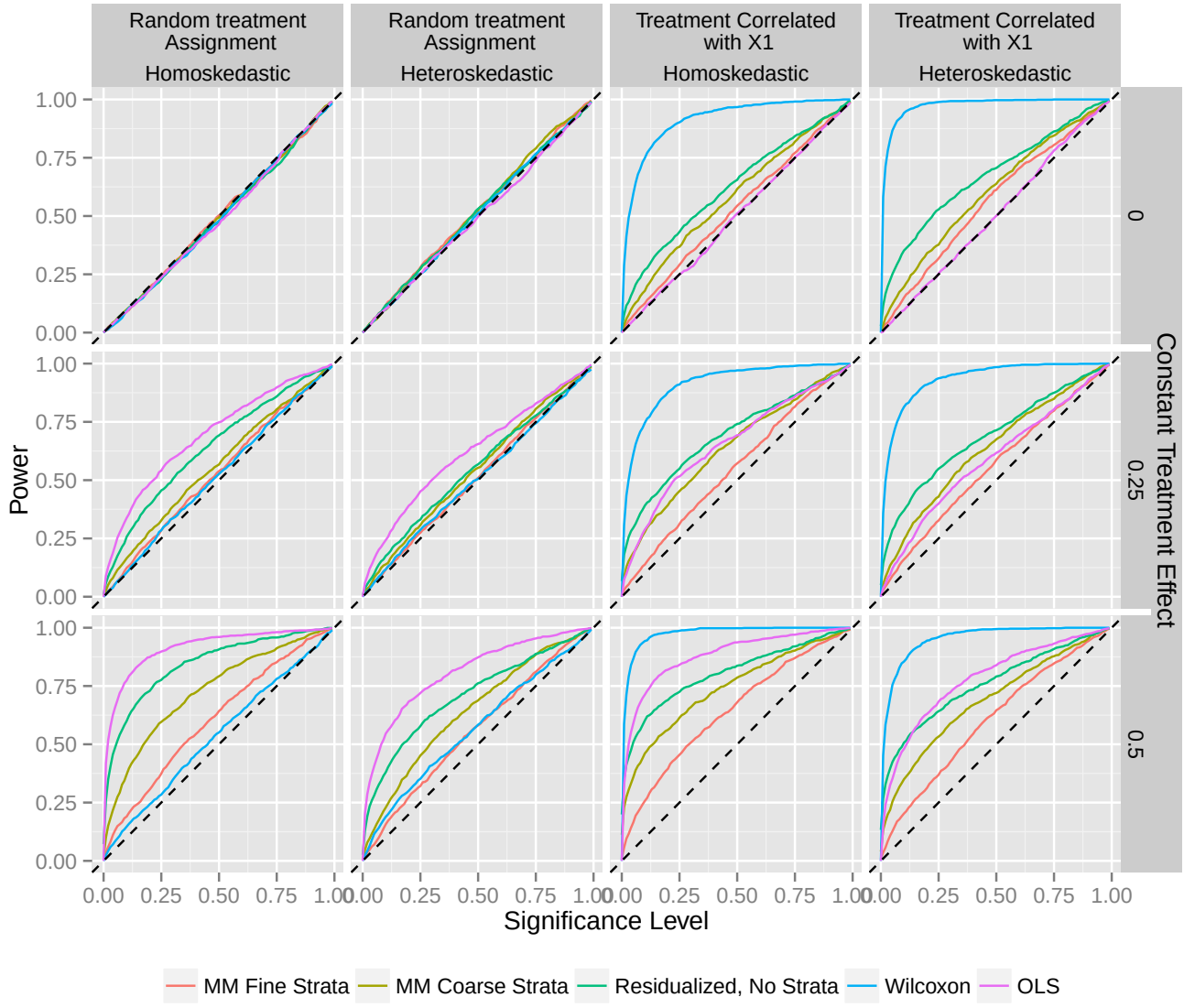

Figure 3: Power curves for increasing constant additive treatment effects

	Random treatment assignment		Treatment Correlated with X_1	
	Homoskedastic	Heteroskedastic	Homoskedastic	Heteroskedastic
MM Fine Strata	0.043	0.061	0.073	0.072
MM Coarse Strata	0.050	0.051	0.115	0.122
Residualized, No Strata	0.045	0.058	0.180	0.251
Wilcoxon	0.038	0.052	0.596	0.867
OLS	0.048	0.059	0.052	0.048

Table 5: Proportion of tests rejected at nominal level 0.05 out of 1000 simulations in the constant treatment effect set-up. If the true level is 0.05, these proportions have a standard error of 0.007.

3.3.2 Heterogeneous Treatment Effects

We follow the set-up in Section 2.3.2. Again, we set the average treatment effect to be $\tau = 0$ (null hypothesis, no treatment effect), 0.25, and 0.5. Figure 4 shows the power curves. Unlike in the previous section, OLS does not always dominate. In fact, the residualization with no strata and the coarse strata tests appear neck-in-neck with OLS. However, only the finely stratified model-based matching test has the correct level when treatment is correlated with X_1, as shown in the top row of plots in Figure 4 and in Table 6.

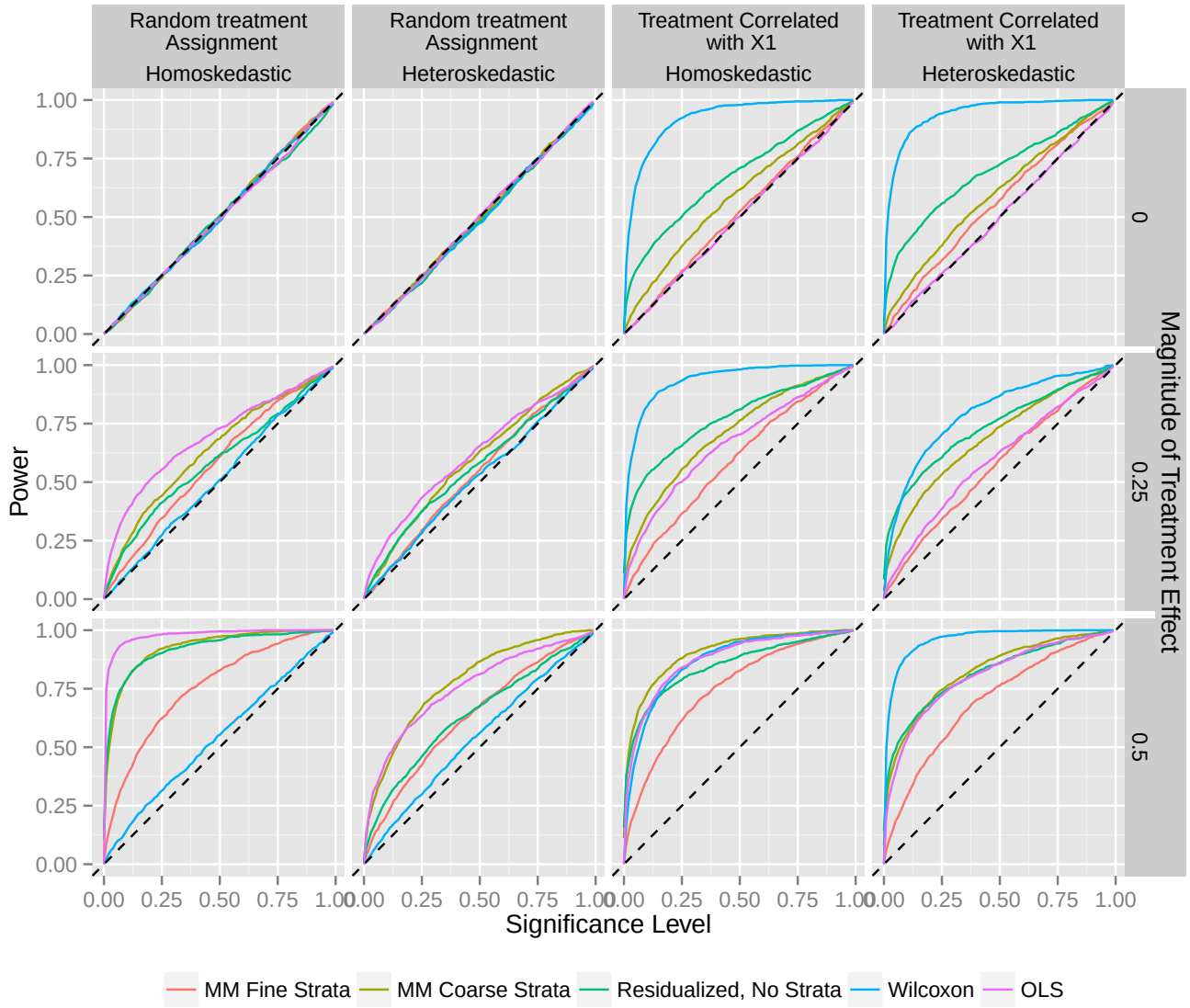

Figure 4: Power curves for increasing magnitude heterogeneous treatment effects

	Random treatment assignment		Treatment Correlated with X_1	
	Homoskedastic	Heteroskedastic	Homoskedastic	Heteroskedastic
MM Fine Strata	0.044	0.053	0.048	0.083
MM Coarse Strata	0.048	0.049	0.116	0.127
Residualized, No Strata	0.043	0.047	0.269	0.312
Wilcoxon	0.052	0.045	0.631	0.713
OLS	0.052	0.043	0.043	0.044

Table 6: Proportion of tests rejected at nominal level 0.05 out of 1000 simulations in the heterogeneous treatment effect set-up. If the true level is 0.05, these proportions have a standard error of 0.007.

3.3.3 Conclusions

The practical usefulness of this model-based matching test may be limited. The trade-off between type I and type II errors is unfavorable. We need a fine stratification to satisfy Assumption 1. If the strata are too big, then individuals are not exchangeable, and the null distribution that we approximate by permutations within strata is wrong. On the other hand, we have very little power to detect an effect when strata are small. Figures 3 and 4 show this: the orange power curve corresponding to the fine stratification test is consistently near the dashed identity line.

4 Discussion

We propose a two-step method for estimation and hypothesis testing of causal effects. Both steps rely on predicting the prognostic score of all individuals. The prognostic score has nice properties: in particular, Hansen [2008] shows that it is a balancing score. Thus, stratifying on the estimated control outcome approximates a random experiment within strata, conditional on Assumption 1. Additionally, subtracting the predicted control outcome from the observed outcome reduces variation in outcomes that comes from covariates. Thus, the two-step procedure of residualizing and stratifying may improve both bias and variance in estimation, and with additional assumptions may render a permutation test possible for observational data.

The key benefit of this method is that it may be less sensitive to non-robust data and better able to pick up on heterogeneous treatment effects. We'd hope that this method would work better than propensity score matching when there is limited overlap in covariates which predict treatment, but there is sufficient overlap in covariates that predict prognostic scores. This way, we may bypass the classical framework of Rosenbaum and Rubin [1983]; this is a different philosophical way to

think of observational studies. In addition, the proposed method improves upon existing methods by eliminating the need for parametric assumptions such as Gaussianity, homoscedastic errors, linearity, and constant treatment effects. Furthermore, the method for hypothesis testing is fully generalizable to any form of treatment and outcome. We focus on the case of binary treatment and continuous outcomes in this paper. A more general version of the method has been applied in a study of the effect of packstock use on an endangered toad population in Yosemite National Park (Matchett et al. [2015]), using correlation between treatment intensity and a continuous outcome. By choosing an appropriate test statistic, the test is customizable to any type of data.

Ideally, we'd like this framework to relate estimation and testing in a clear way. It doesn't. The first issue arises from the residuals. To estimate the ATE, we must use a function $\hat{f}(X) = \hat{Y}$ which predicts the outcome under control. One should estimate $\hat{f}$ using controls only. On the other hand, for hypothesis testing, we must estimate $\hat{f}$ using both treated and controls. Appendix Section 5.2 works through an example to illustrate why this is the case. In short, fitting to controls ensures that no treatment effect enters into the prediction and allows us to estimate the full magnitude of the effect, while fitting to both treated and controls captures some correlation between treatment and covariates that allows for a fair comparison of residuals between the two groups.

The stratification also complicates the relation between testing and estimation. Matching estimators are non-smooth and nonlinear, making it difficult to find their distribution in finite samples. Typically, one approximates complex distributions using asymptotic arguments or by bootstrapping. For instance, Abadie and Imbens [2006] derive the asymptotic distribution of a matching estimator of the average treatment effect on the treated (ATT) for the purpose of making confidence intervals, while Abadie and Imbens [2008] show that the bootstrap fails to correctly estimate the variance of matching estimators in finite samples, making bootstrap confidence intervals invalid. Instead of relying on asymptotics, one may use test inversion to create confidence intervals. However, there is no clear way to invert a stratified permutation test. To do so, one must assume some functional form of the treatment effect, which may vary in complicated ways across strata.

Both the estimator and hypothesis tests rely on the strong assumption that potential outcomes are independent of treatment assignment within strata. Under this assumption, it is as though each stratum approximates a randomized experiment. However, this requires us to identify the "correct" stratification. In our simulations, we simply stratified by quantiles of the observed $\hat{Y}$ and this appeared to work poorly. Furthermore, this assumption hinges on "selection on observables," or incorporating all true confounding variables. In practice, this is an untestable assumption.

More work should be done to study strata selection methods and to study the performance of this framework when the outcome model is misspecified.

5 Appendix

5.1 Estimator

We prove a useful lemma before our main result.

Lemma 5.1. *If* $(Y(1), Y(0)) \perp\!\!\!\perp T \mid S$ *and treatment is assigned independently across units within stratum* s*, then*

$$\mathbb{E}\left(\frac{T_i}{N_{st}} \mid S_i = s, \sum_i \mathbb{I}(S_i = s) = N_s\right) = \frac{1}{N_s}$$

Similarly, $\mathbb{E}\left(\frac{1-T_i}{N_{sc}} \mid S_i = s, \sum_i \mathbb{I}(S_i = s) = N_s\right) = \frac{1}{N_s}$.

Proof.

$$\begin{aligned}
\mathbb{E}\left(\frac{T_i}{N_{st}} \mid S_i = s, \sum_i \mathbb{I}(S_i = s) = N_s\right) &= \mathbb{E}\left(\frac{1}{N_{st}}\mathbb{E}(T_i \mid N_{st}, S_i = s, \sum_i \mathbb{I}(S_i = s) = N_s)\right) \\
&= \mathbb{E}\left(\frac{1}{N_{st}}\frac{N_{st}}{N_s}\right) \\
&= \frac{1}{N_s}
\end{aligned}$$

□

We use the previous lemma to prove the main result:

Proof of Theorem 2.2. Conditional on $N_1, \ldots, N_S$ we have

$$\begin{aligned}
\mathbb{E}(\hat{\tau}^{strat}) &= \mathbb{E}\left[\sum_{s=1}^{S} \frac{N_s}{N}\left(\frac{1}{N_{st}} \sum_{i:T_i=1,S_i=s} (Y_i - \hat{Y}_i) - \frac{1}{N_{sc}} \sum_{i:T_i=0,S_i=s} (Y_i - \hat{Y}_i)\right)\right] \\
&= \mathbb{E}\left[\sum_{s=1}^{S} \frac{N_s}{N}\left(\frac{1}{N_s} \sum_{i:S_i=s} \frac{T_i(Y_i(1) - \hat{Y}_i)}{N_{st}/N_s} - \frac{(1-T_i)(Y_i(0) - \hat{Y}_i)}{N_{sc}/N_s}\right)\right] \\
&= \mathbb{E}\left[\sum_{s=1}^{S} \frac{1}{N}\left(\sum_{i:S_i=s} \mathbb{E}\left(\frac{T_i}{N_{st}/N_s} \mid S_i = s\right) \mathbb{E}(Y_i(1) - \hat{Y}_i \mid S_i = s)\right.\right. \\
&\qquad \left.\left. -\mathbb{E}\left(\frac{1-T_i}{N_{sc}/N_s} \mid S_i = s\right) \mathbb{E}(Y_i(0) - \hat{Y}_i \mid S_i = s)\right)\right] \\
&= \mathbb{E}\left[\sum_{s=1}^{S} \frac{1}{N}\left(\sum_{i:S_i=s} \frac{1/N_s}{1/N_s} \mathbb{E}(Y_i(1) - \hat{Y}_i \mid S_i = s) - \frac{1/N_s}{1/N_s} \mathbb{E}(Y_i(0) - \hat{Y}_i \mid S_i = s)\right)\right] \\
&= \mathbb{E}\left[\sum_{s=1}^{S} \frac{1}{N}\left(\sum_{i:S_i=s} \mathbb{E}(Y_i(1) - \hat{Y}_i \mid S_i = s) - \mathbb{E}(Y_i(0) - \hat{Y}_i \mid S_i = s)\right)\right] \\
&= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left(Y_i(1) - Y_i(0) \mid S_i = s\right)\right] \\
&= ATE
\end{aligned}$$

Then, taking expectations with respect to $N_1, \ldots, N_S$, we have that $\mathbb{E}(\hat{\tau}^{strat}) = ATE$ unconditionally as well. □

Proof of Theorem 2.3. Define $r_i = Y_i - \hat{Y}_i$ to be the observed residualized outcome in the sample. We rewrite the estimator as

$$\hat{\tau}^{adj} = \overline{r_t} - \overline{r_c}$$

where $\overline{r_t}$ and $\overline{r_c}$ are the average residualized outcomes in the treatment and control groups, respectively. The residualized potential outcomes are then defined as $r_i(1) = Y_i(1) - \hat{Y}_i$ and $r_i(0) = Y_i(0) - \hat{Y}_i$. Let $\overline{r_i(1)}$ and $\overline{r_i(0)}$ be the average residualized potential outcomes in the sample and $\overline{R_i(1)}$ and $\overline{R_i(0)}$ be the analogous population quantities: that is, $\mathbb{E}_{SP}(r_i(1)) = \overline{R_i(1)}$ and $\mathbb{E}_{SP}(r_i(0)) = \overline{R_i(0)}$.

$$
\begin{aligned}
\mathrm{Var}(\hat{\tau}^{adj}) &= \mathrm{Var}\left(\overline{r_t} - \overline{r_c}\right) \\
&= \mathbb{E}\left[\left(\overline{r_t} - \overline{r_c} - \mathbb{E}_{SP}\left[r(1) - r(0)\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(\overline{r_t} - \overline{r_c} - \left(\overline{r(1)} - \overline{r(0)}\right) + \left(\overline{r(1)} - \overline{r(0)}\right) - \mathbb{E}_{SP}\left[r(1) - r(0)\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(\overline{r_t} - \overline{r_c} - \left(\overline{r(1)} - \overline{r(0)}\right)\right)^2\right] + \mathbb{E}\left[\left(\left(\overline{r(1)} - \overline{r(0)}\right) - \mathbb{E}_{SP}\left[r(1) - r(0)\right]\right)^2\right] \\
&\qquad + 2\mathbb{E}\left[\left(\overline{r_t} - \overline{r_c} - (\overline{r(1)} - \overline{r(0)})\right)\left((\overline{r(1)} - \overline{r(0)}) - \mathbb{E}_{SP}\left[r(1) - r(0)\right]\right)\right]
\end{aligned}
$$

Note that the expectations above are taken over both the random sampling from the superpopulation and the random assignment of treatments. The third term is zero because, after conditioning on the observed $r_1(1), \ldots, r_N(1), r_1(0), \ldots, r_N(0)$, the expected value of the first factor is 0.

The first term simplifies to

$$
\begin{aligned}
\mathbb{E}\left[\left(\overline{r_t} - \overline{r_c} - \left(\overline{r(1)} - \overline{r(0)}\right)\right)^2\right] &= \mathbb{E}\left[\left((\overline{r_t} - \overline{r(1)}) - (\overline{r_c} - \overline{r(0)})\right)^2\right] \\
&= \mathbb{E}\left[(\overline{r_t} - \overline{r(1)})^2\right] + \mathbb{E}\left[(\overline{r_c} - \overline{r(0)})^2\right] - 2\mathbb{E}\left[(\overline{r_t} - \overline{r(1)})(\overline{r_c} - \overline{r(0)}))\right] \\
&= \frac{\mathrm{Var}(r(1))}{N_t} + \frac{\mathrm{Var}(r(0))}{N_c} - \frac{\mathrm{Var}(r(1) - r(0))}{N}
\end{aligned}
$$

by finite sample results (Imbens and Rubin [2015]). The second term is simply the variance of the difference in means of all potential outcomes. Thus, by definition

$$
\mathbb{E}\left[\left((\overline{r(1)} - \overline{r(0)}) - \mathbb{E}_{SP}\left[r(1) - r(0)\right]\right)^2\right] = \frac{\mathrm{Var}(r(1) - r(0))}{N}
$$

Combining the three terms gives

$$
\mathrm{Var}(\hat{\tau}^{adj}) = \frac{\mathrm{Var}(r(1))}{N_t} + \frac{\mathrm{Var}(r(0))}{N_c} \tag{6}
$$

x

In terms of known quantities:

$$\begin{aligned}\text{Var}(r(1)) &= \text{Var}(Y(1) - \hat{Y}) \\ &= \text{Var}(Y(1)) + \text{Var}(\hat{Y}) - 2\text{Cov}(Y(1), \hat{Y}) \\ &= \sigma_1^2 + \nu^2 - 2\rho_1\sigma_1\nu\end{aligned}$$

where $\nu^2 := \text{Var}(\hat{Y})$ and ρ_1 is the correlation between $Y(1)$ and $\hat{Y}$. This expression can be rearranged as

$$\text{Var}(r(1)) = (\sigma_1 - \nu)^2 - 2(1 - \rho_1)\sigma_1\nu$$

Similarly,

$$\text{Var}(r(0)) = (\sigma_0 - \nu)^2 - 2(1 - \rho_0)\sigma_0\nu$$

where ρ_0 is the correlation between $Y(0)$ and $\hat{Y}$. Plugging these expressions into equation 6 gives the desired result. □

Proof of Theorem 2.4. The stratified estimator variance can be written

$$\text{Var}\left(\hat{\tau}^{strat}\right) = \sum_{s=1}^{S} \frac{N_s^2}{N^2}\text{Var}(\hat{\tau}_s) + 2\sum_{s=1}^{S-1}\sum_{r=s+1}^{S} \frac{N_r N_s}{N^2}\text{Cov}(\hat{\tau}_r, \hat{\tau}_s) \tag{7}$$

If $\text{Var}(\hat{\tau}_s) \leq \text{Var}(\hat{\tau}^{adj})$ for all s, then[4]

$$\begin{aligned}\sum_{s=1}^{S} \frac{N_s^2}{N^2}\text{Var}(\hat{\tau}_s) &\leq \sum_{s=1}^{S} \frac{N_s^2}{N^2}\text{Var}(\hat{\tau}^{adj}) \\ &\leq \text{Var}(\hat{\tau}^{adj}) \left(\sum_{s=1}^{S} \frac{N_s}{N}\right)^2 && \text{by Jensen's inequality} \\ &= \text{Var}(\hat{\tau}^{adj})\end{aligned}$$

Now, for any strata r and s,

[4] This assumption is plausible, since we hope that outcomes are less variable within strata than across strata.

$$\text{Cov}(\hat{\tau}_r, \hat{\tau}_s) = \text{Cov}\left(\frac{1}{N_{rt}}\sum_{i:S_i=r} T_i r_i - \frac{1}{N_{rc}}\sum_{i:S_i=r}(1-T_i)r_i, \frac{1}{N_{st}}\sum_{i:S_i=s} T_i r_i - \frac{1}{N_{sc}}\sum_{i:S_i=s}(1-T_i)r_i\right) \quad (8)$$

Let's focus on the first term. The remaining three may be simplified similarly.

$$\begin{aligned}
\text{Cov}\left(\frac{1}{N_{rt}}\sum_{i:S_i=r} T_i r_i, \frac{1}{N_{st}}\sum_{i:S_i=s} T_i r_i\right) &= \mathbb{E}\left(\frac{1}{N_{rt}N_{st}}\sum_{i:S_i=r}\sum_{j:S_j=s} T_i T_j r_i r_j\right) - \\
&\quad \mathbb{E}\left(\frac{1}{N_{rt}}\sum_{i:S_i=r} T_i r_i\right)\mathbb{E}\left(\frac{1}{N_{st}}\sum_{i:S_i=s} T_i r_i\right) \\
&= \frac{1}{N_{rt}N_{st}}\sum_{i:S_i=r}\sum_{j:S_j=s}\mathbb{E}(T_i T_j) r_i r_j - \\
&\quad \frac{1}{N_{rt}N_{st}}\left(\sum_{i:S_i=r}\mathbb{E}(T_i) r_i\right)\left(\sum_{j:S_j=s}\mathbb{E}(T_j) r_j\right) \\
&= \frac{N_t(N_t-1)}{N(N-1)N_{rt}N_{st}}\sum_{i:S_i=r}\sum_{j:S_j=s} r_i r_j - \frac{N_t^2}{N^2 N_{rt}N_{st}}\left(\sum_{i:S_i=r} r_i\right)\left(\sum_{j:S_j=s} r_j\right) \\
&= \frac{N_t(N_t-1)}{N(N-1)}\overline{r_r(1)}\ \overline{r_s(1)} - \frac{N_t^2}{N^2}\overline{r_r(1)}\ \overline{r_s(1)} \\
&= -\frac{N_t N_c}{N^2(N-1)}\overline{r_r(1)}\ \overline{r_s(1)}
\end{aligned}$$

Thus,

$$\begin{aligned}
\text{Cov}(\hat{\tau}_r, \hat{\tau}_s) &= -\frac{N_t N_c}{N^2(N-1)}\left[\overline{r_r(1)}\ \overline{r_s(1)} + \overline{r_r(1)}\ \overline{r_s(0)} + \overline{r_r(0)}\ \overline{r_s(1)} + \overline{r_r(0)}\ \overline{r_s(0)}\right] \\
&= -\frac{N_t N_c}{N^2(N-1)}\left(\overline{r_r(1)} + \overline{r_r(0)}\right)\left(\overline{r_s(1)} + \overline{r_s(0)}\right)
\end{aligned}$$

The second term of Equation 7 is negative if

$$
\begin{aligned}
0 &> 2\sum_{s=1}^{S-1}\sum_{r=s+1}^{S}\frac{N_r N_s}{N^2}\mathrm{Cov}(\hat{\tau}_r,\hat{\tau}_s) \\
&> 2\sum_{s=1}^{S-1}\sum_{r=s+1}^{S}\frac{N_r N_s}{N^2}\left(-\frac{N_t N_c}{N^2(N-1)}\left(\overline{r_r(1)}+\overline{r_r(0)}\right)\left(\overline{r_s(1)}+\overline{r_s(0)}\right)\right) \\
0 &< 2\sum_{s=1}^{S-1}\sum_{r=s+1}^{S}N_r N_s\left(\overline{r_r(1)}+\overline{r_r(0)}\right)\left(\overline{r_s(1)}+\overline{r_s(0)}\right) \\
&= \sum_{s=1}^{S}\sum_{r=1}^{S}N_r N_s\left(\overline{r_r(1)}+\overline{r_r(0)}\right)\left(\overline{r_s(1)}+\overline{r_s(0)}\right)-\sum_{s=1}^{S}N_s^2\left(\overline{r_s(1)}+\overline{r_s(0)}\right)^2 \\
&= \left[\sum_{s=1}^{S}N_s\left(\overline{r_s(1)}+\overline{r_s(0)}\right)\right]^2-\sum_{s=1}^{S}\left[N_s\left(\overline{r_s(1)}+\overline{r_s(0)}\right)\right]^2 \\
&= \mathrm{Var}_s\left(N_s\left(\overline{r_s(1)}+\overline{r_s(0)}\right)\right)
\end{aligned}
$$

This is the variance across strata of the function $N_s\left(\overline{r_s(1)}+\overline{r_s(0)}\right)$. Since variance is always non-negative, the statement is always true. Therefore, the second term of Equation 7 is always negative, so

$$
\mathrm{Var}\left(\hat{\tau}^{strat}\right)\leq\sum_{s=1}^{S}\frac{N_s^2}{N^2}\mathrm{Var}(\hat{\tau}_s)\leq\mathrm{Var}(\hat{\tau}^{adj})
$$

□

5.2 Fitting the model

What's the difference between fitting our model for matching using all units and fitting the model using controls only? We'll use a simple example in the Neyman-Rubin framework, using linear regression prediction with one covariate to illustrate the bias introduced by fitting the model different ways. To keep things especially simple, we will not do any stratification. Let $\hat{Y}_{ctrl}$ and $\hat{\tau}_{ctrl}$ denote the predictions and the estimate, respectively, using only the controls in the training group, and let $\hat{Y}_{all}$ and $\hat{\tau}_{all}$ denote the predictions and the estimate, respectively, using all units.

Suppose we have a population of N individuals. Suppose without loss of generality that units $i = 1, \ldots, n$ receive the control condition and $i = n+1, \ldots, N$ receive treatment, where $N = 2n$. There is a constant, additive treatment effect. That is, for each individual, $Y_i(1) - Y_i(0) = \Delta$ for some $\Delta \in \mathbb{R}$. Define $\overline{c} = \frac{1}{N}\sum_{i=1}^{N} Y_i(0)$, $\overline{c}_c = \frac{1}{n}\sum_{i=1}^{n} Y_i(0)$, and $\overline{c}_t = \frac{1}{n}\sum_{i=n+1}^{N} Y_i(0)$. Define $\overline{t}, \overline{t}_c$, and $\overline{t}_t$ analogously using $Y_i(1)$ in place of $Y_i(0)$.

Suppose that $Y_i(0) = c_i + x_i$ and $Y_i(1) = c_i + x_i + \Delta$, where c_i is some baseline value and x_i is the value of the covariate. The x_i are not random quantities, but a fixed list of numbers given the population of N units. Let r_c denote the correlation between X and Y in the control group, s_{Y_c} denote the standard deviation of the control outcomes, and s_{X_c} denote the standard deviation of the covariate in the control group. Similarly, let r, s_Y, and s_X be the analogous quantities in the population of N. Using simple linear regression to predict the response without using treatment information, we have

$$\hat{Y}_{i,ctrl} = \overline{c}_c + \overline{x}_c + r_c \frac{s_{Y_c}}{s_{X_c}}(x_i - \overline{x}_c)$$
$$\hat{Y}_{i,all} = \overline{c} + \overline{x} + r \frac{s_Y}{s_X}(x_i - \overline{x})$$

Note, in practice, we would want to create the prediction rules $\hat{Y}_{ctrl}$ and $\hat{Y}_{all}$ based on independent samples. For expository clarity, we use the same sample to create the prediction rule and do the estimation and testing.[5]

[5] In the prediction rule, we would use averages from the training set instead of the test set. Then, the estimators would have additional terms to account for the difference between test and training set averages. If the test and training sets were really random samples from the same population (as they ought to be), then these terms would be small in any particular sample, and equal to zero in expectation over repeated samples.

5.2.1 Estimation

The estimators we use are

$$\hat{\tau}_{ctrl} = \frac{1}{n}\sum_{i=n+1}^{N}(Y_i - \hat{Y}_{i,ctrl}) - \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_{i,ctrl})$$
$$\hat{\tau}_{all} = \frac{1}{n}\sum_{i=n+1}^{N}(Y_i - \hat{Y}_{i,all}) - \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_{i,all})$$

We'll simplify the estimators in terms of known quantities.

$$\begin{aligned}
\hat{\tau}_{ctrl} &= \frac{1}{n}\sum_{i=n+1}^{N}(Y_i - \hat{Y}_{i,ctrl}) - \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_{i,ctrl}) \\
&= \frac{1}{n}\sum_{i=n+1}^{N}(c_i + x_i + \Delta - \overline{c}_c - \overline{x}_c - r_c\frac{s_{Y_c}}{s_{X_c}}(x_i - \overline{x}_c)) - \frac{1}{n}\sum_{i=1}^{n}(c_i + x_i - \overline{c}_c - \overline{x}_c - r_c\frac{s_{Y_c}}{s_{X_c}}(x_i - \overline{x}_c)) \\
&= \Delta + \frac{1}{n}\sum_{i=n+1}^{N}(c_i + x_i - r_c\frac{s_{Y_c}}{s_{X_c}}(x_i - \overline{x}_c)) - \frac{1}{n}\sum_{i=1}^{n}(c_i + x_i - r_c\frac{s_{Y_c}}{s_{X_c}}(x_i - \overline{x}_c)) \\
&= \Delta + \overline{c}_t + \overline{x}_t - r_c\frac{s_{Y_c}}{s_{X_c}}(\overline{x}_t - \overline{x}_c) - \overline{c}_c - \overline{x}_c \\
&= \Delta + (\overline{c}_t - \overline{c}_c) + (1 - r_c\frac{s_{Y_c}}{s_{X_c}})(\overline{x}_t - \overline{x}_c) \qquad (9)
\end{aligned}$$

and similarly,

$$\hat{\tau}_{all} = \Delta + (\overline{c}_t - \overline{c}_c) + (1 - r\frac{s_Y}{s_X})(\overline{x}_t - \overline{x}_c) \qquad (10)$$

The two estimators look similar. The first term in both is Δ, the quantity we want to estimate. The second term is the difference in mean baseline responses between the treatment and control groups. The selection on observables assumption grants us that treatment assignment does not depend on the baseline c_i, and therefore this term equals 0 in expectation. The third term includes the difference in mean covariates between the treatment and control groups. If treatment assignment is random, this term also equals 0 in expectation. If treatment is correlated with X, then it may not be the case that $\mathbb{E}(\overline{x}_t) = \mathbb{E}(\overline{x}_c)$ and so the factor in front of this term matters.

In the first case, the multiplier is $1 - r_c\frac{s_{Y_c}}{s_{X_c}}$. Let Cov and Var denote the sample (as opposed to

population) covariance and variance. Using definitions, we have

$$r_c = \frac{\text{Cov}(X_c, Y_c)}{s_{X_c} s_{Y_c}} = \frac{\text{Cov}(X_c, X_c + C_c)}{s_{X_c} s_{Y_c}} = \frac{\text{Var}(X_c) + \text{Cov}(X_c, C_c)}{s_{X_c} s_{Y_c}} = \frac{s_{X_c}}{s_{Y_c}} + \frac{\text{Cov}(X_c, C_c)}{s_{X_c} s_{Y_c}}$$

However, in any particular control group, there may be some correlation between X_c and C_c, so the second term will be nonzero. However, assuming that $x_i \perp\!\!\!\perp c_i$ in the sample of N, the second term will equal 0 in expectation. We omit the algebra.

$$\mathbb{E}\left((1 - r_c \frac{s_{Y_c}}{s_{X_c}})(\overline{x}_t - \overline{x}_c)\right) = \mathbb{E}\left(-\frac{\text{Cov}(X_c, C_c)}{\text{Var}(X_c)}(\overline{x}_t - \overline{x}_c)\right) = 0$$

This shows that $\hat{\tau}_{ctrl}$ is an unbiased estimate of Δ. On the other hand,

$$\begin{aligned} r &= \frac{\text{Cov}(X, Y)}{s_X s_Y} = \frac{\text{Cov}(X, X + C + \Delta T)}{s_X s_Y} = \frac{\text{Var}(X) + \text{Cov}(X, C) + \Delta \text{Cov}(X, T)}{s_X s_Y} \\ &= \frac{s_X}{s_Y} + \frac{\text{Cov}(X, C)}{s_X s_Y} + \Delta \frac{\text{Cov}(X, T)}{s_X s_Y} \end{aligned}$$

As above, the second term will be small when X and C are unrelated. However, the third term will not vanish unless treatment is orthogonal to X. Note that now, these quantities aren't random: they're based on all N units. This implies

$$\mathbb{E}\left((1 - r \frac{s_Y}{s_X})(\overline{x}_t - \overline{x}_c)\right) = \left(\frac{\text{Cov}(X, C)}{\text{Var}X} + \Delta \frac{\text{Cov}(X, T)}{\text{Var}X}\right) \mathbb{E}(\overline{x}_t - \overline{x}_c) \neq 0$$

since the expected difference in mean covariates won't equal 0 if treatment is correlated with X. Thus, $\hat{\tau}_{all}$ is a biased estimator of Δ. This simple example illustrates why when doing estimation, we want to fit our predictive model $\hat{Y} = f(X_1, \ldots, X_p)$ using the controls only. If we use all of the observations, then we capture some of the effect of the predictors which are correlated with treatment assignment.

5.2.2 Hypothesis testing

We'll use the same test statistic for our tests, either $\hat{\tau}_{ctrl}$ or $\hat{\tau}_{all}$. In the math that follows, we omit the subscript for which model we used to predict $\hat{Y}$. Let $T^* = (T_1^*, \ldots, T_N^*)$ be a permutation of

the treatment assignments $T_1, \ldots, T_N$. Our test statistic can be written

$$\tau(T^*) = \frac{1}{n} \sum_{i:T_i^*=1} (Y_i - \hat{Y}_i) - \frac{1}{n} \sum_{i:T_i^*=0} (Y_i - \hat{Y}_i) \tag{11}$$

Suppose we fix $T_1, \ldots, T_N$ and obtain some permuted treatment vector T^*. Our test statistic simplifies:

$$\begin{aligned}
\tau(T^*) &= \frac{1}{n} \sum_{i=1}^{N} T_i^*(Y_i - \hat{Y}_i) - (1 - T_i^*)(Y_i - \hat{Y}_i) \\
&= \frac{1}{n} \sum_{i=1}^{N} T_i^*(Y_i - \hat{Y}_i) - (1 - T_i^*)(Y_i - \hat{Y}_i) \\
&= \frac{1}{n} \sum_{i=1}^{N} 2T_i^*(Y_i - \hat{Y}_i) - (Y_i - \hat{Y}_i) \\
&= \frac{1}{n} \sum_{i=1}^{N} 2T_i^*(T_i Y_i(1) + (1 - T_i)Y_i(0) - \hat{Y}_i) - (T_i Y_i(1) + (1 - T_i)Y_i(0) - \hat{Y}_i) \\
&= \frac{1}{n} \sum_{i=1}^{N} 2T_i^* T_i(Y_i(1) - Y_i(0)) + 2T_i^*(Y_i(0) - \hat{Y}_i) - T_i(Y_i(1) - Y_i(0)) - (Y_i(0) - \hat{Y}_i) \\
&= \frac{2\Delta}{n} \sum_{i=1}^{N} T_i^* T_i - \Delta + \frac{1}{n} \sum_{i=1}^{N} (2T_i^* - 1)(Y_i(0) - \hat{Y}_i)
\end{aligned}$$

The first term counts the number of units with $T_i = T_i^* = 1$; this is the overlap between the actual and permuted treatment vector. In expectation, the number of such units is $N/4$ (since treatment assignment is independent between the actual and permuted treatments, and each is distributed as Binomial with probability of assignment 1/2), so this term cancels the second term.

The third term varies depending on whether we fit our model to controls only or to all observations. It is the sum of the difference between each individual's potential outcome under control and their predicted outcome, multiplied by either -1 or 1 according to whether T_i^* is 0 or 1. When $\hat{Y}_i = \hat{Y}_{i,ctrl}$, the residuals $Y_i(0) - \hat{Y}_i$ may be systematically smaller in magnitude for individuals with $T_i = 0$ than for individuals with $T_i = 1$. Then the test statistic for the observed data will be

$$\begin{aligned}\tau_{ctrl}(T) &= \frac{2\Delta}{n}\sum_{i=1}^{N} T_i - \Delta + \frac{1}{n}\sum_{i:T_i=1}(Y_i(0) - \hat{Y}_{i,ctrl}) - \frac{1}{n}\sum_{i:T_i=0}(Y_i(0) - \hat{Y}_{i,ctrl}) \\ &= \Delta + \frac{1}{n}\sum_{i:T_i=1}(Y_i(0) - \hat{Y}_{i,ctrl})\end{aligned}$$

The third term drops out because the residuals of a linear regression sum to zero. Intuitively, it seems more likely that this sum will be extremely large in magnitude due to the way the model was fit, rather than any intrinsic treatment effect. The residuals from this model are not exchangeable under the null hypothesis. This will lead to a greater probability of incorrectly rejecting the null hypothesis.

On the other hand, using $\hat{Y}_{i,all}$ will give an observed test statistic of

$$\tau_{all}(T) = \Delta + \frac{1}{n}\sum_{i:T_i=1}(Y_i(0) - \hat{Y}_{i,all}) - \frac{1}{n}\sum_{i:T_i=0}(Y_i(0) - \hat{Y}_{i,all})$$

Under the null of no treatment effect whatsoever, assuming X and T are uncorrelated, the residuals here are exchangeable.[6] Thus, we'd expect the second and third term to roughly cancel each other out under any random treatment assignment, although for any particular treatment assignment, they will differ somewhat.

In summary, we must fit our predictive models differently according to whether we intend to estimate the average treatment effect or to test the strong null hypothesis of no effect whatsoever. The difference lies in the way we use the residuals. Residualizing in estimation helps reduce variation and increase the precision of estimators, but helps detect variation due to the treatment when we do testing. For estimation, we fit our predictive model to controls only to eliminate any leftover correlation between the treatment and other covariates. However, for testing, we want to capture this correlation, so we must fit to all observations.

[6] If X and T are correlated, this is false. They're not exchangeable – there might be some x_i with high leverage, making those residuals artificially small. This problem is exactly why we must assume conditional independence between treatment and potential outcomes within strata, or Assumption 1.

References

Alberto Abadie and Guido W. Imbens. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, 74(1):235–267, January 2006. ISSN 1468-0262. doi: 10.1111/j.1468-0262.2006.00655.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0262.2006.00655.x/abstract.

Alberto Abadie and Guido W. Imbens. On the Failure of the Bootstrap for Matching Estimators. *Econometrica*, 76(6):1537–1557, November 2008. ISSN 1468-0262. doi: 10.3982/ECTA6474. URL http://onlinelibrary.wiley.com/doi/10.3982/ECTA6474/abstract.

Alberto Abadie, Matthew M. Chingos, and Martin R. West. Endogenous Stratification in Randomized Experiments. Working Paper 19742, National Bureau of Economic Research, December 2013. URL http://www.nber.org/papers/w19742.

Peter C. Austin. A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, 33(6):1057–1069, March 2014. ISSN 1097-0258. doi: 10.1002/sim.6004. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.6004/abstract.

Peter C. Austin, Paul Grootendorst, and Geoffrey M. Anderson. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. *Statistics in Medicine*, 26(4):734–753, February 2007.

Alexis Diamond and Jasjeet S. Sekhon. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics*, 95(3):932–945, July 2013. ISSN 0034-6535. doi: 10.1162/REST_a_00318. URL http://dx.doi.org/10.1162/REST_a_00318.

Christiana Drake. Effects of misspecificiation of the propensity score on estimators of treatment effect. *Biometrics*, 49(4):1231–1236, 1993.

Ronald A. Fisher. *Design of Experiments*. New York: Hafner, 1935.

David A. Freedman and Richard A. Berk. Weighting regressions by propensity scores. *Evaluation Review*, 32(4):392–409, August 2008. ISSN 0193-841X. doi: 10.1177/0193841X08317586.

Jens Hainmueller. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, 20(1):25–46, 2012. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpr025. URL `http://pan.oxfordjournals.org/content/early/2011/10/15/pan.mpr025`.

Ben B. Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, June 2008. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asn004. URL `http://biomet.oxfordjournals.org/content/95/2/481`.

Jonathan Hennessy, Tirthankar Dasgupta, Luke Miratrix, Cassandra Pattanayak, and Pradipta Sarkar. A conditional randomization test to account for covariate imbalance in randomized experiments. *arXiv:1510.06817 [stat.ME]*, October 2015.

Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, 71(4):1161–1189, July 2003. ISSN 1468-0262. doi: 10.1111/1468-0262.00442. URL `http://onlinelibrary.wiley.com/doi/10.1111/1468-0262.00442/abstract`.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

J.K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960, October 2004.

J. R. Matchett, Philip B. Stark, Steven M. Ostoja, Roland A. Knapp, Heather C. McKenny, Matthew L. Brooks, William T. Langford, Lucas N. Joppa, and Eric L. Berlow. Detecting the influence of rare stressors on rare species in Yosemite National Park using a novel stratified permutation test. *Scientific Reports*, 5:10702, June 2015. ISSN 2045-2322. doi: 10.1038/srep10702. URL `http://www.nature.com/articles/srep10702`.

Luke W. Miratrix, Jasjeet S. Sekhon, and Bin Yu. Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(2):369–396, March 2013. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2012.01048.x. URL `http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2012.01048.x/abstract`.

Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

Paul R. Rosenbaum. Covariance Adjustment in Randomized Experiments and Observational Studies. *Statistical Science*, 17(3):286–327, August 2002. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1042727942. URL http://projecteuclid.org/euclid.ss/1042727942.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, April 1983. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/70.1.41. URL http://biomet.oxfordjournals.org/content/70/1/41.

Cyrus Samii and Peter M. Aronow. On equivalencies between design-based and regression-based variance estimators for randomized experiments. *Statistics & Probability Letters*, 82 (2):365–370, February 2012. ISSN 0167-7152. doi: 10.1016/j.spl.2011.10.024. URL http://www.sciencedirect.com/science/article/pii/S0167715211003452.

Jerzy Splawa-Neyman, D M Dabrowska, and T P Speed. On the application of probability theory to agricultural experiments. essay on principles. *Statistical Science*, 5(4):465–472, 1990.

Stefan Wager and Susan Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *arXiv:1510.04342 [math, stat]*, October 2015. URL http://arxiv.org/abs/1510.04342. arXiv: 1510.04342.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, December 1945.