Model-based matching

Kellie Ottoboni

Draft April 11, 2016

Abstract

Drawing causal inferences from nonexperimental data is difficult due to the presence of confounders, variables that affect both the selection into treatment groups and the outcome. Post-hoc matching and stratification can be used to group individuals who are comparable with respect to important variables, but commonly used methods often fail to balance confounders between groups. We introduce model-based matching, a nonparametric method which groups observations that would be alike aside from the treatment. We use model-based matching to conduct stratified permutation tests of association between the treatment and outcome, controlling for other variables. Under standard assumptions from the causal inference literature, model-based matching can be used to estimate average treatment effects.

1 Introduction

Observational studies in a range of fields including social sciences, epidemiology, and ecology are used to make inferences about cause and effect. Causal inference can be viewed as a missing data problem: one is only able to see each individual's outcome after treatment or no treatment, but not both (Holland, 1984). To estimate the effect of treatment, one must use a control group as the counterfactual. The treatment effect is obscured by confounders, variables that are entangled with both the treatment and outcome. In an ideal situation, to adjust for the effect of confounders one would estimate the difference in outcomes between cases and controls who are identical with respect to all confounders, then average over the pairs.

In practice, the curse of dimensionality makes this impossible: studies typically account for 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. For example, it is difficult to study the effect of a job-training program on income levels when the participants differ from non-participants on a wide range of socioeconomic variables, as well as potential unmeasured confounders (Dehejia & Wahba, 1999). 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 weighting by propensity score (Rosenbaum & Rubin, 1983), genetic matching (Diamond & Sekhon, 2013), and maximum-entropy weighting (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. When matching and weighting methods fail to achieve balance in all the pretreatment covariates, estimates of the average treatment effect can be severely biased (Freedman & Berk, 2008).

The proposed method improves upon existing methods by eliminating the need for parametric assumptions such as Gaussian and homoscedastic errors, linearity, and adequate support. Observational studies often violate these key assumptions, so model-based matching may better accommodate real-world data from observational studies than traditional methods. Additionally, it is more flexible than traditional methods that assume that the treatment is binary and outcome is continuous; by modifying the statistic used to measure the strength of association, one may use categorical or continuous treatment and outcome variables.

The method has been applied before in a study of the effect of packstock use on the amphibian population in Yosemite National Park (Matchett, Stark, et. al 2015). In this paper, we develop the theory behind the testing method and discuss estimation strategies.

1.1 Notation

Let $Y_i(0)$ and $Y_i(1)$ be individual i's potential outcomes under control and to treatment, respectively. T_i is the treatment assigned to individual i. Then the observed outcome is $Y_i = Y_i(1)T_i + Y_i(0)(1-T_i)$. A standard assumption is exogeneity, or conditional independence: $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$.

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

TO DO: Do a lit review of matching here

In randomized control trials, the potential outcomes are balanced between treatment and control groups by construction: in other words, $(Y(0), Y(1)) \perp\!\!\!\perp T$. In observational studies, this is no longer guaranteed. Rosenbaum and Rubin (1983) achieve a weaker form of this balance by appealing to the propensity score, $p(x) = \mathbb{P}(T = 1 \mid X = x)$. The propensity score is a balancing score, in the sense that $X \perp\!\!\!\perp T \mid p(X)$. Under the assumptions of conditional independence given X and overlap in the distribution of propensity scores (together, called "strong ignorability"), they show that strong ignorability given X implies strong ignorability given $p(X)$, and that by the law of iterated expectations, we can recover the overall average treatment effect.

One issue is that in observational studies, the propensity score is unknown. One typically estimates the propensity score using logistic or probit regression models using a set of observed covariates. When the propensity score estimates are wrong, then estimates of the average treatment effect can be biased.

We'd like to try to achieve balance in the potential outcomes some other way: $(Y(0), Y(1)) \perp\!\!\!\perp T \mid \hat{Y}$, where $\hat{Y}$ is the model-based prediction of Y, absent knowledge of the treatment, for all units.

2.2 Stratified Estimation

Recall that

$$ATE = \mathbb{E}(Y_1 - Y_0)$$

Let $\hat{Y} = f(X)$ be a prediction of the response under the control regime. TO DO: MENTION FITTING F PROBLEM Hansen [2008] defines a prognostic score as a function $\Psi(X)$ such that $Y(0) \perp\!\!\!\perp X \mid \Psi(X)$ for all X. That is, 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 $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$. 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.

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. Instead, we propose stratifying based on $\hat{Y}$. Let $\Psi_s(X)$ be a function that classifies individuals into one of S strata based on their predicted control potential outcome. That is, $\Psi_s(X) = \Pi(\hat{Y}) = \Pi(f(X))$. The stratification rule Ψ_s and estimation procedure for $\hat{Y}$ are defined independently of the observed sample. 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

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

$\hat{\tau}$ is the weighted average of within-stratum estimated treatment effects, where weights are proportional to the stratum sizes. TO DO: CLEAN UP THIS PARAGRAPH For this estimator to "work", we need the stratum-specific estimators to be unbiased. A sufficient condition is conditional independence between potential outcomes and treatment given stratum assignment: $(Y(0), Y(1)) \perp\!\!\!\perp T \mid S$. The stratification collapses redundant information in $\hat{Y}$ but must preserve the relevant information about how covariates relate to treatment assignment. 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. Intuitively, treatment is "as if random" within strata. In a randomized experiment, the analogous condition is that the propensity score is constant within strata, i.e. treatment *is* assigned at random within strata.

We prove a useful lemma before our main result.

Lemma 2.2. *If $(Y(0), Y(1)) \perp\!\!\!\perp T \mid S$ and treatment is assigned independently across units within stratum s, then*

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

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

Proof.

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

□

If $\hat{\tau}$ is unbiased for the ATE:

Theorem 2.3. *If $Y(1), Y(0) \perp\!\!\!\perp T \mid S$ and $0 < N_{st} < N_s$ for $s = 1, \dots, S$, then $\hat{\tau}$ is unbiased for the ATE.*

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

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

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

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 $\sigma_1^2 = \mathrm{Var}_{SP}(Y(1))$, $\sigma_0^2 = \mathrm{Var}_{SP}(Y(0))$, and $\sigma_{01}^2 = \mathrm{Var}_{SP}(Y(1) - Y(0)) = \mathrm{Cov}_{SP}(Y(1), Y(0))$. In addition, define $\nu^2 = \mathrm{Var}_{SP}(\hat{Y})$, $\rho_1 = \frac{\mathrm{Cov}_{SP}(Y(1),\hat{Y})}{\sigma_1 \nu}$, and $\rho_0 = \frac{\mathrm{Cov}_{SP}(Y(0),\hat{Y})}{\sigma_0 \nu}$.

Theorem 2.4.

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

The proof appears in the appendix.

Compare this to $\hat{\tau}^{diff}$, the usual difference in means estimator. In our simulations, we observe that $\text{Var}(\hat{\tau}^{adj})$ is an order of magnitude smaller than $\text{Var}\left(\hat{\tau}^{diff}\right)$. Using the same proof as above but substituting Y for r, 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 $\hat{\tau}$ from Equation 1 depends on the number of strata and how they partition the covariate space. While the within-stratum estimated treatment effects may have smaller variances, their covariance will generally be positive and thus inflate the overall variance. In the next section, we find empirically that TO DO:

2.3 Empirical Results

3 Hypothesis Testing

3.1 Tests of Residuals

Tests of residuals after covariance adjustment appear in various forms in the literature. Rosenbaum (2002) uses residuals after fitting prediction models to stabilize estimates of treatment effects for more powerful randomization tests. Rosenbaum's framework is limited to the case of binary treatment, where individuals are either assigned to treatment or receive the control.

Shah and Bhlmann (2015) use residuals to test for the goodness of fit of high-dimensional linear models by testing for nonlinear signals in the residuals.

3.2 Hypothesis testing in randomized experiments

Suppose we observe N individuals with responses $Y_1, \dots, Y_N$. Assume that their response takes the form

$$Y_i(t) = f(t, X_i) + \varepsilon_i \tag{4}$$

where X_i is a vector of covariates for individual i, t is the treatment received, and ε_i is random error. Assume that the ε_i are independent and identically distributed, with $\mathbb{E}(\varepsilon_i) = 0$. In a randomized experiment, the N individuals are assigned treatment at random. T_i is an indicator for whether individual i received treatment, with $T_i \perp\!\!\!\perp (X_i, \varepsilon_i)$.

Suppose we'd like to test the strong null hypothesis of no treatment effect, which implies that $Y_i(0) = Y_i(1)$. We'd test the null by computing some test statistic of the responses $Y = Y(T) = (Y_1(T_1), \dots, Y_N(T_N))$, and treatment assignment, $T = (T_1, \dots, T_N)$, as $\tau(Y, T)$. Note that under the strong null, Y does not actually depend on T, and so for any permutation T^* of treatment assignments, $Y(T) = Y(T^*)$. Suppose we take K strata of observations. We would generate a null distribution for τ by permuting the treatment assignments within the K strata to obtain an approximate distribution $\tau(Y, T_1^*), \dots, \tau(Y, T_B^*)$ for some large B and compare our observed test statistic to the distribution to obtain a p-value. Extreme values of $\tau(Y, T)$ 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_i. In particular, under the strong null, $f(0,x) = f(1,x)$ for all x. Since the errors ε_i are IID, the values $Y_i(t) - f(t, X_i)$ are also IID (and thus exchangeable). However, f is unobserved so we must estimate it. Since f is constant with respect to treatment, we may estimate it as a function of x alone. We'd like to use the residuals $e_i = Y_i(t) - \hat{f}(X_i)$ in place of the unobserved errors ε_i. Under random assignment and IID assumptions, e_i are also

3.3 Hypothesis testing in observational studies

To ensure that the residuals are exchangeable, we need them to be some exchangeable function of ε_i within each stratum, e.g. if observation i belongs to the kth stratum, then $e_i = g_k(\varepsilon_i)$. What are the conditions on the g_k that ensure e_i are exchangeable? It is sufficient to have g_k be a strictly monotonic function, such as $g_k(\varepsilon) = \alpha_k + \beta_k \varepsilon$.

If exchangeability does not hold, we want to find a transformation ϕ such that $\phi(e_i)$ are exchangeable.

3.4 Empirical Results

4 Discussion

5 Appendix

Proof of Theorem 2.4. 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 (cite the imbens and rubin text). 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

$$\mathrm{Var}(\hat{\tau}^{adj}) = \frac{\mathrm{Var}(r(1))}{N_t} + \frac{\mathrm{Var}(r(0))}{N_c} \tag{5}$$

In terms of known quantities:

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

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

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

Similarly,

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

where ρ_0 is the correlation between $Y(0)$ and $\hat{Y}$. Plugging these expressions into equation 5 gives the desired result. □

References

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`.