Model-based matching

Kellie Ottoboni

Draft February 18, 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. Matching methods can be used to subset the data to groups which are comparable with respect to important variables, but matching often fails to create sufficient balance between groups. Model-based matching is a nonparametric method for matching which groups observations that would be alike if none had received 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 Matching

The great thing about randomized control trials is that 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.

3 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.

4 Theory

4.1 MM for Testing

4.2 MM for Estimation

4.2.1 MM Estimate when T is Binary

Recall that

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

Let $\hat{Y} = f(X_1, \ldots, X_p)$ be a prediction of the response, using the control group the training data. The prediction uses no information on the treatment – it gives our best guess at the response one would have under the control condition. Suppose that we stratify the observations according to their values of $\hat{Y}$ to obtain S strata. 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 using $\hat{\tau}$:

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

If treatment assignment is at random within strata, then $\hat{\tau}$ is unbiased for the ATE:

$$
\begin{aligned}
\mathbb{E}(\hat{\tau}) &= \mathbb{E}\left[\sum_{s=1}^{S} \frac{N_S}{N}\left(\frac{1}{n_{st}} \sum_{i:T_i=1,S_i=s} (Y_i - \hat{Y}_i) - \frac{1}{n_{sc}} \sum_{i:T_i=0,S_i=s} (Y_i - \hat{Y}_i)\right)\right] \\
&= \mathbb{E}\left[\sum_{s=1}^{S} \frac{N_s}{N}\left(\frac{1}{N_s} \sum_{i:S_i=s} \frac{T_i(Y_i - \hat{Y}_i)}{n_{st}/N_s} - \frac{(1-T_i)(Y_i - \hat{Y}_i)}{n_{sc}/N_s}\right)\right] \\
&= \sum_{s=1}^{S} \frac{1}{N}\left(\sum_{i:S_i=s} \frac{\mathbb{E}(T_i)(Y_i(1) - \hat{Y}_i)}{n_{st}/N_s} - \frac{\mathbb{E}(1-T_i)(Y_i(0) - \hat{Y}_i)}{n_{sc}/N_s}\right) \\
&= \sum_{s=1}^{S} \frac{1}{N}\left(\sum_{i:S_i=s} \frac{(n_{st}/N_s)(Y_i(1) - \hat{Y}_i)}{n_{st}/N_s} - \frac{(n_{sc}/N_s)(Y_i(0) - \hat{Y}_i)}{n_{sc}/N_s}\right) \\
&\hspace{12em} \text{assuming } n_{st}, n_{sc} \text{ are fixed within strata} \\
&= \sum_{s=1}^{S} \frac{1}{N}\left(\sum_{i:S_i=s} (Y_i(1) - \hat{Y}_i) - (Y_i(0) - \hat{Y}_i)\right) \\
&= \frac{1}{N} \sum_{i=1}^{N} Y_i(1) - Y_i(0) \\
&= ATE
\end{aligned}
$$

What happens if we don't have random treatment assignment but selection on observables holds? We need some sort of way to show that $\mathbb{E}(T_i \mid X) = n_{st}/N_s$ anyways. Is it the case that $\mathbb{E}(T_i \mid X) = \mathbb{E}(T_i \mid X_i) = \mathbb{E}(T_i \mid \hat{Y}_i)$? In that case, given $\hat{Y}_i$, we know which stratum i belongs to

4.2.2 When T is discrete

We can generalize the previous result when there are more than two treatments.

4.3 Fitting the model

What's the difference between fitting our model for matching using all units and fitting the model using controls only? 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.

4.3.1 Simple linear regression

Suppose we have a population of N individuals. Each individual has two potential outcomes, $Y_i(0)$ and $Y_i(1)$, their responses to the control and the treatment conditions, respectively. Suppose without loss of generality that units $i = 1, \dots, n$ receive the control condition and $i = n+1, \dots, N$ receive treatment, where $N = 2n$. Suppose that there is a constant, additive treatment effect. That is, for each individual, $Y_i(1) - Y_i(0) = \Delta$ for some $\Delta \in \mathbb{R}$. Let $\overline{c} = \frac{1}{N}\sum_{i=1}^{N} Y_i(0)$, $\overline{c}_c = \frac{1}{n}\sum_{i=1}^{n} Y_i(0)$, 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 we have no information on covariates. Then our best guess using controls only is $\hat{Y}_{ctrl} = \overline{c}_c$ and our best guess using all units is $\hat{Y}_{all} = \overline{Y} = \overline{c} + \frac{\Delta}{2}$. Since $\hat{Y}_{ctrl}$ and $\hat{Y}_{all}$ are constant for all units, then $\hat{\tau}_{ctrl} = \hat{\tau}_{all}$ and so using either prediction will give identical estimates and tests.

Suppose now that we have information on one covariate. Suppose that $Y_i(0) = c_i + X_i$, where c_i is some baseline value and X_i is the value of the covariate, and $X_i \perp\!\!\!\perp c_i$ for all i. Then $Y_i(1) = c_i + X_i + \Delta$. 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 overall sample. Using simple linear regression to predict the response without using treatment information, we have

$$\begin{aligned}
\hat{Y}_{i,ctrl} &= \overline{c}_c + \overline{X}_c + r_c \frac{s_{Y_c}}{s_{X_c}}(X_i - \overline{X}_c) \\
\hat{Y}_{i,all} &= \overline{c} + \overline{X} + r \frac{s_Y}{s_X}(X_i - \overline{X})
\end{aligned}$$

Thus, the estimators are

$$
\begin{aligned}
\hat{\tau}_{ctrl} &= \frac{1}{n}\sum_{i=n+1}^{N}(Y_i(1) - \hat{Y}_{i,ctrl}) - \frac{1}{n}\sum_{i=1}^{n}(Y_i(0) - \hat{Y}_{i,ctrl}) \\
&= \frac{1}{n}\sum_{i=n+1}^{N}(c_i + X_i + \Delta - \overline{c}_c - \overline{X}_c - r_c\frac{s_{Y_c}}{s_{X_c}}(X_i - \overline{X}_c)) - \frac{1}{n}\sum_{i=1}^{n}(c_i + X_i - \overline{c}_c - \overline{X}_c - r_c\frac{s_{Y_c}}{s_{X_c}}(X_i - \overline{X}_c)) \\
&= \Delta + \frac{1}{n}\sum_{i=n+1}^{N}(c_i + X_i - r_c\frac{s_{Y_c}}{s_{X_c}}(X_i - \overline{X}_c)) - \frac{1}{n}\sum_{i=1}^{n}(c_i + X_i - r_c\frac{s_{Y_c}}{s_{X_c}}(X_i - \overline{X}_c)) \\
&= \Delta + \overline{c}_t + \overline{X}_t - r_c\frac{s_{Y_c}}{s_{X_c}}(\overline{X}_t - \overline{X}_c) - \overline{c}_c - \overline{X}_c \\
&= \Delta + (\overline{c}_t - \overline{c}_c) + (1 - r_c\frac{s_{Y_c}}{s_{X_c}})(\overline{X}_t - \overline{X}_c) \qquad (1)
\end{aligned}
$$

and similarly,

$$\hat{\tau}_{all} = \Delta + (\overline{c}_t - \overline{c}_c) + (1 - r\frac{s_Y}{s_X})(\overline{X}_t - \overline{X}_c) \qquad (2)$$

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 $c_i \perp\!\!\!\perp T_i$, 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

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

Assuming that $X_i \perp\!\!\!\perp c_i$, the second term will equal 0 in expectation. However, in any particular sample, there may be some correlation between X and c, so the second term will be nonzero. This implies that

This proof is incomplete – how do we get that product of 0 mean things is mean 0?

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

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

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

As above, the second term will drop out in expectation. However, the third will not unless $X \perp\!\!\!\perp T$. This implies

$$\mathbb{E}\left(1 - r\frac{s_Y}{s_X}\right) = \Delta\frac{\text{Cov}(X,T)}{\text{Var}X}$$

but what we're really worried about is $\mathbb{E}((1 - r\frac{s_Y}{s_X})(\overline{X}_t - \overline{X}_c))$

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.

4.3.2 Multivariate regression

Define X_c and Y_c to be the design matrix and responses for controls. Let X_t and Y_t be defined analogously for the treatment group. Then, let X and Y be the design matrix and responses for all observations:

$$X = \left[\begin{array}{c} X_c \\ \hline X_t \end{array}\right], Y = \left[\begin{array}{c} Y_c \\ \hline Y_t \end{array}\right]$$

If we fit our model using only the controls, then the predicted values will be

$$\hat{Y}_{mod1} = X(X_c'X_c)^{-1}X_c'Y \tag{3}$$

If we fit our model using all observations, then the predicted values will be

$$\hat{Y}_{mod2} = X(X'X)^{-1}X'Y \tag{4}$$

Let's expand this out in terms of X_c, X_t, Y_c, and Y_t. First, it's clear that

$$X'Y = X_c'Y_c + X_t'Y_t \tag{5}$$

Now, let's consider the inverse covariance matrix. For notational simplicity, define $C = X_c'X_c$ and $T = X_t'X_t$. Then

$$\begin{aligned}(X'X)^{-1} &= (C+T)^{-1} \\ &= C^{-1} - (I + C^{-1}T)^{-1}C^{-1}TC^{-1} && \text{using a result from } \textbf{?}\end{aligned}$$

Let $N = (I + C^{-1}T)^{-1}$. The predicted responses from the second model are

$$\begin{aligned}X(X'X)^{-1}X'Y &= X\left(C^{-1} - NC^{-1}TC^{-1}\right)(X_c'Y_c + X_t'Y_t) \quad (6)\\ &= X\left(C^{-1}X_c'Y_c + C^{-1}X_t'Y_t - NC^{-1}TC^{-1}(X_c'Y_c + X_t'Y_t)\right) \quad (7)\end{aligned}$$

The difference in predictions from the two models is

$$\hat{Y}_{mod2} - \hat{Y}_{mod1} = XC^{-1}X_t'Y_t - XNC^{-1}TC^{-1}(X_c'Y_c + X_t'Y_t) \tag{8}$$

Our choice of models matters for hypothesis testing. In particular, when we match using the model fit only to controls, $mod1$, we get a permutation test whose level is higher than the nominal level. This comes from additional correlation between the residuals and the treatment.

$$\begin{aligned}\mathrm{Cov}(Y - \hat{Y}_{mod1}, W) &= \mathrm{Cov}(Y, W) - XC^{-1}X_c'\mathrm{Cov}(Y_c, W) \quad (9)\\ \mathrm{Cov}(Y - \hat{Y}_{mod2}, W) &= \mathrm{Cov}(Y, W) - XC^{-1}X_c'\mathrm{Cov}(Y_c, W) - XC^{-1}X_t'\mathrm{Cov}(Y_t, W) \quad (10)\\ &\quad + XNC^{-1}TC^{-1}X'\mathrm{Cov}(Y, W) \quad (11)\end{aligned}$$

Thus the difference in covariances is

$$\text{Cov}(Y-\hat{Y}_{mod1},W)-\text{Cov}(Y-\hat{Y}_{mod2},W) = XC^{-1}X_t'\text{Cov}(Y_t,W)-XNC^{-1}TC^{-1}X'\text{Cov}(Y,W) \quad (12)$$

Now let's compute the two covariance terms.

$$\begin{aligned}\text{Cov}(Y_t,W) &= \text{Cov}(\beta_0+\beta_1X_1+\beta_2X_2+\gamma+\varepsilon,W)\\ &= \beta_1\text{Cov}(X_1,W)+\beta_2\text{Cov}(X_2,W)\end{aligned}$$

$$\begin{aligned}\text{Cov}(Y,W) &= \text{Cov}(WY_t+(1-W)Y_c,W)\\ &= \text{Cov}(WY_t,W)-\text{Cov}(WY_c,W)+\text{Cov}(Y_c,W)\\ &= \text{Cov}(W(\beta_0+\beta_1X_1+\beta_2X_2+\gamma+\varepsilon),W)-\text{Cov}(W(\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon),W)+\text{Cov}(Y_c,W)\\ &= \gamma\text{Var}(W)+\beta_1\text{Cov}(X_1,W)+\beta_2\text{Cov}(X_2,W)\end{aligned}$$

Thus we can rewrite (12) as

$$\begin{aligned}\text{Cov}(Y-\hat{Y}_{mod1},W)-\text{Cov}(Y-\hat{Y}_{mod2},W) = {}&(\beta_1\text{Cov}(X_1,W)+\beta_2\text{Cov}(X_2,W))\times\\ &\big(XC^{-1}X_t'-XNC^{-1}TC^{-1}X'\big)\\ &-\gamma\text{Var}(W)XNC^{-1}TC^{-1}X' \qquad (13)\end{aligned}$$

If treatment assignment is independent of the covariates, then the first term will be 0.

5 Empirical results