Model-based matching

Kellie Ottoboni

Draft February 29, 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 Matching

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.

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

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:

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

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

5 Empirical results