Fitting to controls

Kellie Ottoboni

Draft February 29, 2016

1 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 with a simple linear regression prediction to illustrate the bias introduced by fitting the model different ways. 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.

1.1 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 $i = 1, \ldots, n$ receive the control condition and $i = n+1, \ldots, N$ 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 $Y_i(0) = c_i + x_i$ 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 overall sample. Using simple linear regression to predict the response without using treatment information, we have

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

Thus, the estimators are

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

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 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

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

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? Cauchy-Schwarz gives $|\mathbb{E}(XY)| \leq \sqrt{\text{Var}(X)\text{Var}(Y)}$ for $\mathbb{E}(X) = \mathbb{E}(Y) = 0$

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

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 be small when X and C are unrelated. However, the third term will not vanish unless treatment is assigned using no information on X. 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.

Let's turn our attention to 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

$$\tau(T^*) = \frac{1}{n} \sum_{i:T_i^*=1} (Y_i - \hat{Y}_i) - \frac{1}{n} \sum_{i:T_i^*=0} (Y_i - \hat{Y}_i) \tag{3}$$

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

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

since 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

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

Under the null of no treatment effect whatsoever, the residuals here are exchangeable This is false. They're not exchangeable – there might be some x_i with high leverage. But is it sufficient that the residuals sum to 0?. 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.

1.2 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:

$$X = \left[\frac{X_c}{X_t}\right], Y = \left[\frac{Y_c}{Y_t}\right]$$

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

$$\hat{Y}_{mod1} = X(X_c'X_c)^{-1}X_c'Y \tag{4}$$

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

$$\hat{Y}_{mod2} = X(X'X)^{-1}X'Y \tag{5}$$

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

$$X'Y = X_c'Y_c + X_t'Y_t \tag{6}$$

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

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

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

$$X(X'X)^{-1}X'Y = X\left(C^{-1} - NC^{-1}TC^{-1}\right)(X_c'Y_c + X_t'Y_t) \tag{7}$$

$$= X\left(C^{-1}X_c'Y_c + C^{-1}X_t'Y_t - NC^{-1}TC^{-1}(X_c'Y_c + X_t'Y_t)\right) \tag{8}$$

The difference in predictions from the two models is

$$\hat{Y}_{mod2} - \hat{Y}_{mod1} = XC^{-1}X_t'Y_t - XNC^{-1}TC^{-1}(X_c'Y_c + X_t'Y_t) \tag{9}$$

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.

$$\text{Cov}(Y - \hat{Y}_{mod1}, W) = \text{Cov}(Y, W) - XC^{-1}X_c'\text{Cov}(Y_c, W) \tag{10}$$

$$\text{Cov}(Y - \hat{Y}_{mod2}, W) = \text{Cov}(Y, W) - XC^{-1}X_c'\text{Cov}(Y_c, W) - XC^{-1}X_t'\text{Cov}(Y_t, W) \tag{11}$$

$$+ XNC^{-1}TC^{-1}X'\text{Cov}(Y, W) \tag{12}$$

Thus the difference in covariances is

$$\text{Cov}(Y-\hat{Y}_{mod1},W)-\text{Cov}(Y-\hat{Y}_{mod2},W)=XC^{-1}X_t'\text{Cov}(Y_t,W)-XNC^{-1}TC^{-1}X'\text{Cov}(Y,W) \quad (13)$$

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 (13) as

$$\begin{aligned}\text{Cov}(Y-\hat{Y}_{mod1},W)-\text{Cov}(Y-\hat{Y}_{mod2},W)=&(\beta_1\text{Cov}(X_1,W)+\beta_2\text{Cov}(X_2,W))\times\\&\left(XC^{-1}X_t'-XNC^{-1}TC^{-1}X'\right)\\&-\gamma\text{Var}(W)XNC^{-1}TC^{-1}X' \quad (14)\end{aligned}$$

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