Matching via Mixed Integer Programming

Thomas Wiemann
University of Chicago

TA Discussion # 7
Econ 31740

February 28, 2022

Outline

- Brief Review of Selection on Observables
- Matching Estimators
- Discussion of Zubizarreta (2012)
  - Mixed Integer Programming for Optimal Matching
  - Balancing Penalties and Constraints
  - Empirical Illustration on Post-Surgery Acute Kidney Failure
- Last Comments

The brief review of selection on observables partially relies on Prof. Alex Torgovitsky's notes from his Applied Microeconometrics course (Econ 31720). Of course, all errors in these slides are my own.

Review of Selection on Observables

Consider a binary treatment $\operatorname{supp} D = \{0, 1\}$ with corresponding potential outcomes $Y(0)$ and $Y(1)$ such that

$$Y = DY(1) + (1 - D)Y(0). \tag{1}$$

Additionally, there are other observable variables X. Here, $\operatorname{supp} Y \subset \mathbb{R}$ and $\operatorname{supp} X \subset \mathbb{R}^{d_x}$.

We focus on the setting of Zubizarreta (2012):

- $Y \equiv$ an indicator for acute kidney failure after surgery;
- $D \equiv$ an indicator for being obese (as measured by the BMI);
- $X \equiv$ a vector of patient-hospital characteristics (e.g., gender).

Suppose we observe a sample $(y_i, d_i, x_i) \overset{iid}{\sim} (Y, D, X)$ for $i = \ldots, n$.

Review of Selection on Observables (Contd.)

We consider the selection on observables assumption (SO) that states

$$(Y(0), Y(1)) \perp\!\!\!\perp D|X. \tag{2}$$

Effectively, the assumption states that treatment is "as good as random" given X. (Hence, the random assignment assumption $(Y(0), Y(1)) \perp\!\!\!\perp D$ could be understood as a special case of selection on observables where X is degenerate.)

Given SO, we have

$$\begin{aligned} F_{Y(d)}(y|x) &= P\left(Y(d) \leq y|X = x\right) \\ &= P\left(Y(d) \leq y|D = d, X = x\right) \\ &= P\left(Y \leq y|D = d, X = x\right), \end{aligned} \tag{3}$$

where the second equality follows from SO *and* an overlap condition (OC) that requires $P(D = d|X = x) > 0$ (otherwise the conditional probability is not well-defined). The final term in (3) is identified in the sense of Hurwicz (1950).

Review of Selection on Observables (Contd.)

A key parameter of interest in Zubizarreta (2012) is the average causal effect of obesity on the risk of acute kidney failure after surgery.

For this average treatment effect (ATE), we have

$$\begin{aligned} E[Y(1) - Y(0)] &= E_X[E[Y(1) - Y(0)|X]] \\ &= E_X[E[Y(1)|D = 1, X] - E[Y(0)|D = 0, X]] \\ &= E_X[E[Y|D = 1, X] - E[Y|D = 0, X]] \\ &= E_X[\mu_1(X) - \mu_0(X)], \end{aligned} \tag{4}$$

where $\mu_d(x) := E[Y|D = d, X = x], \forall d \in \{0, 1\}$, and we have again used SO and OC in the second equation.

Equation (4) suggests a imputation estimator for the ATE:

$$\widehat{\text{ATE}} = \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}_1(x_i) - \hat{\mu}(x_i)), \tag{5}$$

where $\hat{\mu}_d$ is an estimator of the conditional expectation function μ_d.

Matching Estimators

The researcher needs to choose estimators for $\hat{\mu}_1$ and $\hat{\mu}_0$. In practice, the most common appear to be linear regression and matching. (Machine learning-based non-parametric estimators may be promising alternatives. See, e.g., Chernozhukov et al., 2018.)

Let $\mathcal{T}$ and $\mathcal{C}$ denote individuals in the treated and control sample, respectively. Matching proceeds as follows:

1. Define a distance δ_{ij} between values x_i and x_j, $\forall i \in \mathcal{T}, j \in \mathcal{C}$.
2. For each i, find Kth smallest element of $\{\delta_{ij} : d_i \neq d_j\}_{j=1}^{n}$.
3. Let $\mathcal{J}_i$ denote the set of K individuals with the smallest distances in the other group.

The matching estimator for $\hat{\mu}_d$ is then given by

$$\hat{\mu}_d(x_i) = \mathbb{1}\{d_i = d\} Y_i + \mathbb{1}\{d_i \neq d\} \frac{1}{|\mathcal{J}_i|} \sum_{j \in \mathcal{J}_i} Y_j. \tag{6}$$

Matching Estimators (Contd.)

There are many practical subtleties to consider when using matching estimators. For example:

- Number of matches K;
- Matching with or without replacement;
- Definition of the distance matrix between covariates;
- Ensuring covariate overlap.

A particularly popular matching estimator is propensity score matching. Here, a reduced form of the treatment indicator on covariates is estimated in a first step. Distances between propensity scores are then used for matching.

Matching Estimators (Contd.)

Notice that the matching procedure outlined above searches for the best set of matches for each individual, but that there is no threshold on how "good" this match should be (i.e., the overlap condition is not enforced).

A potential side effect of this procedure is that the empirical distribution of the covariates in the matched sample of treated individuals is substantially different from that of matched untreated individuals. This may be worrying as we are then integrating over two different distributions of covariates in Equation (6).

Zubizarreta (2012) proposes to use a mixed integer programming approach for more transparent tradeoffs between minimizing distances in covariates and ensuring that the empirical distributions of the covariates are similar.

Matching via Mixed Integer Programming

Zubizarreta (2012) develops mixed integer program for matching w/o replacement that minimizes sum of distances of matches *and* covariate imbalance.

- User-specified distances δ_{ij} (e.g., Mahalanobis).
- Covariate imbalance as a flexible function of empirical distribution.
- Approach allows for enforcing degree of covariate overlap.

Method promises to reduce "guesswork involved in current optimal matching methods" by systematically controlling covariate imbalance.

Throughout, we consider a setting where the number of units in the control sample $\mathcal{C}$ is greater than in the treated sample $\mathcal{T}$.

Matching via Mixed Integer Programming (Contd.)

We begin with a most simple matching program:

$$
\begin{aligned}
\min_{\mathbf{a}} \; & \sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{C}} \delta_{ij} a_{ij} \\
\text{s.t.} \; & \sum_{i \in \mathcal{T}} a_{ij} \leq 1, \; j \in \mathcal{C} \\
& \sum_{j \in \mathcal{C}} a_{ij} = K, \; i \in \mathcal{T} \\
& a_{ij} \in \{0, 1\}, \forall i \in \mathcal{T}, j \in \mathcal{C}
\end{aligned}
\tag{7}
$$

Here, a_{ij} is 1 if treated unit i is assigned to control unit j, and 0 otherwise.

The first set of constraints enforces that each control unit is assigned to at most one treatment unit. The second ensures that each treated unit is matched with K units in the control group.

Matching via Mixed Integer Programming (Contd.)

Zubizarreta (2012) extends the binary program in (7):

$$
\begin{aligned}
\min_{\mathbf{a}} \; & \sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{C}} \delta_{ij} a_{ij} + \sum_{p \in \mathcal{P}} \omega_p \gamma_p(\mathbf{a}) \\
\text{s.t.} \; & \sum_{i \in \mathcal{T}} a_{ij} \leq 1, \; j \in \mathcal{C} \\
& \sum_{j \in \mathcal{C}} a_{ij} = K, \; i \in \mathcal{T} \\
& v_m(\mathbf{a}) \leq \epsilon_m, \; \forall m \in \mathcal{M} \\
& a_{ij} \in \{0, 1\}, \forall i \in \mathcal{T}, j \in \mathcal{C}
\end{aligned}
\tag{8}
$$

- $\omega_p \gamma_p(\mathbf{a})$ defines an imbalance-penalty for the pth covariate in a subset of covariates $\mathcal{P}$;
- $v_m(\mathbf{a}) \leq \epsilon_m$ enforces a maximum level of imbalance on the mth covariate in a subset of covariates $\mathcal{M}$.

Matching via Mixed Integer Programming (Contd.)

There are many options for the imbalance measures $\gamma_p(\mathbf{a})$ and $v_m(\mathbf{a})$. Zubizarreta (2012) suggests five broad categories:

- Balancing univariate moments;
- Balancing multivariate moments;
- Balancing empirical distributions;
- Enforcing exact or near-exact matching;
- Enforcing a specific subset of control units.

These approaches will typically result in introduction of continuous auxiliary decision variables (e.g., to bound absolute values). This then results in a mixed integer program (rather than a binary program).

Balancing Univariate Moments

Balancing covariate means in the treated and control sample can be achieved by specifying

$$\sum_{p\in\mathcal{P}} \omega_p \gamma_p(\mathbf{a}) = \sum_{p\in\mathcal{P}} \omega_p \left| \sum_{i\in\mathcal{T}} \sum_{j\in\mathcal{C}} \frac{x_{j,p} a_{ij}}{m|\mathcal{T}|} - \overline{x}_{\mathcal{T},p} \right|, \tag{9}$$

$\{\omega_p\}_{p\in\mathcal{P}}$ is a penalty (weight) parameter and $\overline{x^r}_{\mathcal{T},p}$ denotes the sample mean of the p covariate among treated individuals.

This can be easily generalized to balancing the rth moments:

$$\sum_{p\in\mathcal{P}} \omega_p \gamma_p(\mathbf{a}) = \sum_{p\in\mathcal{P}} \omega_p \left| \sum_{i\in\mathcal{T}} \sum_{j\in\mathcal{C}} \frac{x^r_{j,p} a_{ij}}{m|\mathcal{T}|} - \overline{x^r}_{\mathcal{T},p} \right|, \tag{10}$$

where $\overline{x^r}_{\mathcal{T},p}$ denotes the rth sample moment of the p covariate among treated individuals.

Balancing Univariate Moments (Contd.)

Notice that instead of penalizing imbalance in the objective function, we could enforce a maximum level of imbalance via the constraints

$$\left| \sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{C}} \frac{x_{j,p}^r a_{ij}}{m|\mathcal{T}|} - \overline{x^r}_{\mathcal{T},p} \right| \leq \epsilon_m, \; \forall m \in \mathcal{M}, \tag{11}$$

for some fixed set of values $\{\epsilon_m\}_{m \in \mathcal{M}}$.

Enforcing a certain maximum level of imbalance is typically a trivial variant of penalizing it. The following discussion will thus typically omit specifying variants of these constraints explicitly.

Balancing Multivariate Moments

Imbalance in multivariate moments can be penalized in similar fashion. For example, we may penalize the imbalance in the correlation of the variables p_1 and p_2 between the matched treated and control sample:

$$
\begin{aligned}
\mu_{p_1,p_2}(\mathbf{a}) = \omega_1 & \left| \sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{C}} \frac{x_{j,p_1} a_{ij}}{m|\mathcal{T}|} - \overline{x}_{\mathcal{T},p_1} \right| \\
& + \omega_2 \left| \sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{C}} \frac{x_{j,p_1}^2 a_{ij}}{m|\mathcal{T}|} - \overline{x^2}_{\mathcal{T},p_1} \right| \\
& + \omega_3 \left| \sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{C}} \frac{x_{j,p_2} a_{ij}}{m|\mathcal{T}|} - \overline{x}_{\mathcal{T},p_2} \right| \\
& + \omega_4 \left| \sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{C}} \frac{x_{j,p_2}^2 a_{ij}}{m|\mathcal{T}|} - \overline{x^2}_{\mathcal{T},p_2} \right| \\
& + \omega_5 \left| \sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{C}} \frac{x_{j,p_1} x_{j,p_2} a_{ij}}{m|\mathcal{T}|} - \overline{x_{j,p_1} x_{j,p_2}}_{\mathcal{T},p_2} \right|.
\end{aligned}
\tag{12}
$$

Balancing empirical distributions

An alternative approach to addressing imbalance is to reduce disreptencies in the empirical distributions of the covariates directly. This can be achieved via a discretized version of the Kolmogorov-Smirnov statistic

$$\sup_{g_p \in \mathcal{G}_p} \left| F_{\mathcal{T},p}(g_p) - F_{\mathcal{C},p}(g_p, \mathbf{a}) \right|, \tag{13}$$

where $\mathcal{G}_p = \{g_{1p}, \ldots, g_{Gp}\}_p$ is a grid of quantiles in the treated group for covariate p. Denote the corresponding cumulative proportions $\{h_{1p}, \ldots, h_{Gp}\}$. Then we can minimize the discretized K-S statistic via introduction of the auxiliary variable z_p that satisfies:

$$\mu_p(\mathbf{a}) = z_p \geq \left| h_{g_p} - \sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{C}} \frac{\mathbb{1}\{x_{j,p} < g_p\} a_{ij}}{m|\mathcal{T}|} \right|, \quad \forall g_p \in \mathcal{G}_p. \tag{14}$$

Exact and Near-Exact Matching

In some settings, we may want to be particularly careful to match exactly (or near-exactly) based on a subset of the covariate. In the setting of Zubizarreta (2012), comparing patients who received surgery at the same hospital may be sensible, for example.

To match exactly for the pth covariate, define the constraints

$$\sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{C}} a_{ij} \mathbb{1}\{x_{i,p} = b \wedge x_{j,p} = b\} = m \sum_{i \in \mathcal{T}} \mathbb{1}\{x_{i,p} = b\}, \quad \forall b \in \operatorname{supp} X_p.$$

In empirical settings, such exact matching is only feasible for discrete covariates, of course. And even then, there may be insufficiently many units in the control sample with the same covariate value.

As an alternative, we may thus also consider

$$\left| \sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{C}} a_{ij} \mathbb{1}\{x_{i,p} = b \wedge x_{j,p} = b\} - m \sum_{i \in \mathcal{T}} \mathbb{1}\{x_{i,p} = b\} \right| \leq \xi_p, \quad \forall b \in \operatorname{supp} X_p,$$

where ξ_p is the maximum deviation allowed.

Including Specific Control Units

When $|\mathcal{T}| < |\mathcal{C}|$, there are typically units in the control sample that are not matched. An additional constraint may be imposed to ensure that certain units in the control sample are always included in the matched sample:

$$\sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{C}} a_{ij} \mathbf{1}\{j \in \mathcal{S}\} = |\mathcal{S}|, \tag{15}$$

for some subset $\mathcal{S} \subset \mathcal{C}$.

This may be a useful alternative to increasing the relative importance of certain covariates, which may also be done (in possibly more complicated manner) via the distance matrix $(\delta_{ij})_{i \in \mathcal{T}, j \in \mathcal{C}}$.

Empirical Application in Zubizarreta (2012)

To illustrate the procedure based on the mixed integer program in (8), Zubizarreta (2012) provides comparisons of the empirical covariate distributions for different balancing approaches.

The excerpts in the paper on the next slides show that discrepancies in the empirical distributions of the covariates between the matched treated and control sample can be substantially reduced.

Figure 1 of Zubizarreta (2012)

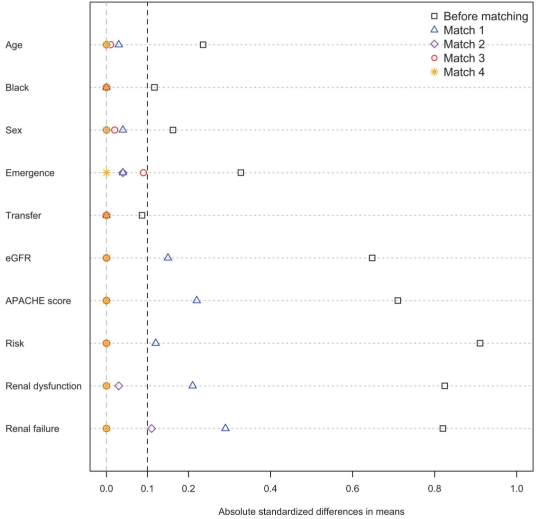

Figure 1. Balancing the means of the observed covariates. Match 1 uses the assignment algorithm. Match 2 uses MIP to directly balance the means with weights $\boldsymbol{\omega} = (0, 0, 0, 0, 0, 10, 10, 10, 10, 10)$. Matches 3 and 4 also use MIP improving upon match 2 with weights $\boldsymbol{\omega} = (0, 0, 0, 0, 10, 10, 10, 10, 25)$ and $\boldsymbol{\omega} = (10, 0, 15, 15, 0, 10, 10, 10, 10, 50)$, respectively. In match 4, the differences in means of all the covariates are reduced to zero. This maximum bias reduction can also be obtained using MIP and forcing the mean imbalances to less or equal than very small tolerances. The online version of this figure is in color.

Figure 2 of Zubizarreta (2012)

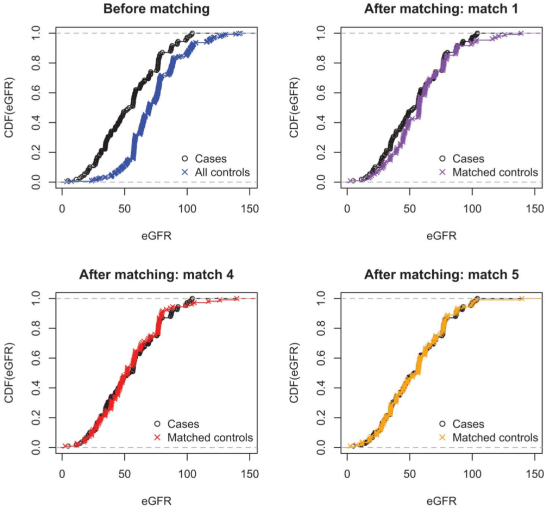

Figure 2. Balancing the K–S statistics for eGFR. Match 1 uses the assignment algorithm. Match 4 uses MIP to directly balance the means of all the covariates through the constraints with tolerances. Match 5 extends match 4 to balance both the means of all the covariates and the K–S statistic for eGFR using the constraint approach. The online version of this figure is in color.

Last Comments

Zubizarreta (2012) suggests an optimal matching approach using mixed integer programming that provides useful tools for covariate balancing.

Unfortunately, the paper does not appear to provide a comparison in the estimates of the parameter of interest – the average causal effect of obesity on the risk of acute kidney failure after surgery – using different matching approaches. Such a comparison could motivate the need for more careful balancing approaches (and is potentially worth trying out yourself!).

Matching via mixed integer programming has applications beyond conventional matching estimators based on the selection on observables assumption. In a setting with instruments, see, for example, Zubizarreta et al. (2013) which extends Baiocchi et al. (2010).

References

Baiocchi, M., Small, D. S., Lorch, S., and Rosenbaum, P. R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association*, 105(492):1285–1296.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1).

Hurwicz, L. (1950). Generalization of the concept of identification. *Statistical inference in dynamic economic models*, 10:245–57.

Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500):1360–1371.

Zubizarreta, J. R., Small, D. S., Goyal, N. K., Lorch, S., and Rosenbaum, P. R. (2013). Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *The Annals of Applied Statistics*, pages 25–50.