OLS, Matching, and Preprocessing: Models Versus Design

Charlie Gibbons
Political Science 239

February 6, 2009

Outline

Outline

These notes draw from:

Freedman, David A. 2008. "On Regression Adjustments to Experimental Data." *Advances in Applied Mathematics.* 40: 180–193.

Ho, David E., Kosuke Imai, Gary King, and Elizabeth A. Stewart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis.* 15: 199–236.

Sekhon, Jasjeet. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." Working paper.

OLS assumptions

Recall the assumptions of ordinary least squares (OLS):

1. $Y_i = x_i'\beta + \epsilon_i$
2. $\mathbb{E}(\epsilon_i | X_i) = 0$
3. X is non-stochastic
4. X is non-singular
5. $\mathbb{E}(\epsilon_i) = \sigma^2$

OLS assumptions

Recall the assumptions of ordinary least squares (OLS):

1. $Y_i = x_i'\beta + \epsilon_i$ About *modeling*
2. $\mathbb{E}(\epsilon_i | X_i) = 0$ About *design*

OLS assumptions

Recall the assumptions of ordinary least squares (OLS):

1. $Y_i = x_i'\beta + \epsilon_i$ About *modeling*
2. $\mathbb{E}(\epsilon_i | X_i) = 0$ About *design*

The response by many methodologists is: *if you have a good, "as if randomized" design, then matching is a preferred model.*

Matching assumptions

Recall the assumptions of matching:

1. Selection on observables:
   $(Y_i(0), Y_i(1)) \perp T_i | X_i$
   Heckman showed that this can be weakened to conditional mean independence:
   $\mathbb{E}\left(Y_i(t)|X, T = 1\right) = \mathbb{E}\left(Y_i(t)|X, T = 0\right)$ for $t \in 0, 1$
2. Common support on covariates:
   $0 < \Pr(T = 1|X = x) < 1 \; \forall \; x \in X$
3. Stable Unit Treatment Value Assumption (SUTVA):
   $(Y_i(0), Y_i(1)) \perp T_j \; \forall \; j \neq i.$

Matching assumptions

Recall the assumptions of matching:

1. Selection on observables:
   $(Y_i(0), Y_i(1)) \perp T_i | X_i$
   Heckman showed that this can be weakened to conditional mean independence:
   $\mathbb{E}\left(Y_i(t)|X, T = 1\right) = \mathbb{E}\left(Y_i(t)|X, T = 0\right)$ for $t \in 0, 1$
2. Common support on covariates:
   $0 < \Pr(T = 1|X = x) < 1 \; \forall \; x \in X$
3. Stable Unit Treatment Value Assumption (SUTVA):
   $(Y_i(0), Y_i(1)) \perp T_j \; \forall \; j \neq i.$

These are all *design* assumptions because matching is non-parametric. Note, too, that SUTVA is an often unstated, yet important assumption in OLS as well. Also, for both cases, X is assumed to contain pre-treatment covariates only.

Framework

Ho et al. making the following claims:

1. We don't know the *theoretically* right model for our data and
2. We can't *empirically* determine the correct model for our data,
3. But most causal estimates are *model dependent*,
4. Thereby implying that we don't know the right causal estimate.

Framework

Ho et al. making the following claims:

1. We don't know the *theoretically* right model for our data and
2. We can't *empirically* determine the correct model for our data,
3. But most causal estimates are *model dependent*,
4. Thereby implying that we don't know the right causal estimate.

They also note that "we cannot logically even ask whether an estimator has desirable properties, such as unbiasedness, consistency, efficiency, mean squared error, etc., since a unique estimator must exist before it can be evaluated."

Framework

They start with the experimental benchmark, which has the following qualities:

- Random selection of units into the potential treatment pool
- Random assignment of treatment
- Large sample sizes

In this case, there can be no omitted variable bias (in expectation) confounding our results.

Framework

To move to observational data, they assume selection on observables. Additionally, for all cases, they assume "SUTVA" and a homogeneous treatment effect.

Framework

To move to observational data, they assume selection on observables. Additionally, for all cases, they assume "SUTVA" and a homogeneous treatment effect.

Note that a homogeneous treatment effect implies that:

1. Treatment does not interact with other covariates
2. "Coefficients" on treatment do not vary across individuals

This assumption already restricts the space of potential parametric models.

Preprocessing as a solution

To solve the preceding modeling problems, they propose creating a matched data set as an initial "preprocessing" step, followed by a standard modeling approach.

Their goals are:

1. To reduce or eliminate the relationship between X and T
2. Add little bias or inefficiency due to subsequent modeling assumptions

They claim that causal estimates based upon a matched data set are more robust to different model choices.

Preprocessing as a solution

To solve the preceding modeling problems, they propose creating a matched data set as an initial "preprocessing" step, followed by a standard modeling approach.

Their goals are:

1. To reduce or eliminate the relationship between X and T
2. Add little bias or inefficiency due to subsequent modeling assumptions

They claim that causal estimates based upon a matched data set are more robust to different model choices.

Note that perfect balance implies 0 covariance between treatment and the covariates.

Variance and causal estimates

They assume selection on observables, common support, and SUTVA, all the assumptions necessary for matching. They question now becomes, since the matching estimator is consistent, *why do post-matching parametric modeling at all?*

Variance and causal estimates

They assume selection on observables, common support, and SUTVA, all the assumptions necessary for matching. They question now becomes, since the matching estimator is consistent, *why do post-matching parametric modeling at all?*

Their first answer: to permit interpolation and (slight) extrapolation.

Variance and causal estimates

They assume selection on observables, common support, and SUTVA, all the assumptions necessary for matching. They question now becomes, since the matching estimator is consistent, *why do post-matching parametric modeling at all?*

Their first answer: to permit interpolation and (slight) extrapolation.

But how this fitting occurs is certainly model dependent. They would respond that, when treatment is (almost) uncorrelated with the other covariates, a linear approximation may not be so bad.

Variance and causal estimates

Their second answer: to reduce variance.

Variance and causal estimates

Their second answer: to reduce variance.

We have already seen (see, *e.g.*, PS 236 Problem Set 4) that running OLS on randomized data reduces mean squared error, which may help to identify statistically significant causal effects. Can it help in as-if randomizations generated through matching?

Variance and causal estimates

Wait a minute! Freedman (2008) shows that, in an experimental setting, the standard intention to treat (ITT) estimator that is generated by matching (without observing unit take-up) is identical to the estimate from a simple regression model of outcome on an intercept and a treatment dummy.

Variance and causal estimates

But he also shows that:

1. The multiple regression estimator is biased; the bias tends to 0 as the number of subjects increases.
2. Asymptotically, the multiple regression estimator may perform worse than the simple regression estimator.
3. "Nominal" standard errors (computed from the usual formulas) can be severely biased.
4. The nominal standard error for the simple regression estimator may differ from the nominal standard error for the intention-to-treat estimator—even though the two estimators coincide.

Variance and causal estimates

Why? *Because randomization does not justify the linear model.*

Variance and causal estimates

Why? *Because randomization does not justify the linear model.*

Specifically, the "coefficients" need not be the same for all units or independent of the covariates (*i.e.*, heterogeneous treatment effects), the responses need not be linear, and the errors need not be heteroskedastic.

Variance and causal estimates

That was for randomization (the "simple" case), but what about matching?

Variance and causal estimates

That was for randomization (the "simple" case), but what about matching?

We know (see, *e.g.*, PS 236 Problem Set 3) that naive post-matching modeling that uses "nominal" standard errors rather than the Abadie-Imbens standard errors *does not produce correct coverage*. Thus, Ho et al.'s primary logic for performing post-matching bias adjustment is flawed.

Variance and causal estimates

Why should we use post-matching bias adjustment?

Variance and causal estimates

Why should we use post-matching bias adjustment?

Since a model on top of the matched data incorporates any remaining covariance between treatment status and the covariates in estimation, it may demonstrate that our non-parametric estimate is sensitive to modeling assumptions, thereby implying that important sources of confounding may remain in our matched sample; *i.e.*, balance is not sufficient to generate selection on observables and to produce a reliable estimate.

Variance and causal estimates

Why should we use post-matching bias adjustment?

Since a model on top of the matched data incorporates any remaining covariance between treatment status and the covariates in estimation, it may demonstrate that our non-parametric estimate is sensitive to modeling assumptions, thereby implying that important sources of confounding may remain in our matched sample; *i.e.*, balance is not sufficient to generate selection on observables and to produce a reliable estimate.

Of course, a different bias-adjusted result could just be biased itself, so we cannot truly verify anything with this practice.

Focus on design

As the preceding makes clear, relying on models can be very misleading. Hence, more thought ought to be placed on creating a randomized design.

Focus on design

In his paper, Jas makes many of these points and offers the following suggestions:

- The ITT should always be reported, and going beyond ITT should only be done with care.
- All data analysis should leverage the experimental design as much as possible.
- Our belief should be that selection on observables and other identifying assumptions not guaranteed by the design are incorrect unless compelling evidence to the contrary is provided.
- Placebo tests should be conducted whenever possible, and observational studies without them should be marked down.