Model-based matching for causal inference in observational studies

Kellie Ottoboni
with Philip B. Stark and Jas Sekhon

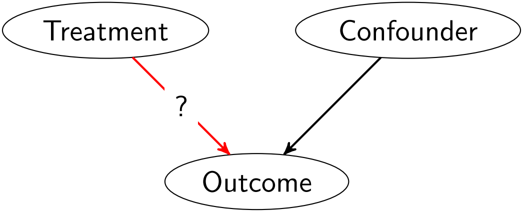
Department of Statistics, UC Berkeley
Berkeley Institute for Data Science

March 10, 2016

Observational Studies vs Experiments

TO DO: Check that bibliography entries look okay

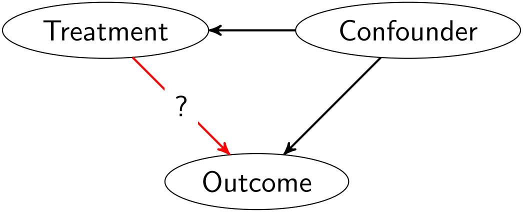
- **Problem:** Estimate the causal effect of a treatment on outcome of interest
- In randomized experiments, treatment is assigned to individuals at random.
- In observational studies, the way individuals select into treatment groups is unknown.



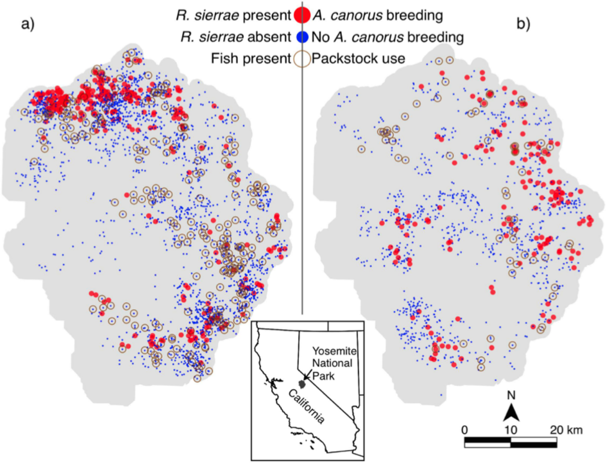
Observational Studies vs Experiments

TO DO: Check that bibliography entries look okay

- **Problem:** Estimate the causal effect of a treatment on outcome of interest
- In randomized experiments, treatment is assigned to individuals at random.
- In observational studies, the way individuals select into treatment groups is unknown.



Motivating Example: Toads and Packstock



J. R. Matchett, Philip B. Stark, Steven M. Ostoja, Roland A. Knapp, Heather C. McKenny, Matthew L. Brooks, William T. Langford, Lucas N. Joppa, and Eric L. Berlow. Detecting the influence of rare stressors on rare species in Yosemite National Park using a novel stratified permutation test. Scientific Reports, 5: 10702, June 2015.

Motivating Example: Toads and Packstock

- The response is rare (few meadows have toads).
- The treatment is rare (few meadows are used by packstock).
- Randomized experiment is impossible, and toad/packstock presence is not random across meadows.
- We're interested in detecting any effect, no matter how small. If treatment effect varies across meadows, then averages might not be informative.

Goal

Goal: test the **strong null hypothesis** of no treatment effect whatsoever.

$$H_0 : Y_i(1) = Y_i(0) \text{ for all } i$$
$$H_1 : Y_i(1) \neq Y_i(0) \text{ for some } i$$

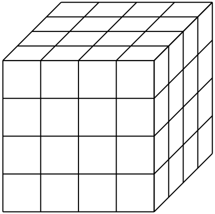
We'd like our test to have power to detect

- non-constant effects
- non-linear effects
- effects with non-constant sign

Matching

How can we estimate the counterfactual for treated individuals?

- **Ideal:** group individuals by X_i to estimate subgroup treatment effects and then average over subgroups
- **Reality:** many covariates, perhaps continuous, make it difficult to stratify



- **Solution:** use a one-dimensional score to match or group individuals

Propensity score matching

$p(x)$ is usually unknown and estimated by $\hat{p}(x)$ using logistic or probit regressions

- Assumes a simple functional form for relationship between covariates and treatment
- Assumes that probability of treatment takes same form for all individuals
- May actually worsen balance if estimated incorrectly [Diamond and Sekhon, 2012]

Matching complicates inference

- Standard errors are difficult to compute for matching estimators [Abadie and Imbens, 2006, 2008]
- Rarely used in hypothesis testing procedures
- There's no "optimal" way to match [Austin, 2014]

Model-based Matching

Idea: Instead of modeling the propensity score, model the outcome

Computing $\hat{Y}$, the "best" prediction of the outcome based on all covariates except for the treatment, buys us two things:

- $\hat{Y}$ is a score on which to stratify observations
- Using residuals $Y - \hat{Y}$ improves precision by removing variation due to X [Rosenbaum, 2002]

Model-based Matching

Suppose that outcomes have the form

$$Y_i(t) = f(t, X_i) + \varepsilon_i$$

for $i = 1, \ldots, N$ and $t = 0, 1$. Let X_i be fixed and suppose that the ε_i are IID with $\mathbb{E}(\varepsilon_i) = 0$.

Under the strong null hypothesis, $f(0, X_i) = f(1, X_i)$ for each i.

Thus, our best guess of Y_i needn't involve the treatment:

$$\hat{Y}_i = \hat{f}(X_i)$$

Model-based Matching

Stratify or match units on their $\hat{Y}_i = \hat{f}(X_i)$.

- Let $S_i = j$ if unit i is in stratum j, where $j \in \{1, \ldots, J\}$. (For now, don't worry about how to select J strata.)

- **Under the null**, we expect units in the same strata to have the similar responses.

- **Under the alternative**, the treatment adds additional information about the responses beyond $\hat{f}$.

  The residuals will capture some of the effect of treatment:

$$Y_i - \hat{Y}_i \not\perp T_i$$

Permutation tests

We will use the average difference in means across strata as our test statistic:

$$\tau(Y,T) = \frac{N_j}{N}\sum_{j=1}^{J}\left|\frac{n_j}{N_j}\sum_{\substack{i:S_i=j\\T_i=1}}\left(Y_i-\hat{Y}_i\right) - \frac{N_j-n_j}{N_j}\sum_{\substack{i:S_i=j\\T_i=0}}\left(Y_i-\hat{Y}_i\right)\right|$$

NB: we can use any other test statistic that measures association between $Y_i - \hat{Y}_i$ and T_i, e.g. correlation

Permutation tests

TO DO: REWRITE THE BASIC IDEA IN PLAIN LANGUAGE **Basic idea:** If, under the null hypothesis, the probability distribution of the data is invariant under permutation of treatment assignments, then once we observe the actual data, we know other possible data sets that are equally likely.

Suppose that the jth stratum contains N_j units, n_j of which are treated. Then there are

$$\prod_{j=1}^{J} \binom{N_j}{n_j}$$

equally likely assignments to treatment, conditional on the number of treated units in each stratum.

Permutation tests

We approximate the null distribution using this invariance principle.

- Within strata, permute treatment assignments to obtain new treatment vector T^*_1.
- Compute the test statistic $\tau(Y, T^*_1)$.
- Repeat a large number B times to get a distribution $\tau(Y, T^*_1), \ldots, \tau(Y, T^*_B)$.
- The p-value of the test is

$$p = \mathbb{P}(\tau(Y, T) \geq \tau(Y, t)) \approx \frac{\sum_{i=1}^{B} \mathbb{I}(\tau(Y, T^*_b) \geq \tau(Y, T))}{B}$$

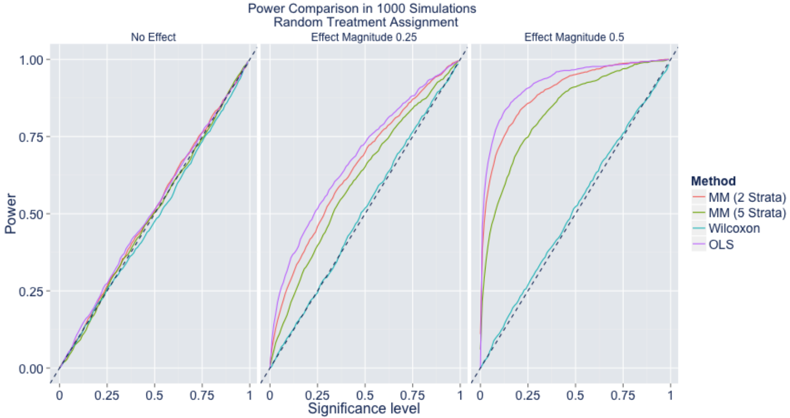
Simulation set-up

$$Y_i = 1 + 2X_{1i} + 4X_{2i} + \tau_i T_i + \varepsilon_i, \qquad i = 1, \ldots, 100$$

- X_{1i}, X_{2i} are independent $N(0,1)$
- $\varepsilon_i \sim N(0,1)$ (unless specified otherwise)
- T_i assigned various ways
    - Random, independent of everything
    - Correlated with X_1: $T_i = \nu X_{1i} + \delta_i$, with $\delta_i \sim N(0,1)$
    - Correlated with X_1 and X_2: $T_i = \nu X_{1i} + X_{1i}X_{2i} + \delta_i$, with $\delta_i \sim N(0,1)$
- We vary τ_i and the method of generating T_i

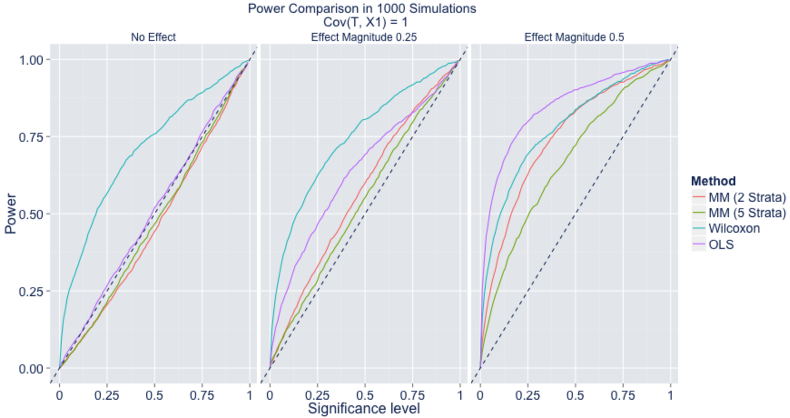
Results

Model-based matching tests have correct level



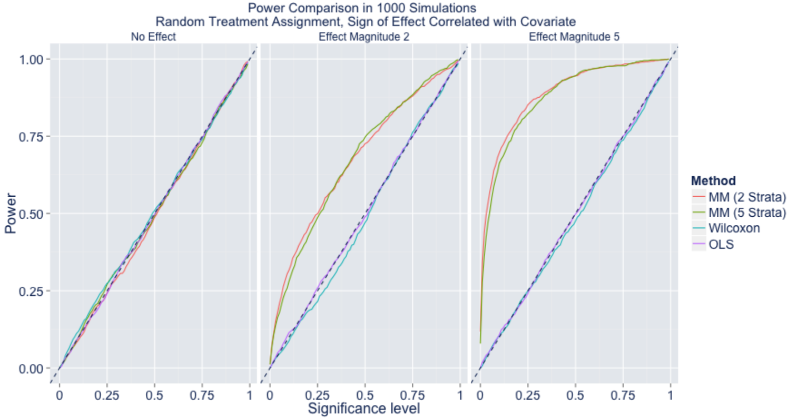
Results

Model-based matching tests have correct level



Results

Model-based matching tests have higher power when treatment effects are non-constant



Future Directions

- Do different test statistics give greater power? Under what circumstances?
- What is the optimal way to stratify?
- How to estimate effects and quantify uncertainty – standard errors and confidence intervals?

Stratification

There are two competing forces that determine optimal strata:

- Power: we need enough variation in treatment within strata
- Precision: we want small enough strata to capture variation in treatment effects across strata

TO DO: FLESH OUT

Estimation

Approach 1: direct estimation

If selection on observables holds and we fit $\hat{f}$ using only the controls, then an unbiased estimate of ATE τ is

$$\hat{\tau} = \frac{1}{N_t} \sum_{i:T_i=1} (Y_i - \hat{Y}_i) - \frac{1}{N_c} \sum_{i:T_i=0} (Y_i - \hat{Y}_i)$$

How can we put a standard error on this? Asymptotics...

Estimation

Approach 2: inverting hypothesis tests

Let A_{τ_0} be the acceptance region of a level-α test of the hypothesis $\tau = \tau_0$.

$S(X) = \{\tau \in \mathbb{R} : X \in A_\tau\}$ is a $1 - \alpha$ confidence set for τ.

An estimate of τ is the value which minimizes the probability of rejecting the null (i.e. maximizes the p-value).

$$\tilde{\tau} = \operatorname*{argmax}_{\tau \in \mathbb{R}} \mathbb{P}_\tau(X \in A_\tau)$$

Estimation

Approach 2: inverting hypothesis tests

Under $H_0 : \tau = 0$, we know both potential outcomes. For $\tau \neq 0$, we don't.

We must assume some form for the treatment effect.

- Typically, one assumes constant additive effect
- We can generalize to $Y(1) = g(Y(0), \tau)$ where g satisfies certain assumptions
- How can we let effects vary across strata?

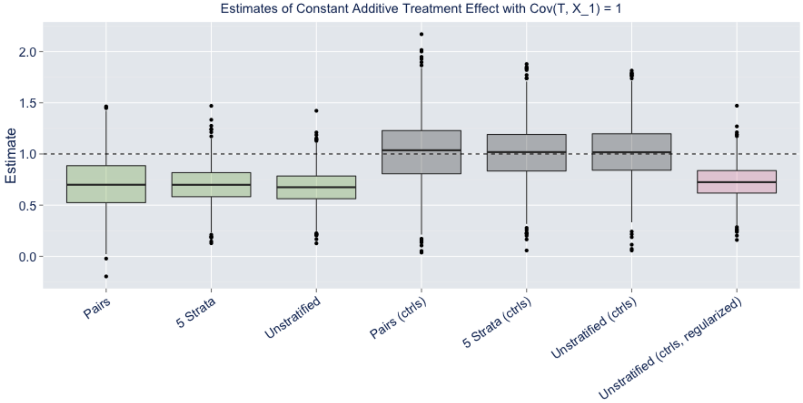
Estimation

Several questions arise:

- What is the model of treatment effects under the alternative hypothesis?
- Are we interested in ATE? What about
  - $\mathbb{E}(Y(1) - Y(0) \mid Y(0))$
  - $\mathbb{E}(Y(1) - Y(0) \mid X)$
  - $\max\{Y(1) - Y(0)\}$
- Back to the original problem of how to fit $\hat{f}$
  - Fitting to controls only gives a test with incorrect level
  - Fitting to all observations biases estimated ATE

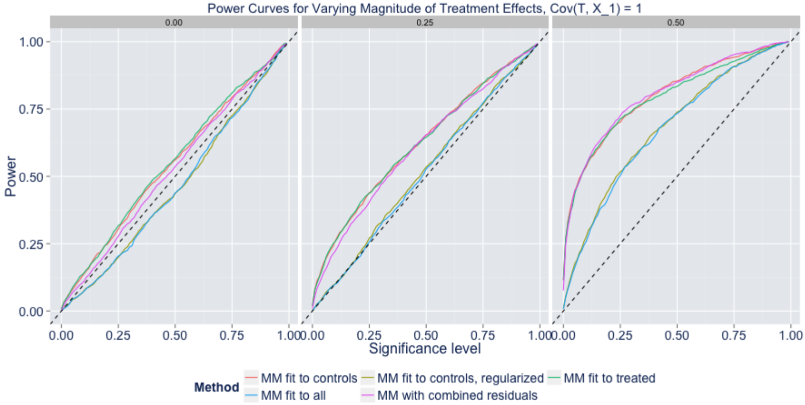
Fitting method

Estimation is unbiased when we fit to controls



Fitting method

Testing has higher than nominal level when we fit to controls



References

Alberto Abadie and Guido W. Imbens. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, 74(1):235–267, January 2006. ISSN 1468-0262. doi: 10.1111/j.1468-0262.2006.00655.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0262.2006.00655.x/abstract.

Alberto Abadie and Guido W. Imbens. On the Failure of the Bootstrap for Matching Estimators. *Econometrica*, 76 (6):1537–1557, November 2008. ISSN 1468-0262. doi: 10.3982/ECTA6474. URL http://onlinelibrary.wiley.com/doi/10.3982/ECTA6474/abstract.

Peter C. Austin. A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, 33(6): 1057–1069, March 2014. ISSN 1097-0258. doi: 10.1002/sim.6004. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.6004/abstract.

Alexis Diamond and Jasjeet S. Sekhon. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics*, 95(3): 932–945, October 2012. ISSN 0034-6535. doi: 10.1162/REST_a_00318. URL http://dx.doi.org/10.1162/REST_a_00318.

Paul R. Rosenbaum. Covariance Adjustment in Randomized Experiments and Observational Studies. *Statistical Science*, 17(3):286–327, August 2002. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1042727942. URL http://projecteuclid.org/euclid.ss/1042727942.