September 2019 **Potsdam**

Causal Inference and Machine Learning

Guido Imbens, imbens@stanford.edu

Course Description

The course will cover topics on the intersection of causal inference and machine learning. There will be particular emphasis on the use of machine learning methods for estimating causal effects. In addition there will be some discussion of basic machine learning methods that we view as useful tools for empirical economists.

Lectures

There will be six lectures.

Background Reading

We strongly recommend that participants read these articles in preparation for the course.

- Athey, Susan, and Guido W. Imbens. "The state of applied econometrics: Causality and policy evaluation." Journal of Economic Perspectives 31.2 (2017): 3-32.

Course Outline

1. Monday September 9th, 14.30-16.00: Introduction to Causal Inference
   (a) Holland, Paul W. "Statistics and causal inference." Journal of the American statistical Association 81.396 (1986): 945-960.
   (b) Imbens, Guido W., and Donald B. Rubin. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, 2015.
   (c) Imbens, Guido W., and Jeffrey M. Wooldridge. "Recent developments in the econometrics of program evaluation." Journal of economic literature 47.1 (2009): 5-86.
2. Monday, September 9th 16.30-18.00: Introduction to Machine Learning Concepts
   (a) S. Athey (2018, January) "The Impact of Machine Learning on Economics," *Sections 1-2.* `http://bit.ly/2EENtvy`
   (b) H. R. Varian (2014) "Big data: New tricks for econometrics." *The Journal of Economic Perspectives*, 28 (2):3-27. `http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.28.2.3`
   (c) S. Mullainathan and J. Spiess (2017) "Machine learning: an applied econometric approach" *Journal of Economic Perspectives*, 31(2):87-106 `http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.31.2.87`

(d) L. Breiman, J. Friedman, C. J. Stone R. A. Olshen (1984) "Classification and regression trees," CRC press.

(e) Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. No. 10. New York, NY, USA:: Springer series in statistics, 2001.

(f) I. Goodfellow, Y. Bengio, and A. Courville (2016) "Deep Learning." MIT Press.

3. Tuesday, September 10th, 10.30-12.00: Causal Inference: Average Treatment Effects with Many Covariates

   (a) A. Belloni, V. Chernozhukov, and C. Hansen (2014) "High-dimensional methods and inference on structural and treatment effects." *The Journal of Economic Perspectives*, 28(2):29-50. `http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.28.2.29`

   (b) V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2017, December) "Double/Debiased Machine Learning for Treatment and Causal Parameters." `https://arxiv.org/abs/1608.00060`.

   (c) Athey, Susan, Guido W. Imbens, and Stefan Wager. "Approximate residual balancing: debiased inference of average treatment effects in high dimensions." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80.4 (2018): 597-623.

   (d) S. Athey, G. Imbens, and S. Wager (2016) "Estimating Average Treatment Effects: Supplementary Analyses and Remaining Challenges." `http://arXiv/abs/1702.01250`. Forthcoming, *Journal of the Royal Statistical Society-Series B.*

4. Tuesday, September 10th, 13.15-14.45: Causal Inference: Heterogeneous Treatment Effects

   (a) S. Wager and S. Athey (2017) "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* `http://arxiv.org/abs/1510.04342`

   (b) S. Athey, Tibshirani, J., and S. Wager (2017, July) "Generalized Random Forests" `http://arxiv.org/abs/1610.01271`

5. Tuesday, September 10th, 15.15-16.45pm: Causal Inference: Experimental Design and Multi-armed Bandits

   (a) S. Athey and S. Wager (2017) "Efficient Policy Learning." `http://arXiv.org/abs/1702.02896`.

   (b) M. Dudik, D. Erhan, J. Langford, and L. Li, (2014) "Doubly Robust Policy Evaluation and Optimization" *Statistical Science*, Vol 29(4):485-511.

   (c) S. Scott (2010), "A modern Bayesian look at the multi-armed bandit," *Applied Stochastic Models in Business and Industry*, vol 26(6):639–658.

   (d) M. Dimakopoulou, S. Athey, and G. Imbens (2017). "Estimation Considerations in Contextual Bandits." `http://arXiv.org/abs/1711.07077`.

6. Wednesday, September 11th, 10.00-11.30: Synthetic Control Methods and Matrix Completion

   (a) S. Athey, M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2017) "Matrix Completion Methods for Causal Panel Data Models." `http://arXiv.org/abs/1710.10251`.

   (b) J. Bai (2009), "Panel data models with interactive fixed effects." *Econometrica*, 77(4): 1229–1279.

   (c) E. Candès and B. Recht (2009) "Exact matrix completion via convex optimization." *Foundations of Computational mathematics*, 9(6):717-730.

Causal Inference

and Machine Learning

Guido Imbens – Stanford University

Lecture 1:

Introduction to Causal Inference

Potsdam Center for Quantitative Research

Monday September 9th, 14.30-16.00

Outline

1. Causality: Potential Outcomes, Multiple Units, and the Assignment Mechanism

2. Fisher Randomization Tests

3. Neyman's Repeated Sampling Approach

4. Stratified Randomized Experiments

1. Causality: Potential Outcomes, Multiple Units, and the Assignment Mechanism

Three key notions underlying the general approach to causality. First, *potential outcomes*, each corresponding to the various levels of a *treatment* or manipulation.

Second, the presence of *multiple units*, and the related *stability* assumption.

Central role of the *assignment mechanism*, which is crucial for inferring causal effects and serves as the organizing principle.

1.1 Potential Outcomes

Given a unit and a set of actions, we associate each action/unit pair with a potential outcome: "potential" because only one will ultimately be realized and therefore possibly observed: the potential outcome corresponding to the action actually taken at that time.

The causal effect of the action or treatment involves the **comparison** of these potential outcomes, some realized (and perhaps observed) and others not realized and thus not observed.

$Y(0)$ denotes the outcome given the control treatment,
$Y(1)$ denotes the outcome given the active treatment.
$W \in \{0, 1\}$ denotes indicator for treatment,
observe W and $Y^{\text{obs}} = Y(W) = W \cdot Y(1) + (1 - W) \cdot Y(0)$.

Is this useful?

• Potential outcome notion is consistent with the way economists think about demand functions: quantities demanded at different prices.

• some causal questions become more tricky: causal effect of race on economic outcomes. One solution is to make manipulation precise: change names on cv for job applications (Bertrand and Mullainathan).

• what is causal effect of physical appearance, height, or gender, on earnings, obesity on health? Strong statistical correlations, but what do they mean? Many manipulations possible, probably all with different causal effects.

1.2 Multiple Units

Because we cannot learn about causal effects from a single observed outcome, we must rely on multiple units exposed to different treatments to make causal inferences.

By itself, however, the presence of multiple units does not solve the problem of causal inference. Consider a drug (aspirin) example with two units—you and I—and two possible treatments for each unit—aspirin or no aspirin.

There are now a total of four treatment levels: you take an aspirin and I do not, I take an aspirin and you do not, we both take an aspirin, or we both do not.

In many situations it may be reasonable to assume that treatments applied to one unit do not affect the outcome for another (*Stable Unit Treatment Value Assumption*, Rubin, 1978).

• In agricultural fertilizer experiments, researchers have taken care to separate plots using "guard rows," unfertilized strips of land between fertilized areas.

• In large scale job training programs the outcomes for one individual may well be affected by the number of people trained when that number is sufficiently large to create increased competition for certain jobs (Crepon, Duflo et al)

• In the peer effects / social interactions literature these interaction effects are the main focus.

Six Observations from the GAIN Experiment in Los Angeles

Individual	Potential Outcomes $Y_i(0)$	$Y_i(1)$	Actual Treatment	Observed Outcome Y_i^{obs}
1	66	?	0	66
2	0	?	0	0
3	0	?	0	0
4	?	0	1	0
5	?	607	1	607
6	?	436	1	436

Note: $(Y_i(0), Y_i(1))$ fixed for $i = 1, \ldots, 6$. $(W_1, \ldots, W_6)$ is stochastic.

1.3 The Assignment Mechanism

The key piece of information is *how* each individual came to receive the treatment level received: in our language of causation, the **assignment mechanism**.

$$\Pr(\mathbf{W}|\mathbf{Y}(0),\mathbf{Y}(1),\mathbf{X})$$

Known, no dependence on $\mathbf{Y}(0),\mathbf{Y}(1)$: **randomized experiment** (first three lectures)

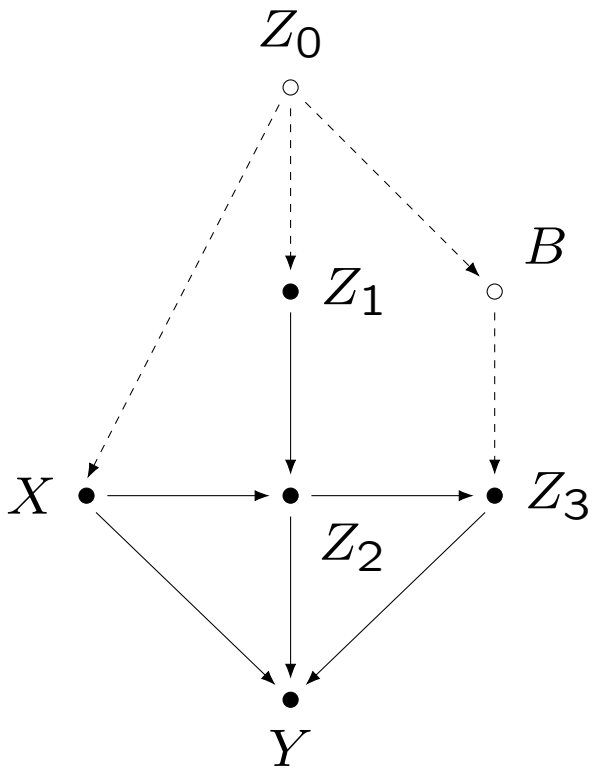
Unknown, no dependence on $\mathbf{Y}(0),\mathbf{Y}(1)$: **Unconfounded assignment / Selection on Observables** (later in course)

• Compare with conventional focus on distribution of outcomes given explanatory variables. Here, other way around, e.g.,

$$Y^{\text{obs}}|W_i \sim \mathcal{N}(\alpha+\beta W_i,\sigma^2)$$

1.4 Graphical Models for Causality

In graphical models the causal relationships are captured by arrows. (Pearl, 1995, 2000)



Differences between Directed Acyclical Graphs (DAGs) and Potential Outcome Framework

- DAGs are all about identification, not about estimation.

- Causes need not be manipulable. in DAGs

- No special role for randomized experiments

- Difficult to capture shape restrictions, e.g., monotonicity, convexity, that are common in economics, for example in instrumental variables.

- Pearl views DAG assumptions as more accessible then potential outcome assumptions.

2. Randomized Experiments: Fisher Exact P-values

Given data from a randomized experiment, Fisher was interested testing **sharp null hypotheses**, that is, null hypotheses under which all values of the potential outcomes for the units in the experiment are either observed or can be inferred.

Notice that this is distinct from the question of whether the **average** treatment effect across units is zero.

The null of a zero average is a much weaker hypothesis because the average effect of the treatment may be zero even if for some units the treatment has a positive effect, as long as for others the effect is negative.

2.1 Basics

Because the null hypothesis is **sharp** we can determine the distribution of any test statistic T (a function of the stochastic assignment vector, $\mathbf{W}$, the observed outcomes, $\mathbf{Y}^{\text{obs}}$, and pretreatment variables, $\mathbf{X}$) generated by the randomization of units across treatments.

The test statistic is stochastic solely through the stochastic nature of the assignment vector, leading to the **randomization distribution** of the test statistic.

Using this distribution, we can compare the observed test statistic, T^{obs}, against its distribution under the null hypothesis.

The Fisher exact test approach entails two choices: (*i*) the choice of the sharp null hypothesis, (*ii*) the choice of test statistic.

We will test the sharp null hypothesis that the program had absolutely no effect on earnings, that is:

$$H_0: \quad Y_i(0) = Y_i(1) \qquad \text{for all} \quad i = 1, \ldots, 6.$$

Under this null hypothesis, the unobserved potential outcomes are equal to the observed outcomes for each unit. Thus we can fill in all six of the missing entries using the observed data.

This is the first key point of the Fisher approach: under the sharp null hypothesis all the missing values can be inferred from the observed ones.

Six Observations from the GAIN Experiment in Los Angeles

Individual	Potential Outcomes $Y_i(0)$	$Y_i(1)$	Actual Treatment	Observed Outcome Y_i
1	66	(66)	0	66
2	0	(0)	0	0
3	0	(0)	0	0
4	(0)	0	1	0
5	(607)	607	1	607
6	(436)	436	1	436

Now consider testing this null against the alternative hypothesis that $Y_i(0) \neq Y_i(1)$ for some units, based on the test statistic:

$$T_1 = T(\mathbf{W}, \mathbf{Y}^{\text{obs}}) = \frac{1}{3}\sum_{i=1}^{6} W_i \cdot Y_i^{\text{obs}} - \frac{1}{3}\sum_{i=1}^{6}(1-W_i)\cdot Y_i^{\text{obs}}$$

$$= \frac{1}{3}\sum_{i=1}^{6} W_i \cdot Y_i(1) - \frac{1}{3}\sum_{i=1}^{6}(1-W_i)\cdot Y_i(0).$$

For the observed data the value of the test statistic is $(Y_4^{\text{obs}} + Y_5^{\text{obs}} + Y_6^{\text{obs}} - Y_1^{\text{obs}} - Y_2^{\text{obs}} - Y_3^{\text{obs}})/3 = 325.6$.

Suppose for example, that instead of the observed assignment vector $\mathbf{W}^{\text{obs}} = (0,0,0,1,1,1)'$ the assignment vector had been $\tilde{\mathbf{W}} = (0,1,1,0,1,0)$. Under this assignment vector the test statistic would have been $(-Y_4^{\text{obs}} + Y_5^{\text{obs}} - Y_6^{\text{obs}} - Y_1^{\text{obs}} + Y_2^{\text{obs}} + Y_3^{\text{obs}})/3 = 35$.

Randomization Distribution for six observations from GAIN data

W_1	W_2	W_3	W_4	W_5	W_6	levels	ranks
0	**0**	**0**	**1**	**1**	**1**	**325.6**	**1.00**
0	0	1	0	1	1	325.6	1.67
0	0	1	1	0	1	-79.0	-1.67
0	0	1	1	1	0	35.0	-1.00
0	1	0	0	1	1	325.6	2.33
0	1	0	1	0	1	-79.0	-1.00
0	1	0	1	1	0	35.0	-0.33
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	1	0	0	0	325.6	-1.00

Given the distribution of the test statistic, how unusual is this observed average difference 325.6), assuming the null hypothesis is true?

One way to formalize this question is to ask how likely it is (under the randomization distribution) to observe a value of the test statistic that is as large in absolute value as the one actually observed.

Simply counting we see that there are twelve vectors of assignments with at least a difference in absolute value of 325.6 between treated and control classes, out of a set of twenty possible assignment vectors. This implies a p-value of $8/20 = 0.40$.

2.2 The Choice of Null Hypothesis

The first question when considering a Fisher Exact P-value calculation is the choice of null hypothesis. Typically the most interesting sharp null hypothesis is that of no effect of the treatment: $Y_i(0) = Y_i(1)$ for all units.

Although Fisher's approach cannot accommodate a null hypothesis of an average treatment effect of zero, it can accommodate sharp null hypotheses other than the null hypothesis of no effect whatsoever, e.g.,

$$H_0: \quad Y_i(1) = Y_i(0) + c_i, \quad \text{for all } i = 1, \ldots, N,$$

for known c_i.

2.3 The Choice of Statistic

The second decision, the choice of test statistic, is typically more difficult than the choice of the null hypothesis. First let us formally define a statistic:

A statistic T *is a known function* $T(\mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X})$ *of assignments,* $\mathbf{W}$*, observed outcomes,* $\mathbf{Y}^{\text{obs}}$*, and pretreatment variables,* $\mathbf{X}$*.*

Any statistic that satisfies this definition is valid for use in Fisher's approach and we can derive its distribution under the null hypothesis.

The most standard choice of statistic is the difference in average outcomes by treatment status:

$$T = \frac{\sum W_i Y_i^{\text{obs}}}{\sum W_i} - \frac{\sum (1 - W_i) Y_i^{\text{obs}}}{\sum (1 - W_i)}.$$

An obvious alternative to the simple difference in average outcomes by treatment status is to transform the outcomes before comparing average differences between treatment levels, e.g., by taking logarithms, leading to the following test statistic:

$$T = \frac{\sum W_i \ln(Y_i^{\text{obs}})}{\sum W_i} - \frac{\sum (1 - W_i) \ln(Y_i^{\text{obs}})}{\sum 1 - W_i}.$$

An important class of statistics involves transforming the outcomes to **ranks** before considering differences by treatment status. This improves robustness.

We also often subtract $(N+1)/2$ from each to obtain a normalized rank that has average zero in the population:

$$R_i(Y_1^{\text{obs}}, \ldots, Y_N^{\text{obs}}) = \sum_{j=1}^{N} 1_{Y_j^{\text{obs}} \leq Y_i^{\text{obs}}} - \frac{N+1}{2}.$$

Given the ranks R_i, an attractive test statistic is the difference in average ranks for treated and control units:

$$T = \frac{\sum W_i R_i}{\sum W_i} - \frac{\sum (1-W_i) R_i}{\sum 1 - W_i}.$$

2.4 Computation of p-values

The p-value calculations presented so far have been exact. With both N and M sufficiently large, it may therefore be unwieldy to calculate the test statistic for every value of the assignment vector.

In that case we rely on numerical approximations to the p-value.

Formally, randomly draw an N-dimensional vector with $N - M$ zeros and M ones from the set of assignment vectors. Calculate the statistic for this draw (denoted T_1). Repeat this process $K - 1$ times, in each instance drawing another vector of assignments and calculating the statistic T_k, for $k = 2, \ldots, K$. We then approximate the p-value for our test statistic by the fraction of these K statistics that are more extreme than T^{obs}.

Comparison to p-value based on normal approximation to distribution of t-statistic:

$$t = \frac{\overline{Y}_1 - \overline{Y}_0}{\sqrt{s_0^2/(N-M) + s_1^2/M}}$$

where

$$s_0^2 = \frac{1}{N-M-1} \sum_{i:W_i=0} (Y_i^{\text{obs}} - \overline{Y}_0)^2, \qquad s_1^2 = \frac{1}{M-1} \sum_{i:W_i=1} (Y_i^{\text{obs}} - \overline{Y}_1$$

and

$$p = 2 \times \Phi(-|t|) \quad \text{where} \quad \Phi(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

P-values for Fisher Exact Tests: Ranks versus Levels

Prog	Loc	sample size controls	sample size treated	p-values t-test	p-values FET (levels)	p-values FET (ranks)
GAIN	AL	601	597	0.835	0.836	0.890
GAIN	LA	1400	2995	0.544	0.531	0.561
GAIN	RI	1040	4405	0.000	0.000	0.000
GAIN	SD	1154	6978	0.057	0.068	**0.018**
WIN	AR	37	34	0.750	0.753	0.805
WIN	BA	260	222	0.339	0.339	0.286
WIN	SD	257	264	0.136	0.137	**0.024**
WIN	VI	154	331	0.960	0.957	0.249

Exact P-values: Take Aways

• Randomization-based p-values underly tests for treatment effects.

• In practice using t-statistic based p-values is often similar to exact p-values based on difference in averages test.

• With very skewed distributions rank-based tests are much better.

• See recent Alwyn Young papers on inference and leverage.

3. Randomized Experiments: Neyman's Repeated Sampling Approach

During the same period in which Fisher was developing his p-value calculations, Jerzey Neyman was focusing on methods for estimating **average** treatment effects.

His approach was to consider an estimator and derive its distribution under repeated sampling by drawing from the randomization distribution of $\mathbf{W}$, the assignment vector.

- $\mathbf{Y}(0)$, $\mathbf{Y}(1)$ still fixed in repeated sampling thought experiment.

3.1 Unbiased Estimation of the Ave Treat Effect

Neyman was interested in the population average treatment effect:

$$\tau = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)) = \overline{Y}(1) - \overline{Y}(0).$$

Suppose that we observed data from a completely randomized experiment in which M units were assigned to treatment and $N - M$ assigned to control. Given randomization, the intuitive estimator for the average treatment effect is the difference in the average outcomes for those assigned to the treatment versus those assigned to the control:

$$\widehat{\tau} = \frac{1}{M}\sum_{i:W_i=1} Y_i^{\text{obs}} - \frac{1}{N-M}\sum_{i:W_i=0} Y_i^{\text{obs}} = \overline{Y}_1^{\text{obs}} - \overline{Y}_0^{\text{obs}}.$$

To see that this measure, $\overline{Y}_1 - \overline{Y}_0$, is an unbiased estimator of τ, consider the statistic

$$T_i = \left(\frac{W_i \cdot Y_i^{\text{obs}}}{M/N} - \frac{(1 - W_i) \cdot Y_i^{\text{obs}}}{(N - M)/N} \right).$$

The average of this statistic over the population is equal to our estimator, $\widehat{\tau} = \sum_i T_i / N = \overline{Y}_1^{\text{obs}} - \overline{Y}_0^{\text{obs}}$:

Using the fact that Y_i^{obs} is equal to $Y_i(1)$ if $W_i = 1$ and $Y_i(0)$ if $W_i = 0$, we can rewrite this statistic as:

$$T_i = \left(\frac{W_i \cdot Y_i(1)}{M/N} - \frac{(1 - W_i) \cdot Y_i(0)}{(N - M)/N} \right).$$

The only element in this statistic that is random is the treatment assignment, W_i, with $\mathbb{E}[1 - W_i] = (1 - \mathbb{E}[W_i])$, is equal to $(N - M)/N$.

Using these results we can show that the expectation of T_i is equal to the unit-level causal effect, $Y_i(1) - Y_i(0)$:

$$\mathbb{E}[T_i] = \left(\frac{\mathbb{E}[W_i] \cdot Y_i(1)}{M/N} - \frac{(1 - \mathbb{E}[W_i]) \cdot Y_i(0)}{(N - M)/N} \right) = Y_i(1) - Y_i(0)$$

3.2 The Variance of the Unbiased Estimator $\overline{Y}_1^{\text{obs}} - \overline{Y}_0^{\text{obs}}$

Neyman was also interested in the variance of this unbiased estimator of the average treatment effect

This involved two steps: first, deriving the variance of the estimator for the average treatment effect; and second, developing unbiased estimators of this variance.

In addition, Neyman sought to create confidence intervals for the population average treatment effect which also requires an appeal to the central limit theorem for large sample normality.

Consider a completely randomized experiment of N units, M assigned to treatment. To calculate the variance of $\overline{Y}_1^{\text{obs}} - \overline{Y}_0^{\text{obs}}$, we need the second and cross moments of the random variable W_i, $\mathbb{E}[W_i^2]$ and $\mathbb{E}[W_i \cdot W_j]$.

$$\mathbb{E}[W_i^2] = \mathbb{E}[W_i] = M/N.$$

$$\mathbb{E}[W_i \cdot W_j] = \text{Pr}(W_i = 1) \cdot \text{Pr}(W_j = 1 | W_i = 1)$$

$$= (M/N) \cdot (M-1)/(N-1) \neq \mathbb{E}[W_i] \cdot \mathbb{E}[W_j],$$

for $i \neq j$, since conditional on $W_i = 1$ there are $M - 1$ treated units remaining out of $N - 1$ total remaining.

The variance of $\overline{Y}_1^{\text{obs}} - \overline{Y}_0^{\text{obs}}$ is equal to:

$$\text{Var}(\overline{Y}_1^{\text{obs}} - \overline{Y}_0^{\text{obs}}) = \frac{S_0^2}{N-M} + \frac{S_1^2}{M} - \frac{S_{01}^2}{N}, \tag{1}$$

where S_w^2 is the variance of $Y_i(w)$ in the population, defined as:

$$S_w^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i(w) - \bar{Y}(w))^2,$$

for $w = 0, 1$, and S_{01}^2 is the population variance of the unit-level treatment effect, defined as:

$$S_{01}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i(1) - Y_i(0) - \tau)^2.$$

The numerator of the first term, the population variance of the potential control outcome vector, $\mathbf{Y}(0)$, is equal to

$$S_0^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i(0) - \overline{Y}(0))^2.$$

An unbiased estimator for S_0^2 is

$$s_0^2 = \frac{1}{N-M-1} \sum_{i:W_i=0} (Y_i^{\text{obs}} - \overline{Y}_0^{\text{obs}})^2.$$

The third term, S^2_{01} (the population variance of the unit-level treatment effect) is more difficult to estimate because we cannot observe both $Y_i(1)$ and $Y_i(0)$ for any unit.

We have no direct observations on the variation in the treatment effect across the population and cannot directly estimate S^2_{01}.

As noted previously, if the treatment effect is additive ($Y_i(1) - Y_i(0) = c$ for all i), then this variance is equal to zero and the third term vanishes.

Under this circumstance we can obtain an unbiased estimator for the variance as:

$$\hat{\mathbb{V}}\left(\overline{Y}_1^{\text{obs}} - \overline{Y}_0^{\text{obs}}\right) = \frac{s_0^2}{N-M} + \frac{s_1^2}{M}. \tag{2}$$

This estimator for the variance is widely used, even when the assumption of an additive treatment effect is inappropriate. There are two main reasons for this estimator's popularity.

First, confidence intervals generated using this estimator of the variance will be *conservative* with actual coverage at least as large, but not necessarily equal to, the nominal coverage.

The second reason for using this estimator for the variance is that it is always unbiased for the variance of $\widehat{\tau} = \overline{Y}_1^{\text{obs}} - \overline{Y}_0^{\text{obs}}$ when this statistic is interpreted as the estimator of the average treatment effect in the super-population from which the N observed units are a random sample. (we return to this interpretation later)

Confidence Intervals

Given the estimator $\widehat{\tau}$ and the variance estimator $\widehat{\mathbb{V}}$, how do we think about confidence intervals?

Let's consider the case where $\mathbb{E}[W_i] = 1/2$, and define

$$D_i = 2W_i - 1, \quad \text{so that} \quad \mathbb{E}[D_i] = 0, \quad D_i^2 = 1.$$

Write

$$\widehat{\tau} = \overline{Y}_1 - \overline{Y}_0 = \frac{1}{N/2} \sum_{i=1}^{N} W_i Y_i(1) - \frac{1}{N/2} \sum_{i=1}^{N} (1 - W_i) Y_i(0)$$

$$= \frac{1}{N} \sum_{i=1}^{N} \Big(Y_i(1) - Y_i(0) \Big) + \frac{1}{N} \sum_{i=1}^{N} D_i \Big(Y_i(1) + Y_i(0) \Big)$$

The stochastic part, normalized by the sample size, is

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} D_i \Big(Y_i(1) + Y_i(0) \Big)$$

It has mean zero and variance

$$\mathbb{V} = \frac{1}{N} \sum_{i=1}^{N} \Big(Y_i(1) + Y_i(0) \Big)^2.$$

Under conditions on the sequence of $\sigma_i^2 = (Y_i(1) + Y_i(0))^2$, we can use a central limit theorem for independent but not identically distributed random variables to get

$$\frac{\frac{1}{\sqrt{N}} \sum_{i=1}^{N} D_i \Big(Y_i(1) + Y_i(0) \Big)}{\sqrt{\frac{1}{N} \sum_{i=1}^{N} \sigma_i^2}} \xrightarrow{d} \mathcal{N}(0, 1)$$

Neyman Repeated Sampling Thought Experiments

- Basis for estimating causal effects
- Finite population argument
- Uncertainty based on assignment mechanism, not sampling.

4. Stratified Randomized Experiments

- Suppose we have N units, we observe some covariates on each unit, and wish to evaluate a binary treatment.

- Should we randomize the full sample, or should we stratify the sample first, or even pair the units up?

Recommendation In Literature:

- In large samples, and if the covariates are strongly associated with the outcomes, definitely stratify or pair.

- In small samples, with weak association between covariates and outcomes, the literature offers mixed advice.

Quotes from the Literature

Snedecor and Cochran (1989, page 101) write, comparing paired randomization and complete randomization:

> "If the criterion [the covariate used for constructing the pairs] has no correlation with the response variable, **a small loss in accuracy results from the pairing due to the adjustment for degrees of freedom.** A substantial loss may even occur if the criterion is badly chosen so that member of a pair are negatively correlated."

Box, Hunter and Hunter (2005, page 93) also suggest that there is a tradeoff in terms of accuracy or variance in the decision to pair, writing:

> "Thus you would gain from the paired design **only** if the reduction in variance from pairing outweighed the effect of the decrease in the number of degrees of freedom of the t distribution."

Klar and Donner (1997) raise additional issues that make them concerned about pairwise randomized experiments (in the context of randomization at the cluster level):

> "We shown in this paper that there are also several **analytic limitations** associated with pair-matched designs. These include: the restriction of prediction models to cluster-level baseline risk factors (for example, cluster size), the inability to test for homogeneity of odds ratios, and difficulties in estimating the intracluster correlation coefficient. These limitations lead us to present arguments that favour stratified designs in which there are more than two clusters in each stratum."

Imai, King and Nall (2009) claim there are no tradeoffs at all between pairing and complete randomization, and summarily dismiss all claims in the literature to the contrary:

> "Claims in the literature about problems with matched-pair cluster randomization designs are misguided: **clusters should be paired prior to randomization** when considered from the perspective of efficiency, power, bias, or robustness."

and then exhort researchers to randomize matched pairs.

> "randomization by cluster without prior construction of matched pairs, when pairing is feasible, is an exercise in selfdestruction."

How Do We Reconcile These Statements?

- Be careful and explicit about goals: precision of estimators versus power of tests.

- Be careful about estimands: population versus sample, average over clusters or average over individuals.

4.1 Expected Squared Error Calculations for Completely Randomized vs Stratified Randomized Experiments

Suppose we have a single binary covariate $X_i \in \{f, m\}$. Define

$$\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \middle| X_i = x\right]$$

where the expectations denote expectations taken over the superpopulation.

The estimand we focus on is the (super-)population version of the the finite sample average treatment effect,

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[\tau(X_i)]$$

Notation

$$\mu(w,x) = \mathbb{E}\left[Y_i(w) \middle| W_i = w, X_i = x\right],$$

$$\sigma^2(w,x) = \mathbb{V}\left(Y_i(w) \middle| W_i = w, X_i = x\right),$$

for $w = 0, 1$, and $x \in \{f, m\}$, and

$$\sigma^2_{01}(x) = \mathbb{E}\left[\left(Y_i(1) - Y_i(0) - (\mu(1,x) - \mu(0,x))\right)^2 \middle| X_i = x\right],$$

Three Estimators: $\widehat{\tau}_{\text{dif}}$, $\widehat{\tau}_{\text{reg}}$, and $\widehat{\tau}_{\text{strata}}$

First, simple difference:

$$\widehat{\tau}_{\text{dif}} = \overline{Y}_1^{\text{obs}} - \overline{Y}_0^{\text{obs}}$$

Second, use the regression function

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + \beta \cdot \mathbf{1}_{X_i=f} + \varepsilon_i.$$

Then estimate τ by least squares regression. This leads to $\widehat{\tau}_{\text{reg}}$.

The third estimator we consider is based on first estimating the average treatment effects within each stratum, and then weighting these by the relative stratum sizes:

$$\widehat{\tau}_{\text{strata}} = \frac{N_{0f} + N_{1l}}{N} \cdot (\overline{Y}_{1f}^{\text{obs}} - \overline{Y}_{0f}^{\text{obs}}) + \frac{N_{0m} + N_{1m}}{N} \cdot (\overline{Y}_{1m}^{\text{obs}} - \overline{Y}_{0m}^{\text{obs}}).$$

Large (infinitely large) superpopulation.

We draw a stratified random sample of size $4N$ from this population, where N is integer. Half the units come from the $X_i = f$ subpopulation, and half come from the $X_i = m$ subpopulation.

Two experimental designs. First, a **completely randomized design** ($\mathcal{C}$) where $2N$ units are randomly assigned to the treatment group, and the remaining $2N$ are assigned to the control group.

Second, a **stratified randomized design** ($\mathcal{S}$) where N are randomly selected from the $X_i = f$ subsample and assigned to the treatment group, and N units are randomly selected from the $X_i = m$ subsample and assigned to the treatment group.

In both designs the conditional probability of a unit being assigned to the treatment group, given the covariate, is the same: $\text{pr}(W_i = 1|X_i) = 1/2$, for both types, $x = f, m$.

$$\mathbb{V}_{\mathcal{S}} = \mathbb{E}\left[(\widehat{\tau}_{\mathsf{dif}} - \tau)^2 \middle| \mathcal{S}\right]$$

$$= \frac{q}{N} \cdot \left(\frac{\sigma^2(1,f)}{p} + \frac{\sigma^2(0,f)}{1-p}\right) + \frac{1-q}{N} \cdot \left(\frac{\sigma^2(1,m)}{p} + \frac{\sigma^2(0,m)}{1-p}\right)$$

$$\mathbb{V}_{\mathcal{C}} = \mathbb{E}\left[(\widehat{\tau}_{\mathsf{dif}} - \tau)^2 \middle| \mathcal{C}\right] = q(1-q)(\mu(0,f) - \mu(0,m))^2$$

$$+\frac{q\sigma^2(0,f)}{(1-p)N} + \frac{(1-q)\sigma^2(0,m)}{(1-p)N}$$

$$+q(1-q)(\mu(1,f) - \mu(1,m))^2 + \frac{q\sigma^2(1,f)}{pN} + \frac{(1-q)\sigma^2(1,m)}{pN}$$

$$\mathbb{V}_{\mathcal{C}} - \mathbb{V}_{\mathcal{S}} =$$

$$q(1-q) \cdot \Big((\mu(0,f) - \mu(0,m))^2 + (\mu(1,f) - \mu(1,m))^2\Big) \geq 0$$

Comment 1:

Stratified randomized design has **lower** expected squared error than completely randomized design.

Strictly lower if the covariate predict potential outcomes in population.

- True irrespective of sample size

Comment 2: For this result it is important that we compare the **marginal** variances, not **conditional** variances. There is no general ranking of the conditional variances

$$\mathbb{E}\left[(\widehat{\tau}_{\text{dif}}-\tau)^2\Big|\,\mathbf{Y}(0),\mathbf{Y}(1),\mathbf{X},\mathcal{C}\right]$$

versus

$$\mathbb{E}\left[(\widehat{\tau}_{\text{dif}}-\tau)^2\Big|\,\mathbf{Y}(0),\mathbf{Y}(1),\mathbf{X},\mathcal{S}\right].$$

It is possible that stratification leads to larger variances because of negative correlations within strata in a finite sample (Snedecor and Cochran quote). That is not possible on average, that is, over repeated samples.

In practice it means that if the primary interest is in the most precise estimate of the average effect of the treatment, **stratification dominates complete randomization**, even in small samples.

Comment 3: Under a stratified design the three estimators $\widehat{\tau}_{\text{reg}}$, $\widehat{\tau}_{\text{strata}}$, and $\widehat{\tau}_{\text{dif}}$ are identical, so their variances are the same.

Under a completely randomized experiment, the estimators are generally different. In sufficiently large samples, if there is some correlation between the outcomes and the covariates that underly the stratification, the regression estimator $\widehat{\tau}_{\text{reg}}$ will have a lower variance than $\widehat{\tau}_{\text{dif}}$.

However, for any fixed sample size, if the correlation is sufficiently weak, the variance of $\widehat{\tau}_{\text{reg}}$ will actually be strictly higher than that of $\widehat{\tau}_{\text{dif}}$.

Think through analyses in advance

Thus for *ex post* adjustment there is a potentially complicated tradeoff: in small samples one should not adjust, and in large samples one should adjust if the objective is to minimize the expected squared error.

If one wishes to adjust for differences in particular covariates, do so by design: randomize in a way such that $\widehat{\tau}_{\text{dif}} = \widehat{\tau}_{\text{reg}}$ (e.g., stratify, or rerandomize).

4.2 Analytic Limitations of Pairwise Randomization

Compare two designs with $4N$ units.

- N strata with 4 units each ($\mathcal{S}$).
- $2N$ pairs with 2 units each ($\mathcal{P}$).

What are costs and benefits of $\mathcal{S}$ versus $\mathcal{P}$?

Benefits of Pairing

- The paired design will lead to lower expected squared error than stratified design in finite samples. (similar argument as before.)

- In sufficiently large sample power of paired design will be higher (but not in very small samples, similar argument as before).

Difference with Stratified Randomized Experiments

- Suppose we have a stratum with size ≥ 4 and conduct a randomized experiment within the stratum with ≥ 2 treated and ≥ 2 controls.

- Within each stratum we can estimate the average effect **and** its variance (and thus intraclass variance). The variance may be imprecisely estimated, but we can estimate it without bias.

- Suppose we have a stratum (that is, a pair) with 2 units. We can estimate the the average effect in each pair (with the difference in outcomes by treatment status), but we **can not** estimate the variance.

Difference with Stratified Randomized Experiments (ctd)

- From data on outcomes and pairs alone we cannot establish whether there is heterogeneity in treatment effects.

- We can establish the presence of heterogeneity if we have data on covariates used to create pairs (compare "similar" pairs).

- Efficiency gains from going from strata with 4 units to strata with 2 units is likely to be small.

Recommendation

- Use small strata, rather than pairs (but not a big deal either way)

- Largely agree with Klar & Donner

4.3 Power Comparisons for t-statistic Based Tests

The basic calculation underlying the concern with pairwise randomization is based on calculation of t-statistics.

Randomly sample N units from a large population. Covariate $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$. We then draw another set of N units, with exactly the same values for the covariates. **Assume covariates are irrelevant.**

The distribution of the potential control outcome is

$$Y_i(0)|X_i = \mathcal{N}(\mu, \sigma^2) \quad \text{and} \quad Y_i(1) = Y_i(0) + \tau$$

Completely randomized design ($\mathcal{C}$): randomly pick N units to receive the treatment.

Pairwise randomized design ($\mathcal{P}$): pair the units by covariate and randomly assign one unit from each pair to the treatment.

The estimator for τ under both designs is

$$\widehat{\tau} = \overline{Y}_1^{\text{obs}} - \overline{Y}_0^{\text{obs}}.$$

Its distribution under the two designs is the same as well (because covariate is independent of outcomes):

$$\widehat{\tau}|\mathcal{C} \sim \mathcal{N}\left(\tau, \frac{2 \cdot \sigma^2}{N}\right) \quad \text{and} \quad \widehat{\tau}|\mathcal{P} \sim \mathcal{N}\left(\tau, \frac{2 \cdot \sigma^2}{N}\right)$$

The natural estimator for the variance for the estimator given the pairwise randomized experiment is

$$\widehat{\mathbb{V}}_{\mathcal{P}} = \frac{1}{N \cdot (N-1)} \sum_{i=1}^{N} (\widehat{\tau}_i - \widehat{\tau})^2 \sim \frac{2 \cdot \sigma^2}{N} \cdot \frac{\mathcal{X}^2(N-1)}{N-1}$$

The variance estimator for the completely randomized design, exploiting homoskedasticity, is

$$\widehat{\mathbb{V}}_{\mathcal{C}} = \frac{2}{N} \left(\frac{(N-1) \cdot s^2(0) + (N-1) \cdot s^2(1)}{2N-2} \right) \sim \frac{2 \cdot \sigma^2}{N} \cdot \frac{\mathcal{X}^2(2 \cdot N - 2)}{2 \cdot N - 2}$$

Under the normality the expected values of the varianc estimators are the same

$$\mathbb{E}\left[\widehat{\mathbb{V}}_{\mathcal{P}}\right] = \mathbb{E}\left[\widehat{\mathbb{V}}_{\mathcal{C}}\right] = \frac{2 \cdot \sigma^2}{N}$$

but their variances differ:

$$\mathbb{V}\left(\widehat{\mathbb{V}}_{\mathcal{P}}\right) = 2 \cdot \mathbb{V}\left(\widehat{\mathbb{V}}_{\mathcal{C}}\right) = \frac{8 \cdot \sigma^4}{N^2 \cdot (N-1)}$$

This leads to the t-statistics

$$t_{\mathcal{P}} = \frac{\widehat{\tau}}{\sqrt{\widehat{\mathbb{V}}_{\mathcal{P}}}}, \quad \text{and} \quad t_{\mathcal{C}} = \frac{\widehat{\tau}}{\sqrt{\widehat{\mathbb{V}}_{\mathcal{C}}}}.$$

If we wish to test the null hypothesis of $\tau = 0$ against the alternative of $\tau \neq 0$ at level α, we would reject the null hypothesis if $|t|$ exceeds the critical value c_α (different for the two designs)

$$c_\alpha^{\mathcal{P}} = q_{1-\alpha/2}^t(N-1), \quad c_\alpha^{\mathcal{C}} = q_{1-\alpha/2}^t(2N-2)$$

For any $\tau \neq 0$, and for any $N \geq 2$ the power of the test based on the t-statistic $t_{\mathcal{C}}$ is strictly greater than the power based on the t-statistic $t_{\mathcal{P}}$. (assuming covariates are irrelevant.)

(at $N = 1$ we cannot test the hypothesis without knowledge of the variances)

By extension, the power for the test based on the completely randomized design is still greater than the power based on the pairwise randomized experiment if the association between the covariate and the potential outcomes is weak, at least in small samples.

This is the formal argument against doing a pairwise (or by extension) a stratified randomized experiment if the covariates are only weakly associated with the potential outcomes.

Limitations

- Test comparison relies on normality. Without normality we cannot directly rank the power, and the actual size of the tests need not even be equal to the nominal size.

- Homoskedastic case is most favorable to completely randomized experiment (but features most often in power comparisons). In the case of heteroskedasticity, the loss in power for pairwise randomized experiment is less.

Conclusion

• Stratify, with small strata, but at least two treated and two control units.

• Dont worry about power, use variance estimator that takes into account stratification.

Causal Inference

and Machine Learning

Guido Imbens – Stanford University

Lecture 2:

Introduction to Machine Learning Concepts

Potsdam Center for Quantitative Research

Monday September 9th, 16.30-18.00

Outline

1. Nonparametric Regression
2. Regression Trees
3. Multiple Covariates/Features
4. Pruning
5. Random Forests
6. Boosting

7. Neural Nets

8. Generative Adverserial Nets

1. Nonparametric Regression

Data:

$$(X_i, Y_i), \quad i = 1, \ldots, N, \text{ i.i.d.}$$

where $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$, or $Y_i \in \{0, 1\}$

Define

$$g(x) = \mathbb{E}[Y_i | X_i = x]$$

Goal: estimate $g(x)$, minimize

$$\mathbb{E}\left[(g(X_i) - \widehat{g}(X_i))^2\right]$$

The regression/prediction problem is special:

Suppose we put one randomly chosen observation aside: (Y_i, X_i), and use the rest of the sample to estimate $g(\cdot)$ as $\widehat{g}_{(i)}(\cdot)$.

Then we can assess the quality of the estimator by calculating the squared error

$$\left(Y_i - \widehat{g}_{(i)}(X_i)\right)^2$$

We can use this out-of-sample **cross-validation** to rank different estimators $\widehat{g}_1(\cdot)$ and $\widehat{g}_2(\cdot)$.

Not true directly for estimators of average causal effects, or when we want to estimate the regression function at a point, $g(x_0)$.

Many methods satisfy:

$$\widehat{g}(x) = \sum_{i=1}^{N} \omega_i Y_i, \qquad \text{often with } \sum_{i=1}^{N} \omega_i = 1, \quad \text{sometimes } \omega_i \geq 0.$$

- Question: how to choose the weights ω_i?

- Is it important or not to do inference on $g(\cdot)$ (confidence intervals / standard errors)?

- How well do estimators perform in terms of **out-of-sample** mean squared error (as opposed to **in-sample** fit)?

- What to do if dim(X_i) is high (relative to N)?

First: scalar case, $X_i \in [0,1]$

1. Define **knots** $\kappa_{jJ} = j/J$, for $j = 1, \ldots, J-1$, $J = 2, 3, \ldots$, and

$$\widehat{g}_J(x) = \sum_{j=1}^{N} \mathbf{1}_{x \in [\kappa_{j-1J}, \kappa_{jJ}]} \widehat{c}_{jJ}$$

where $\widehat{c}_{jJ}$ is the average outcome within the interval $[\kappa_{j-1J}, \kappa_{jJ}]$:

$$\widehat{c}_{jJ} = \sum_{j=1}^{N} \mathbf{1}_{X_i \in [\kappa_{j-1J}, \kappa_{jJ}]} Y_i \Big/ \sum_{j=1}^{N} \mathbf{1}_{X_i \in [\kappa_{j-1J}, \kappa_{jJ}]}$$

Also define number of observations in each interval:

$$N_{jJ} = \sum_{j=1}^{N} \mathbf{1}_{X_i \in [\kappa_{j-1J}, \kappa_{jJ}]}$$

For fixed x the bias of this estimator depends on the derivative of $g(\cdot)$ around x, and the density of of X_i around x. Given some smoothness, the bias-squared is approximately equal to to the square of $g'(x)/J$ and the variance is equal to $V(Y_i|X_i = x)/(Nf(x)/J)$.

So as a function of the number of intervals J:

$$\text{Bias}^2(J) + \text{Var}(J) = \frac{g'(x)^2}{J^2} + \frac{JV(Y_i|X_i = x)}{Nf(x)}$$

Optimal choice for J is

$$J^{\text{opt}} = N^{1/3} \left(\frac{2g'(x)^2}{V(Y_i|X_i = x)} \right)^{1/3}$$

If we let J increase slightly slower than proportional to $N^{1/3}$ we get asymptotic normality without bias, without under-smoothing no valid confidence intervals.

(You can do better than this by centering the interval at x, which lowers the bias, and then the optimal rate is $J \propto N^{1/5}$.)

How can we modify this to improve properties?

1. Use more sophisticated ways of averaging:

   (a) Use weights that give more weight to nearby observations (kernel estimation)

   (b) Instead of using means within intervals use polynomial approximation within interval (e.g., local linear regression, splines)

2. Choose knots in data dependent way, but need to give up easy asymptotic properties (regression trees)

2. Regression Trees

Define for a, b the set of indices such that X_i is in $[a, b)$:

$$I_{a,b} = \{i = 1, \ldots, N | X_i \in [a, b)\}$$

Define the average within an interval:

$$\overline{Y}_{a,b} = \sum_{i \in I_{a,b}} Y_i \Big/ \sum_{i \in I_{a,b}} 1$$

Define the sum of squared deviations from means:

$$Q(x) = \sum_{i: X_i \in I_{0,x}} \left(Y_i - \overline{Y}_{0,x}\right)^2 + \sum_{i: X_i \in I_{x,1}} \left(Y_i - \overline{Y}_{x,1}\right)^2$$

Find the split point that minimizes the sum of squared deviations:

$$c_1 = \arg \min_{x \in [0,1]} Q(x)$$

Then:

$$\widehat{g}(x) = \begin{cases} \overline{Y}_{0,c_1} & \text{if} x \leq c_1, \\ \overline{Y}_{c_1,1} & \text{if} x > c_1, \end{cases}$$

- This is a tree with two leaves: $[0, c_1]$ and $[c_1, 1]$.

- $\widehat{g}(\cdot)$ is step function.

We can do this again: consider all possible split points $c_2 \in [0,1]$ and calculate the sum of squares as a function of the additional split point. For example, for $c_2 \in (c_1, 1)$, the sum of squares is

$$Q(c_1, c_2) = \sum_{i:X_i \in I_{0,c_1}} \left(Y_i - \overline{Y}_{0,c_1}\right)^2 + \sum_{i:X_i \in I_{c_1,c_2}} \left(Y_i - \overline{Y}_{c_1,c_2}\right)^2$$

$$+ \sum_{i:X_i \in I_{c_2,1}} \left(Y_i - \overline{Y}_{c_2,1}\right)^2$$

Now we have a tree with three leaves. $\widehat{g}(\cdot)$ is still a step function:

$$\widehat{g}(x) = \begin{cases} \overline{Y}_{0,c_1} & \text{if} x \in [0, c_1), \\ \overline{Y}_{c_1,c_2} & \text{if} x \in [c_1, c_2) \\ \overline{Y}_{c_2,1} & \text{if} x \in [c_2, 1] \end{cases}$$

Note:

• We can keep doing this, each time adding a leaf to the tree.

• For every new potential split the sum of squares is lower than what it is without the additional split, until we have only the same value of X_i within each interval.

1. Given J splits, this looks very similar to just dividing the interval $[0, 1]$ into J equal subintervals.

2. It is more **adaptive**: it will be more likely to divide for values of x where

   (a) there are more observations (where the variance is smaller – nearest neighbor estimators also do that)

   (b) the derivative of $g(x)$ is larger (where the bias is bigger)

In both cases (simple dividing $[0, 1]$ into J equal intervals, or tree with J leaves), we need to choose the smoothing parameter J.

• **leave-one-out cross-validation**: leave out observation i, re-estimate model with J pieces/leaves, predict Y_i as $\widehat{g}_{J,(i)}(X_i)$, and calculate error $Y_i - \widehat{g}_{J,(i)}(X_i)$.

Minimize over J:

$$CV(J) = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \widehat{g}_{J,(i)}(X_i) \right)^2$$

To make this computationally easier, do 10-fold cross-validation: partition sample into ten subsamples, and estimate 10 times on the samples of size $N \times 0.9$ and validate on 10% samples.

This is how cross-validation is often done for kernel and nearest neighbor type regression estimators. Note: this means bias-squared and variance are balanced, and so confidence intervals are **not** valid.

Cross-validation is not implemented this way for regression trees, partly for computational reasons, and partly because this is not necessarily unimodal in J.

Instead the criterion is, to choose tree $\mathcal{T}$ that minimizes the sum of squared deviations **plus a penalty term**, typically a constant times the number of leaves in the tree:

$$Q(\mathcal{T}) + \lambda|\mathcal{T}|$$

Now the penalty parameter λ is choosen through cross-validation, say 10-fold cross-validation.

3. Multiple Covariates

Suppose we have multiple covariates or features, say $x = (x_1, x_2, \ldots, x_p) \in [0,1]^p$.

Suppose X_i has uniform distribution.

Suppose we want to estimate $\mathbb{E}[Y_i | X_i = (0, 0, \ldots, 0)]$ by average over nearby observations:

$$\widehat{g}(0) = \sum_{i=1}^{N} Y_i \mathbf{1}_{X_{ik} \leq \varepsilon} \Big/ \sum_{i=1}^{N} \mathbf{1}_{X_{ik} \leq \varepsilon}.$$

Problem is that the number of observations close by,

$$\mathbb{E}\left[\sum_{i=1}^{N} \mathbf{1}_{X_{ik} \leq \varepsilon}\right] = N\varepsilon^p \quad \textbf{curse of dimensionality}$$

For kernel methods we typically use multivariate kernels that are simply the product of univariate kernels:

$$K(x_1, x_2) = K_0(x_1) \times K_0(x_2),$$

possibly with different bandwidths, but similar rates for the different bandwidths.

This works poorly in high dimensions - rate of convergence declines rapidly with the dimension of X_i.

Trees deal with multiple covariates differently.

Now, for the first split, we consider all subsets of $[0,1] \times [0,1]$ of the form

$$[0,c) \times [0,1], \quad \text{split on } x_1$$

or

$$[0,1] \times [0,c), \quad \text{split on } x_2$$

Repeat this after the first split.

• This means that some covariates may never be used to split the sample - the method will deal better with cases where the regression function is flat in some covariates (sparsity).

• It can deal with high dimensional covariates, as long as the regression function does not depend too much on too many of them. (will not perform uniformly well, but well in important parts of parameter space)

This difference in the way trees (and forests) deal with multiple covariates compared to kernel methods is important in practice. There is some tension there:

• for asymptotic properties (focus of much of econometrics literature) it is key that eventually the leaves are small in all dimensions. Kernel type methods do this automatically. With trees and forests it can be imposed by forcing the splits to depend on any covariate with probability bounded away from zero (or even equal probability).

• But for finite sample properties with many covariates (focus of much of machine learning literature) you dont want to split very often on covariates that do not matter much.

Comparison with linear additive models:

- Trees allow for complex nonlinearity and non-monotonicity.

- With social science data conditional expectations are often **monotone**, so linear additive models may provide good fit. If conditional mean of Y_i is increasing in X_{i2} given $X_{i1} < c$, it is likely to be increasing in X_{i2} given $X_{i1} \geq c$. Trees do not exploit this. You could do linear models within leaves, but then need to be careful with many covariates.

4. Pruning

If we grow a tree as just described, we may stop too early and miss important features of the joint distribution.

Suppose $(x_1, x_2) \in \{(-1,-1),(-1,1),(1,-1),1,1)\}$, and

$$g(x_1, x_2) = x_1 \times x_2$$

No first split (either on x_1 or on x_2) improves the expected squared error compared to no-split, but two or three splits improve the expected squared error substantially.

How do we get there if the first split delivers no benefit?

• First grow a "big" tree, with many leaves, even if they do not improve the sum of squared errors enough given λ, up to the point that the leaves are all very small in terms of the number of observations per leave.

• Then "prune" the tree: consider dropping splits (and combining all the subsequent leaves) to see if that improves the criterion function.

5. Random Forests

Trees are step function approximations to the true regression function. They are not smooth, and a single observation may affect the tree substantially. We may want smoother estimates, and ones that are more robust to single observations.

Random forests achieve this by introducing two modifications that introduce randomness in the trees.

Random Forests

1. Create B trees based on bootstrap samples. Start by constructing a bootstrap sample of size N from the original sample. Grow a tree on the bootstrap sample (this part is known as **bagging**), and leads to smoother estimates.

2. For each split (in each tree) only a subset of size m of the p covariates are considered in the split (typically $m = \sqrt{p}$, or $m = p/3$ - heuristic, no formal result).

3. Average estimates $\widehat{g}_b(\cdot)$ for each of the B bootstrap sample based trees.

Flexible, simple and effective out-of-the-box method in many cases. Not a lot of tuning to be done.

6. Gradient Boosting

Initial estimate $\widehat{G}_0(x) = 0$.

First estimate $g(\cdot)$ using a very simple method (a simple **base learner**). For example, a tree with a single split on $(Y_i - \widehat{G}_0(X_i), X_i)$. Call this estimate $\widehat{g}_1(x)$, and define $\widehat{G}_1(x) = \widehat{G}_0(x) + \widehat{g}_1(x)$

Then calculate the residual $\widehat{\varepsilon}_{1i} = Y_i - \widehat{G}_1(X_i)$.

Apply the same simply method again to $\widehat{\varepsilon}_{1i}$, with estimator $\widehat{g}_2(x)$. The estimator for $g(x)$ is now $\widehat{G}_2(x) = \widehat{G}_1(x) + \widehat{g}_2(x)$.

Apply the same simply method again to $\widehat{\varepsilon}_{2i} = Y_i - \widehat{G}_2(X_i)$, with estimator $\widehat{g}_3(x)$. The estimator for $g(x)$ is now $\widehat{G}_3(x) = \widehat{G}_2(x) + \widehat{g}_3(x)$.

What does this do?

Each $\widehat{g}_k(x)$ depends only on a single element of x (single covariate/feature).

Hence $\widehat{g}(x)$ is always an additive function of $x_1, \ldots, x_p$.

What if we want the approximation to allow for some but not all higher order interactions?

If we want only first order interactions, we can use a base learner that allows for two splits. Then the approximation allows for the sum of general functions of two variables, but not more.

Boosting refers to the repeated use of a simple basic estimation method, repeatedly applied to the residuals.

Can use methods other than trees as base learners.

For each split, we can calculate the improvement in mean squared error, and assign that to the variable that we split on.

Sum this up over all splits, and over all trees.

This is informative about the importance of the different variables in the prediction.

Modification

Three tuning parameters: number of trees B, depth of trees d, and shrinkage factor $\varepsilon \in (0, 1]$.

Initial estimate $\widehat{G}_0(x) = 0$, for all x.

First grow tree of depth d on $(Y_i - \widehat{G}_0(X_i), X_i)$, call this $\widehat{g}_1(x)$.

New estimate: $\widehat{G}_1(x) = \widehat{G}_0(x) + \varepsilon \widehat{g}_1(x)$.

Next, grow tree of depth d on $(Y_i - \widehat{G}_b(X_i), X_i)$, call this $\widehat{g}_{b+1}(x)$.

$\varepsilon = 1$ is regular boosting. $\varepsilon < 1$ slows down learning, spreads importance around more variables.

Generalized Boosting

We can do this in more general settings. Suppose we are interested in estimating a binary response model, with a high-dimensional covariate. Start again with

$\widehat{G}_0(x) = 0$ specify parametric model $g(x;\gamma)$

Minimize over γ : $\sum_{i=1}^{N} L(Y_i, \widehat{G}_0(X_i) + g(X_i;\gamma))$

and update $\widehat{G}_{k=1}(x) = \widehat{G}_k(x) + \varepsilon g(x;\widehat{\gamma})$

$L(\cdot)$ could be log likelihood with g log odds ratio:

$$L(y,g) = y\,(g - \ln(1+\exp(g))) - (1-y)\ln(1+\exp(g))$$

7. Neural Nets

Scalar outcome Y_i, p-dimsional vector of covariates/inputs/features X_i, j^{th} element equal to X_{ij}.

Let's focus on the case where Y_i is ordered (not discrete unordered).

Interest in conditional mean

$$h(x) = \mathbb{E}[Y_i | X_i = x]$$

Linear Model:

$$f(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

OLS estimator: minimize

$$\sum_{i=1}^{N} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2$$

- Does not work well if p large relative to N.
- Restrictive if $p << N$

Let's make this more flexible:

Single index model:

$$h(x) = g\left(\sum_{j=1}^{p} \beta_j x_j\right)$$

Estimate p parameters β_j and single function $g(\cdot)$

Additive model:

$$h(x) = \sum_{j=1}^{p} g_j(x_j)$$

Estimate p functions $g_j(\cdot)$

Projection Pursuit:

$$h(x) = \sum_{l=1}^{L} g_l \left(\sum_{j=1}^{p} \beta_{lj} x_j \right)$$

Estimate L functions $g_l(\cdot)$ and $L \times p$ parameters β_{lj}

Neural net with **single hidden layer**:

$$f(x) = \beta_0^{(2)} + \sum_{l=1}^{L} \beta_l^{(2)} g\left(\sum_{j=1}^{p} \beta_{lj}^{(1)} x_j\right)$$

Fix $g(\cdot)$ and estimate $L \times p$ parameters $\beta_{lj}^{(1)}$ and $L+1$ parameters β_l^2.

(Note that (1) and (2) index **layers**, does not mean "to the power of")

General neural net with K **hidden** layers, one **observed input** layer, one **observed output** layer, $K + 2$ layers total.

Observe y and $x_1, \ldots, x_p$. $z_k^{(m)}$ are hidden variables.

Layer k: p_k input variables, p_{k+1} output variables. $p_1 = p$, $p_{K+2} = 1$.

K and p_k, $k = 2, \ldots, K + 1$ are tuning parameters.

First layer: p_2 hidden elements, $l = 1, \ldots, p_2$. Model:

$$z_l^{(2)} = \omega_{l0}^{(1)} + \sum_{j=1}^{p} \omega_{lj}^{(1)} x_j,$$

Transformation of output variables:

$$\alpha_l^{(2)} = g\left(z_l^{(2)}\right)$$

Layer k, p_{k+1} hidden elements, p_k hidden inputs, for layer $k = 2, \ldots, K+1$, $l = 1, \ldots, p_k$. Model:

$$z_l^{(k+1)} = \omega_{l0}^{(k)} + \sum_{j=1}^{p_k} \omega_{lj}^{(k)} \alpha_j^{(k)}$$

Transformation:

$$\alpha_l^{(k+1)} = g\left(z_l^{(k)}\right)$$

Final layer (layer $K+2$), output layer, with single, observed output variable, $p_{K+2} = 1$. Model:

$$y = \omega_0^{(K+1)} + \sum_{j=1}^{p_{K+1}} \omega_j^{(K+1)} \alpha_j^{K+1}$$

Naive approach: minimize

$$\sum_{i=1}^{N} (Y_i - f(X_i; \omega))^2$$

This is badly behaved. Multiple solutions, numerically unstable.

Instead, **regularize**, and minimize:

$$\sum_{i=1}^{N} (Y_i - f(X_i; \omega))^2 + \lambda \sum_{k=1}^{K} \sum_{j=1}^{p_k} \sum_{l=1}^{p_{k-1}} \left(\omega_{jl}^{(k)}\right)^2$$

over all $\omega_{lj}^{(k)}$.

Choose penalty factor λ through cross-validation.

Common choices for transformation $g(\cdot)$ (pre-specified, not choosen by optimization):

1. sigmoid: $g(a) = (1 + \exp(-a))^{-1}$

2. tanh: $g(a) = (\exp(a) - \exp(-a))/(\exp(a) + \exp(-a))$

3. rectified linear $g(a) = a\mathbf{1}_{a>0}$

4. leaky rectified linear $g(a) = a(\mathbf{1}_{a>0} + \gamma\mathbf{1}_{a<0})$

Important to have nonlinearity in the transformation, but exact nature of nonlinearity appears to be less important.

Lost of complexity allowed for in neural nets, but comes with lots of choices.

Not easy to use out-of-the-box, but very successful in complex settings.

Computationally tricky because of multi-modality.

• can approximate smooth functions accurately (**universal approximator**) with many layers and many hidden units.

Interpretation

We can think of the layers up to the last one as constructing regressors: $z^{(K+1)} = h(\omega, x)$

Alternative is to choose functions of regressors, e.g., polynomials $z_{ij} = x_{i1} \times x_{i4} \times x_{i7}^2$.

In what sense is this better? Is this a statement about the type of functions we encounter?

Multiple layers versus multiple hidden units

"We observe that shallow models [models with few layers] in this context overfit at around 20 millions parameters while deep ones can benefit from having over 60 million. This suggests that using a deep model expresses a useful preference over the space of functions the model can learn."

Goodfellow, Bengio, and Courville, *Deep Learning*

Convolutional Neural Nets

Recall model for layer k:

$$z_l^{(k+1)} = \omega_{l0}^{(k)} + \sum_{j=1}^{p_k} \omega_{lj}^{(k)} \alpha_j^{(k)}$$

We can set some of the $\omega_{lj}^{(k)}$ equal to zero. This obviously simplifies computations and makes estimation more precise. But, how do we choose restrictions?

Example Digit recognition: x_j are black/white scale measures on pixels. Suppose we have 16 by 16 pixels, 256 total. So, x_{ij}, $i = 1, \ldots, 16$, $j = 1, \ldots, 16$. We could make the nodes in the first hidden layer functions of only sets of pixels close together:

$$z_1^{(2)} = \omega_{l0}^{(1)} + \sum_{i=1}^{3} \sum_{j=1}^{3} \omega_{lij}^{(1)} x_{ij}$$

$$z_2^{(2)} = \omega_{20}^{(1)} + \sum_{i=4}^{6} \sum_{j=1}^{3} \omega_{2ij}^{(1)} x_{ij}$$

et cetera.

Estimating the Parameters of a Neural Network: Back-propagation

Define objective function (without regularization), for single observation:

$$J_i(\omega, x, y) = (y_i - f(x_i; \omega))^2$$

For N observations:

$$J(\omega, \mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} J_i(\omega, x_i, y_i) = \sum_{i=1}^{N} (y_i - f(x_i; \omega))^2$$

We wish to minimize this over ω.

Recall: K hidden layers, one input and output layer, $K+2$ layers total.
First layer: p_1 observed elements,

$$z_l^{(1)} = x_l, \quad \alpha_l^{(1)} = g^{(1)}\left(z_l^{(1)}\right) = z_l^{(1)} \quad l = 1, \ldots, p_1$$

Hidden layer k, p_k hidden elements, for $k = 2, \ldots, K+1$,

$$z_l^{(k)} = \omega_{l0}^{(k-1)} + \sum_{j=1}^{p_{k-1}} \omega_{lj}^{(k-1)} \alpha_j^{(k-1)}, \; \alpha_l^{(k)} = g^{(k)}\left(z_l^{(k)}\right) = g\left(z_l^{(k)}\right)$$

Final layer (layer $K+2$) with $p_{K+2} = 1$:

$$z^{(K+2)} = \omega_0^{(K+1)} + \sum_{j=1}^{p_{K+1}} \omega_j^{(K+1)} \alpha_j^{K+1},$$

$$f(x) = g^{(K+2)}\left(z^{(K+2)}\right) = z^{(K+2)}$$

We can write

$$J(\omega, \mathbf{x}, \mathbf{y}) = J\left(\omega^{(1)}, \omega^{(2)}, \ldots, \omega^{(K+1)}, \mathbf{x}, \mathbf{y}\right)$$

We can also write the vector

$$\mathbf{z}^{(k)} = h^{(k)}\left(\omega^{(k-1)}, \mathbf{z}^{(k-1)}\right)$$

This function $h^{(k)}(\cdot)$ does not depend on ω beyond $\omega^{(k-1)}$. By further substitution we can write this as

$$\mathbf{z}^{(k)} = \tilde{h}^{(k)}\left(\omega^{(k-1)}, \ldots, \omega^{(1)}, \mathbf{x}\right)$$

Now start with the last layer.

Write

$$f(x) = g^{(K+2)}\left(z^{(K+2)}\right) = z^{(K+2)}$$

Define

$$\delta_i^{K+2} = \frac{\partial}{\partial z_i^{(K+2)}}\left(y_i - g^{(K+2)}\left(z_i^{(K+2)}\right)\right)^2$$

$$= -2\left(y_i - g\left(z_i^{(K+2)}\right)\right) g^{(K+2)\prime}\left(z_{il}^{(K+2)}\right)$$

$$= -2\left(y_i - z_i^{(K+2)}\right)$$

(this is just the scaled residual).

We can write:

$$z_i^{(K+2)} = \omega_0^{(K+1)} + \sum_{j=1}^{p_{K+1}} \omega_j^{(K+1)} g\left(z_{ji}^{(K+1)}\right)$$

so

$$\frac{\partial}{\partial z_{ji}^{(K+1)}} z_i^{(K+2)} = \omega_j^{(K+1)} g'\left(z_{ji}^{(K+1)}\right)$$

Now consider writing the objective function in terms of $z_i^{(K+1)}$:

$$(y_i - f(x_i))^2 = \left(y_i - z_i^{(K+2)}\right)^2$$

$$= \left(y_i - \omega_0^{(K+1)} - \sum_{j=1}^{p_{K+1}} \omega_j^{(K+1)} g\left(z_{ji}^{(K+1)}\right)\right)^2$$

Then define:

$$\delta_{li}^{K+1} = \frac{\partial}{\partial z_{li}^{(K+1)}} \left(y_i - f(x_i)\right)^2$$

$$= \frac{\partial}{\partial z_i^{(K+2)}} \left(y_i - f(x_i)\right)^2 \times \frac{\partial}{\partial z_i^{(K+1)}} z_i^{(K+1)}$$

$$= \delta_i^{(K+2)} \omega_l^{(K+1)} g' \left(z_{li}^{(K+1)}\right)$$

Go down the layers:

Define

$$\delta_{li}^{(k)} = \left(\sum_{j=1}^{p_k} \omega_{jl}^{(k)} \delta_{li}^{(k+1)} \right) g' \left(z_{li}^{(k)} \right)$$

Then the derivatives are

$$\frac{\partial}{\partial \omega_{lj}^{(k)}} J_i(\omega, x_i, y_i) = g \left(z_{li}^{(k)} \right) \delta_{li}^{(k+1)}$$

Given the derivatives, iterate over

$$\omega_{m+1} = \omega_m - \alpha \times \frac{\partial}{\partial \omega} J(\omega, \mathbf{x}, \mathbf{y})$$

α is the "learning rate" often set at 0.01.

• often **stochastic gradient descent**: Instead of calculating

$$\frac{\partial}{\partial \omega} J(\omega, \mathbf{x}, \mathbf{y})$$

use

$$\sum_{i=1}^{N} R_i \frac{\partial}{\partial \omega} J(\omega, x_i, y_i) \Big/ \sum_{i=1}^{N} R_i$$

for a random selection of units ($R_i \in \{0, 1\}$, eg, $\overline{R} = 0.01$) because it is faster.

Regularization:

- add penalty term, $\sum_{jlk} \left(\omega_{jl}^{(k)}\right)^2$

- early stopping rule: stop iterating when test error deteriorates.

8. Generative Adverserial Nets (GANs)

Given data set X_i, $i = 1, \ldots, N$

- Generate data that are indistinguishable from real data
- Use two stage procedure:
  - Generate artificial data
  - Use classifier/discriminator/critic to see if it is possible to distinguish between real and artificial data
- Successful in generating fake pictures

General set up:

- Real observations $X_1, \ldots, X_{N_R}$, with empirical distribution $\hat{F}_X(\cdot)$, in $\mathbb{X}$
- Noise distribution $F_Z(\cdot)$, e.g., multivariate normal, in $\mathbb{Z}$.
- Generator $g_\theta : \mathbb{Z} \mapsto \mathbb{X}$, can be quite flexible.
- Discriminator/critic to tell fake and real data apart, can be quite flexible.

Goal is to find θ so that $g_\theta(Z) \sim \hat{F}_X$ according to discriminator/critic.

Distances and Divergences Consider two dist $f(\cdot)$ and $g(\cdot)$

- Kullback-Leibler

$$KL(f,g) = \int \ln\left(\frac{f(x)}{g(x)}\right) f(x) d\mu(x)$$

- Jannson-Shannon

$$JS(f,g) = KL\left(f, \frac{f+g}{2}\right) + KL\left(g, \frac{f+g}{2}\right)$$

- Earth-Mover / Wasserstein Distance

$$W(f,g) = \inf_{\gamma \in \Pi(f,g)} \mathbb{E}_{(X,Y)\sim\gamma} \|Y - X\|$$

where $\Pi(f, g)$ is set of joint distrs with marginals f and g.

Original GAN proposal (Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, & Bengio, 2014)

KL divergence with discriminator $D_\phi : \mathbb{X} \mapsto [0, 1]$,

$$\inf_\theta \sup_\phi \left\{ \sum_{i:Y_i=F} \ln D_\phi(g_\theta(Z_i)) + \sum_{i:Y_i=R} \ln(1 - D_\phi(X_i)) \right\}$$

Awkward if support of X_i and $g_\theta(Z_i)$ do not agree.

Wasserstein GAN (WGAN, Arjovsky, Chintala, & Bottou, 2017):

WGAN uses Earth Mover distance, through a function $f_\phi : \mathbb{X} \to \mathbb{R}$, parametrized by ϕ, called critic.

• Find a function $f_\phi(x)$ so that the difference between the expected value of of $f_\phi(X_i)$ and $f_\phi(g_\theta(Z_i))$ is maximized.

• Then choose θ to minimize this difference.

Formally:

$$\inf_{\theta} \sup_{\phi,\ \|f_\phi\|_L \leq 1} \left\{ \frac{1}{N_F} \sum_{i:Y_i=F} f_\phi(g_\theta(Z_i)) - \sum_{i:Y_i=R} \frac{1}{N_R} f_\phi(X_i) \right\}$$

(where $\|f\|_L$ denotes the Lipschitz constant of f)

Lalonde Data: Summary Statistics

	NSW Treated		NSW Controls		CPS Controls	
	mean	(s.d.)	mean	(s.d.)	mean	(s.d.)
Black	0.84	(0.36)	0.83	(0.38)	0.07	(0.26)
Hisp	0.06	(0.24)	0.11	(0.31)	0.07	(0.26)
Age	25.8	(7.2)	25.1	(7.1)	33.2	(11.0)
Married	0.19	(0.39)	0.15	(0.36)	0.71	(0.45)
Nodegree	0.71	(0.46)	0.83	(0.37)	0.30	(0.46)
Education	10.3	(2.0)	10.1	(1.6)	12.0	(2.9)
E'74	2.10	(4.89)	2.11	(5.69)	14.02	(9.57)
U'74	0.71	(0.46)	0.75	(0.43)	0.12	(0.32)
E'75	1.53	(3.22)	1.27	(3.10)	13.65	(9.27)
U'74	0.60	(0.49)	0.68	(0.47)	0.11	(0.31)
E'78	6.35	(7.87)	4.55	(5.48)	14.85	(9.65)
U'78	0.24	(0.43)	0.35	(0.48)	0.16	(0.34)

Simulating the Lalonde Data

For the generator we use 11-dimensional normally distributed noise.

Three hidden layers:

1. 11 inputs, 64 outputs, rectified linear

2. 64 inputs, 128 outputs, rectified linear

3. 138 inputs, 256 outputs, rectified linear

Final layer, 256 inputs, 11 outputs: For binary variables, use sigmoid, for censored variables use rectified linear, for continuous variables use linear.

For the discriminator, three hidden layers

1. 11 inputs, 256 outputs, rectified linear

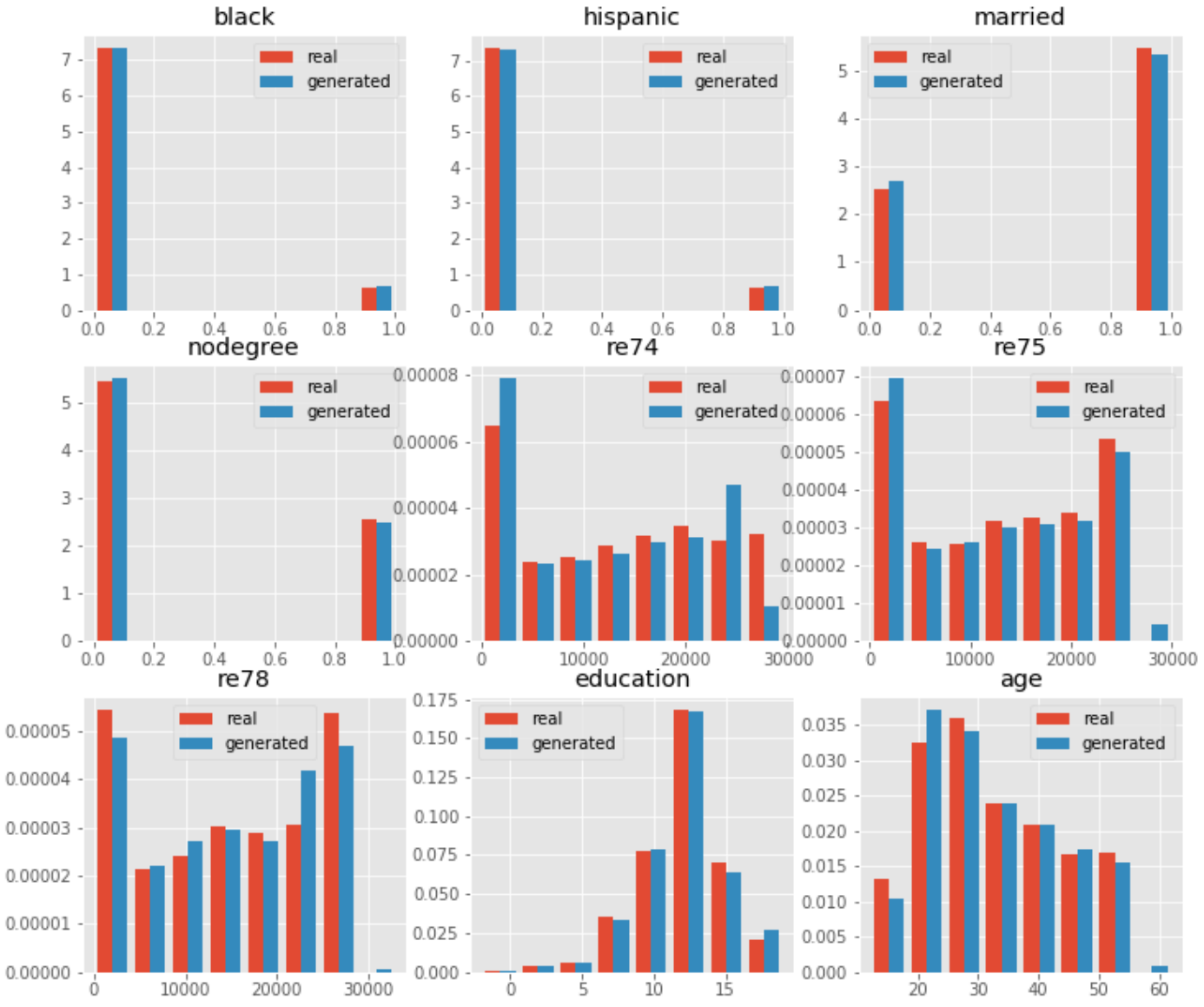
2. 256 inputs, 128 outputs, rectified linear

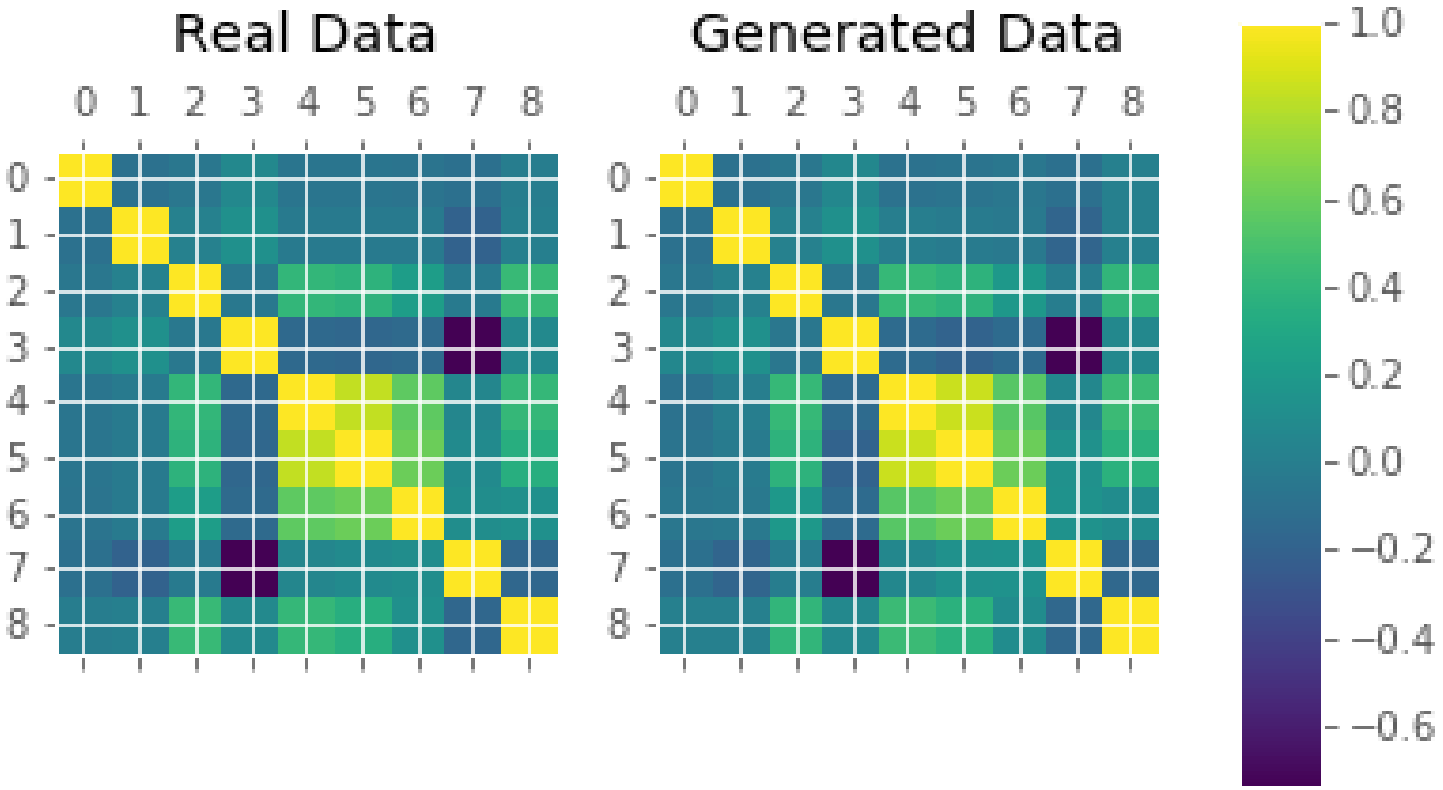
3. 128 intputs, 64 outputs, rectified linear

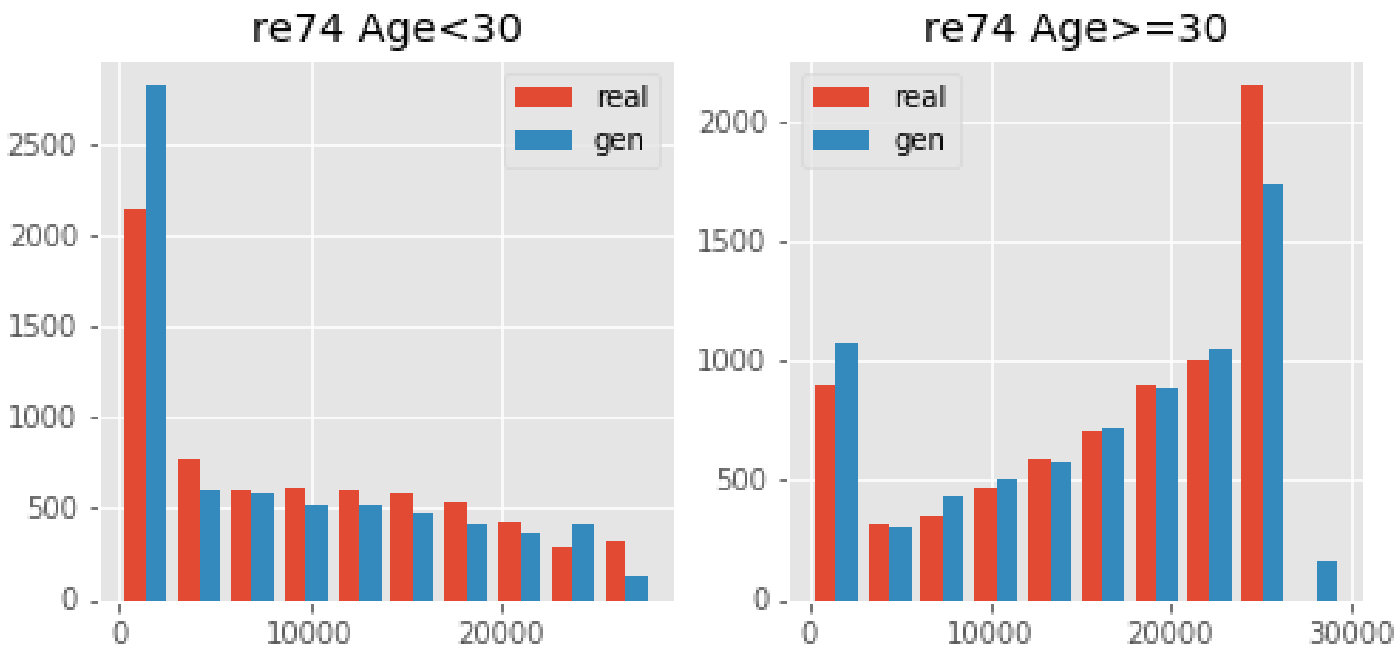
Final layer, 64 inputs, 1 output, linear.

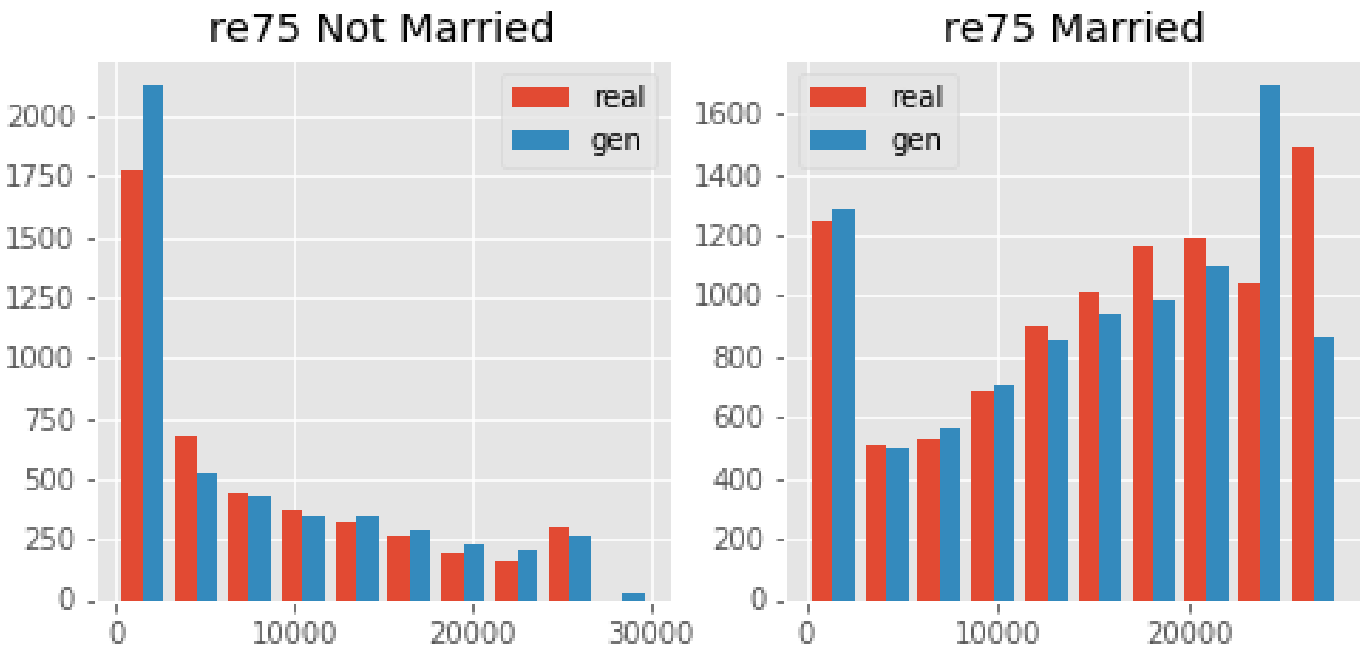
How do the generated data compare to the actual data?

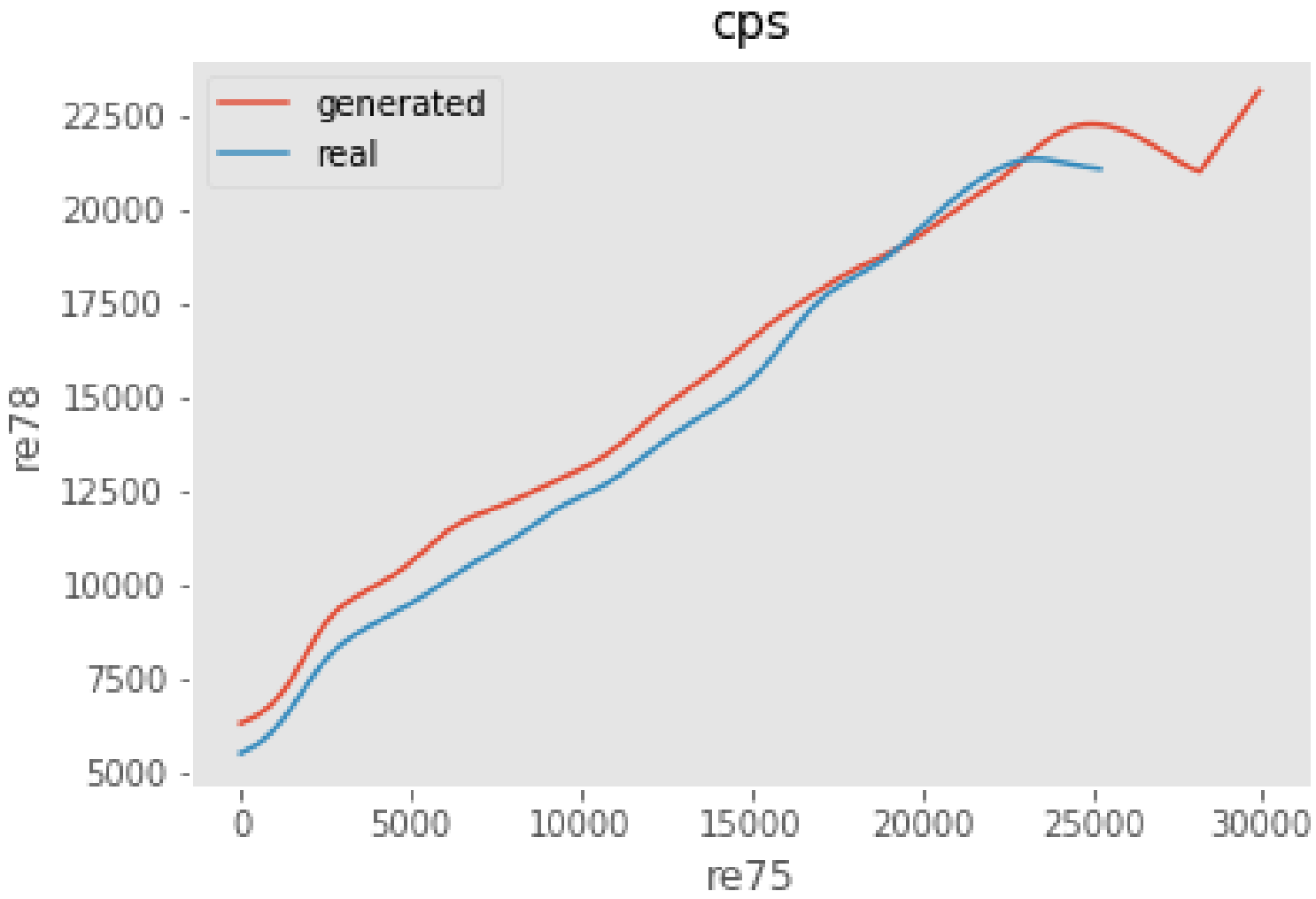
- Marginal distributions are close
- Correlations are close
- Conditional distributions (conditional on one variable at a time) are close


black
real
generated
hispanic
real
generated
married
real
generated
nodegree
real
generated
re74
real
generated
re75
real
generated
re78
real
generated
education
real
generated
age
real
generated


Real Data
Generated Data
0 1 2 3 4 5 6 7 8
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


re74 Age<30
real
gen
2500
2000
1500
1000
500
0
0
10000
20000
re74 Age>=30
real
gen
2000
1500
1000
500
0
0
10000
20000
30000


re75 Not Married
real
gen
2000
1750
1500
1250
1000
750
500
250
0
0
10000
20000
30000
re75 Married
real
gen
1600
1400
1200
1000
800
600
400
200
0
0
10000
20000


cps
generated
real
22500
20000
17500
15000
12500
10000
7500
5000
re78
0
5000
10000
15000
20000
25000
30000
re75

Causal Inference

and Machine Learning

Guido Imbens – Stanford University

Lecture 3:

Average Treatment Effects with Many Covariates

Potsdam Center for Quantitative Research

Tuesday September 10th, 10.30-12.00

Outline

1. Unconfoundedness

2. Efficiency Bound

3. Outcome Modeling, Propensity Score Modeling, and Double Robust Methods

4. Many Covariates

5. Efficient Score Methods

6. Balancing Methods

7. Comparisons of Estimators

1. Unconfoundedness Set up:

Treatment indicator: $W_i \in \{0, 1\}$

Potential Outcomes $Y_i(0), Y_i(1)$

Covariates X_i

Observed outcome: $Y_i^{\text{obs}} = W_i \cdot Y_i(1) + (1 - W_i) \cdot Y_i(0)$.

Estimand: average effect for treated:

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1]$$

Key Assumptions: unconfoundedness:

$$W_i \perp \Big(Y_i(0), Y_i(1)\Big) \Big| X_i.$$

Overlap

$$\text{pr}(W_i = 1|X_i = x) \in (0, 1)$$

If there are concerns with overlap, we may need to time sample based on propensity score

$$e(x) = \text{pr}(W_i = 1 | X_i = x) \quad \text{propensity score}$$

Trim if $e(x) \notin [0.1, 0.9]$

See Crump, Hotz, Imbens & Mitnik (Biometrika, 2008) for optimal trimming.

Important in practice.

Define the conditional mean of potential outcomes

$$\mu_w(x) = \mathbb{E}[Y_i(w)|X_i = x]$$

and the conditional variance

$$\sigma_w^2(x) = \mathbb{V}[Y_i(w)|X_i = x]$$

Under unconfoundedness the conditional potential outcome mean is equal to conditional mean of observed outcome:

$$\mu_w(x) = \mathbb{E}[Y_i^{\text{obs}}|W_i = w, X_i = x]$$

2. Semi-parametric efficiency bound for average treatment effect

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$$

under unconfoundedness

$$\mathbb{E}\left[\frac{\sigma_1^2(x)}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1-e(X_i)} + (\mu_1(X_i) - \mu_0(X_i) - \theta)^2\right]$$

$$= \mathbb{E}\left[(\psi(Y_i, W_i, X_i))^2\right]$$

where the efficient influence function is

$$\psi(y, w, x) = \mu_1(x) - \mu_0(x) + w\frac{y - \mu_1(x)}{e(x)} - (1-w)\frac{y - \mu_0(x)}{1-e(x)} - \tau$$

How can we estimate τ efficiently?

Let $\widehat{\mu}_w(x)$ and $\widehat{e}(x)$ be nonparametric estimators for $\mu_w(x)$ and $e(x)$. Then the following three estimators are efficient for τ:

A. based on estimation of regression function

$$\widehat{\tau}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} \Big(\widehat{\mu}_1(X_i) - \widehat{\mu}_1(X_i) \Big)$$

B. based on estimation of the propensity score

$$\widehat{\tau}_{\text{ipw}} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{W_i \cdot Y_i^{\text{obs}}}{\widehat{e}(X_i)} - \frac{(1 - W_i) \cdot Y_i^{\text{obs}}}{1 - \widehat{e}(X_i)} \right)$$

C. based on estimation of efficient score

$$\widehat{\tau}_{\mathsf{es}} = \frac{1}{N}\sum_{i=1}^{N}$$

$$\left(\frac{W_i \cdot (Y_i - \widehat{\mu}_1(X_i))}{\widehat{e}(X_i)} - \frac{(1-W_i) \cdot (Y_i - \widehat{\mu}_0(X_i))}{1-\widehat{e}(X_i)} + \left\{(\widehat{\mu}_1(X_i - \widehat{\mu}_0(X_i)\right\}\right)$$

• Single nearest neighbor matching also possible, but not efficient.

• Estimators seem very different.

• How should we think about choosing between them and what are their properties?

$\widehat{\tau}_{\text{reg}}$, $\widehat{\tau}_{\text{ipw}}$, and $\widehat{\tau}_{\text{es}}$ are efficient in the sense that they achieve the semiparametric efficiency bound, for fixed dimension of the covariates, but irrespective of what that dimension is.

Define:

$$\widehat{\tau}^{\text{infeasible}} = \frac{1}{N}\sum_{i=1}^{N}$$

$$\left(\frac{W_i \cdot (Y_i - \mu_1(X_i))}{e(X_i)} - \frac{(1-W_i)\cdot(Y_i - \mu_0(X_i))}{1-e(X_i)} + \{(\mu_1(X_i - \mu_0(X_i)\}\right)$$

Then:

$$\widehat{\tau}_{\text{reg}} = \widehat{\tau}_{\text{ipw}} + o_p(N^{-1/2}) = \widehat{\tau}_{\text{es}} + o_p(N^{-1/2})$$

$$= \widehat{\tau}^{\text{infeasible}} + o_p(N^{-1/2})$$

Why are these estimators first order equivalent?

Suppose single binary regressor: $X_i \in \{0, 1\}$

Simple non-parametric estimators are available for $e(x)$ and $\mu_w(x)$:

$$\widehat{e}(x) = x\frac{\sum_{i=1}^{N} \mathbf{1}_{W_i=1, X_i=x}}{\sum_{i=1}^{N} \mathbf{1}_{X_i=x}}$$

$$\widehat{\mu}_w(x) = \frac{\sum_{i=1}^{N} Y_i \mathbf{1}_{W_i=w, X_i=x}}{\sum_{i=1}^{N} \mathbf{1}_{W_i=w, X_i=x}}$$

Then all estimators are identical:

$$\widehat{\tau}_{\text{reg}} = \widehat{\tau}_{\text{ipw}} = \widehat{\tau}_{\text{es}}$$

How do they do this with continuous covariates?

- Assume lots of smoothness of the conditional expectations $\mu_w(x)$ and $e(x)$ (existence of derivatives up to high order)

- Use bias reduction techniques: higher order kernels, or local polynomial regression. The order of the kernel required is related to the dimension of the covariates.

- Regression estimator based on series estimator for $\mu_w(x)$.

Suppose X_i is an element of a compact subset of $\mathbb{R}^d$ We can approximate $\mu_w(x)$ by a polynomial series with including all terms up to x_j^k, where x_j is the jth element of $x \in \mathbb{R}^d$. (Other series are possible.)

The approximation error is small if $\mu_w(\cdot)$ has many derivatives relative to the dimension of x.

- Regression estimator based on kernel estimator for $\mu_w(x)$.

$$\widehat{\mu}_w(x) = \sum_{i=1}^{N} \mathbf{1}_{W_i=w} Y_i K\left(\frac{X_i - x}{h}\right) \Big/ \sum_{i=1}^{N} \mathbf{1}_{W_i=w} K\left(\frac{X_i - x}{h}\right)$$

This estimator is consistent under weak conditions, but to make the bias vanish from the asymptotic distribution we need to use higher order kernels (kernels with negative weights).

4. What do we do with many covariates?

Kernel regression and series methods do not work well in high dimensions.

A. Propensity score methods. Estimate $e(\cdot)$ using machine learning methods, e.g., LASSO, random forests, deep learning methods, to minimize something

$$\mathbb{E}\left[(\hat{e}(X_i) - e(X_i))^2\right]$$

leading to $\hat{e}(\cdot)$. Then use inverse propensity score weighting:

$$\hat{\tau} = \frac{1}{N_t} \sum_{i:W_i=1} Y_i - \sum_{i:W_i=0} \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)} Y_i \Bigg/ \sum_{i:W_i=0} \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}$$

Problem is that this does not select covariates that are highly correlated with Y_i

B. Regression methods. Estimate $\mu_0(x) = \mathbb{E}[Y_i|W_i = 0, X_i = x]$ using machine learning methods, e.g., LASSO, random forests, deep learning methods, to minimize something

$$\mathbb{E}\left[(\widehat{\mu}_0(X_i) - \mu_0(X_i))^2\right]$$

leading to $\widehat{e}(\cdot)$. Then use regression difference:

$$\widehat{\tau} = \frac{1}{N_t} \sum_{i:W_i=1} (Y_i - \widehat{\mu}_0(X_i))$$

Problem is that this does not select covariates that are highly correlated with W_i

Recall omitted variable bias:

$$Y_i = \alpha + \tau W_i + \beta^\top X_i + \varepsilon_i$$

Omitted X_i from regression leads to bias in τ that is proportional to β and correlation between W_i and X_i.

Selecting covariates **only** on basis of correlation with Y_i, or **only** on the basis of correlation with W_i is not effective.

• As in case with few covariates, it is better to work both with the correlations between W_i and X_i **and** the correlations between $Y_i(w)$ and X_i.

First improvement, use selection methods that select covariates that are correlated with W_i **or** Y_i (double selection, Belloni et al, 2012).

E.g., use lasso to select covariates that predict Y_i. Use lasso to select covariates that predict W_i.

Take union of two sets of covariates, and then regress Y_i on that set of covariates.

- works better than single selection methods.

5. Efficient Score Methods and Double Robustness (Robins & Rotnitzky, 1996; Van Der Laan and Rubin (2006), Imbens and Rubin (2015) and others.

We do not need $e(\cdot)$ to be estimated consistently as long as $mu_0(\cdot)$ and $\mu_1(\cdot)$ are estimated consistently because

$$\mathbb{E}\left[W_i\frac{Y_i-\mu_1(X_i)}{a(X_i)}-(1-W_i)\frac{Y_i-\mu_0(X_i)}{1-a(X_i)}+\mu_1(X_i)-\mu_0(X_i)\right]=\tau$$

for any function $a(\cdot)$

Also, we do not need $\mu_0(\cdot)$ and $\mu_1(\cdot)$ to be estimated consistenly, as long as $e(\cdot)$ is estimated consistently because

$$\mathbb{E}\left[W_i\frac{Y_i-c(X_i)}{e(X_i)}-(1-W_i)\frac{Y_i-b(X_i)}{1-e(X_i)}+c(X_i)-b(X_i)\right]=\tau$$

for any functions $b(\cdot)$ and $c(\cdot)$

But, we can improve on these etimators: (e.g., Chernozhukov *et al*, 2016):

Split the sample randomly into two equal parts, $i = 1, \ldots, N/2$ and $i = N/2 + 1, \ldots, N$.

Estimate $\mu_0(\cdot)$, $\mu_1(\cdot)$ and $e(\cdot)$ on the first subsample, and then estimate τ on the second subsample as

$$\widehat{\tau}_1 = \frac{1}{N/2} \sum_{i=N/2+1}^{N}$$

$$\left(W_i \frac{Y_i - \widehat{\mu}_1^{(1)}(X_i)}{\widehat{e}^{(1)}(X_i)} - (1 - W_i) \frac{Y_i - \widehat{\mu}_0^{(1)}(X_i)}{1 - \widehat{e}^{(1)}(X_i)} + \widehat{\mu}_1^{(1)}(X_i) - \widehat{\mu}_0^{(1)}(X_i) \right)$$

This is consistent, but not efficient.

Do the reverse to get

$$\widehat{\tau}_2 = \frac{1}{N/2} \sum_{i=1}^{N/2}$$

$$\left(W_i \frac{Y_i - \widehat{\mu}_1^{(2)}(X_i)}{\widehat{e}^{(2)}(X_i)} - (1 - W_i) \frac{Y_i - \widehat{\mu}_0^{(2)}(X_i)}{1 - \widehat{e}^{(2)}(X_i)} + \widehat{\mu}_1^{(2)}(X_i) - \widehat{\mu}_0^{(2)}(X_i) \right)$$

Finally, combine:

$$\widehat{\tau} = \frac{\widehat{\tau}_1 + \widehat{\tau}_2}{2}$$

Key Assumptions

Estimators for $\mu_0(\cdot)$, $\mu_1(\cdot)$ and $e(\cdot)$ need to converge fast enough, e..g., faster than $N^{-1/4}$ rate.

That is not as fast as parametric models, which converge at $N^{-1/2}$ rate, but still faster than simple nonparametric (non-negative) kernel estimators that converge at a rate that depends on the dimension of X_i. Using kernel estimators one would need to use higher order kernels. Other methods, e.g., random forests, deep neural nets, may work, but no easy interpretable assumptions available.

6. Balancing Methods

Suppose we are interested in $\tau = \mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1]$, so that we need to estimate

$$\mathbb{E}[Y_i|W_i = 0, X_i]|W_i = 1]$$

Note that, with $e(\cdot)$ the propensity score,

$$\mathbb{E}\left[\frac{e(X_i)}{1 - e(X_i)}(1 - W_i)Y_i\right] = \mathbb{E}[Y_i|W_i = 0, X_i]|W_i = 1]$$

So, we could estimate $e(\cdot)$ as $\widehat{e}(\cdot)$, and then

$$\frac{1}{N_1}\sum_{i=1}^{N}(1 - W_i)Y_i\gamma_i, \quad \text{where} \quad \gamma_i = \frac{e(X_i)}{1 - e(X_i)}$$

The key insight is that for any function $h : \mathbb{X} \mapsto \mathbb{R}^p$,

$$\mathbb{E}\left[\frac{e(X_i)}{1-e(X_i)}(1-W_i)h(X_i)\right] = \mathbb{E}[h(X_i)|W_i = 1]$$

including for $h(X_i) = X_i$:

$$\mathbb{E}\left[\frac{e(X_i)}{1-e(X_i)}(1-W_i)X_i\right] = \mathbb{E}[X_i|W_i = 1]$$

Zubizarreta (2012) suggests directly focusing on the balance in covariates. Find weights γ_i that solve

$$\min_{\gamma_1,\ldots,\gamma_N} \sum_{i=1}^{N_c} \gamma_i^2, \quad \text{subject to } \sum_{i=1}^{N} (1-W_i)\gamma_i X_i = \overline{X}_t$$

See also Hainmueller (2010), and Abadie, Diamond and Hainmueller (2012) in a different context.

$\gamma_i = e(X_i)/(1-e(X_i))$ solves the restriction in expectation, but not in sample.

We may get better balance directly focusing on balance in sample than by propensity score weighting.

Athey, Imbens and Wager (2015) combine this with a linear regression for the potential outcomes.

In their setting there are too many covariates to balance the averages exactly: there is no solution for γ that solves

$$\sum_{i=1}^{N}(1-W_i)\gamma_i X_i = \overline{X}_t$$

So, the objective function for γ is

$$\min_{\gamma_1,\ldots,\gamma_N} \zeta \times \frac{1}{N_c}\sum_{i=1}^{N_c}\gamma_i^2 + (1-\zeta)\times\left|\frac{1}{N_c}\sum_{i=1}^{N}(1-W_i)\gamma_i X_i - \overline{X}_t\right|^2$$

where $\zeta \in (0,1)$ is a tuning parameter, e.g., 1/2.

Suppose that the conditional expectation of $Y_i(0)$ given X_i is linear:

$$\mu_0(x) = \beta^\top x$$

AIW estimate β using lasso or elastic nets:

$$\min_{\beta} \sum_{i:W_i=0} \left(Y_i - \beta^\top X_i\right)^2 + \lambda \left(\alpha \sum_{k=1}^{p} |\beta_p| + (1-\alpha) \sum_{k=1}^{p} |\beta_p|^2 \right)$$

A standard estimator for the average effect for the treated would be

$$\widehat{\tau} = \overline{Y}_t - \overline{X}_t^\top \widehat{\beta}$$

A simple weighting estimator would be

$$\widehat{\tau} = \overline{Y}_t - \sum_{i=1}^{N} (1 - W_i)\gamma_i Y_i$$

The residual balancing estimator for the average effect for the treated is

$$\widehat{\tau} = \overline{Y}_t - \left(\overline{X}_t^\top \widehat{\beta} + \sum_{i=1}^{N} (1 - W_i)\gamma_i \left(Y_i - X_i^\top \widehat{\beta} \right) \right)$$

- does not require estimation of the propensity score.

- relies on approximate linearity of the regression function.

7. Comparison of Estimators

1. Methods based on Outcome Modeling

   (a) Generalized Linear Models (Linear and Logistic Models)

   (b) Random Forests

   (c) Neural Nets

2. Methods based on Propensity Score Modeling

3. Doubly Robust Methods

	Experimental		CPS		PSID	
	est	s.e.	est	s.e.	est	s.e.
DIM	1.79	(0.67)	-8.50	(0.58)	-15.20	(0.66)
BCM	1.90	()	2.35	()	1.47	()
Outcome Models						
L	1.00	(0.57)	0.69	(0.60)	0.79	(0.60)
RF	1.73	(0.58)	0.92	(0.6)	0.06	(0.63)
NN	2.07	(0.59)	1.43	(0.59)	2.12	(0.59)
Propensity Score Weighting						
L	1.81	(0.83)	1.18	(0.77)	1.26	(1.13)
RF	1.78	(0.94)	0.65	(0.77)	-0.46	(1.00)
NN	1.92	(0.87)	1.26	(0.93)	0.10	(1.28)
Double Robust Methods						
L	1.80	(0.67)	1.27	(0.65)	1.50	(0.97)
RF	1.84	(0.8)	1.46	(0.63)	1.34	(0.85)
NN	2.15	(0.74)	1.52	(0.75)	1.14	(1.08)

Estimator	rmse	rmse rank	bias	sdev	coverage
Difference in Means	0.62	9	-0.29	0.55	0.90
Bias Corrected Matching	0.64	10	-0.08	0.64	
Outcome Models					
Linear	0.56	2	-0.06	0.56	0.90
Random Forest	0.58	4	-0.15	0.56	0.89
Neural Nets	0.65	11	-0.17	0.63	0.85
Propensity Score Weighting					
Linear	0.56	3	-0.04	0.56	0.99
Random Forest	0.60	7	-0.17	0.58	0.99
Neural Nets	0.59	5	-0.11	0.58	0.99
Double Robust Methods					
Linear	0.56	1	-0.04	0.56	0.95
Random Forest	0.60	8	-0.08	0.60	0.95
Neural Nets	0.59	6	-0.09	0.59	0.95

Estimator	rmse	rank	bias	sdev	coverage
Difference in Means	10.50	11	-10.49	0.40	0.00
Bias Corrected Matching	0.71	7	-0.37	0.61	0.00
Outcome Models					
Linear	0.77	8	-0.62	0.45	0.67
Random Forest	0.80	9	-0.67	0.44	0.62
Neural Nets	0.51	3	-0.10	0.50	0.89
Propensity Score Weighting					
Linear	0.66	6	-0.47	0.46	0.95
Random Forest	0.89	10	-0.77	0.45	0.86
Neural Nets	0.52	4	0.09	0.51	0.98
Double Robust Methods					
Linear	0.64	5	-0.45	0.45	0.84
Random Forest	0.50	1	-0.13	0.48	0.92
Neural Nets	0.50	2	0.01	0.50	0.96

	185 treated, 2,490 Controls					945 Treated, 12,450			
Estimator	rmse	rank	bias	sdev	cov	rmse	rank	bias	s
DIM	15.18	12	-15.17	0.48	0.00	15.17	11	-15.17	0
BCM	0.88	8	0.42	0.77	0.00	0.51	7	0.38	0
Outcome Models									
L	0.57	1	0.09	0.56	0.88	0.27	1	0.09	0
RF	0.97	10	-0.79	0.57	0.52	0.63	10	-0.57	0
NN	1.20	11	0.85	0.85	0.48	0.62	9	0.50	0
Propensity Score Weighting									
L	0.67	2	-0.01	0.67	0.98	0.29	2	-0.02	0
RF	0.91	9	-0.65	0.64	0.94	0.53	8	-0.44	0
NN	0.83	7	-0.40	0.73	0.96	0.31	3	0.06	0
Double Robust Methods									
L	0.72	4	0.27	0.67	0.93	0.38	6	0.23	0
RF	0.70	3	0.11	0.69	0.91	0.33	4	0.07	0
NN	0.76	6	0.35	0.67	0.90	0.34	5	0.17	0

	CPS Original Estimator			Ave Simulations			Standard Dev	
	RMSE	Bias	sdev	RMSE	Bias	sdev	RMSE	Bias
DIM	10.50	-10.49	0.40	10.46	-10.45	0.38	0.52	0.52
BCM	0.71	-0.37	0.61	0.78	-0.11	0.59	0.15	0.55
Outcome Models								
L	0.77	-0.62	0.45	0.89	-0.66	0.42	0.40	0.61
RF	0.80	-0.67	0.44	0.66	-0.50	0.40	0.19	0.24
NN	0.51	-0.10	0.50	0.50	-0.10	0.46	0.05	0.17
Propensity Score Weighting								
L	0.66	-0.47	0.46	0.66	-0.44	0.43	0.24	0.35
RF	0.89	-0.77	0.45	0.74	-0.60	0.40	0.23	0.27
NN	0.52	0.09	0.51	0.46	0.06	0.45	0.05	0.06
Double Robust Methods								
L	0.64	-0.45	0.45	0.66	-0.44	0.43	0.23	0.34
RF	0.50	-0.13	0.48	0.45	-0.09	0.43	0.05	0.13
NN	0.50	0.01	0.50	0.45	-0.02	0.44	0.05	0.05

Causal Inference

and Machine Learning

Guido Imbens – Stanford University

Lecture 4:

Heterogenous Treatment Effects

Potsdam Center for Quantitative Research

Tuesday September 10th, 13.15-14.45

Heterogenous Treatment Effects

• Given experimental data with binary treatment, how can we flexibly estimate the average effect conditional on pretreatment variables, in settings with a large number of pretreatment variables, and large samples?

• Adapt machine learning / supervised learning methods designed for prediction.

• Focus mainly on tree methods, because they lead to the partitioning of the covariate space into interpretable subsets with approximately constant conditional treatment effects.

Potential Outcome Set Up for Causal Inference

Binary treatment $W_i \in \{0, 1\}$, randomly assigned.

Pair of potential outcomes $(Y_i(0), Y_i(1))$

Vector-valued pre-treatment variable X_i

Observe for a random sample from a large population the triple $(W_i, Y_i^{\text{obs}}, X_i)$ where

$$Y_i^{\text{obs}} = Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

Unit-level treatment effect and conditional average treatment effect are

$$\tau_i = Y_i(1) - Y_i(0), \quad \tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x]$$

Application: effect of placement of answers to search queries on screen on click rates.

- Units are search queries.

- Treatment is moving the answer that is rated first by the search engine algorithm from the first place on the screen to the third place on the screen.

- Outcome is indicator that the answer that is rated first by search is clicked on ("click rate").

- Pre-treatment variables are characteristics of the search queries. Many of these are binary indicators, e.g., is the query about consumer electronics, celebrities, clothes and shoes, is it from the Safari operating system, is it about movie times, is it for images, etc.

Application (ctd)

- 499,486 search queries.

- 60 pre-treatment variables

- Experimental estimate of overall average effect on click-through rate: $\widehat{\tau} = \overline{Y}_t - \overline{Y}_c = -0.13$.

Moving answer down the list lowers substantially the click-through rate.

Question: are there search queries where the effect is small or large relative to this?

• If I search for "ebay" (and my algorithm ranks "ebay.com" first) it probably does not make much difference whether I put ebay.com on the first line or the fifth line, people know where they want to go.

• If I search for "econometrics textbook" and the algorithm ranks 'mostly harmless" first, it probably does make a difference for the click-through rate whether I actually put "mostly harmless" on the first line or on the fifth line.

Naive Solution

• If all we had was one or two binary covariates, we would just partition the sample by covariate values and estimate the average causal effects for each subsample and we would be done.

• Too many covariates to do that: 2^{60} different cells.

• No strong prior beliefs on where ranking matters.

Approach: be flexible / nonparametric about estimating $\tau(x)$.

Regression Trees (conventional, non-causal, set up, e.g., Breiman book - also possible to use lasso or other flexible prediction methods)

Suppose we have a random sample of (X_i, Y_i), $i = 1, \ldots, N$, and we wish to estimate $\mu(x) = \mathbb{E}[Y_i | X_i = x]$.

Trees recursively partition covariate space into "leaves." Within a leaf the average outcome is estimated as the subsample average (could do something more sophisticated, model-based, within leaves).

- trees are easy to interpret.

Start with single "leaf." Predicted outcome is $\widehat{\mu}(x) = \overline{Y}$. Average in-sample squared error is

$$Q(\widehat{\mu}) = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \widehat{\mu}(X_i))^2$$

Next we consider splitting this leaf into two leaves, in order to optimize in-sample fit.

We need to choose **which covariate** we split on, and **what threshold** we split at. Split covariate k, at threshold c.

Leave 1: $X_{ik} \leq c$

Leave 2: $X_{ik} > c$

Consider splitting leaf into two parts, depending on whether kth covariate is below or above threshold c: Now the predicted outcome is

$$\widehat{\mu}(x) = \begin{cases} \overline{Y}_L & \text{if } x_k \leq c \\ \overline{Y}_H & \text{if } x_k > c, \end{cases}$$

where

$$\overline{Y}_L = \frac{1}{N_L} \sum_{i:X_{ik} \leq c} Y_i, \quad \overline{Y}_H = \frac{1}{N_H} \sum_{i:X_{ik} > c} Y_i$$

Choose the covariate to split on and the threshold to minimize

$$Q(\widehat{\mu}) = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \widehat{\mu}(X_i))^2$$

Next, consider splitting either of the two leaves, and find the optimal threshold, the optimal covariate and the optimal leaf to split.

Keep doing this to minimize

$$Q(\widehat{\mu}) + \alpha \cdot M$$

where M is the number of leaves.

The penalty rate α is choosen by out-of-sample cross-validation to avoid over-fitting.

• many variations on simple trees, boosting, bagging, random forests. All tend to work better in terms of out-of-sample performance than kernel regression.

Alternative Representation of Goodness-of-Fit Measure

We compare two possible splits leading to estimates $\widehat{\mu}_1$ and $\widehat{\mu}_2$ by comparing $Q(\widehat{\mu}_1)$ and $Q(\widehat{\mu}_2)$, where

$$Q(\widehat{\mu}) = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \widehat{\mu}(X_i))^2$$

If the models include an intercept, as they usually do, most estimation methods would ensure that the average of $(Y_i - \widehat{\mu}(X_i)) \cdot \widehat{\mu}(X_i)$ would be equal to zero.

Then we get an identical ranking by comparing $\tilde{Q}(\widehat{\mu}_1)$ and $\tilde{Q}(\widehat{\mu}_2)$, where

$$\tilde{Q}(\widehat{\mu}) = -\frac{1}{N} \sum_{i=1}^{N} \widehat{\mu}(X_i)^2.$$

Trees for Causal Effects

• We would like to construct trees for $\tau(x)$

• Problem 1: we do not observe $\tau_i = Y_i(1) - Y_i(0)$, so cannot directly use standard tree methods. **Need other estimation methods.**

• Problem 2: given two candidate estimators, trees or lasso or otherwise, we cannot directly use out-of-sample comparisons between methods because we do not observe $\tau_i = Y_i(1) - Y_i(0)$. **Need other validation methods.**

Simple, Non-causal, Solutions to First Problem

Solution I (single tree): use conventional tree methods to construct tree for $\mu(w, x) = \mathbb{E}[Y_i^{\text{obs}} | W_i = w, X_i = x]$ and estimate

$$\widehat{\tau}(x) = \widehat{\mu}(1, x) - \widehat{\mu}(0, x)$$

May never split on treatment w. E.g., Imai and Ratkovic (2013), Foster, Tailor, Ruberg (2011)

Solution II (two trees): construct separate trees for $\mu_0(x) = \mathbb{E}[Y_i^{\text{obs}}|W_i = 0, X_i = x]$ and $\mu_1(x) = \mathbb{E}[Y_i^{\text{obs}}|W_i = 1, X_i = x]$, and estimate

$$\widehat{\tau}(x) = \widehat{\mu}_1(x) - \widehat{\mu}_0(x)$$

$\mu_w(x)$ may vary much more with pretreatment variables than $\tau(x)$.

Still second problem: How do we compare the two methods in test sample?

Insight

Define

$$Y_i^* = Y_i^{\text{obs}} \cdot \frac{W_i - \mathbb{E}[W_i]}{\mathbb{E}[W_i] \cdot (1 - \mathbb{E}[W_i])}$$

Y_i^* is unbiased for treatment effect $Y_i(1) - Y_i(0)$ (based on single observation!), but quite noisy.

Then

$$\tau(x) = \mathbb{E}[Y_i^* | X_i = x]$$

Generalization to observational studies with unconfoundedness:

$$Y_i^* = Y_i^{\text{obs}} \cdot \frac{W_i - e(X_i)}{e(X_i) \cdot (1 - e(X_i))}$$

where $e(x)$ is the propensity score:

$$e(x) = \text{pr}(W_i = 1 | X_i = x)$$

This suggests an out-of-sample goodness-of-fit measure:

$$Q_1(\widehat{\tau}) = \frac{1}{N} \sum_{i=1}^{N} (Y_i^* - \widehat{\tau}(X_i))^2$$

Alternative out-of-sample goodness-of-fit measure based on matching. Replace τ_i by

$$\widehat{\tau}_i = (2 \cdot W_i - 1) \cdot \left(Y_i^{\text{obs}} - Y_{\ell(i)}^{\text{obs}} \right),$$

where $\ell(i)$ is closest match:

$$\ell(i) = \arg \min_{j: W_j \neq W_i} \|X_i - X_j\|$$

Then use

$$Q_2(\widehat{\tau}) = \frac{1}{N} \sum_{i=1}^{N} (\widehat{\tau}_i - \widehat{\tau}(X_i))^2$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left((2 \cdot W_i - 1) \cdot \left(Y_i^{\text{obs}} - Y_{\ell(i)}^{\text{obs}} \right) - \widehat{\tau}(X_i) \right)^2$$

Solution III (transformed outcome tree): Use conventional tree methods to construct tree based on (X_i, Y_i^*) data (discarding W_i).

(Not necessarily efficient: suppose $\mathbb{V}(Y_i(w)|X_i = x)$ is very small, but treatment effect is substantial and heterogenous. Then Solutions I and II will be better than Solution III.)

Solution IV (causal tree 1):

Start with a single leaf. Consider splitting it based on a particular covariate and a particular threshold, leading to two potential new leaves.

Estimate within each potential leaf the average treatment effect, as well as the overall average treatment effect:

$$\widehat{\tau} = \overline{Y}_1^{\text{obs}} - \overline{Y}_0^{\text{obs}}, \quad \widehat{\tau}_H = \overline{Y}_{H,1}^{\text{obs}} - \overline{Y}_{H,0}^{\text{obs}}, \quad \widehat{\tau}_L = \overline{Y}_{L,1}^{\text{obs}} - \overline{Y}_{L,0}^{\text{obs}},$$

$$\overline{Y}_w^{\text{obs}} = \frac{1}{N} \sum_{i:W_i=w} Y_i^{\text{obs}}, \quad \overline{Y}_{L,w}^{\text{obs}} = \frac{1}{N_L} \sum_{i:W_i=w, X_i \leq c} Y_i^{\text{obs}}, \quad \overline{Y}_{H,w}^{\text{obs}} = \frac{1}{N_H}$$

If $N_{L,w}$ or $N_{H,w}$ is zero for $w = 0, 1$, we do not consider this potential split.

To assess the improvement of goodness of fit we would like to calculate the difference

$$\sum_{i=1}^{N} (\tau_i - \widehat{\tau})^2 - \left(\sum_{i:X_i \leq c} (\tau_i - \widehat{\tau}_L)^2 + \sum_{i:X_i > c} (\tau_i - \widehat{\tau}_H)^2 \right).$$

This is not feasible because we do not observe τ_i. We replace τ_i by Y_i^*, which is unbiased for τ_i, and calculate the difference

$$Q_1(\widehat{\tau}) = \sum_{i=1}^{N} (Y_i^* - \widehat{\tau})^2 - \left(\sum_{i:X_i \leq c} (Y_i^* - \widehat{\tau}_L)^2 + \sum_{i:X_i > c} (Y_i^* - \widehat{\tau}_H)^2 \right)$$

The difference with Solution III is that $\widehat{\tau}_H$ and $\widehat{\tau}_L$ are not calculated as the average of Y_i^* within the leafs, but as the difference in average outcomes by treatment status.

Solution V (causal tree 2):

Same approach to leaf splitting, but now with modified criterion, along the lines of

$$\tilde{Q}(\hat{\mu}) = -\frac{1}{N}\sum_{i=1}^{N}\hat{\mu}(X_i)^2.$$

Now we choose the split to minimize

$$\tilde{Q}_1(\hat{\tau}) = \frac{1}{N}\cdot\left(N\cdot\hat{\tau}^2 - \left(N_L\cdot\hat{\tau}_L^2 + N_H\cdot\hat{\tau}_H^2\right)\right)$$

Does not rely on transformed outcome, less noisy.

Application

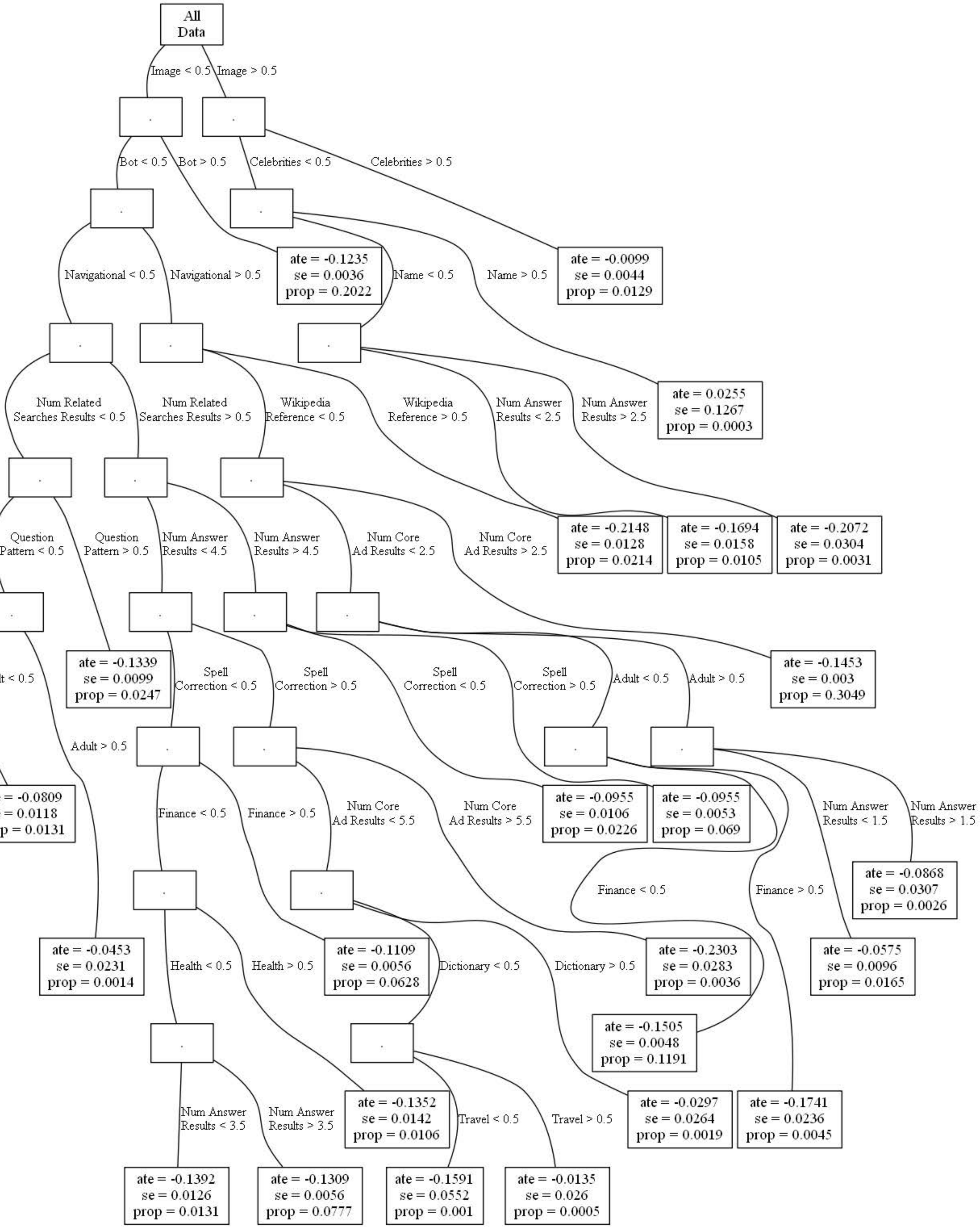
Training sample, $N_{\text{train}} = 249,742$, Test sample, $N_{\text{test}} = 249,742$.

	Single Tree	Two Trees	Transf. Outc. Tree	Causal Tree	Modified Causal Tree
OOS I	0.8053	0.8053	0.8046	0.8048	**0.8044**
OOS II	0.3111	0.3111	0.3107	0.3106	**0.3105**
# Leaves	52	36 26	21	21	24

Correlation Between Predictions and Y_i^* in Test Sample

	Y_i^*	Single Tree	Two Trees	Transf. Outc.	Causal Tree	Mod. Causal
Y_i^*	1.000	0.026	0.027	0.034	0.030	0.037
Single Tree	0.026	1.000	0.963	0.638	0.729	0.664
Two Trees	0.027	0.963	1.000	0.671	0.734	0.685
Transf. Outc.	0.034	0.638	0.671	1.000	0.733	0.864
Causal Tree	0.030	0.729	0.734	0.733	1.000	0.791
Mod Causal	0.037	0.664	0.685	0.864	0.791	1.000

leaf	training sample			test sample		
	est	se	share	est	se	share
1	-0.1235	0.0036	0.2022	-0.1236	0.0036	0.2018
2	-0.1339	0.0099	0.0247	-0.1349	0.0102	0.0240
3	-0.0099	0.0044	0.0129	-0.0073	0.0044	0.0132
4	-0.2148	0.0128	0.0214	-0.2467	0.0126	0.0216
5	-0.1453	0.0030	0.3049	-0.1480	0.0030	0.3044
6	-0.1109	0.0056	0.0628	-0.1100	0.0055	0.0635
7	-0.2303	0.0283	0.0036	-0.2675	0.0284	0.0037
8	-0.0575	0.0096	0.0165	-0.0324	0.0095	0.0168
9	-0.0868	0.0307	0.0026	-0.0559	0.0294	0.0025
10	-0.1505	0.0048	0.1191	-0.1693	0.0047	0.1191
11	-0.1741	0.0236	0.0045	-0.1682	0.0239	0.0046
12	0.0255	0.1267	0.0003	0.2857	0.1235	0.0002
13	-0.0297	0.0264	0.0019	-0.0085	0.0250	0.0022
14	-0.1352	0.0142	0.0106	-0.1139	0.0147	0.0100
15	-0.1591	0.0552	0.0010	-0.1432	0.0526	0.0011
16	-0.0135	0.0260	0.0005	0.0080	0.0502	0.0004
17	-0.0809	0.0118	0.0131	-0.0498	0.0124	0.0132
18	-0.0453	0.0231	0.0014	-0.0454	0.0208	0.0014
19	-0.1694	0.0158	0.0105	-0.1997	0.0162	0.0106
20	-0.2072	0.0304	0.0031	-0.2790	0.0305	0.0030
21	-0.0955	0.0106	0.0226	-0.0834	0.0108	0.0223


All Data
Image < 0.5
Image > 0.5
Bot < 0.5
Bot > 0.5
Celebrities < 0.5
Celebrities > 0.5
ate = -0.1235
se = 0.0036
prop = 0.2022
ate = -0.0099
se = 0.0044
prop = 0.0129
Navigational < 0.5
Navigational > 0.5
Name < 0.5
Name > 0.5
Num Related Searches Results < 0.5
Num Related Searches Results > 0.5
Wikipedia Reference < 0.5
Wikipedia Reference > 0.5
Num Answer Results < 2.5
Num Answer Results > 2.5
ate = 0.0255
se = 0.1267
prop = 0.0003
ate = -0.2148
se = 0.0128
prop = 0.0214
ate = -0.1694
se = 0.0158
prop = 0.0105
ate = -0.2072
se = 0.0304
prop = 0.0031
Question Pattern < 0.5
Question Pattern > 0.5
Num Answer Results < 4.5
Num Answer Results > 4.5
Num Core Ad Results < 2.5
Num Core Ad Results > 2.5
ate = -0.1339
se = 0.0099
prop = 0.0247
ate = -0.1453
se = 0.003
prop = 0.3049
Adult < 0.5
Adult > 0.5
Spell Correction < 0.5
Spell Correction > 0.5
Spell Correction < 0.5
Spell Correction > 0.5
Adult < 0.5
Adult > 0.5
ate = -0.0809
se = 0.0118
prop = 0.0131
ate = -0.0955
se = 0.0106
prop = 0.0226
ate = -0.0955
se = 0.0053
prop = 0.069
Finance < 0.5
Finance > 0.5
Num Core Ad Results < 5.5
Num Core Ad Results > 5.5
Num Answer Results < 1.5
Num Answer Results > 1.5
ate = -0.0868
se = 0.0307
prop = 0.0026
ate = -0.0453
se = 0.0231
prop = 0.0014
ate = -0.1109
se = 0.0056
prop = 0.0628
ate = -0.2303
se = 0.0283
prop = 0.0036
ate = -0.0575
se = 0.0096
prop = 0.0165
Health < 0.5
Health > 0.5
Dictionary < 0.5
Dictionary > 0.5
ate = -0.1505
se = 0.0048
prop = 0.1191
Num Answer Results < 3.5
Num Answer Results > 3.5
Travel < 0.5
Travel > 0.5
ate = -0.1352
se = 0.0142
prop = 0.0106
ate = -0.0297
se = 0.0264
prop = 0.0019
ate = -0.1741
se = 0.0236
prop = 0.0045
ate = -0.1392
se = 0.0126
prop = 0.0131
ate = -0.1309
se = 0.0056
prop = 0.0777
ate = -0.1591
se = 0.0552
prop = 0.001
ate = -0.0135
se = 0.026
prop = 0.0005

Substantial variation in conditional treatment effects

Let's look at two specific leaves out of the 24.

Leaf 3: -0.0073 (s.e. 0.0044), proportion 0.0132
What is the query in this leaf: Image & Celebrity

If I search for "image of Chuck Manski" it does not matter whether the image is ranked first or third.

Leaf 4: -0.2467 (s.e. 0.0126), proportion 0.0216
Not image & not search bot & navigation & wikipedia reference.

If I search for "machine learning" or "instrumental variables" the ranking may be very important.

• leaves defined through interactions, no through simple main effects, even second order effects may not be sufficient to capture all effects of interest.

• interpretable leaves.

Simulations

$$Y_i(w) = \sum_{k=1}^{5} X_{ik} \cdot \beta_{kw} + \varepsilon_{iw}$$

$$\beta_0 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \qquad \beta_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\varepsilon_{i0} = \varepsilon_{i1} \sim \mathcal{N}(0, 0.4^2)$$

$$X_{ik} \sim \mathcal{B}(0, 0.5) \qquad W_i \sim \mathcal{B}(0, 0.4)$$

$$N^{\text{train}} = N^{\text{test}} = 100$$

Simulation Results

	Single Tree	Two Trees	Transformed Outcome Tree	Causal Tree	Modified Causal Tree
OOS I	1.31	1.29	1.28	1.25	**1.24**
OOS II	0.45	0.43	0.42	0.40	**0.39**
Oracle	0.14	0.12	0.13	0.10	**0.07**
# Leaves	12.7	5.1 8.4	5.6	6.7	5.8

Causal Inference

and Machine Learning

Guido Imbens – Stanford University

Lecture 5:

Experimental Design and Multi-armed Bandits

Potsdam Center for Quantitative Research

Tuesday September 10th, 15.15-16.45

Outline

1. Re-randomization

2. Experiments in Networks

3. Multi-armed Bandits

1. Re-Randomization

Sometimes researchers randomize assignment to treatment, then assess the (im)balance the specific assignment would generate, and decide to re-randomize if the initial assignment failed to lead to sufficient balance.

What to make of that?

Re-randomization can improve precision of estimates and power of tests considerably, but needs to be done carefully to maintain ability to do inference.

Re-randomization is conceptually similar to completely randomized experiment:

Consider a sample of $2N$ units.

Randomize treatment to each unit by flipping a fair coin.

Re-randomize till the number of treated units is exactly equal to N.

This leads to the same design as randomly selecting N units for assignment to treatment in a completely randomized experiment.

Formal Analysis of Re-Randomization

Suppose we have $2N$ units. We observe a K-vector of covariates X_i. Without taking into account the covariate values, N units are randomly selected to receive the treatment, and the remaining units are assigned to the control group.

Calculate

$$\overline{X}_w = \frac{1}{N} \sum_{i:W_i=w} X_i, \qquad t_X = \frac{\overline{X}_1 - \overline{X}_0}{\sqrt{s^2_{X,0}/N + s^2_{X,1}/N}}$$

What to do if $|t_X|$ is large, if discovered before assignment is implemented?

Two Cases

- Decide *a priori* to randomize M times, and implement assignment vector that minimizes some criterion e.g., minimize the maximum of the t-statistics for the K covariates.

- Re-randomize until the criterion meets some threshold: e.g., with two covariates, until both t-statistics are below 1.

  (need to be careful here: the threshold should be feasible).

Key:
1. articulate strategy **a priori**, so randomization inference is possible.

2. Do **not** search over all assignments for optimal value for criterion because then there is little randomness left.

Cautionary Note

- Suppose with $2N$ units, X_i earnings, $2N-1$ units have $X_i \in [0, 10]$, and one unit has $X_i = 1000$.

- Minimizing t-statistic leads to one treatment group containing individual with $X_i = 1000$ and $N-1$ individual with lowest earnings, and other group containing N richest individuals after very richest individual.

- Irrespective of design estimation of ave effect is difficult.

- Rank-based tests may still have substantial power.

- Maybe remove outlier unit for estimation purposes.

Conclusion

• Instead of re-randomization, lay out acceptable set of random assignments.

2. Experiments in Networks

• We are interested in testing complex hypotheses on causal effects in settings where individuals are connected through a network and there may be spillovers.

Bond, Fariss, Jones, Kramer, Marlow, Settle and Fowler ("A 61-million-person experiment ...", 2012) write:

> "Furthermore, the messages not only influenced the users who received them but also the users friends, and friends of friends."

Christakis and Fowler (2008, "the spread of obesity in a large social network") claim changes in weight spread beyond friends.

How can we test such claims, in the presence of unrestricted homophily, in a single network?

Clearly there is some evidence in the data

Compare two individuals, both in the control group, both with one friend, one with a treated friend, one with a friend in the control group.

Finding a correlation between outcomes for egos and treatment for alters is evidence of spillovers.

- Does that rely on large sample approximations?

- Can we test hypotheses about friends of friends?

2.1 Causal Effects and Potential Outcomes

We have a finite population $\mathbb{P}$ with N units. These units may be linked through a network with adjacency matrix $\mathbf{A}$. We also measure covariates on the individuals, with $\mathbf{X}$ the matrix of covariates.

The units are exposed to a treatment $\mathbf{W}$, where $\mathbf{W}$ is an N-vector with ith element W_i. $\mathbf{W}$ takes on values in $\mathbb{W}$.

For each unit there is a set of potential outcomes $Y_i(\mathbf{w})$, one for each $\mathbf{w} \in \mathbb{W}$. We observe $Y_i^{\text{obs}} = Y_i(\mathbf{W})$.

Causal effects are comparisons $Y_i(\mathbf{w}) - Y_i(\mathbf{w}')$ for any pair $\mathbf{w} \neq \mathbf{w}' \in \mathbb{W}$

Most of the literature: Stable-unit-treatment-value-assumption (sutva, Rubin), so that $Y_i(\mathbf{w})$ depends only on own treatment w_i.

Without sutva there are lots of causal effects.

• Here we focus on exact tests whether these spillovers are present and detectable.

• Ultimately tests are not that interesting on their own, but it demonstrates that the randomization allows researcher to detect these effects.

2.2 Three Null Hypotheses of Interest

No treatment effects:
$Y_i(\mathbf{w}) = Y_i(\mathbf{w}')$ for all units i, and all pairs of assignments $\mathbf{w}, \mathbf{w}' \in \mathbb{W}$.
(straightforward because this hypothesis is sharp)

No spillover effects: (but own treatment effects)
$Y_i(\mathbf{w}) = Y_i(\mathbf{w}')$ for all units i, and all pairs of assignment vectors $\mathbf{w}, \mathbf{w}' \in \mathbb{W}$ such that $w_i = w_i'$.

No higher order effects: (but effects of own treatment and friends' treatment)
$Y_i(\mathbf{w}) = Y_i(\mathbf{w}')$ for all units i, and for all pairs of assignment vectors $\mathbf{w}, \mathbf{w}' \in \mathbb{W}$ such that $w_j = w_j'$ for all units j such that $d(i,j) < 2$ (distance in network).

Problem with second and third null hypothesis is that they are not sharp:

We see one pair $(\mathbf{w}, \mathbf{Y}(\mathbf{w}))$, based on that we cannot infer the value of $\mathbf{Y}(\mathbf{w}')$ for some other $\mathbf{w}'$ under the null.

Using the standard approach, we cannot calculate exact p-values without that.

Aronow (2012) shows how to calculate p-values for second hypothesis.

We develop general way to deal with hypotheses like the third one, as well as others.

2.3 What Not To Do

Bond et al (2012) focus on the statistic that averages ego outcomes over friendships with treated alter and control alter:

$$T = \frac{1}{N_1} \sum_{i,j:G_{ij}=1} Y_i^{\text{obs}} \cdot W_j - \frac{1}{N_0} \sum_{i,j:G_{ij}=1} Y_i^{\text{obs}} \cdot (1 - W_j)$$

They then calculate p-values **as if** the null hypothesis is that of no treatment effects whatsoever (which is sharp so they can evaluate the distribution of the statistic).

This is not valid: the rejection rates under the null for a 5% test can be much larger than 5%.

Randomization inference:

Specify sharp null hypothesis of direct effects, no spillovers:

$$Y_i(w_i, w) = Y_i(0, 0) + \tau w_i,$$

Possible alternative hypothesis that is also sharp (for some value of β):

$$Y_i(w_i, w) = Y_i(0, 0) + \tau w_i + \tilde{\mathbf{A}}_i' w \beta,$$

2.4 Artificial Experiments

Think of an experiment $\mathcal{E}$ as a combination of a set of treatment values $\mathbb{W}$, a population $\mathbb{P}$ of units characterized by potential outcomes, and a distribution of assignments, $p : \mathbb{W} \mapsto (0, 1)$.

We will analyze a different, artificial, experiment.

Take a subset of units $\mathbb{P}_F$, the *focal* units. We will only use outcome data for these individuals.

Now for individual i in the focal group, given the actual treatment $\mathbf{W}$, figure out the set of treatments $\mathbb{W}_i(\mathbf{W}, H_0)$ that would lead to the same outcome under the null outcome.

If the null is no effect of the treatment whatsoever, then $\mathbb{W}_i(\mathbf{W}, H_0) = \mathbb{W}$. If H_0 allows for own treatment effects, but no spillovers, $\mathbb{W}_i(\mathbf{W}, H_0) = \{\mathbf{w} \in \mathbb{W} | w_i = W_i\}$.

Take the union over all focal individuals:

$$\mathbb{W}_R = \cup_{i \in \mathbb{P}_F} \mathbb{W}_i(\mathbf{W}, H_0).$$

The new assignment probability is

$$p'(\mathbf{w}) = \frac{p(\mathbf{w})}{\sum_{\mathbf{w}' \in \mathbb{W}_R} p(\mathbf{w}')}$$

Analyse the experiment $\mathcal{E}' = (\mathbb{W}_R, \mathbb{P}_F, p'(\cdot))$

This artificial experiment is **valid**, because of the randomization underlying $\mathcal{E}$, and the null hypothesis is now **sharp**.

Example I H_0 is no treatment effect whatsoever. $\mathbb{P}_F = \mathbb{P}$, $\mathcal{E}' = (\mathbb{W}, \mathbb{P}, p(\cdot)) = \mathcal{E}$.

Example II H_0 is no spillover effects. Choose $\mathbb{P}_F \subset \mathbb{P}$, arbitrarily. Then the restricted set of assignments is

$$\mathbb{W}_R(\mathbf{W}, \mathbb{P}_F) = \{\mathbf{w} \in \mathbb{W} | w_i = W_i \text{ for all } i \in \mathbb{P}_F\},$$

with assignment probabilities

$$p'(\mathbf{w}) = \frac{p(\mathbf{w})}{\sum_{\mathbf{w}' \in \mathbb{W}_R} p(\mathbf{w}')}$$

The artificial experiment is

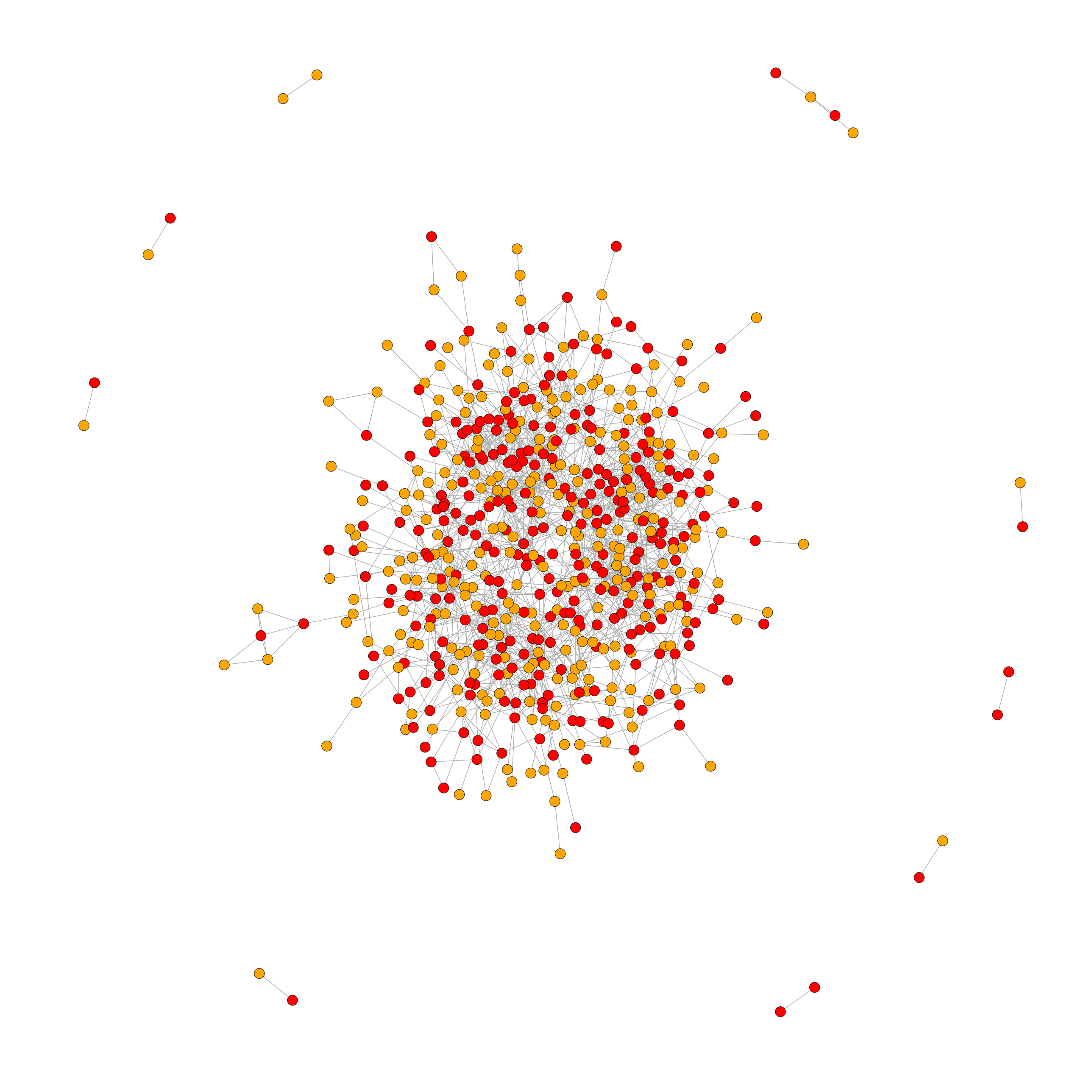
$$\mathcal{E}' = (\mathbb{W}_R, \mathbb{P}_F, p'(\cdot))$$

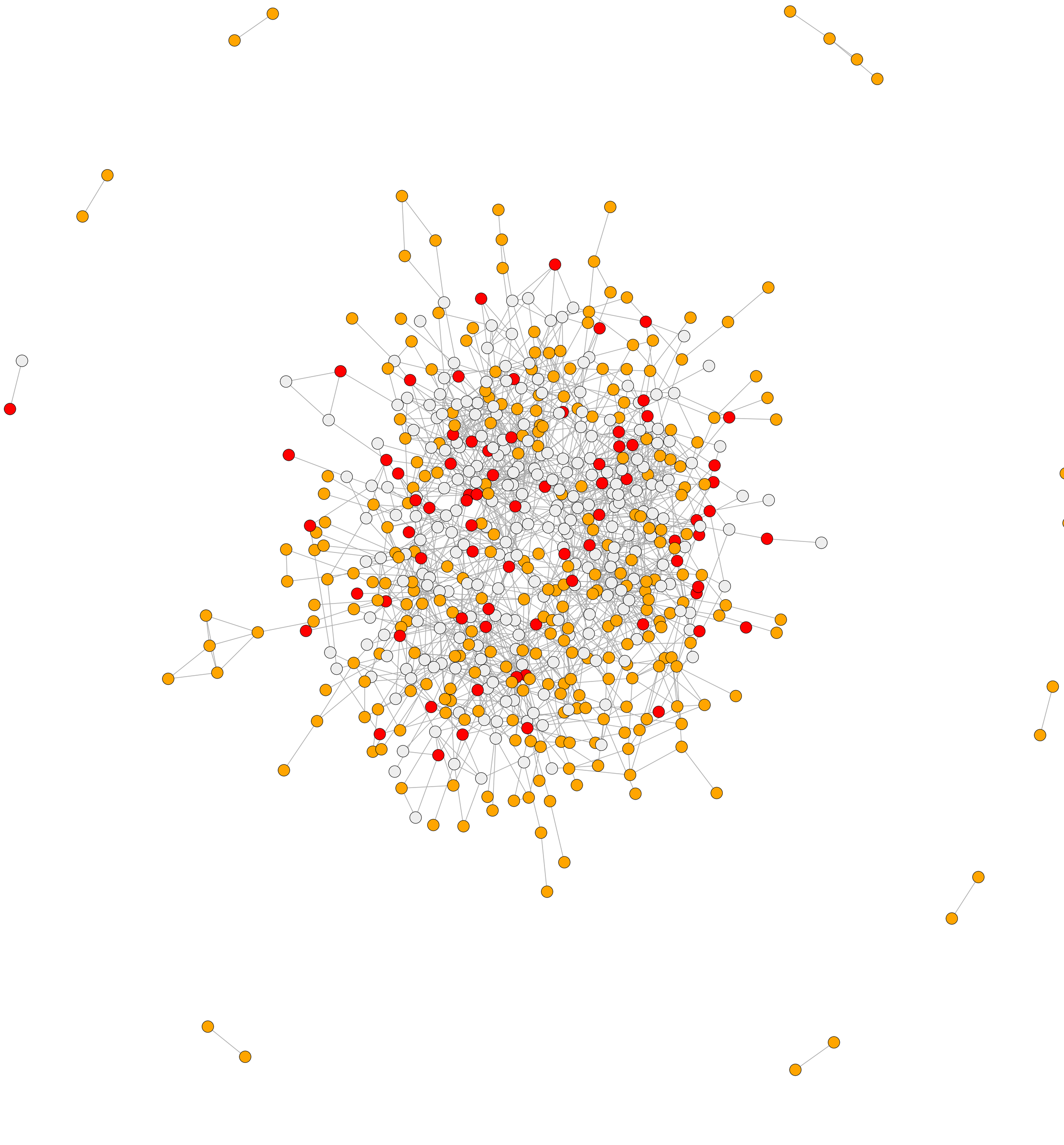
The artificial experiment takes a set of focal units, and looks at randomization distribution induced by changing assignment for a set of auxiliary units.

• for conventional null of no effects, set of auxiliary units is identical to set of focal units

• in aronow case of test for no spillovers population partitions in set of focal units and set of auxiliary units

• with high order spillover null, the set of auxiliary units is a subset of the complement of set of focal units: the population partitions into set of focal units, set of auxiliary units, and the rest.





Statistics

Any statistic $T : \mathbb{W}_R \times \mathbb{Y}^N \mapsto \mathbb{R}$ is valid as statistic.

Edge-level-contrast: average outcome for focal egos with non-focal treated alters minus average outcome for focal egos with non-focal control alters (where F_i is indicator for being focal $i \in \mathbb{P}_F$):

$$T = \frac{1}{N_1} \sum_{i,j,G_{ij}=1} Y_i^{\text{obs}} \cdot F_i \cdot (1 - F_j) \cdot W_j$$

$$-\frac{1}{N_0} \sum_{i,j,G_{ij}=1} Y_i^{\text{obs}} \cdot F_i \cdot (1 - F_j) \cdot (1 - W_j)$$

Score statistic based on linear model

$$Y_i^{\text{obs}} = \alpha_0 + \alpha_w \cdot W_i + \alpha_y \cdot \overline{Y}_{(i)}^{\text{obs}} + \varepsilon_i$$

$\overline{\mathbf{G}}$ is the row-normalized adjacency matrix

$$T_{\text{score}} = \frac{1}{N_F} \sum_{i \in \mathbb{P}_F} \left\{ \left(Y_i^{\text{obs}} - \overline{Y}_{F,0}^{\text{obs}} - W_i \cdot \left(\overline{Y}_{F,1}^{\text{obs}} - \overline{Y}_{F,0}^{\text{obs}} \right) \right) \right.$$

$$\left. \times \sum_{j=1}^{N} \left(\overline{G}_{ij} \cdot W_j - \overline{\overline{G} \cdot W} \right) \right\}$$

T_A is average of indicator of having at least one treated friend.

2.5 Some Simulations

We took a network of high school friends from AddHealth (599) individuals.

$$Y_i(\mathbf{w}_0) \sim \mathcal{N}(0,1), \quad \text{independent across all units}$$

$$Y_i(\mathbf{w}) = Y_i(\mathbf{w}_0) + w_i \cdot \tau_{\text{direct}} + \frac{K_{i,1}}{K_i} \cdot \tau_{\text{spill}}.$$

K_i and $K_{i,1}$ are number of friends and number of treated friends.

Focal units are selected at random, to maximize number of contrasts between focal and auxiliary units, or based on epsilon nets.

Network	Statistic	Own Effect	Spillover Effect	Focal Node Selection: Random	ε-net	$\delta_{N,i}$
AddHealth	T_{score}	0	0	0.059	0.056	0.045
	T_{elc}	0	0	0.058	0.054	0.044
	T_A	0	0	0.059	0.039	0.046
	T_{score}	4	0	0.056	0.053	0.051
	T_{elc}	4	0	0.051	0.048	0.059
	T_A	4	0	0.050	0.053	0.051
	T_{score}	0	0.4	0.362	0.463	0.527
	T_{elc}	0	0.4	0.174	0.299	0.413
	T_A	0	0.4	0.141	0.296	0.327
	T_{score}	4	0.4	0.346	0.461	0.529
	T_{elc}	4	0.4	0.083	0.102	0.123
	T_A	4	0.4	0.069	0.088	0.116

- Also looked at test for second order spillover effect.

- There power may be very low.

- lots of design questions: proportion of focal individuals, distribution of focal individuals through network.

Rejection Rates of Null Hypothesis of No Spillovers
Beyond the First Order Spillovers from the Sparsified Network

Network	Statistic	α_w	α_{spill}	λ	Proportion of Links Dropped $q = 0.9$	$q = 0.5$
AddHealth	T_{corr}	4	0.4	0	0.047	0.046
	T_{elc}	4	0.4	0	0.048	
	T_{corr}	4	0.1	0	0.050	0.049
	T_{elc}	4	0.1	0	0.046	
	T_{corr}	4	0.4	0.5	0.216	0.120
	T_{elc}	4	0.4	0.5	0.059	
	T_{corr}	0	0.4	0.5		
	T_{elc}	0	0.4	0.5	0.123	0.087
	T_{corr}	4	0.1	0.5	0.059	0.061

3. Multi-armed Bandits

• In many cases we wish to evaluate multiple treatments: putting the button on the left or on the right, making it green or red, making it big or small.

• We could run experiments with multiple treatments and test various null hypotheses.

• This is cumbersome, and not effective for answering the question: which is the best treatment out of a set.

Suppose there are K treatments, with binary outcomes $Y_i \sim \mathcal{B}(1, p_k)$.

We are interested in identifying the treatment arm k with the highest value of p_k.

Suppose we start by observing 100 draws for each arm, and get $\widehat{p}_k$ for each arm. Then our best guess is the arm with the highest $\widehat{p}_k$.

Now suppose we have the opportunity to allocate another 1000 units to these K treatment arms, how should we do that?

E.g., $\widehat{p}_1 = 0.10$, $\widehat{p}_2 = 0.80$, $\widehat{p}_3 = 0.81$, $\widehat{p}_4 = 0.70$

Allocating a lot of units to treatment arm 1 does not serve much of a purpose: it is unlikely that arm 1 is the best arm.

To learn about the optimal arm, we should assign more units to treatment arms 2, 3 and 4.

But: how many units to each?

Should we assign a lot to arm 4?

Thompson Sampling and Upper Confidence Bound Methods

Two approaches to determining assignment for next unit.

In both cases we assign more units to arms that look promising, in slightly different ways.

1. Thompson sampling: calculate posterior probability that arm k is the optimal arm, and assign to this arm with probability proportional to that.

2. Upper Confidence Bound method. Calculate confidence intervals for each p_k, with confidence level α_N (N is the total sample size so far, $\alpha_N \to 1$ as $N \to \infty$.

Thompson Sampling

• Calculate the posterior distribution of $p_1, \ldots, p_K$ given prior (say flat prior). Easy here because these are Beta posterior distributions. In other cases this may require some numerical approximations.

• Allocate to each arm proportional to the probability that $p_k = \max_{m=1}^{K} p_m$. Easy to implement by drawing p_k from its posterior for each k, and then assign to arm with highest p_k.

This balances **exploration**: learn more about the arms by allocating units to them, and **exploitation**: send more units to arms that do well.

In example, arm 1 does very poorly, dont send more units to that arm. We are not sure about the other arms, so we send units to all of them, but more to 2 and 3 than to 4.

Very effective way to experiment in settings with many treatments, and with sequential assignment.

Consider a case with two arms, and $p_1 = 0.04$, $p_2 = 0.05$.

Consider a classical experiment with testing at the 0.05 level, for 95% power.

We need 22,000 observations for this.

The regret is $11000 \times (0.05 - 0.04) = 111$.

Suppose we get 100 observations per day, the experiment will take 220 days.

How bad is equal allocation?

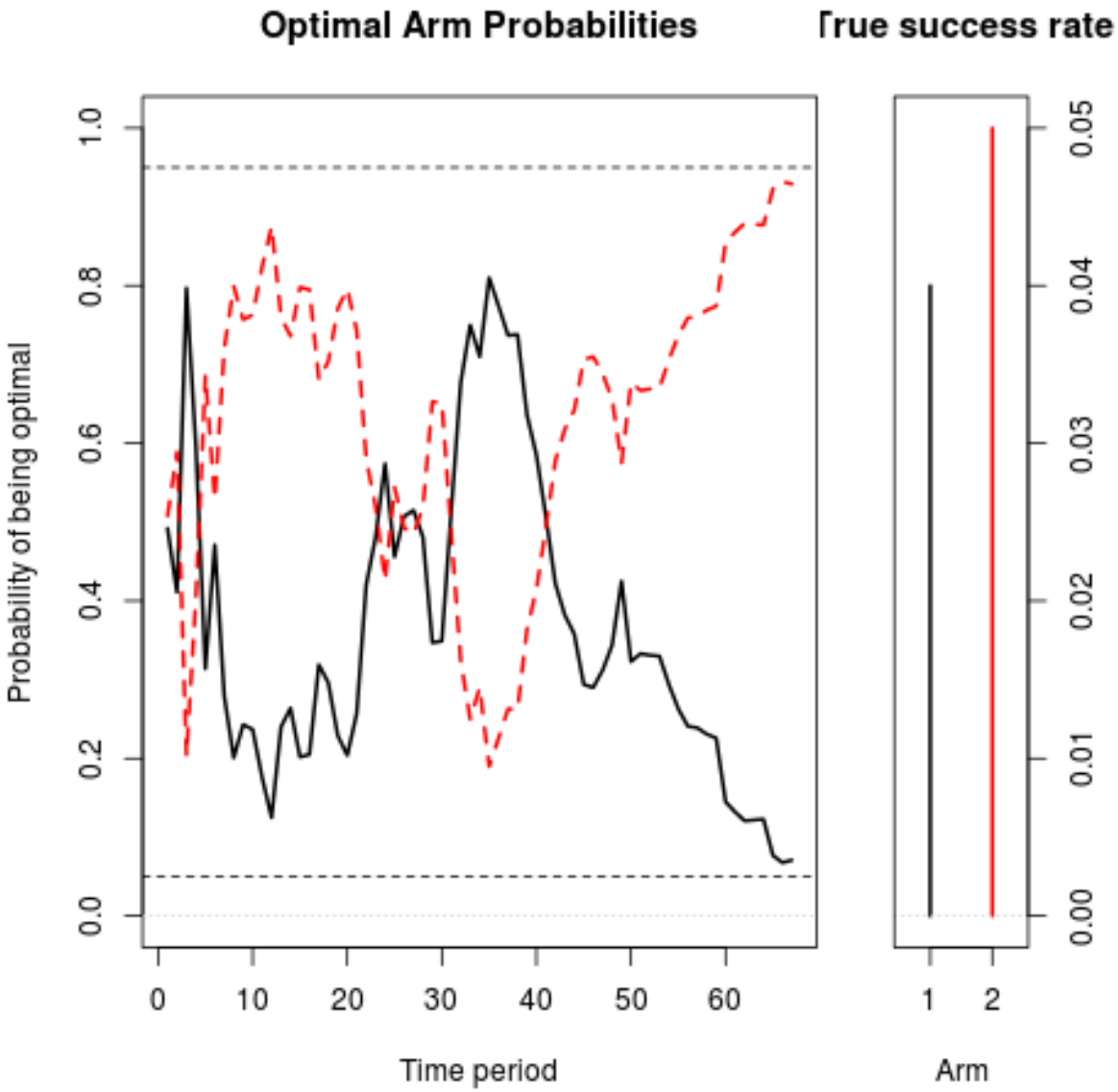
- Consider two arms: $\theta_1 = .04$, and $\theta_2 = .05$.
- Plan a classical experiment to detect this change with 95% power at 5% significance.

```
> power.prop.test(p1 = .04, p2 = .05, power = .95)
              n = 11165.99
NOTE: n is number in *each* group
```

- We need over 22,000 observations.
- Regret is $11,165 \times .01 = 111$ lost conversions.
- At 100 visits per day, the experiment will take over 220 days.

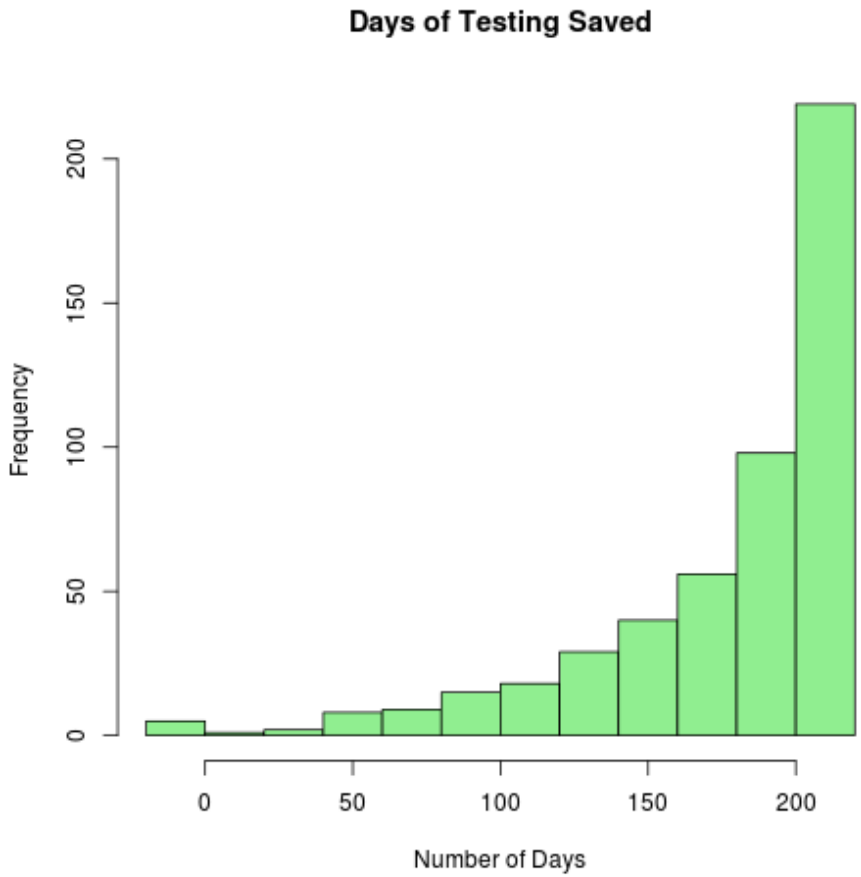
Two-armed experiment

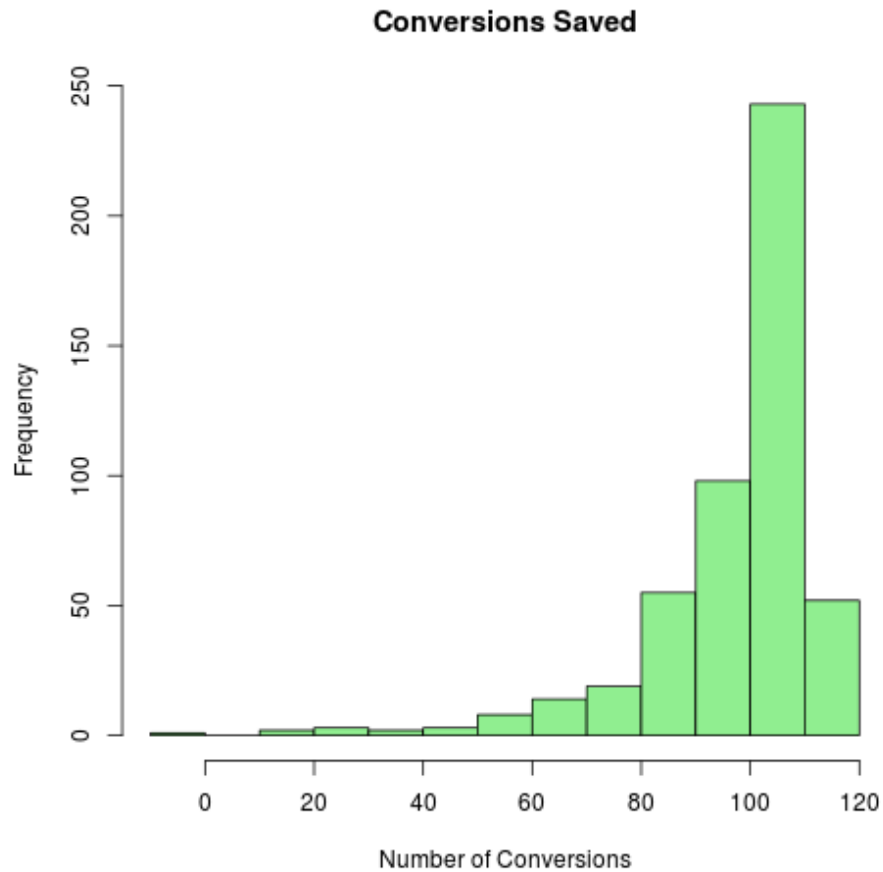
Bandit shown 100 visits per day



Two armed experiment

Savings vs equal allocation in terms of time and conversions





Source: https://support.google.com/analytics/answer/2844870?hl=en

Bandits' advantage grows with experiment size

Now consider 6 arms (formerly the limit of GA Content Experiments).

- Compare the original arm to the "best" competitor.
- Bonferroni correction says divide significance level by 5.

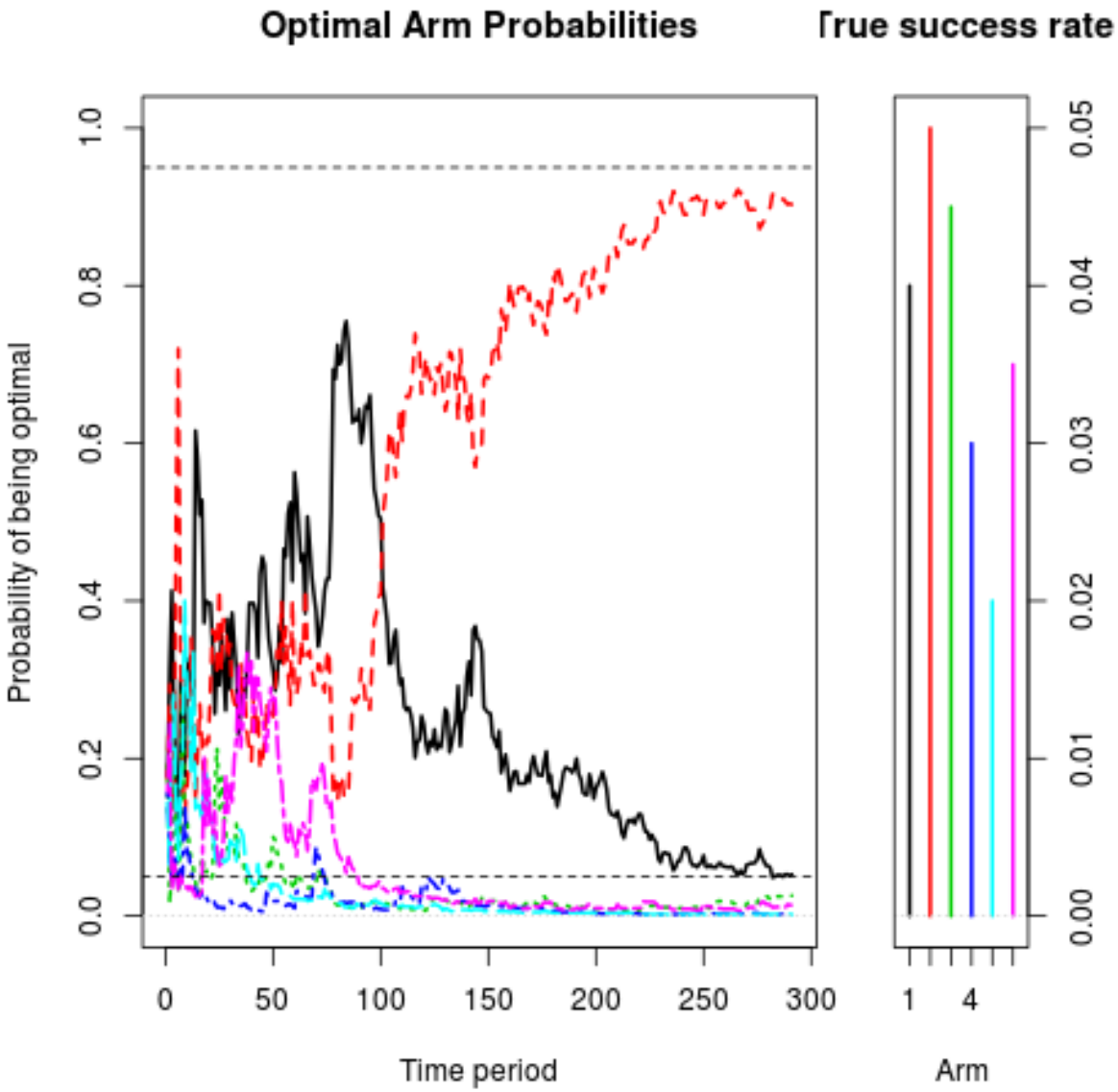
```
> power.prop.test(p1 = .04, p2 = .05, power = .95,
                  sig.level=.01)
             n = 15307.8
NOTE: n is number in *each* group
```

- In theory we only need this sample size in the largest arm, but we don't know ahead of time which arm that will be.
- Experiment needs 91848 observations.
- At 100 per day that is 2.5 years.

Google

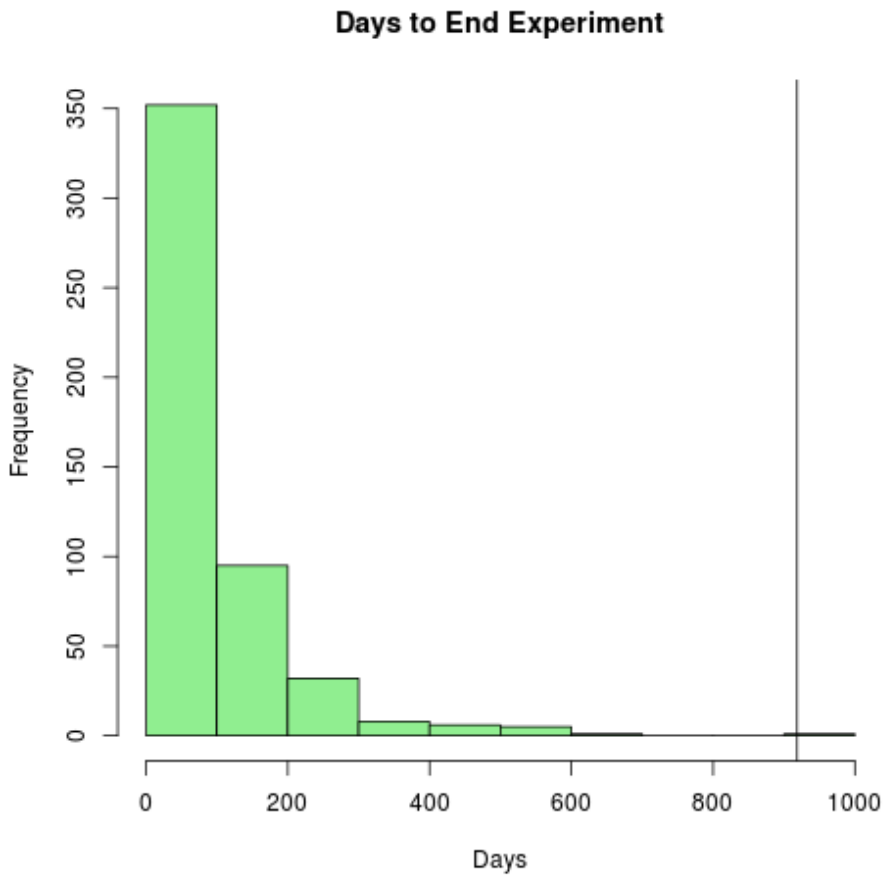
6-arm experiment

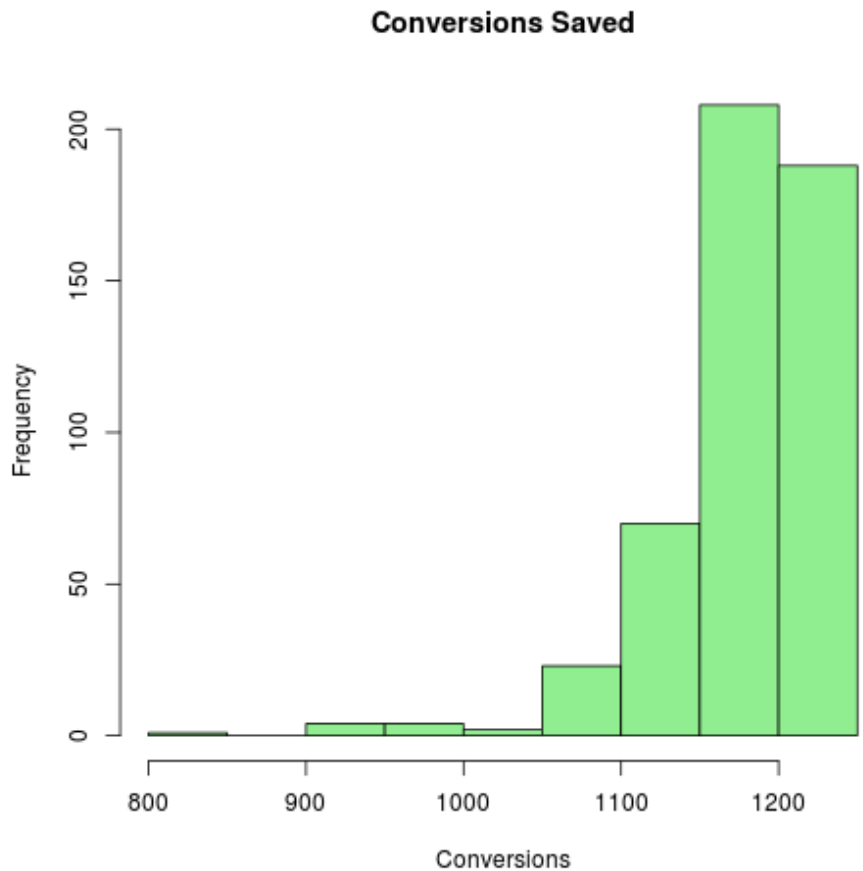
Still 100 observations per day



Huge savings vs equal allocation

Partly due to ending early, and partly due to lower cost per day.





Now suppose we have 6 arms, best arm is 0.05, second best is 0.04. We now need to test each comparison at 1% level for Bonferroni correction because we do 5 tests.

Need 90,000 observations, will take 2.5 years.

Huge savings!

Upper Confidence Bounds

Construct confidence bound for p_k with confidence level α_N. Let α_N go to 1 slowly.

Pick arm with the highest upper confidence limit, and assign next unit to that arm.

• if that is a poor arm, the upper confidence bound will shrink relative to the others, and it will get less traffic subsequently.

Contextual Bandits

Suppose we also have covariates X_i for each unit, and want to find the function that assigns each unit to the treatment with the highest expected return as a function of the covariates.

• Given a parametric model for the expected return, we can directly use Thompson sampling.

• We may wish to build increasingly flexible models to avoid basing assignment on a misspecified model $\Longrightarrow$ can use random forests, but need to account for variation in propensity score.

Causal Inference

and Machine Learning

Guido Imbens – Stanford University

Lecture 6:

Synthetic Control and Matrix Completion Methods

Potsdam Center for Quantitative Research

Wednesday September 11th, 10.00-11.30

Based on:

Doudchenko, Nikolay, and Guido W. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. No. w22791. National Bureau of Economic Research, 2016.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. Synthetic difference in differences, 2019.

Athey, Susan, Mohsen Bayati, Mohsen, Nick Doudchenko, Guido Imbens, and Khashayar Khosravi, (2018). Matrix completion methods for causal panel data models.

• California's anti-smoking legislation (Proposition 99) took effect in 1989.

• **What is the causal effect of the legislation on smoking rates in California in 1989?**

• We **observe** smoking rates in California in 1989 given the legislation. We need to **impute** the **counterfactual** smoking rates in California in 1989 had the legislation not been enacted.

• We have data in the absence of smoking legislation in California prior to 1989, and for other states both before and in 1989. (and other variables, but not of essence)

Set Up: we observe (in addition to covariates):

$$\mathbf{Y} = \begin{pmatrix} Y_{11} & Y_{12} & Y_{13} & \dots & Y_{1T} \\ Y_{21} & Y_{22} & Y_{23} & \dots & Y_{2T} \\ Y_{31} & Y_{32} & Y_{33} & \dots & Y_{3T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{N1} & Y_{N2} & Y_{N3} & \dots & Y_{NT} \end{pmatrix} \quad \text{(realized outcome).}$$

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & 0 & \dots & 1 & 1 \end{pmatrix} \quad \text{(binary treatment).}$$

• rows of $\mathbf{Y}$ and $\mathbf{W}$ correspond to units (e.g., states), columns correspond to time periods (years).

In terms of potential outcome matrices $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$:

$$\mathbf{Y}(0) = \begin{pmatrix} \checkmark & \checkmark & \dots & \checkmark & \checkmark \\ \checkmark & \checkmark & \dots & \checkmark & \checkmark \\ \checkmark & \checkmark & \dots & \checkmark & \checkmark \\ \vdots & \vdots & \ddots & \vdots & \\ \checkmark & \checkmark & \dots & ? & ? \\ \checkmark & \checkmark & \dots & ? & ? \end{pmatrix} \qquad \mathbf{Y}(1) = \begin{pmatrix} ? & ? & \dots & ? & ? \\ ? & ? & \dots & ? & ? \\ ? & ? & \dots & ? & ? \\ \vdots & \vdots & \ddots & \vdots & \\ ? & ? & \dots & \checkmark & \checkmark \\ ? & ? & \dots & \checkmark & \checkmark \end{pmatrix}.$$

$Y_{it} = (1 - W_{it})Y_{it}(0) + W_{it}Y_{it}.$

In order to estimate the average treatment effect for the treated, (or other average, e.g., overall average effect)

$$\tau = \frac{\sum_{i,t} W_{it}\Big(Y_{it}(1) - Y_{it}(0)\Big)}{\sum_{it} W_{it}},$$

we **impute** the missing potential outcomes in $\mathbf{Y}(0)$.

Alternative Possible Structures on $\mathbf{W}$:

Staggered adoption (e.g., adoption of technology, Athey and Stern, 1998)

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 & \text{(never adopter)} \\ 0 & 0 & 0 & 0 & \dots & 1 & \text{(late adopter)} \\ 0 & 0 & 0 & 0 & \dots & 1 & \\ 0 & 0 & 1 & 1 & \dots & 1 & \\ 0 & 0 & 1 & 1 & \dots & 1 & \text{(medium adopter)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 1 & 1 & 1 & \dots & 1 & \text{(early adopter)} \end{pmatrix}$$

Part of the talk I will focus on case with a single treated unit/time-period

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

Challenge:

Trying to predict $Y_{NT}(0)$ based on observed values $Y_{it}(0)$ for $(i,t) \neq (N,T)$.

In empirical studies there is a wide range of values for

- N_0, the number of control units
- N_1, the number of treated units
- T_0, the number of pre-treatment periods
- T_1, the number of post-treatment periods

This is important for guiding choice of analyses.

1. Mariel Boatlift Study (Card,1990), $N_1 = 1, N_0 = 44, T_0 = 7, T_1 = 6$

2. Minimum wage study (Card-Krueger 1994), $N_1 = 321, N_0 = 78, T_0 = 1, T_1 = 1$

3. California smoking example (Abadie, Diamond, Hainmueller, 2010) $N_1 = 1, N_0 = 29, T_0 = 17, T_1 = 13$

4. German unification (Abadie, Diamond, Hainmueller, 2014) $N_1 = 1, N_0 = 16, T_0 = 30, T_1 = 14$

5. Lalonde study (1986) $N_1 = 185, N_0 = 15992, T_0 = 2, T_1 = 1$

Three related literatures on causal inference for this setting:

1. causal literature with unconfoundedness / horizontal regression

2. synthetic control literature / vertical regression

3. difference-in-differences and factor models

Here: **doubly robust** methods that combine **weighting** and **outcome modeling**

Unconfoundedness Methods / Horizontal Regression

Typical setting: N_0 and N_1 large, T_0 modest, $T_1 = 1$.

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Linear Model

$$\widehat{\tau}_{\mathsf{UNC}} = \frac{1}{N_1} \sum_{i:W_{iT}=1} \Big(Y_{iT}(1) - \widehat{Y}_{iT}(0) \Big)$$

where

$$\widehat{Y}_{iT}(0) = \widehat{\alpha} + \sum_{t=1}^{T-1} \widehat{\lambda}_t Y_{it}$$

and $\widehat{\alpha}$ and $\widehat{\lambda}$ are estimated by least squares:

$$\min_{\alpha,\lambda} \sum_{i=1}^{N_0-1} \left(Y_{iT} - \alpha - \sum_{t=1}^{T-1} \lambda_t Y_{it} \right)^2 \qquad \text{"horizontal" regression}$$

Note: regression with N_0 observations, and T_0 regressors. May need regularization if T_0 is big.

Fancier methods:

Matching: for each treated unit i with $W_{iT} = 1$, find the closest match $j(i)$:

$$j(i) = \arg \min_{j:W_j=0} \left\| \mathbf{Y}_{i,1:T_0} - \mathbf{Y}_{j,1:T_0} \right\|$$

Then:

$$\widehat{Y}_{iT}(0) = Y_{j(i),T}$$

Double robust methods:

Estimate the propensity score

$$e(\mathbf{y}) = \mathsf{pr}(W_{iT} = 1 | \mathbf{Y}_{i,1:T_0} = \mathbf{y})$$

Estimate conditional mean for controls:

$$\mu(\mathbf{y}) = \mathbb{E}[Y_{iT} | W_{iT} = 0, \mathbf{Y}_{i,1:T_0} = \mathbf{y}]$$

Then for all treated units:

$$\widehat{Y}_{iT}(0) = \mu(\mathbf{Y}_{i,1:T_0}) + \frac{1}{N_0} \sum_{j:W_{jT}=0} \frac{e(X_j)}{1 - e(X_j)} \Big(Y_{jT} - \mu(\mathbf{Y}_{i,1:T_0}) \Big)$$

Abadie-Diamond-Hainmueller Synthetic Control Method

Typical setting: T_0 and T_1 modest, N_0 small, $N_1 = 1$.

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 1 & \dots & 1 \end{pmatrix}$$

For simplicity focus on case with $T_1 = 1$, $T_0 = T - 1$.

ADH suggest using a weighted average of outcomes for other states:

$$\widehat{Y}_{NT}(0) = \sum_{j=1}^{N-1} \omega_i Y_{jT}$$

ADH restrict the weights ω_j to be non-negative, and restrict them to sum to one.

- $\mathbf{Y}_{i,1:T_0}$ is the lagged values $Y_{i,t}$ for $t \leq T_0$.

- X_i are other covariates, unit specific.

Let Z_i be the vector of functions of covariates X_i, including possibly some lagged outcomes Y_{it} for pre-T periods. Let the norm be $\|a\|_{\mathbf{V}} = a'\mathbf{V}^{-1}a$, for positive semi-definite square matrix $\mathbf{V}$.

ADH first solve, for given $\mathbf{V}$

$$\omega(\mathbf{V}) = \arg\min_{\omega} \left\| Z_N - \sum_{i=1}^{N-1} \omega_i \cdot Z_i \right\|_{\mathbf{V}}$$

This finds, for a given weight matrix $\mathbf{V}$ the optimal weights ω.

• But: how do we choose $\mathbf{V}$? Equal weights is not right, would not be invariant to linear transformations of covariates.

ADH find the optimal positive semi-definite $\mathbf{V}$ by minimizing

$$\widehat{\mathbf{V}} = \arg\min_{\mathbf{V}} \left\| \mathbf{Y}_{N,1:T_0} - \sum_{i=1}^{N-1} \omega_i(\mathbf{V}) \cdot \mathbf{Y}_{i,1:T_0} \right\|$$

Then they use the optimal weights ω based on that $\widehat{\mathbf{V}}$:

$$\omega^* = \omega(\widehat{\mathbf{V}}) = \arg\min_{\omega} \left\| Z_N - \sum_{i=1}^{N-1} \omega_i \cdot Z_i \right\|_{\widehat{\mathbf{V}}}$$

Doudchenko-Imbens:

$$\widehat{\tau}_{\text{DI}} = Y_{NT} - \widehat{Y}_{NT}(0), \quad \widehat{Y}_{NT}(0) = \alpha + \sum_{i=1}^{N-1} \omega_i Y_{iT}$$

where

$$\min_{\alpha,\omega} \sum_{t=1}^{T-1} \left(Y_{Nt} - \alpha - \sum_{i=1}^{N-1} \omega_i Y_{it} \right)^2 \quad \text{"vertical" regression}$$

Regularization is important here if N is large relative to T, partly because of lack of restrictions on ω

Note: regression with T_0 observations, and N_0 regressors.

Comparison Unconfoundedness vs Synthetic Controls in Case with $N_1 = T_1 = 1$

- Unconfoundedness req. $N_0 > T_0 \Longrightarrow$ horizontal regression

- Synthetic Control requires $N_0 < T_0 \Longrightarrow$ vertical regression

But, with **regularization** on regression coefficients we can use either unconfoundedness or synthetic control methods, irrespective of relative magnitude of N_0 and T_0.

Difference-In-Differences / Factor Models

Model $Y_{it}(0)$:

$$Y_{it}(0) = \alpha_i + \gamma_t + \varepsilon_{it}$$

leading to

$$\min_{\alpha,\gamma} \sum_{i=1}^{N} \sum_{t=1}^{T} (1 - W_{it}) \left(Y_{it} - \gamma_t - \alpha_i\right)^2$$

$$\widehat{\tau} = \frac{1}{N_1 T_1} \sum_{i=N_0+1}^{N} \sum_{t=T_0+1}^{T} Y_{it} - \frac{1}{N_1 T_0} \sum_{i=N_0+1}^{N} \sum_{t=1}^{T_0} Y_{it}$$

$$- \left(\frac{1}{N_0 T_1} \sum_{i=1}^{N_0} \sum_{t=T_0+1}^{T} Y_{it} - \frac{1}{N_0 T_0} \sum_{i=1}^{N_0} \sum_{t=1}^{T_0} Y_{it} \right)$$

More general, factor models:

$$Y_{it}(0) = \sum_{r=1}^{R} \gamma_{tr}\alpha_{ir} + \varepsilon_{it}$$

(Athey, Bayati, Doudchenko, Imbens, Khosravi, 2018)

$$\arg\min_{\alpha,\gamma,\mathbf{L}} \sum_{i=1}^{N}\sum_{t=1}^{T} (1 - W_{it}) \left(Y_{it} - \alpha_i - \gamma_t - L_{it}\right)^2 + \lambda\|\mathbf{L}\|$$

with **nuclear normal** regularization on $\mathbf{L}$ to lead to low rank solution.

- Challenge: How to choose between these methods (vertical/horizontal regression, factor models), or how to tie them together?

- Relative merits of these methods

Comparison of

1. unconfoundedness (horizontal) regression with elastic net regularization (EN-H)

2. synthetic control (vertical) regression with elastic net regularization and no restrictions (EN-V)

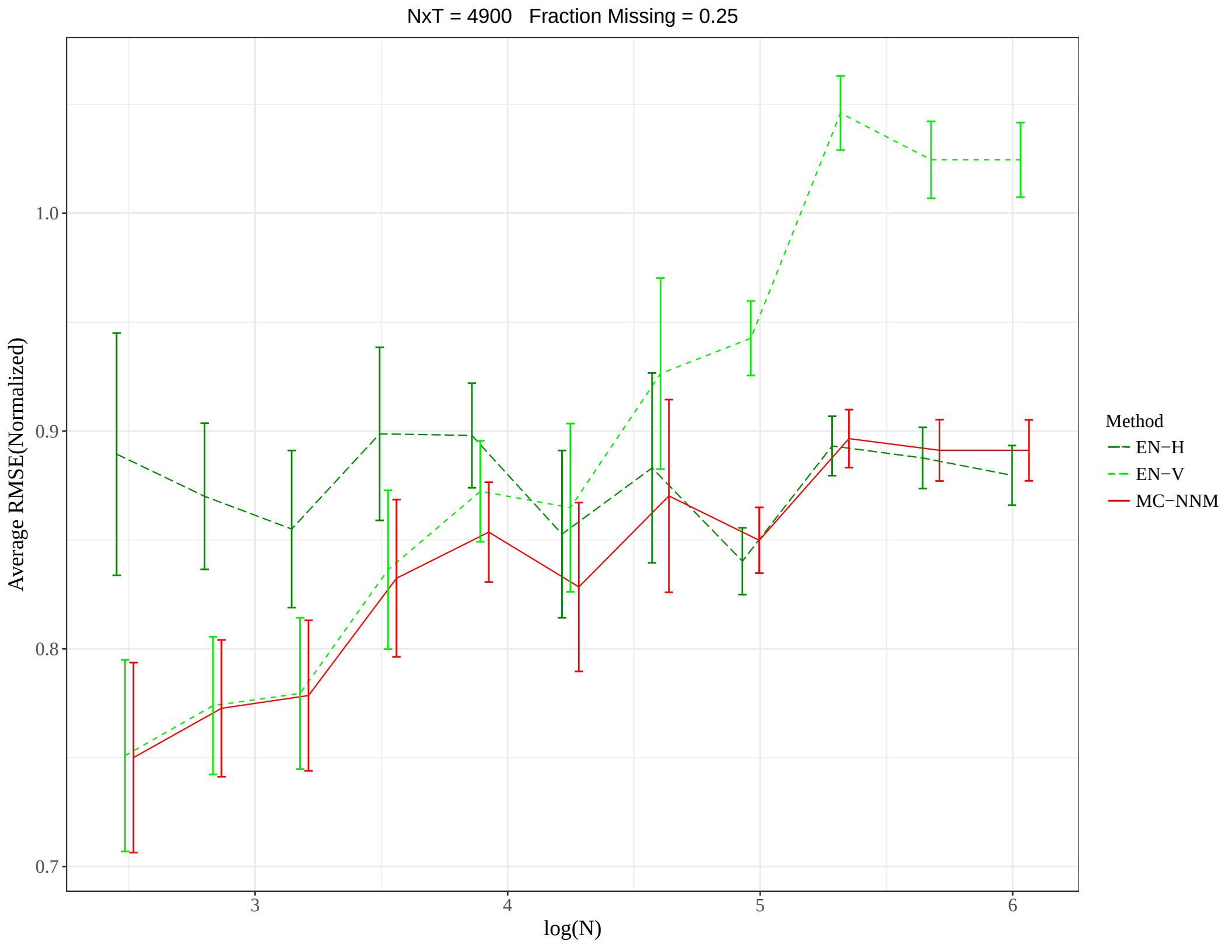
3. matrix completion with nuclear normal (MC-NNM)

Illustration: Stock Market Data

We use daily returns for 2453 stocks over 10 years (3082 days). We create sub-samples by looking at the first T daily returns of N randomly sampled stocks for pairs of (N, T) such that $N \times T = 4900$, ranging from fat to thin:
$(N, T) = (10, 490), \ldots, (70, 70), \ldots, (490, 10)$.

Given the sample, we pretend that half the stocks are treated at the mid point over time, so that 25% of the entries in the matrix are missing.

$$\mathbf{Y}_{N\times T} = \begin{pmatrix} \checkmark & \checkmark & \checkmark & \checkmark & \ldots & \checkmark \\ \checkmark & \checkmark & \checkmark & \checkmark & \ldots & \checkmark \\ \checkmark & \checkmark & \checkmark & \checkmark & \ldots & \checkmark \\ \checkmark & \checkmark & \checkmark & ? & \ldots & ? \\ \checkmark & \checkmark & \checkmark & ? & \ldots & ? \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \checkmark & \checkmark & \checkmark & ? & \ldots & ? \end{pmatrix}$$


NxT = 4900 Fraction Missing = 0.25
Average RMSE(Normalized)
1.0
0.9
0.8
0.7
3
4
5
6
log(N)
Method
EN−H
EN−V
MC−NNM

Results

• MC-NNM does better than EN-H and EN-V, adapts to shape of matrix

• ADH restrictions (non-negativity of weights, and summing to one, and no intercept) sometimes improve things relative to Elastic-Net estimator, more so for the vertical regressions than for the horizontal regressions.

Combining Synthetic Control Methods and Matrix Completion: Observation I

Synthetic Control is weighted linear regression without unit fixed effects:

$$\widehat{\tau}^{\text{ADH}} = \arg\min_{\tau,\gamma} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \gamma_t - \tau W_{it})^2 \times \omega_i^{\text{ADH}}$$

• regression with time fixed effects and ADH weights (easy to include covariates).

• under some conditions standard errors can be based on regression interpretation taking weights as given (even though the weights depend on outcome data).

Combining Synthetic Control Methods and Matrix Completion: Observation II

DID is unweighted regression with unit and time fixed effects:

$$\widehat{\tau}^{\text{DID}} = \arg\min_{\tau,\gamma,\alpha} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \gamma_t - \alpha_i - \tau W_{it})^2$$

• regression with time fixed effects and unit fixed effects, no weights.

Synthetic Difference In Differences

$$\widehat{\tau}^{\text{SDID}} = \arg\min_{\tau,\gamma,\alpha} \sum_{i=1}^{N}\sum_{t=1}^{T} (Y_{it} - \gamma_t - \alpha_i - \tau W_{it})^2 \times \omega_i^{\text{ADH}} \times \lambda_t^{\text{ADH}}$$

Regression with unit and time fixed effects, and with unit and time weights.
Time weights satisfy:

$$\lambda = \arg\min_{\lambda} \sum_{i=1}^{N-1} \left(Y_{iT} - \sum_{t=1}^{T-1} \lambda_t Y_{it} \right)^2 + \text{regularization term},$$

subject to

$$\lambda_t \geq 0, \quad \sum_{t=1}^{T-1} \lambda_t = 1.$$

(or down-weight observations from distant past.)

Generalization: Synthetic Factor Models (SFM)

$$\widehat{\tau}^{\mathsf{SFM}} =$$

$$\arg \min_{\mathbf{L},\alpha,\gamma,\tau} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \alpha_i - \gamma_t - L_{it} - \tau W_{it})^2 \, \omega_i^{\mathsf{ADH}} \lambda_t^{\mathsf{ADH}}$$

$$+\lambda \|\mathbf{L}\|,$$

Double Robustness

- If a factor model holds, but the weights are good (e.g., ADH weights), SDID is consistent.

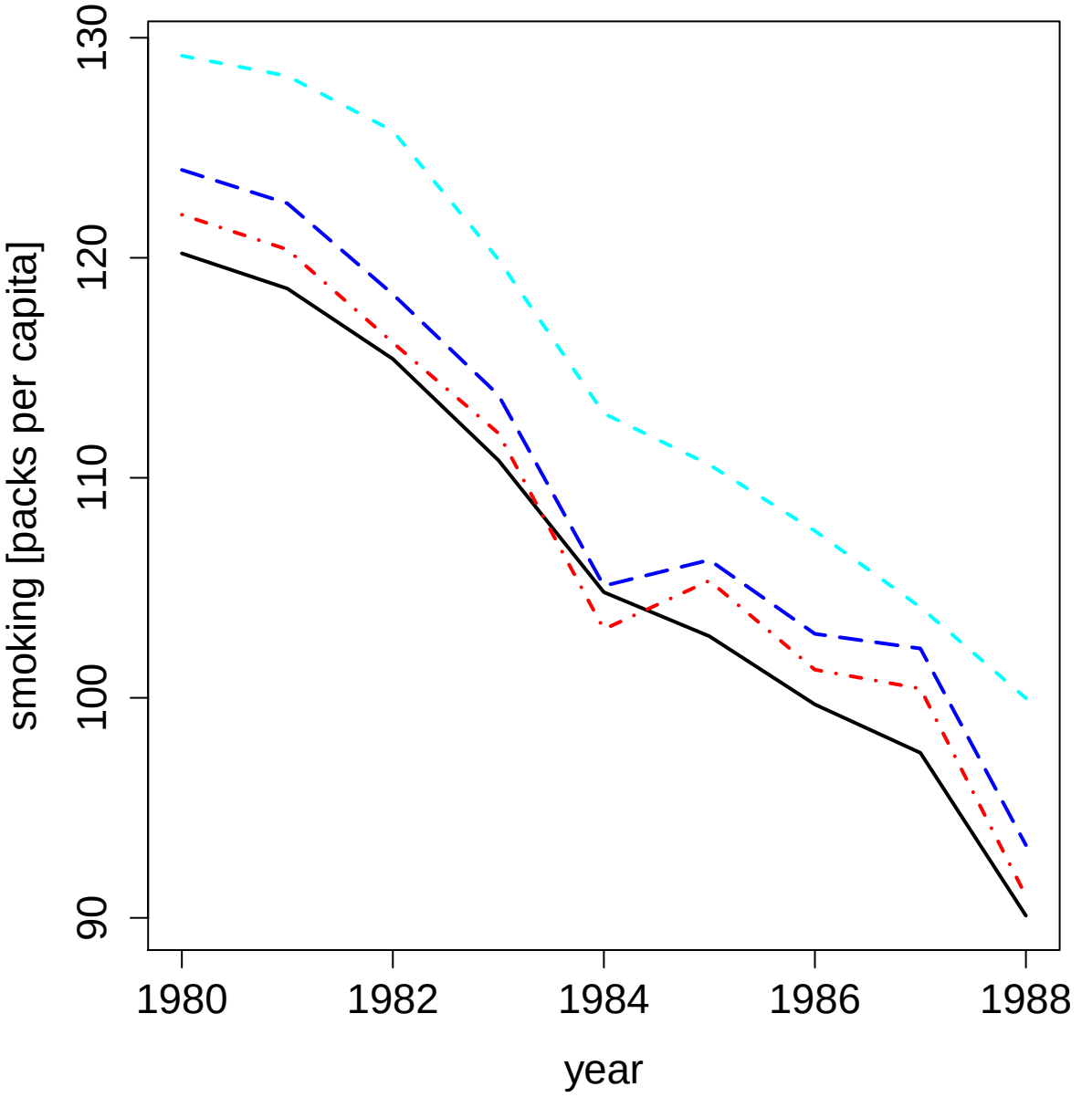
- If the DID model holds, but we use arbitrary weights, SDID is consistent.

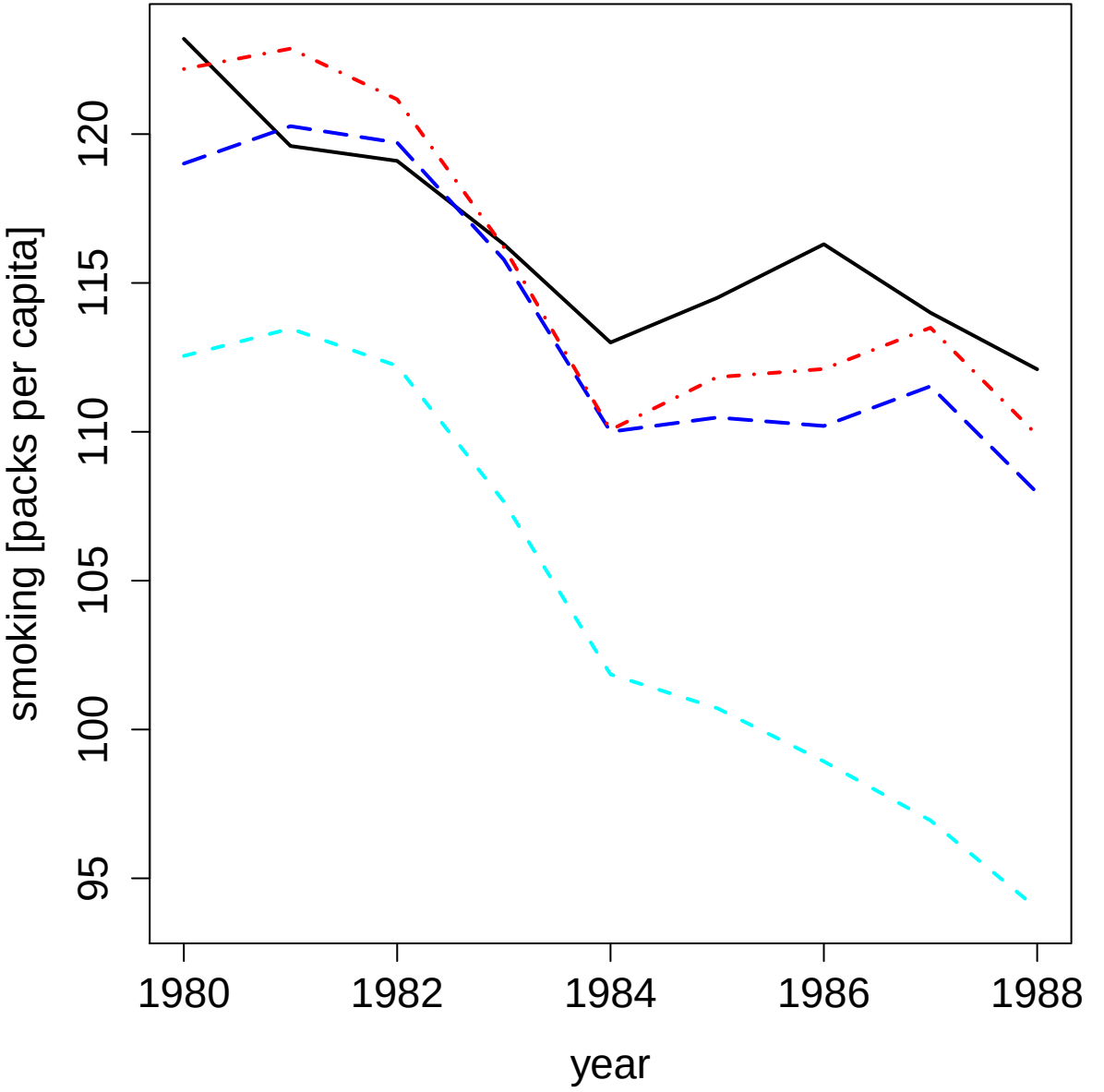
California smoking data calculations

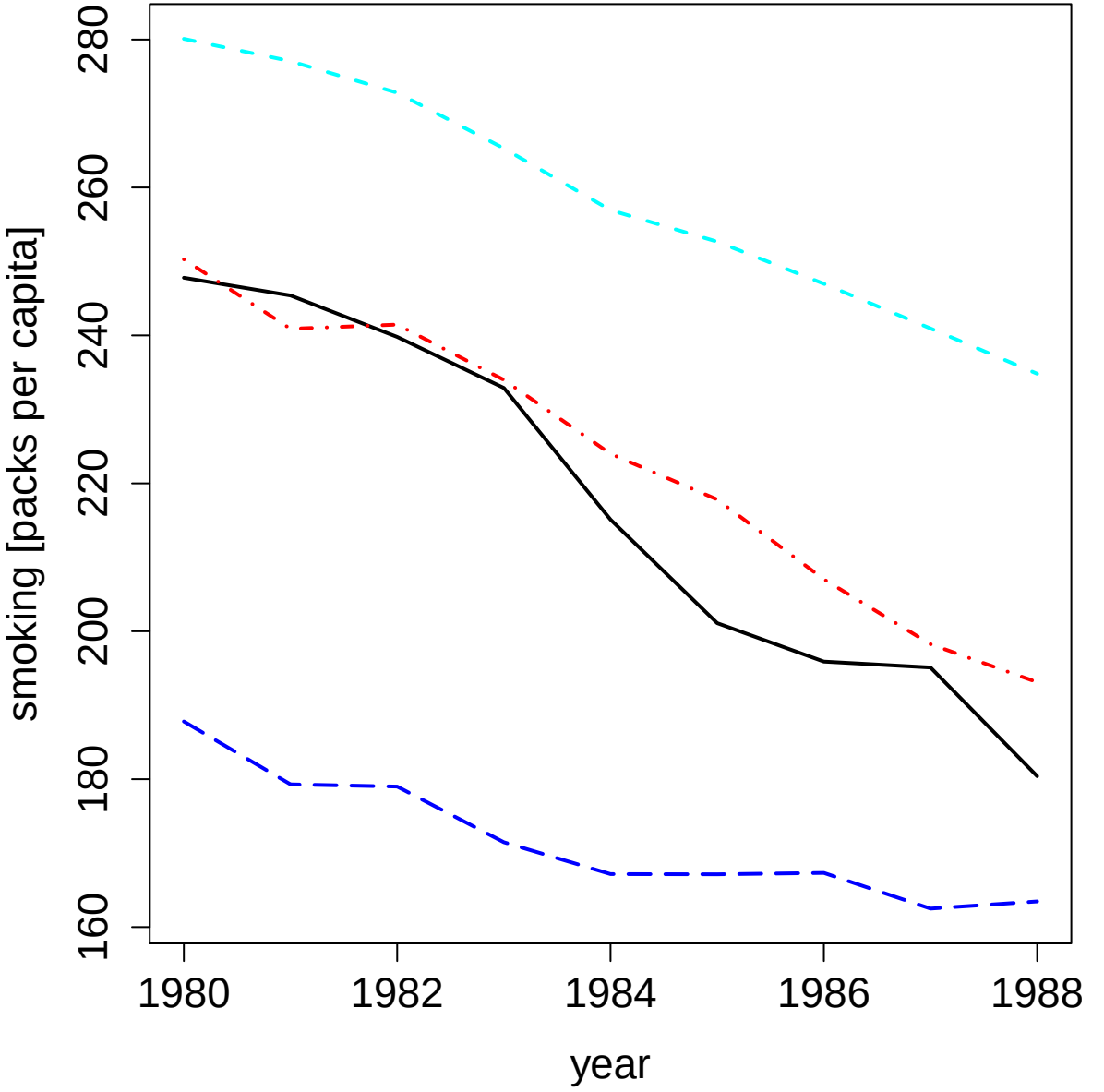
Take pre-1988 data for all states, so we observe all $Y_{it}(0)$ for all unit/time pairs.

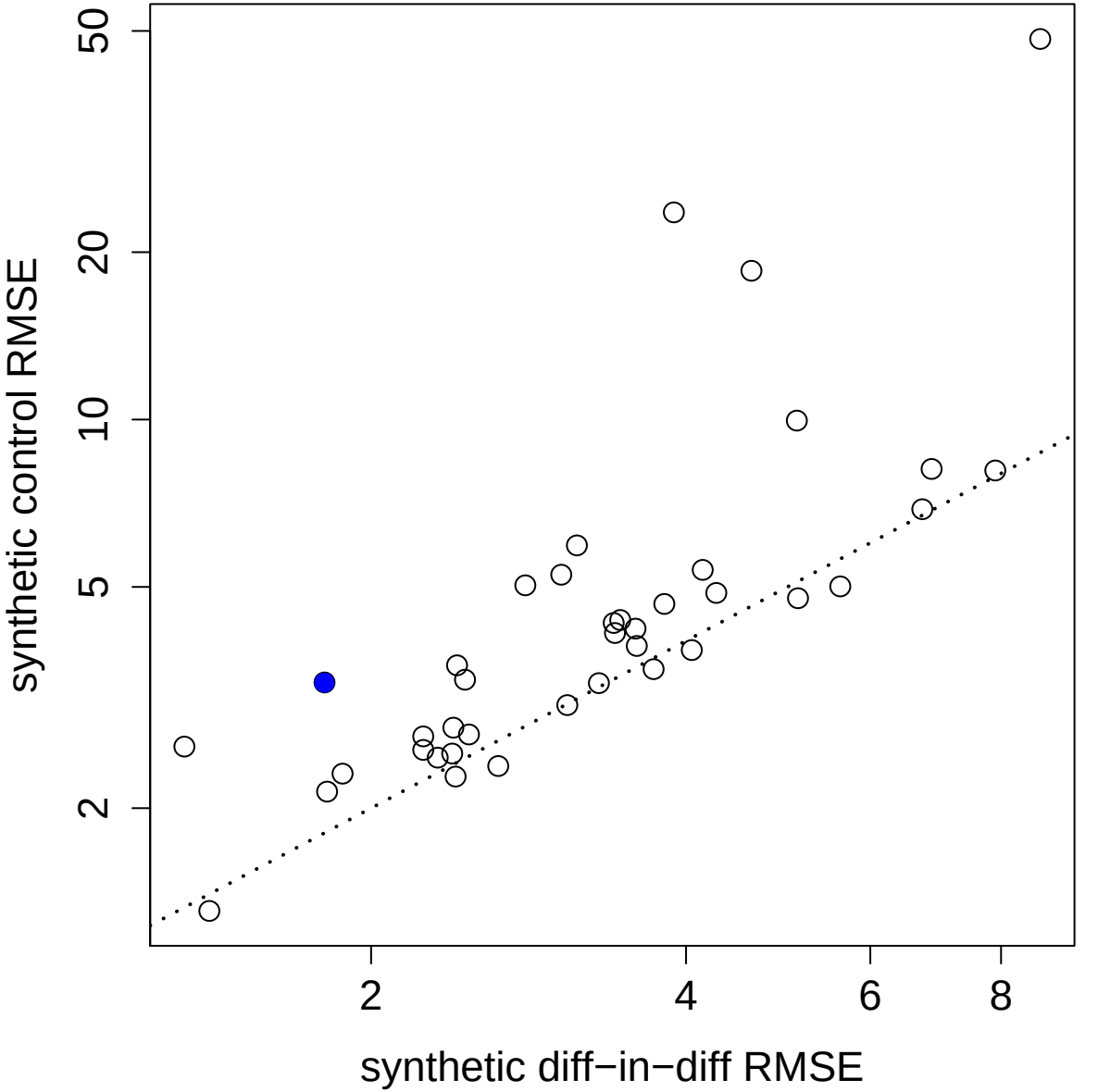
We **pretend** unit i was treated in periods $T_0+1, \ldots, T$, impute the "missing" values and compare them to actual values using SC (blue), DID (teal), SDID (red).

We average squared error by state for 8 periods ($T - T_0 = 8$) to get RMSEs for each state.


smoking [packs per capita]
90
100
110
120
130
1980
1982
1984
1986
1988
year


smoking [packs per capita]
95
100
105
110
115
120
1980
1982
1984
1986
1988
year


smoking [packs per capita]
160
180
200
220
240
260
280
1980
1982
1984
1986
1988
year


synthetic control RMSE
50
20
10
5
2
2
4
6
8
synthetic diff−in−diff RMSE