Advanced Statistical Methods for Observational Studies

LECTURE 01

introduction

this class

- Website
- Expectations
- Questions

observational studies

- The world of “observational studies” is kind of hard to get into because it grew up in several distinct, but overlapping, disciplines:
  - Epidemiology
  - Demography
  - Economics (econometrics)
  - Political Science
  - Sociology
  - Biostatistics
  - Statistics
  - Psychology (psychometrics)
  - Computer Science

a small bit about me

- I do causal inference:
  - Observational studies of: cardiothoracic interventions, neonates, and criminology.
  - Randomized studies: six interventions here at Stanford to improve educational outcomes educational outcomes, two large trials of a sexual assault prevention program.

an aside…

- You can call me “Mike”
- If you want to use my last name, Baiocchi, totally feel free to… if you say it this way I’ll definitely know you’re talking to me:

bye-oh-key

study design vs. inference

Don Rubin: For objective causal inference, design trumps analysis

study design vs. inference

- 90% of statistics classes are about inference
- Why?
  - It's useful, getting you those confidence intervals and p-values.
  - The math is pretty cool.
  - It feels hard.
  - Because many of us don't really know much about the real world…

design

RANDOMIZATION AND SAMPLING

where does the data come from?

- We design trials.
  - Assign groups that are similar at baseline
  - Construct most informative contrast groups
- We also design sampling schemes.
  - Representative groups
  - Understand population from subsets of those populations
- Both use elements of control and randomness

an example: randomization

- Want to study a pill.
- Design the study
  - Uniform randomization
  - Matched pairs randomization
  - Crossover design
  - Cluster-randomized
- Inference
  - t-test
  - Matched-pairs t-test
  - Repeated measures model
  - Generalized linear mixed model
  - But… maybe all of those could be GLMM.

an example: randomization

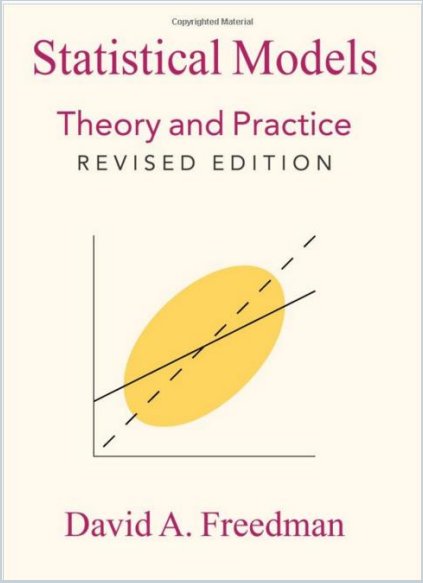
- Want to study a pill.
- Design the study
  - Uniform randomization
  - **Matched pairs randomization**
  - Crossover design
  - Cluster-randomized
- Inference
  - t-test
  - **Matched-pairs t-test**
  - Repeated measures model
  - Generalized linear mixed model
  - But… maybe all of those could be GLMM.

an example: sampling

- Want to study an election.
- Design the study
  - Simple random sample
  - Stratified sampling
  - Snowball sampling
- Inference
  - t-test
  - Inverse probability weighting
  - Generalized linear mixed model
  - But… maybe all of those could be GLMM.

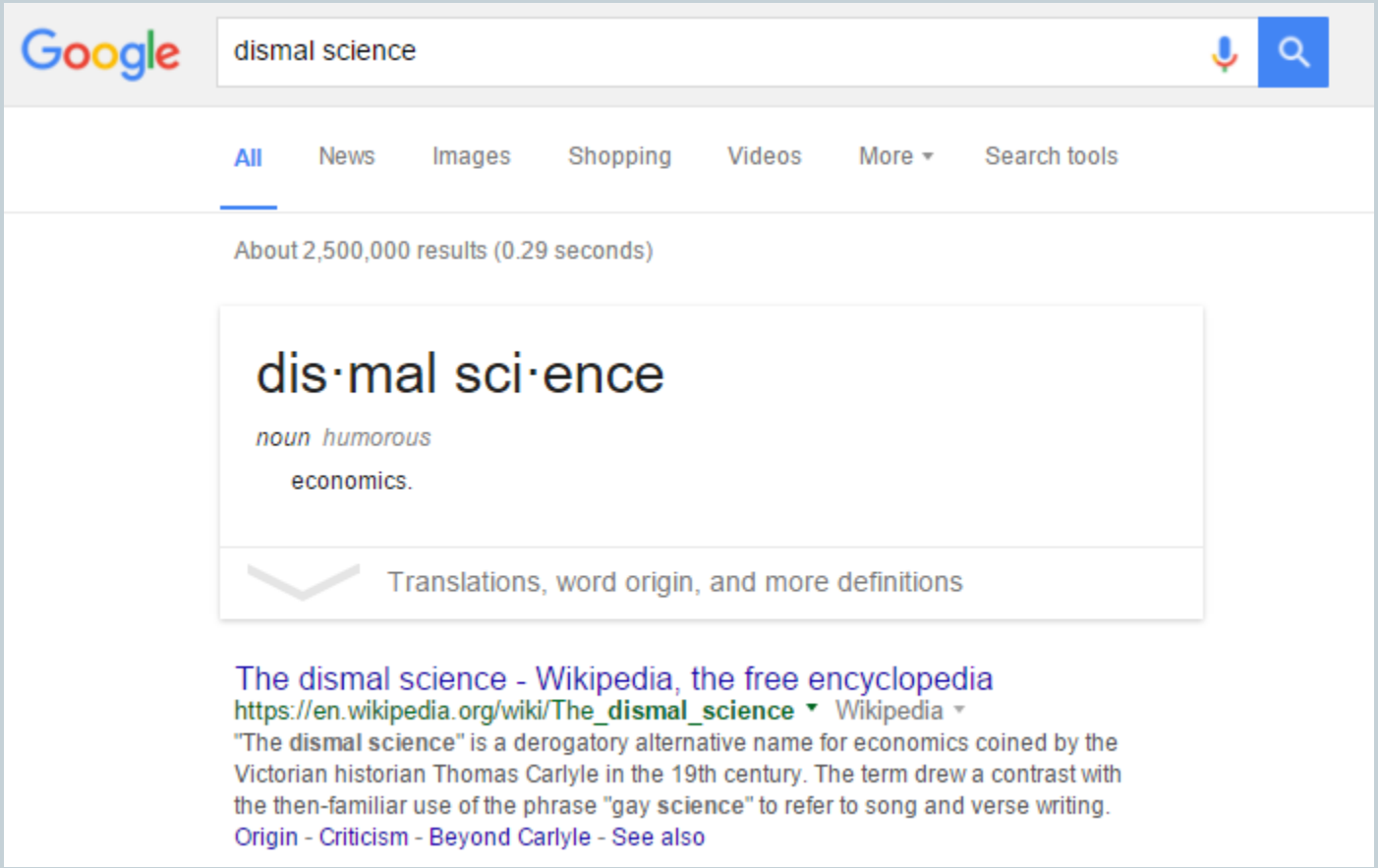
different beliefs about where data come from

- RCT and sampling
  - True (in the world) by construction
- Structural equation modeling
  - $y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i$
- If you want to be disabused of SEM spend some time reading



where data come from

- If you'd like to be abused by SEM please see



inference

picking inference

- Inference requires assumptions
- Linear regression:
  - Linearity and additivity
  - Independent errors
  - Homoskedastiticity
  - Normality of errors
- Permutation test:
  - Known assignment mechanism to T or C
- “Fancier” methods tend to have more assumptions… and thus leave you open to more lines of attack.
- These attacks can be obviated by careful preparation during the design phase.

picking inference

- Use the simplest method that gets the job done.
- If you want to accomplish more, collect more data or do additional analyses. (“If have to use something more complicated than a t-test then someone messed up…”)
- The fewer assumptions there are, the easier it will be to perform a “sensitivity analysis” – build an argument to beat back the haters.

picking inference

- Another option: Proof by intimidation

This paper presents a breakthrough in rhetorical logic, a promising field of science, of great values to those writing research proposals. It provides new, and utterly convincing tools for closing embarrassing gaps in your reasoning, without having to resort to "brute-force" methods such as actually thinking about the problem in the first place. The Craske-Trump Theorem Conjecture will allow researchers in any field to use the technique of "Proof by Intimidation" fully.

- Michael Wilkinson (Annals of Improbable Research 2000)

observational study design

NEONATAL INTENSIVE CARE UNITS

Application: Regionalization

- Hospitals vary in their ability to care for premature infants.
- The American Academy of Pediatrics recognizes levels: 1, 2, 3A, 3B, 3C, 3D and Regional Centers.
- ***Regionalization of care*** refers to a policy that suggests or requires that high-risk mothers deliver at hospitals with greater levels of capabilities.

2%

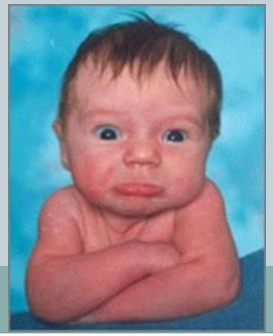
1%

Outcome

Outcome

Outcome

Outcome

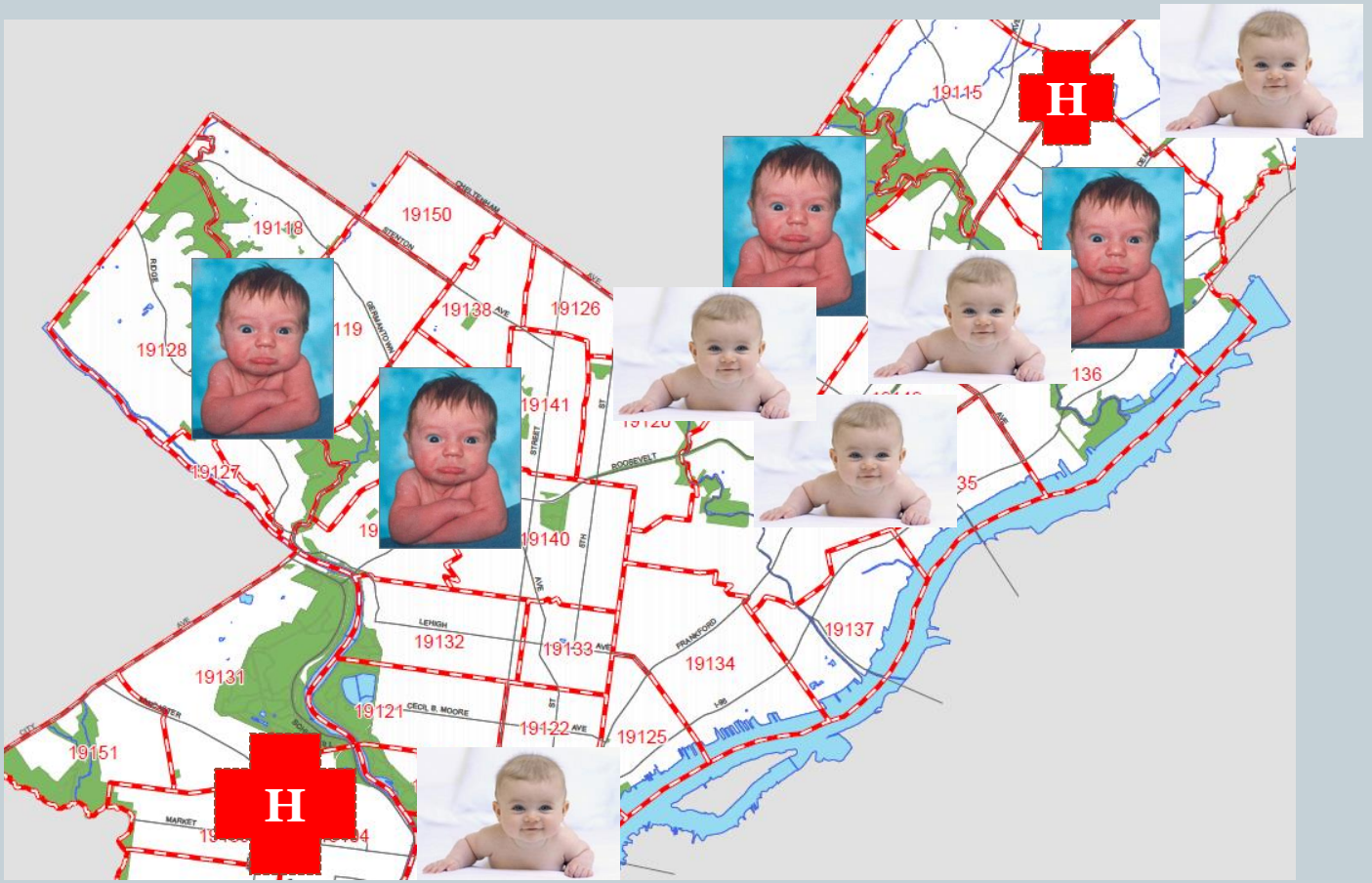


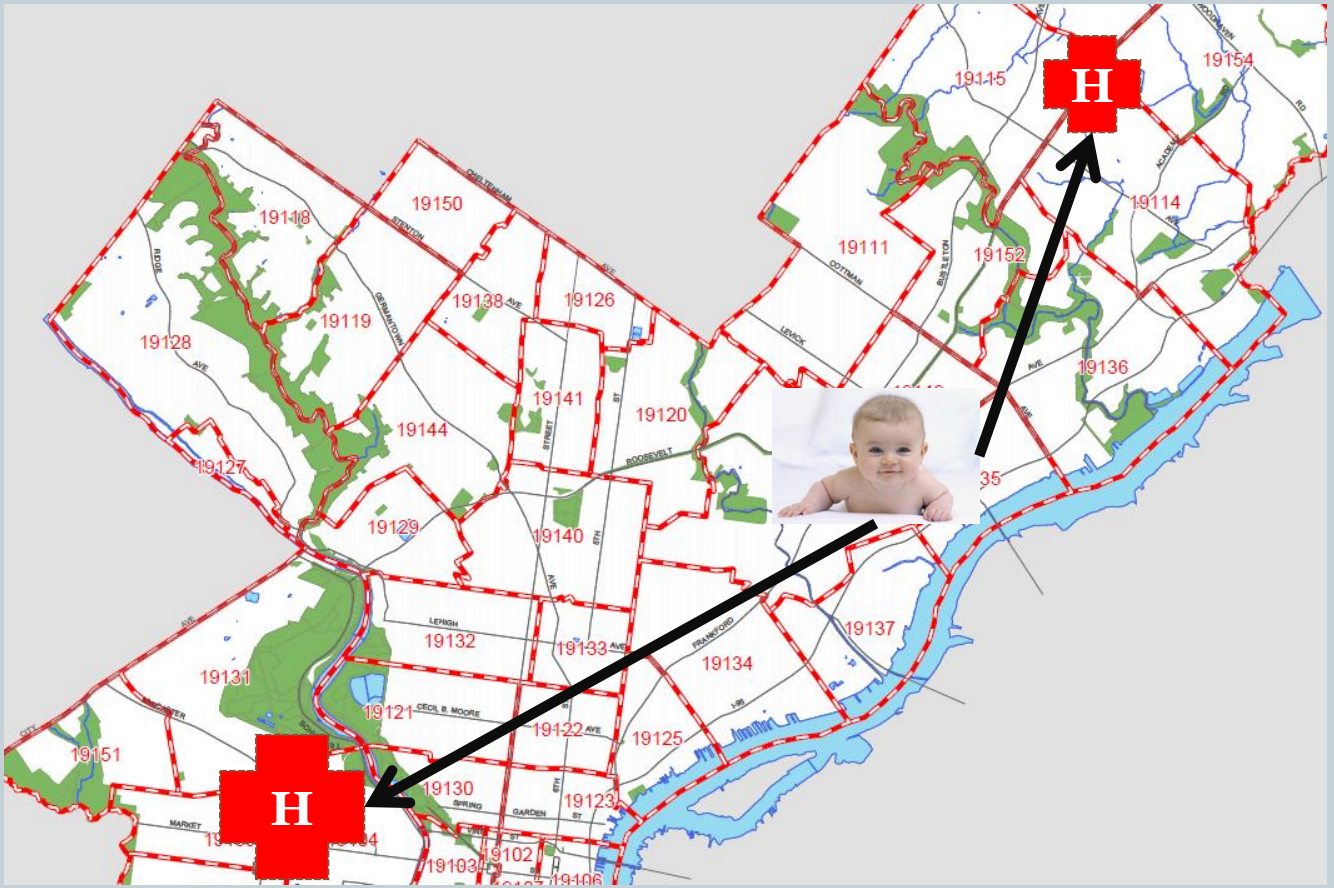
The data

- Every baby delivered in a 10+ year period
  - California
  - Pennsylvania
  - Missouri
- Mothers' information
  - ICD9 codes
    - Delivery
    - Post-delivery complications
    - Some pre-delivery
  - Some SES information
  - Zip code of residence
- Birth/death certificates
- Census information
  - PA and MO have zip code level
  - CA will have block group

Summary of Problem

- Want to quantify effect of level of NICU on rate of death
- Observational data
- Sorting bias
- Some sorting variables are **unobserved**


19118
19150
19138 AVE
19126
19128
19141
19127
19140
19132
19133 AVE
19131
19121
CECIL B. MOORE
19122 AVE
19125
19151
19115
19137
19134
136
35
H
H


H
H
19154
19145
19114
19111
19152
19136
19150
19118
19119
19138
19126
19128
19141
19120
19144
19127
19129
19140
19135
19137
19132
19133
19134
19131
19121
19122
19125
19151
19130
19123
19102
19103
19106

a matched study

outline of a study

- Outline of a study:
  - Introduction
  - Methods
  - Results
  - Discussion
- Reporting standards:
  - The CONSORT Statement
  - The STROBE Statement

Outcome

Outcome

	Variable Type	High NICU	Low NICU	sd	Δ/sd
Mortality	Outcome	2.26%	1.25%	13.33%	0.08
Difference in Travel Time	Instrument	4.57	19.00	17.18	-0.84
% attending high level NICU	Treatment	100.0%	0.0%	49.7%	2.01
Birth weight	Preemie covariates	2,454.07	2,693.24	739.27	-0.32
Gestational age		34.61	35.69	2.80	-0.39
GI	% of preemies with type of congenital disorders	0.9%	0.6%	8.7%	0.04
GU		0.9%	0.8%	9.0%	0.01
CNS		0.9%	0.4%	8.3%	0.05
Pulmonary		0.8%	0.7%	8.8%	0.01
Cardio		1.4%	0.7%	10.5%	0.06
Skeletal		0.7%	0.9%	9.0%	-0.02
Skin		0.0%	0.0%	0.0%	0.00
Chromosomes		0.4%	0.3%	6.3%	0.02
Other_Anomaly		0.8%	0.1%	7.0%	0.09
Gestational_DiabetesM	Mother covariates	4.9%	4.3%	21.0%	0.03
Mother's education		3.76	3.58	1.19	0.16
Insurance - Fee for service		24.0%	24.5%	42.8%	-0.01
Insurance - HMO		32.3%	27.8%	46.0%	0.10
Insurance - Government		23.5%	24.2%	42.6%	-0.02
Insurance - Other		16.8%	21.4%	39.1%	-0.12
Uninsured		2.2%	1.6%	13.7%	0.04
Prenatal care		2.51	2.37	1.30	0.11
Single birth (y/n)		79.0%	86.1%	38.3%	-0.18
Parity		2.08	2.09	1.31	-0.01
Mother's age		28.41	27.71	6.25	0.11
Median income	Census level covariates	41,484.25	40,258.92	14,587.24	0.08
Median home value		97,663.00	95,083.15	48,762.43	0.05
% completed high school		79.9%	80.0%	9.7%	-0.01
% completed college		22.2%	19.4%	13.1%	0.21
% renting		31.4%	27.9%	12.8%	0.28
% below poverty line		13.4%	11.8%	9.9%	0.16

the data

Unadjusted comparison of T vs C

n = 180,000

	High NICU	Low NICU	sd	Δ/sd
death	2.26%	1.25%	13.67%	0.07
birth weight (g)	2,454	2,693	739	-0.32
gestational age (months)	34.61	35.69	2.76	-0.39

Matched comparison of T vs C

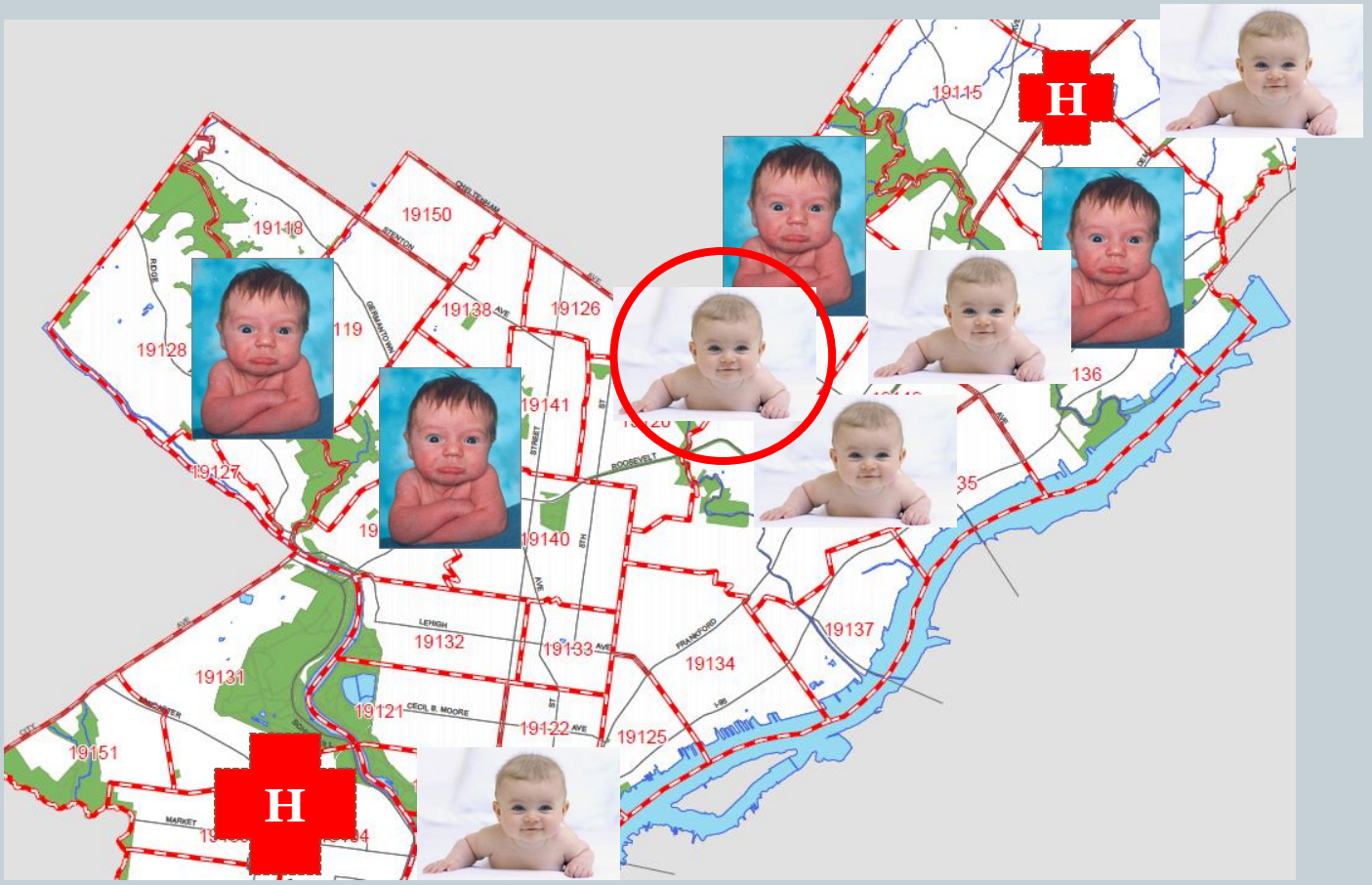
n = 120,000

	Matched High NICU	Matched Low NICU	sd	Δ/sd
death	1.55%	1.94%	13.67%	-0.03
birth weight (g)	2,584	2,581	739	0.00
gestational age (months)	35.14	35.13	2.76	0.00

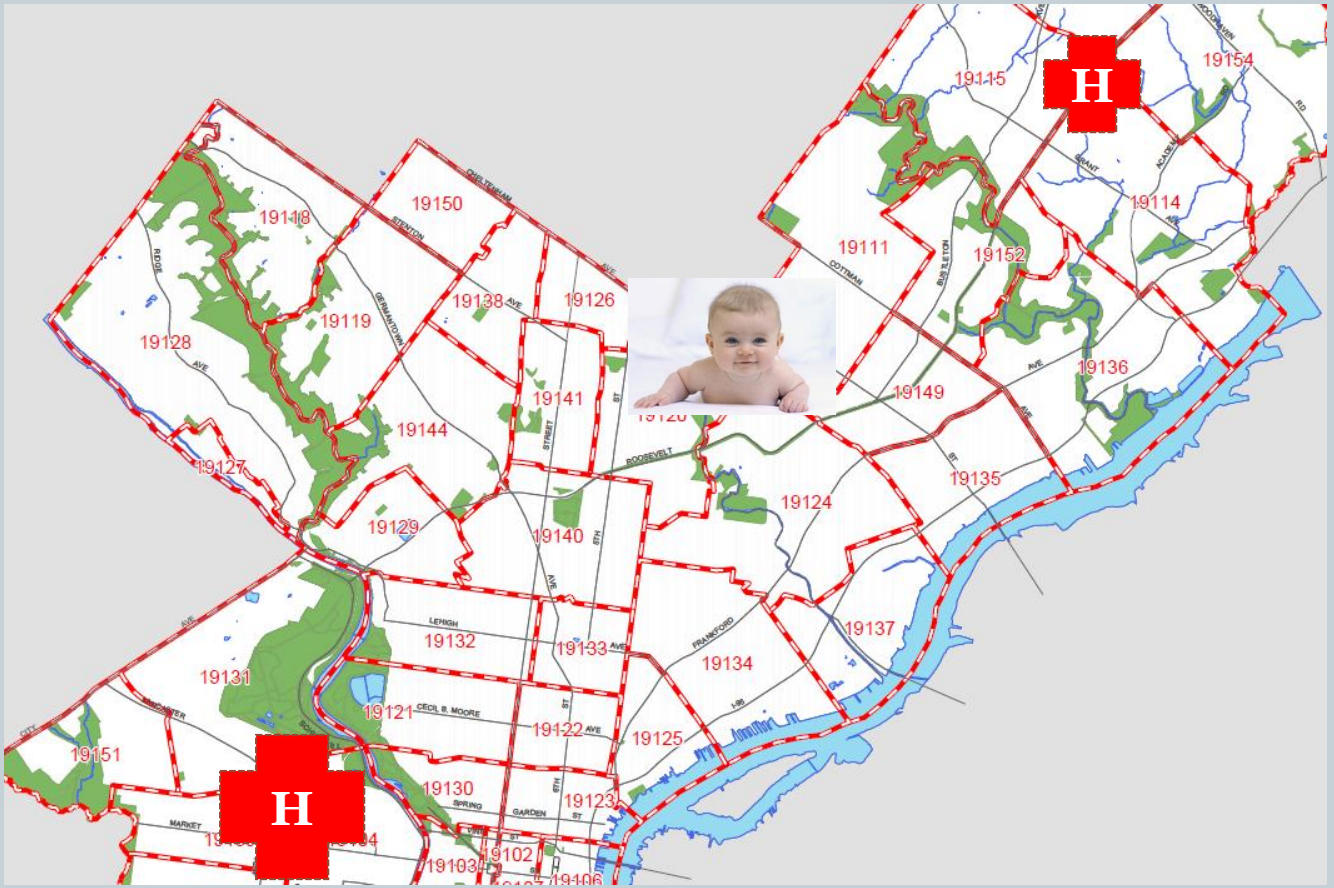
matching two observations

- Exact matching.
- Exact matching would be awesome, but consider how unlikely it is to be achievable.
- Wonderful to have some summary of how “close” two observations are to each other.

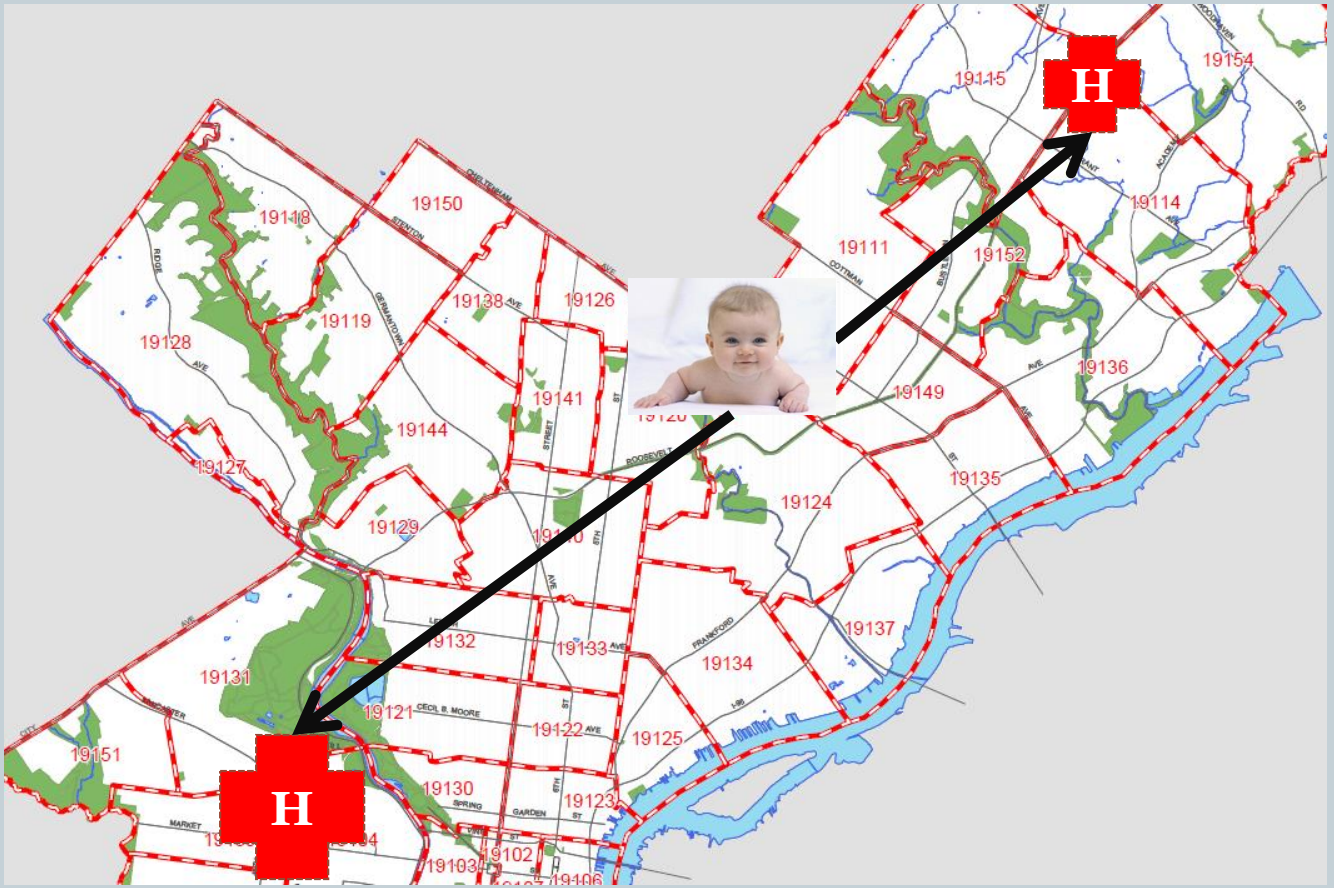
the propensity score



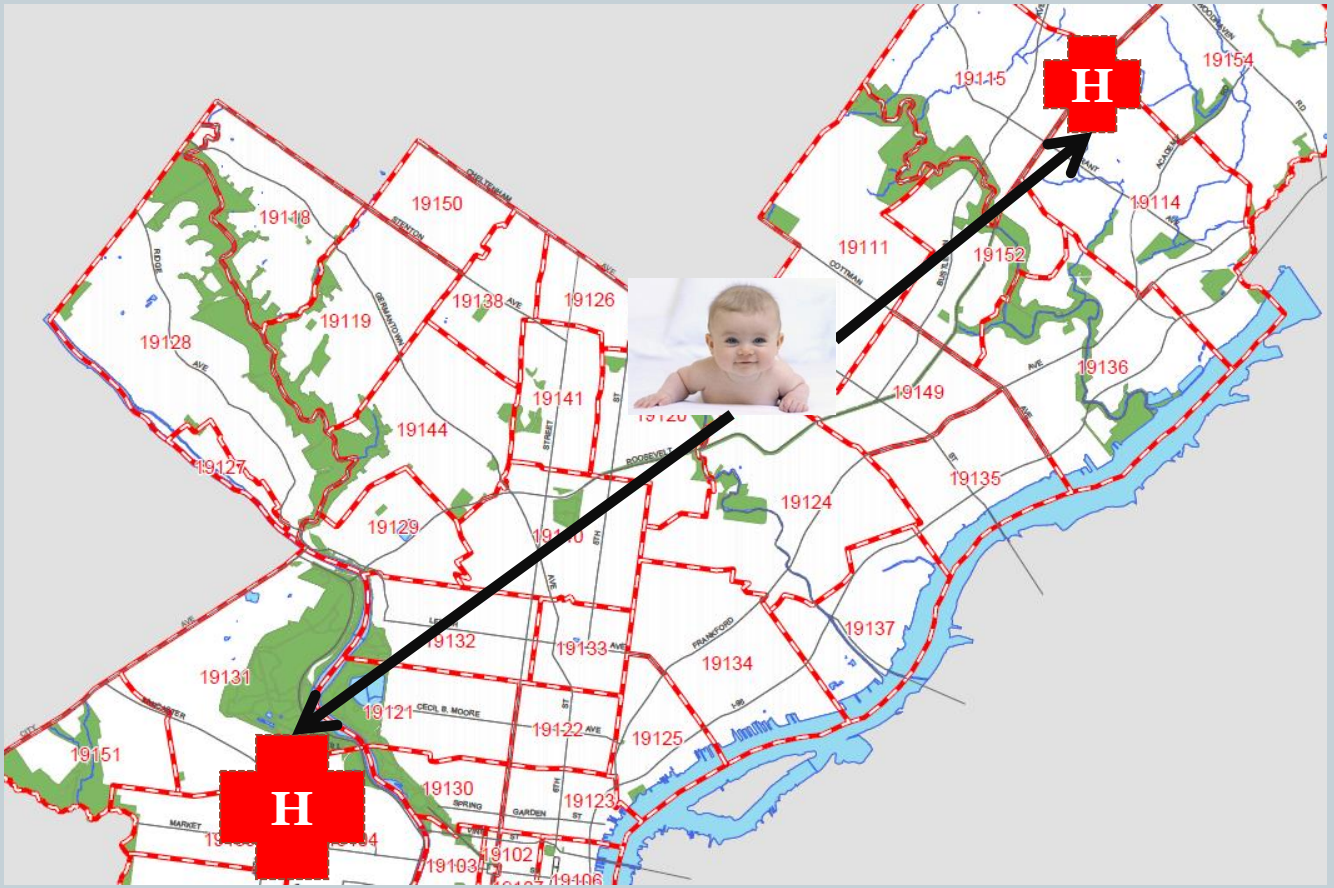
the propensity score



the propensity score

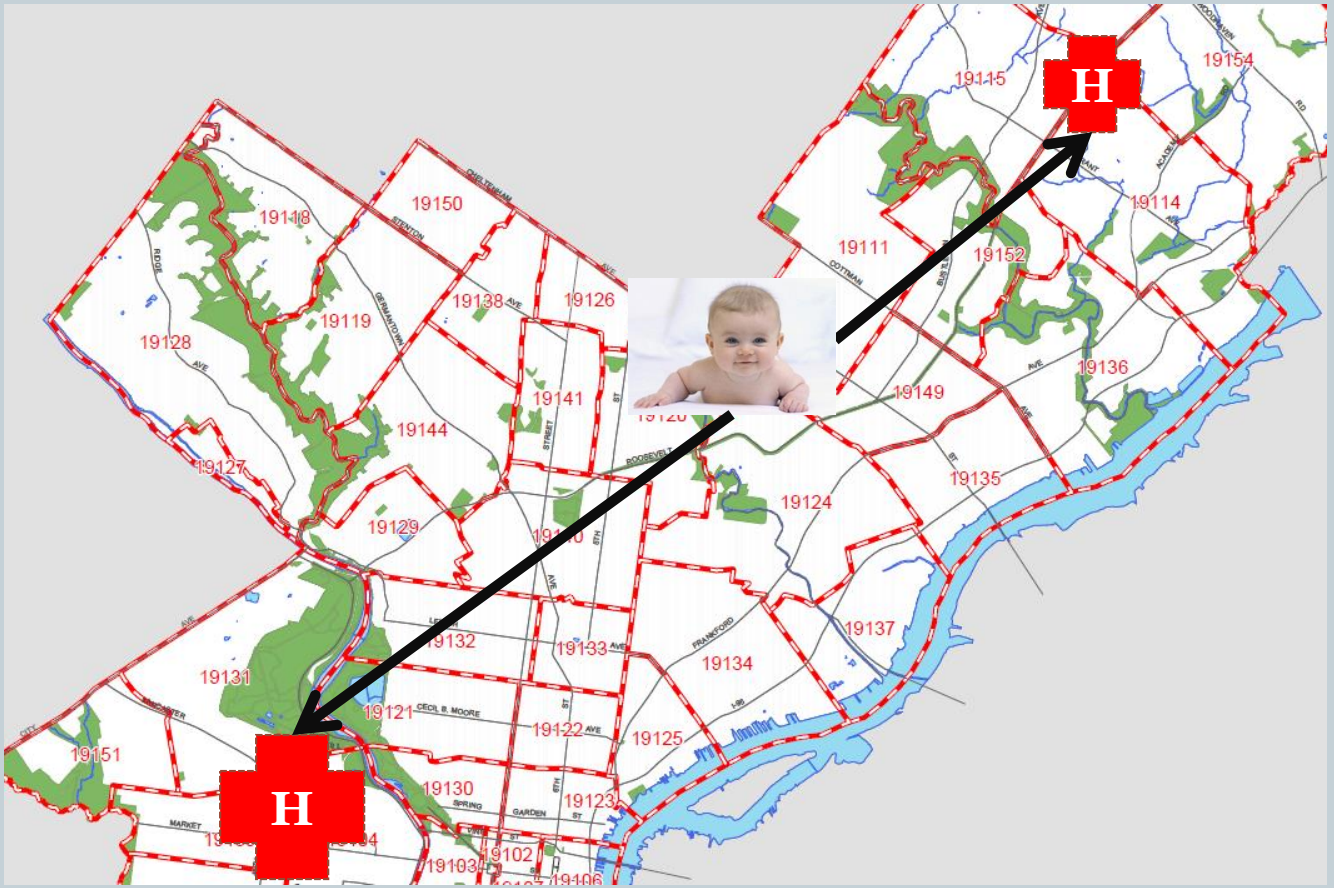


the propensity score



$$e(\boldsymbol{x}) = \Pr(d = 1|\boldsymbol{x}) = f(\mathit{birth\ weight}, \mathit{gestational\ age}, \mathit{mother's\ insurance}, \ldots)$$

the propensity score



$e(\boldsymbol{x}) = \Pr(d = 1|\boldsymbol{x}) = f(\boldsymbol{x})$ = feed in a high dimensional vector that describes the observations at baseline; get out a one-dimensional summary

results: “table 1”

Unadjusted comparison of T vs C

n = 180,000

	High NICU	Low NICU	sd	Δ/sd
death	2.26%	1.25%	13.67%	0.07
birth weight (g)	2,454	2,693	739	-0.32
gestational age (months)	34.61	35.69	2.76	-0.39

Matched comparison of T vs C

n = 120,000

	Matched High NICU	Matched Low NICU	sd	Δ/sd
death	1.55%	1.94%	13.67%	-0.03
birth weight (g)	2,584	2,581	739	0.00
gestational age (months)	35.14	35.13	2.76	0.00

Unlike a Table 1 from an RCT, we aren’t verifying that a randomization seems to have worked. Instead, we are obviating a “you obviously stacked the deck” argument.

the results

- There should be strong effort to show the two groups are similar.
  - Inclusion/exclusion
  - Observational units that may be completely missing
  - Missing data
  - Imbalances in observed data
  - Imbalances in unobserved data
- If the reader is willing to accept this then we move on to the analysis of the groups.
- Inference can be done in many different forms (largely driven by your upbringing). The use of matching in the study design phase meshes well with randomization test type inference.

takeaways

- Matching on observables is possible (details next)
- Table 1 can be assessed without needing to understand the matching technique
  - Mean differences
  - Distribution of variables
  - Which variables were matched on (and which weren't…)
  - Don't need to understand how we got there
- The analysis at the end is "simple"
  - Easier to get buy in
  - The naïve model is our foundation for doing these forms of analyses
  - Easier to build a sensitivity model

basic tools

PAIR MATCHING

Design of Observational Studies: chapter 8.1-8.4

tools

- Propensity scores
- Distance matrices
- Calipers and penalty functions
- Optimal matching
- Matching with multiple controls
- Full matching
- Efficiency of matching

example: NICU

treated

obs	b_weight	gest_age	dose	death
1	2412	36	1	0
2	2205	29	1	1
3	2569	36	1	0
4	2443	34	1	0
5	2569	36	0	0
6	2436	35	0	0
7	2461	34	0	0
8	2759	32	0	0
9	2324	27	0	1
10	2667	34	0	0

example: NICU

obs	b_weight	gest_age	dose	death
1	2412	36	1	0
2	2205	29	1	1
3	2569	36	1	0
4	2443	34	1	0
5	2569	36	0	0
6	2436	35	0	0
7	2461	34	0	0
8	2759	32	0	0
9	2324	27	0	1
10	2667	34	0	0

control

example: NICU

- Exact matching
  - When all (observed) baseline covariates are identical within a set.
  - There is only one exact match in our example.

example: NICU

obs	b_weight	gest_age	dose	death
1	2412	36	1	0
2	2205	29	1	1
3	2569	36	1	0
4	2443	34	1	0
5	2569	36	0	0
6	2436	35	0	0
7	2461	34	0	0
8	2759	32	0	0
9	2324	27	0	1
10	2667	34	0	0

example: NICU

- Exact matching
  - When all (observed) baseline covariates are identical within a set.
  - There is only one exact match in our example.
  - Usually only occurs when you have a few binary variables, or categorical variables with few categories.
- Exact matching is probably not possible:
  - If you have 40 binary covariates then you have $2^{40} \cong 1.1 * 10^{12}$.
  - Continuous variables make exact matching even harder.
  - We quickly get into questions about “close enough.”
  - We also get into the idea that not all covariates are equally important in determining which observations are “similar.”

propensity scores

- Uniform randomization has ½ by construction
- That's not the case here
  - Younger more likely to go to high NICU.
- Propensity score
  - Propensity: assignment to treatment (Fisher's inference)
  - Score: creates two similar groups on average

propensity scores

- Strongly Ignorable Treatment Assignment: Those that look alike (in our data set) are alike

$$\pi_i = \Pr(Z_i = 1 | r_{Ti}, r_{Ci}, \boldsymbol{x}_i, u_i) = \Pr(Z_i = 1 | \boldsymbol{x}_i)$$

and

$$0 < \pi_i < 1 \text{ for all } i = 1, 2, \ldots, n$$

- If two subjects have the same propensity score, then their values of $\boldsymbol{x}$ may be different.
- By SITA, if these two subjects have the same $e(\boldsymbol{x})$ then the differences in their $\boldsymbol{x}$ are not predictive of treatment assignment (i.e., $\boldsymbol{x} \perp\!\!\!\perp Z | e(\boldsymbol{x})$).
- Therefore the mismatches in $\boldsymbol{x}$ will be due to chance and will tend to balance. (more details)

propensity scores

- Dimensional reduction technique
  - Not guaranteed to match two people who look alike
  - Histograms of covariates
  - Histograms of propensity scores
  - “Table 1” (discussed below)

propensity scores

- How does one estimate the propensity score: logistic model

$$logit(\widehat{dose}) = \hat{\beta}_0 + \hat{\beta}_1 * \text{b\_weight} + \hat{\beta}_2 * \text{gest\_age}$$

obs	b_weight	gest_age	dose	death	e^(x)
1	2412	36	1	0	0.54
2	2205	29	1	1	0.43
3	2569	36	1	0	0.57
4	2443	34	1	0	0.53
5	2569	36	0	0	0.57
6	2436	35	0	0	0.54
7	2461	34	0	0	0.53
8	2759	32	0	0	0.57
9	2324	27	0	1	0.43
10	2667	34	0	0	0.57

$$logit(\widehat{dose}) = -0.3 + 0.0002 * \text{b\_weight} + 0.01 * \text{gest\_age}$$

propensity scores

- Why logistic?
- I've got two answers:

(i) It's what we do. [*cultural*]

(ii) Consider how a regression tree might produce different results. [*technical*]

propensity scores

- Question: if you knew the propensity score would you want to use it in lieu of the estimated propensity score?
- There are (at least) two valid answers:
  (i) no, for balance purposes (1 and 2) and
  (ii) yes, because of invalid inference (1 and 2).

takeaways: propensity scores

- Two key features:
  - Propensity: used for inference
  - Score: used for creating two groups
- Dimensional reduction technique:
  - Not guaranteed to match two people who look alike
  - Histograms of covariates
  - Histograms of propensity scores
- Reasons to use propensity score:

(i) matching on $e(\boldsymbol{x})$ is often practical even when there are many covariates in $\boldsymbol{x}$ because $e(\boldsymbol{x})$ is a single variable,

(ii) matching on $e(\boldsymbol{x})$ tends to balance all of $\boldsymbol{x}$, and

(iii) failure to balance $e(\boldsymbol{x})$ implies that $\boldsymbol{x}$ is not balanced.

distance matrices

distance matrices

- How do we summarize?
- If you're matching treated to control, then it's a matrix with:
  - A row for each treated
  - A column for each control
  - Each entry in the matrix represents a "distance" between a treated and control unit
- We sum up the entries of those that are matched to get an overall metric of quality of match. The algorithms are targeted toward minimizing these sums.

distance matrices

difference: $\hat{e}(\boldsymbol{x}_i) - \hat{e}(\boldsymbol{x}_j)$

obs	e^(x)
1	0.54
2	0.43
3	0.57
4	0.53
5	0.57
6	0.54
7	0.53
8	0.57
9	0.43
10	0.57

	5	6	7	8	9	10
1	-0.03	0.00	0.01	-0.03	0.11	-0.03
2	-0.14	-0.11	-0.10	-0.14	0.00	-0.14
3	0.00	0.04	0.04	0.00	0.14	0.00
4	-0.05	-0.01	0.00	-0.04	0.09	-0.04

We can describe a distance in pretty much any way we want.

distance matrices

absolute difference: $|\hat{e}(\boldsymbol{x}_i) - \hat{e}(\boldsymbol{x}_j)|$

obs	e^(x)
1	0.54
2	0.43
3	0.57
4	0.53
5	0.57
6	0.54
7	0.53
8	0.57
9	0.43
10	0.57

	5	6	7	8	9	10
1	0.03	0.00	0.01	0.03	0.11	0.03
2	0.14	0.11	0.10	0.14	0.00	0.14
3	0.00	0.04	0.04	0.00	0.14	0.00
4	0.05	0.01	0.00	0.04	0.09	0.04

We can describe a distance in pretty much any way we want.

distance matrices

quadratic difference: $C^*\left(\hat{e}(\boldsymbol{x}_i) - \hat{e}(\boldsymbol{x}_j)\right)^2$

obs	e^(x)
1	0.54
2	0.43
3	0.57
4	0.53
5	0.57
6	0.54
7	0.53
8	0.57
9	0.43
10	0.57

	5	6	7	8	9	10
1	0.20	0.00	0.02	0.17	2.32	0.19
2	4.08	2.26	2.05	3.96	0.00	4.06
3	0.00	0.27	0.35	0.00	3.86	0.00
4	0.41	0.01	0.00	0.37	1.76	0.40

We can describe a distance in pretty much any way we want.

distance matrices

- According to the estimated propensity score, 3 has adequate matches with 5, 8 and 10

obs	b_weight	gest_age	dose	death	e^(x)
1	2412	36	1	0	0.54
2	2205	29	1	1	0.43
3	2569	36	1	0	0.57
4	2443	34	1	0	0.53
5	2569	36	0	0	0.57
6	2436	35	0	0	0.54
7	2461	34	0	0	0.53
8	2759	32	0	0	0.57
9	2324	27	0	1	0.43
10	2667	34	0	0	0.57

distance matrices

- According to the estimated propensity score, 3 has adequate matches with 5, 8 and 10.

obs	b_weight	gest_age	dose	death	e^(x)
3	2569	36	1	0	0.57
5	2569	36	0	0	0.57
8	2759	32	0	0	0.57
10	2667	34	0	0	0.57

- While it's clear that 5 is best, because of the exact match, we can probably rank the match quality as 5, 10, followed by 8.
- It'd be nice to have a method for taking account of both the p-score and individual level covariates.

distance matrices

- One solution: use Mahalanobis distance
- If we write the sample covariance matrix of $\boldsymbol{x}$ as $\hat{\Sigma}$ and $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ as the covariate vectors for observations i and j, then the Mahalanobis distance between i and j is:

$$\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^T \hat{\Sigma}^{-1} \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)$$

- Intuition:
  - Trying to weight each variable equally
  - It was one of the first distances used by the matching community
  - Mahalanobis distance was created with iid Normals in mind
  - Not great at dealing with highly correlated covariates
  - Not great at dealing with non-symmetric data

distance matrices

obs	b_weight	gest_age
1	2412	36
2	2205	29
3	2569	36
4	2443	34
5	2569	36
6	2436	35
7	2461	34
8	2759	32
9	2324	27
10	2667	34

Mahalanobis distance

	5	6	7	8	9	10
1	0.37	0.16	0.31	1.12	1.02	0.74
2	0.88	0.68	0.62	1.16	0.45	0.96
3	0.00	0.27	0.26	0.80	0.97	0.40
4	0.28	0.13	0.04	0.88	0.76	0.52

distance matrices

- Better to default to the rank-based Mahalanobis distance.

- Big picture:
  - Why don't we use just Mahalanobis?
  - Why not just the pscore?
  - IF we take pscore and Mahalanobis together…

distance matrices

Why Propensity Scores Should Not Be Used for Matching*

Gary King† Richard Nielsen‡

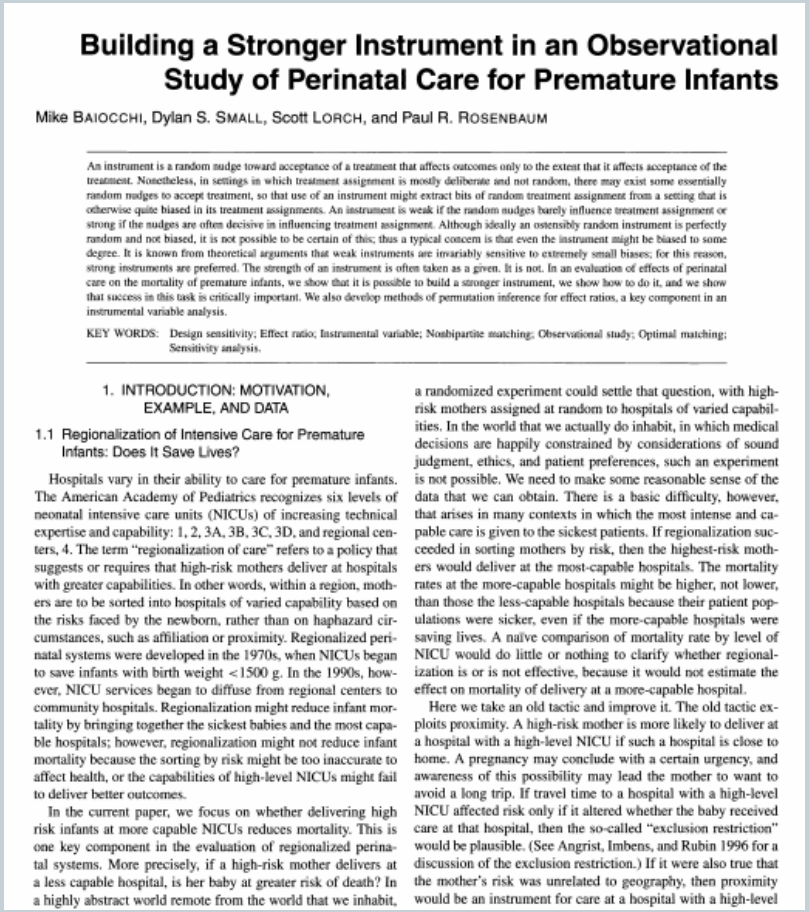
February 28, 2016

Abstract

We show that propensity score matching (PSM), an enormously popular method of preprocessing data for causal inference, often accomplishes the opposite of its intended goal — increasing imbalance, inefficiency, model dependence, and bias. PSM supposedly makes it easier to find matches by projecting a large number of covariates to a scalar propensity score and applying a single model to produce an unbiased estimate. However, in observational analysis the data generation process is rarely known and so users typically try many models before choosing one to present. The weakness of PSM comes from its attempts to approximate a completely randomized experiment, rather than, as with other matching methods, a more efficient fully blocked randomized experiment. PSM is thus uniquely blind to the often large portion of imbalance that can be eliminated by approximating full blocking with other matching methods. Moreover, in data balanced enough to approximate complete randomization, either to begin with or after pruning some observations, PSM approximates random matching which, we show, increases imbalance even relative to the original data. Although these results suggest that researchers replace PSM with one of the other available methods when performing matching, propensity scores have many other productive uses.

(link)

distance matrices



Building a Stronger Instrument in an Observational Study of Perinatal Care for Premature Infants

Mike BAIOCCHI, Dylan S. SMALL, Scott LORCH, and Paul R. ROSENBAUM

An instrument is a random nudge toward acceptance of a treatment that affects outcomes only to the extent that it affects acceptance of the treatment. Nonetheless, in settings in which treatment assignment is mostly deliberate and not random, there may exist some essentially random nudges to accept treatment, so that use of an instrument might extract bits of random treatment assignment from a setting that is otherwise quite biased in its treatment assignments. An instrument is weak if the random nudges barely influence treatment assignment or strong if the nudges are often decisive in influencing treatment assignment. Although ideally an ostensibly random instrument is perfectly random and not biased, it is not possible to be certain of this; thus a typical concern is that even the instrument might be biased to some degree. It is known from theoretical arguments that weak instruments are invariably sensitive to extremely small biases; for this reason, strong instruments are preferred. The strength of an instrument is often taken as a given. It is not. In an evaluation of effects of perinatal care on the mortality of premature infants, we show that it is possible to build a stronger instrument, we show how to do it, and we show that success in this task is critically important. We also develop methods of permutation inference for effect ratios, a key component in an instrumental variable analysis.

KEY WORDS: Design sensitivity; Effect ratio; Instrumental variable; Nonbipartite matching; Observational study; Optimal matching; Sensitivity analysis.

1. INTRODUCTION: MOTIVATION, EXAMPLE, AND DATA

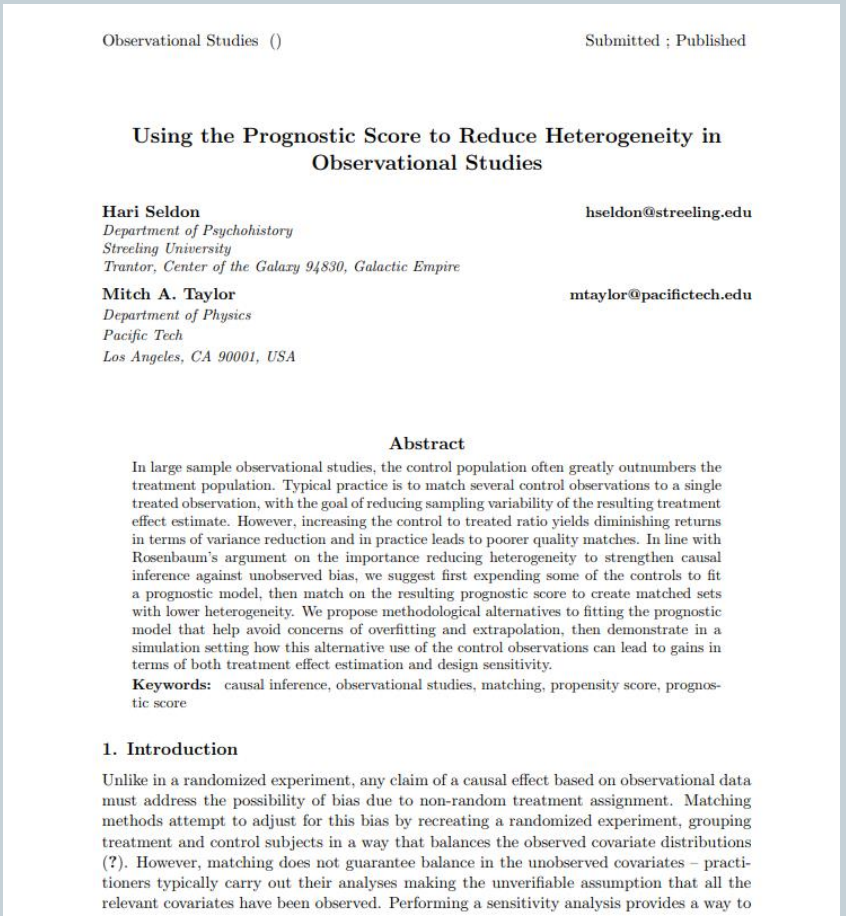
1.1 Regionalization of Intensive Care for Premature Infants: Does It Save Lives?

Hospitals vary in their ability to care for premature infants. The American Academy of Pediatrics recognizes six levels of neonatal intensive care units (NICUs) of increasing technical expertise and capability: 1, 2, 3A, 3B, 3C, 3D, and regional centers, 4. The term "regionalization of care" refers to a policy that suggests or requires that high-risk mothers deliver at hospitals with greater capabilities. In other words, within a region, mothers are to be sorted into hospitals of varied capability based on the risks faced by the newborn, rather than on haphazard circumstances, such as affiliation or proximity. Regionalized perinatal systems were developed in the 1970s, when NICUs began to save infants with birth weight <1500 g. In the 1990s, however, NICU services began to diffuse from regional centers to community hospitals. Regionalization might reduce infant mortality by bringing together the sickest babies and the most capable hospitals; however, regionalization might not reduce infant mortality because the sorting by risk might be too inaccurate to affect health, or the capabilities of high-level NICUs might fail to deliver better outcomes.

In the current paper, we focus on whether delivering high risk infants at more capable NICUs reduces mortality. This is one key component in the evaluation of regionalized perinatal systems. More precisely, if a high-risk mother delivers at a less capable hospital, is her baby at greater risk of death? In a highly abstract world remote from the world that we inhabit, a randomized experiment could settle that question, with high-risk mothers assigned at random to hospitals of varied capabilities. In the world that we actually do inhabit, in which medical decisions are happily constrained by considerations of sound judgment, ethics, and patient preferences, such an experiment is not possible. We need to make some reasonable sense of the data that we can obtain. There is a basic difficulty, however, that arises in many contexts in which the most intense and capable care is given to the sickest patients. If regionalization succeeded in sorting mothers by risk, then the highest-risk mothers would deliver at the most-capable hospitals. The mortality rates at the more-capable hospitals might be higher, not lower, than those the less-capable hospitals because their patient populations were sicker, even if the more-capable hospitals were saving lives. A naïve comparison of mortality rate by level of NICU would do little or nothing to clarify whether regionalization is or is not effective, because it would not estimate the effect on mortality of delivery at a more-capable hospital.

Here we take an old tactic and improve it. The old tactic exploits proximity. A high-risk mother is more likely to deliver at a hospital with a high-level NICU if such a hospital is close to home. A pregnancy may conclude with a certain urgency, and awareness of this possibility may lead the mother to want to avoid a long trip. If travel time to a hospital with a high-level NICU affected risk only if it altered whether the baby received care at that hospital, then the so-called "exclusion restriction" would be plausible. (See Angrist, Imbens, and Rubin 1996 for a discussion of the exclusion restriction.) If it were also true that the mother's risk was unrelated to geography, then proximity would be an instrument for care at a hospital with a high-level

(near-far matching)



Observational Studies ()

Submitted ; Published

Using the Prognostic Score to Reduce Heterogeneity in Observational Studies

Hari Seldon hseldon@streeling.edu
Department of Psychohistory
Streeling University
Trantor, Center of the Galaxy 94830, Galactic Empire

Mitch A. Taylor mtaylor@pacifictech.edu
Department of Physics
Pacific Tech
Los Angeles, CA 90001, USA

Abstract

In large sample observational studies, the control population often greatly outnumbers the treatment population. Typical practice is to match several control observations to a single treated observation, with the goal of reducing sampling variability of the resulting treatment effect estimate. However, increasing the control to treated ratio yields diminishing returns in terms of variance reduction and in practice leads to poorer quality matches. In line with Rosenbaum's argument on the importance reducing heterogeneity to strengthen causal inference against unobserved bias, we suggest first expending some of the controls to fit a prognostic model, then match on the resulting prognostic score to create matched sets with lower heterogeneity. We propose methodological alternatives to fitting the prognostic model that help avoid concerns of overfitting and extrapolation, then demonstrate in a simulation setting how this alternative use of the control observations can lead to gains in terms of both treatment effect estimation and design sensitivity.

Keywords: causal inference, observational studies, matching, propensity score, prognostic score

1. Introduction

Unlike in a randomized experiment, any claim of a causal effect based on observational data must address the possibility of bias due to non-random treatment assignment. Matching methods attempt to adjust for this bias by recreating a randomized experiment, grouping treatment and control subjects in a way that balances the observed covariate distributions (?). However, matching does not guarantee balance in the unobserved covariates – practitioners typically carry out their analyses making the unverifiable assumption that all the relevant covariates have been observed. Performing a sensitivity analysis provides a way to

(ask Dylan Greaves)

penalty functions

- Add on a caliper using the pscore

- Penalty functions for pscore

penalty functions

obs	b_weight	gest_age
1	2412	36
2	2205	29
3	2569	36
4	2443	34
5	2569	36
6	2436	35
7	2461	34
8	2759	32
9	2324	27
10	2667	34

Mahalanobis distance

	5	6	7	8	9	10
1	0.37	0.16	0.31	1.12	1.02	0.74
2	0.88	0.68	0.62	1.16	0.45	0.96
3	0.00	0.27	0.26	0.80	0.97	0.40
4	0.28	0.13	0.04	0.88	0.76	0.52

penalty functions

obs	b_weight	gest_age
1	2412	36
2	2205	29
3	2569	36
4	2443	34
5	2569	36
6	2436	35
7	2461	34
8	2759	32
9	2324	27
10	2667	34

Mahalanobis distance

	5	6	7	8	9	10
1	0.37	0.16	0.31	1.12	1.02	0.74
2	0.88	0.68	0.62	1.16	0.45	0.96
3	0.00	0.27	0.26	0.80	0.97	0.40
4	0.28	0.13	0.04	0.88	0.76	0.52

penalty functions

obs	b_weight	gest_age
1	2412	36
2	2205	29
3	2569	36
4	2443	34
5	2569	36
6	2436	35
7	2461	34
8	2759	32
9	2324	27
10	2667	34

Mahalanobis distance

	5	6	7	8	9	10
1	0.37	0.16	0.31	1.12	∞	0.74
2	∞	∞	∞	∞	0.45	∞
3	0.00	0.27	0.26	0.80	∞	0.40
4	0.28	0.13	0.04	0.88	0.76	0.52

This caliper sets the distance matrix to infinity if $\left|\hat{e}(\boldsymbol{x}_i) - \hat{e}(\boldsymbol{x}_j)\right| \geq 0.10$

penalty functions

- Penalty functions for pscore
  - Add a penalty to the existing distance matrix

$$M_M + M_p = M_d$$

where M_M is from the Mahalanobis M_p is a penalty based on the pscores and M_d is the distance matrix that will be used to match.

penalty functions

obs	b_weight	gest_age
1	2412	36
2	2205	29
3	2569	36
4	2443	34
5	2569	36
6	2436	35
7	2461	34
8	2759	32
9	2324	27
10	2667	34

Mahalanobis distance

	5	6	7	8	9	10
1	0.37	0.16	0.31	1.12	1.02	0.74
2	0.88	0.68	0.62	1.16	0.45	0.96
3	0.00	0.27	0.26	0.80	0.97	0.40
4	0.28	0.13	0.04	0.88	0.76	0.52

You can add in a penalty, such as $C^*\left(\hat{e}(\boldsymbol{x}_i) - \hat{e}(\boldsymbol{x}_j)\right)^2$, when $|\hat{e}(\boldsymbol{x}_i) - \hat{e}(\boldsymbol{x}_j)|$ are outside of some acceptable range.

penalty functions: covariates

obs	b_weight	gest_age
1	2412	36
2	2205	29
3	2569	36
4	2443	34
5	2569	36
6	2436	35
7	2461	34
8	2759	32
9	2324	27
10	2667	34

Mahalanobis distance

	5	6	7	8	9	10
1	0.37	0.16	0.31	1.12	1.02	0.74
2	0.88	0.68	0.62	1.16	0.45	0.96
3	0.00	0.27	0.26	0.80	0.97	0.40
4	0.28	0.13	0.04	0.88	0.76	0.52

You can add in a penalty, such as $C^*\left(x_{i,gest\_age} - x_{j,gest\_age}\right)^2$, when $|x_{i,gest\_age} - x_{j,gest\_age}|$ are outside of some acceptable range.

penalty functions: covariates

obs	b_weight	gest_age
1	2412	36
2	2205	29
3	2569	36
4	2443	34
5	2569	36
6	2436	35
7	2461	34
8	2759	32
9	2324	27
10	2667	34

Mahalanobis distance

	5	6	7	8	9	10
1	0.37	0.16	0.31	1.12	1.02	0.74
2	0.88	0.68	0.62	1.16	0.45	0.96
3	0.00	0.27	0.26	0.80	0.97	0.40
4	0.28	0.13	0.04	0.88	0.76	0.52

You can also do a one-side penalty function to nudge in one direction. For example when $(x_{i,gest\_age} - x_{j,gest\_age}) \leq -1$.

takeaways

- The distance matrix is the core of how you describe the acceptability of a pair to be matched together.
- Think about both the individual and the group level.
  - Individual level for matched-pairs randomization
  - Group level for the comparability of the two groups (e.g., “Table 1”)

takeaways

takeaways

- The potential outcomes framework helps organize our thinking on counterfactuals
- Design comes in two flavors
  (actually, three… but the third one is not very healthy)
- In prospective studies
  - design is an obvious consideration
  - and one that MUST be passed through in order to obtain data
- In retrospective studies,
  - design is a less obvious consideration
  - but one that MUST be passed through… unfortunately without much attention paid

fin.

CHECK OUT THE WEBSITE.