Weighting

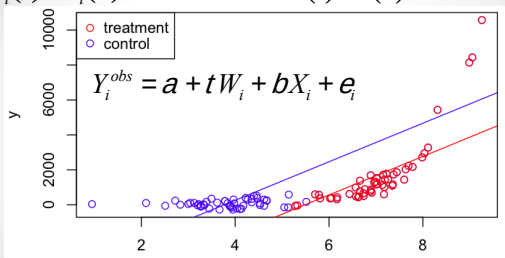
STA 320
Design and Analysis of Causal Studies
Dr. Kari Lock Morgan and Dr. Fan Li
Department of Statistical Science
Duke University

Homework 2

- Unconfounded
- Describe imbalance – direction matters
- Dummy variables and interactions
- Rubin 2007 common comments:
  - Full probability model on the science?
  - Throwing away units => bias?
  - Regression?

Regression

$Y_i(1) = Y_i(0)$ for all i $\quad t = \bar{Y}(1) - \bar{Y}(0) = 0$




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4150.0      574.2  -7.228 1.13e-10 ***
x             1102.4      143.1   7.706 1.12e-11 ***
w            -1899.6      515.2  -3.687 0.000374 ***
```

Regression

- Regression models are okay when covariate balance is good between the groups (randomized experiment, after subclassification or matching)
- If covariate balance bad (after unadjusted observational study), regression models rely heavily on extrapolation, and can be very sensitive to departures from linearity

Decisions

- Matching:
  - with or without replacement?
  - 1:1, 2:1, … ?
  - distance measure?
  - what to match on?
  - how to weight variables?
  - exact matching for any variable(s)?
  - calipers for when a match is "acceptable"?
  - …
- Let's try it on the Lalonde data…

Weighting

- An alternative to either subclassification or matching is to use weighting
- Each unit gets a weight, and estimates are a weighted average of the observed outcomes within each group

$$\hat{t} = \sum_{i=1}^{N} I_i W_i Y_i(1) - \sum_{i=1}^{N} I_i (1 - W_i) Y_i(0)$$

- weights sum to 1 within each group

Observed Difference in Means

- For the simple estimator observed difference in means:

$$\hat{t} = \frac{\sum_{i=1}^{N} W_i Y_i(1)}{N_T} - \frac{\sum_{i=1}^{N} (1 - W_i) Y_i(0)}{N_C}$$

- treated units weighted by $1/N_T$ and control units weighted by $1/N_C$

Subclassification

Sample size	Subclass 1	Subclass 2
Control	8	4
Treatment	2	6

Weights	Subclass 1	Subclass 2
Control	(1/8)/2	(1/4)/2
Treatment	(1/2)/2	(1/6)/2

Propensity Score Weighting

- Various different choices for the weights, most based on the propensity score, e(x)
- The following create covariate balance:

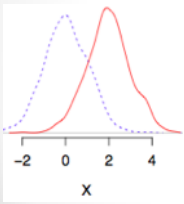
Weights	Treatment	Control
Horvitz-Thompson (ATE)	$\frac{1}{e(x)}$	$\frac{1}{1 - e(x)}$
ATT	1	$\frac{e(x)}{1 - e(x)}$
Overlap	$1 - e(x)$	$e(x)$

Weighting Methods

- Horvitz-Thompson (ATE, average treatment effect): Weights sample to look like covariate distribution of entire sample (like subclassification)
- ATT (Average treatment effect for the treated): Weights sample to look like covariate distribution of treatment group (like matching)
- Overlap: Weights sample to emphasize the "overlap" between the treatment and control groups (new: Profs Li and Morgan)

Toy Example

- Control simulated from N(0,1)
- Treatment simulated from N(2, 1)

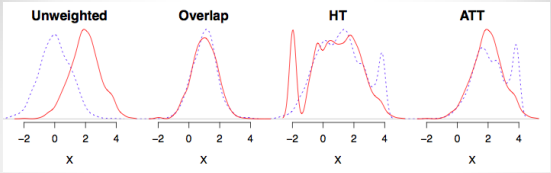


Weighted Means	Control	Treatment
Unweighted	0.03	1.98
HT (ATE)	1.19	0.74
ATT	2.22	1.98
Overlap	1.01	1.01

Weighting Methods

- Don Rubin does not like weighting estimators – he says you should use subclassification or matching
- Why?
- The weights everyone uses (HT, ATT) don't work!
- Prof Li and I working together to develop the overlap weight to address this problem

Weighting Methods



- Problem with propensity scores in the denominator: weights can blow up for propensity scores close to 0 or 1, and extreme units can dominate

Toy Example

$$Y_i^{obs} \sim N(X_i, 1) + tW_i$$

$$t = 1$$

	Unweighted	Overlap	HT	ATT
$\hat{\tau}$	2.945	1.000	0.581	0.640
$SE(\hat{\tau})$	0.054	0.038	0.386	0.402

Racial Disparity

- Goal: estimate racial disparity in medical expenditures (causal inference?)
- The Institute of Medicine defines disparity as "a difference in treatment provided to members of different racial (or ethnic) groups that is not justified by the underlying health conditions or treatment preferences of patients"
- Balance health status and demographic variables; observe difference in health expenditures

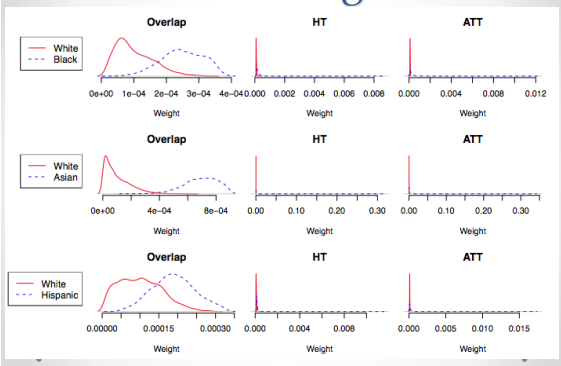
MEPS Data

- Medical Expenditure Panel Survey (2009)
- Adults aged 18 and older
  - 9830 non-Hispanic Whites
  - 4020 Blacks
  - 1446 Asians
  - 5280 Hispanics
- 3 different comparisons: non-Hispanic whites with each minority group

Propensity Scores

- 31 covariates (5 continuous, 26 binary), mix of health indicators and demographic variables
- Logistic regression with all main effects
- Separate propensity score model for each minority group

MEPS Weights



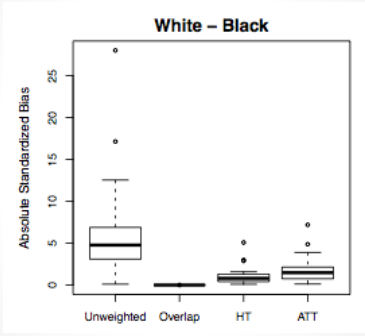
One High Asian Weight

- One Asian has a weight of 30%! (out of 1446 Asians)
- 78 year old Asian lady with a BMI of 55.4 (very high, highest Asian BMI in data)
- In the sample, White people are older and fatter, so this obese old Asian lady has a very high propensity score (0.9998)
- Unnormalized weight: 1/(1-0.9998) = 5000
- This is too extreme – will mess up weights with propensity score in the denominator

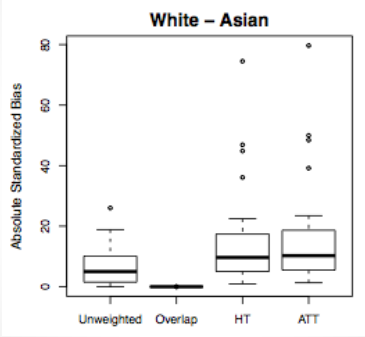
Truncate?

- In practice people truncate these extreme weights to avoid this problem, but then estimates can become very dependent on truncation choice

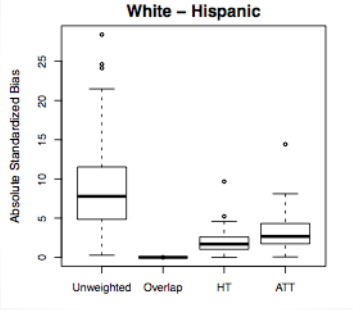
MEPS Imbalance



MEPS Imbalance



MEPS Imbalance



MEPS Imbalance

- Using the overlap weights, largest imbalance was a difference in means 0.0000003 standard errors apart
- For the Asian and Hispanic comparison, HT and ATT gave some differences in means over 74 standard errors apart!!!! (BMI)
- Initial imbalance can make propensity scores in the denominator of weights mess up horribly

Overlap Weights

- The overlap weights provide better balance and avoid extreme weights
- Automatically emphasizes those units who are comparable (could be in either treatment group)
- Pros: Better balance, better estimates, lower standard errors
- Con: estimand not clear

Disparity Estimates

- Estimated disparities for total health care expenditure, after weighting:

	Unweighted	Overlap	HT	ATT
White - Black	786.2 (222.4)	824.30 (184.7)	855.8 (200.3)	851.0 (219.7)
White - Asian	2763.9 (209.5)	1226.7 (204.8)	2167.4 (640.1)	2310.3 (711.1)
White - Hispanic	2598.9 (173.7)	1212.1 (170.7)	596.3 (323.3)	200.2 (445.4)

Weighting

- Weighting provides a convenient and flexible way to do causal inference
- Easy to incorporate into other models
- Easy to incorporate survey weights
- However, conventional weighting methods that place the propensity score in the denominator can go horribly wrong
- The overlap weighting method is a new alternative developed to avoid this problem

Review

Not everything you need to know, but some of the main points…

Causality

- Causality is tied to an action (treatment)
- Potential outcomes represent the outcome for each unit under treatment and control
- A causal effect compares the potential outcome under treatment to the potential outcome under control for each unit
- In reality, only one potential outcome observed for each unit, so need multiple units to estimate causal effects

SUTVA, Framework

- SUTVA assumes no interference between units and only one version of each treatment
- Y, W, Y_i^{obs}, Y_i^{mis}
- The assignment mechanism (how units are assigned to treatments) is important to consider
- Using only observed outcomes can be misleading (e.g. Perfect Doctor)
- Potential outcome framework and Rubin's Causal Model can help to clarify questions of causality (e.g. Lord's Paradox)

Assignment Mechanism

- Covariates (pre-treatment variables) are often important in causal inference
- Assignment probabilities:
  - $\Pr(\mathbf{W} \mid \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0))$
  - $p_i(\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)) = \Pr(W_i \mid \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0))$
- Properties of the assignment mechanism:
  - individualistic
  - probabilistic
  - unconfounded
  - known and controlled

Randomized Experiments

- We'll cover four types of classical randomized experiments:
  - Bernoulli randomized experiment
  - Completely randomized experiment
  - Stratified randomized experiment
  - Paired randomized experiment
- Increasingly restrictive regarding possible assignment vectors

Fisher Inference

- Sharp null of no treatment effect allows you to fill in missing potential outcomes under the null
- Randomization distribution: distribution of the statistic due to random assignment, assuming the null
- p-value: Proportion of statistics in the randomization distribution as extreme as observed statistic

Neyman's Inference (Finite Sample)

1. Define the estimand: $t \equiv \overline{Y(1)} - \overline{Y(0)}$
2. unbiased estimator of the estimand:
$$\hat{t} \equiv \overline{Y}_T^{obs} - \overline{Y}_C^{obs}$$
3. true sampling variance of the estimator
$$\text{var}(\hat{t}) = \frac{S_T^2}{N_T} + \frac{S_C^2}{N_C} - \frac{S_{TC}^2}{N}$$
4. unbiased estimator of the true sampling variance of the estimator

   (IMPOSSIBLE!) Overestimate: $\widehat{\text{var}}(\hat{t}) = \frac{s_T^2}{N_T} + \frac{s_C^2}{N_C}$
5. Assume approximate normality to obtain p-value and confidence interval

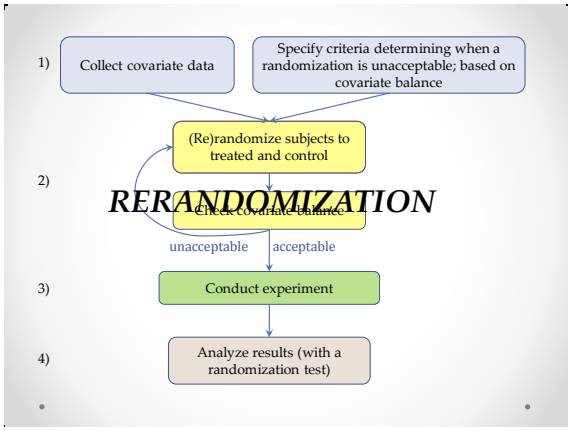
Slide adapted from Cassandra Pattanayak, Harvard

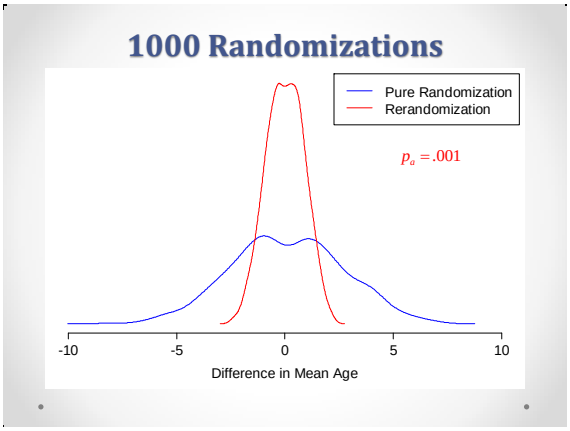
Fisher vs Neyman

Fisher	Neyman
Goal: testing	Goal: estimation
Considers only random assignment	Considers random assignment and random sampling
H_0: no treatment effect	H_0: average treatment effect = 0
Works for any test statistic	Difference in means
Exact distribution	Approximate, relies on large n
Works for any known assignment mechanism	Only derived for common designs

Summary: Using Covariates

- By **design**:
  - stratified randomized experiments
  - paired randomized experiments
  - rerandomization
- By **analysis**:
  - outcome: gain scores
  - separate analyses within subgroups
  - regression
  - imputation





Select Facts about Classical Randomized Experiments

Timing of treatment assignment clear

Design and Analysis separate by definition: design automatically "prospective," without outcome data

Unconfoundedness, probabilisticness by definition

Assignment mechanism – and so propensity scores – known

Randomization of treatment assignment leads to expected balance on covariates

("Expected Balance" means that the joint distribution of covariates is the same in the active treatment and control groups, on average)

Analysis defined by protocol rather than exploration

Slide by Cassandra Pattanayak

Select Facts about Observational Studies

Timing of treatment assignment may not be specified

Separation between design and analysis may become obscured, if covariates and outcomes arrive in one data set

Unconfoundedness, probabilisticness not guaranteed

Assignment mechanism – and therefore propensity scores – unknown

Lack of randomization of treatment assignment leads to imbalances on covariates

Analysis often exploratory rather than defined by protocol

Slide by Cassandra

40

Best Practices for Observational Studies

Timing of treatment assignment may not be specified

Separation between design and analysis may become obscured, if covariates and outcomes arrive in one data set

Unconfoundedness, probabilisticness not guaranteed

Assignment mechanism – and therefore propensity scores – unknown

Lack of randomization of treatment assignment leads to imbalances on covariates

Analysis often exploratory rather than defined by protocol

Slide by Cassandra

41

Best Practices for Observational Studies

1. Determine timing of treatment assignment relative to measured variables

Separation between design and analysis may become obscured, if covariates and outcomes arrive in one data set

Unconfoundedness, probabilisticness not guaranteed

Assignment mechanism – and therefore propensity scores – unknown

Lack of randomization of treatment assignment leads to imbalances on covariates

Analysis often exploratory rather than defined by protocol

Slide by Cassandra

42

Best Practices for Observational Studies

1. Determine timing of treatment assignment relative to measured variables

2. Hide outcome data until design phase is complete

Unconfoundedness, probabilisticness not guaranteed

Assignment mechanism – and therefore propensity scores – unknown

Lack of randomization of treatment assignment leads to imbalances on covariates

Analysis often exploratory rather than defined by protocol

Slide by Cassandra Pattanayak 43

Best Practices for Observational Studies

1. Determine timing of treatment assignment relative to measured variables

2. Hide outcome data until design phase is complete

3. Identify key covariates likely related to outcomes and/or treatment assignment. If key covariates not observed or very noisy, usually better to give up and find a better data source.

4. Remove units not similar to any units in opposite treatment group

Assignment mechanism – and therefore propensity scores – unknown

Lack of randomization of treatment assignment leads to imbalances on covariates

Analysis often exploratory rather than defined by protocol

Slide by Cassandra Pattanayak 44

Best Practices for Observational Studies

1. Determine timing of treatment assignment relative to measured variables

2. Hide outcome data until design phase is complete

3. Identify key covariates likely related to outcomes and/or treatment assignment. If key covariates not observed or very noisy, usually better to give up and find a better data source.

4. Remove units not similar to any units in opposite treatment group

Assignment mechanism – and therefore propensity scores – unknown

Lack of randomization of treatment assignment leads to imbalances on covariates

Analysis often exploratory rather than defined by protocol

Slide by Cassandra Pattanayak 45

Best Practices for Observational Studies

1. Determine timing of treatment assignment relative to measured variables

2. Hide outcome data until design phase is complete

3. Identify key covariates likely related to outcomes and/or treatment assignment. If key covariates not observed or very noisy, usually better to give up and find a better data source.

4. Remove units not similar to any units in opposite treatment group

5. Estimate propensity scores, as a way to…

6. Find subgroups (subclasses or pairs) in which the active treatment and control groups are balanced on covariates (not always possible; inferences limited to subgroups where balance is achieved)

Analysis often exploratory rather than defined by protocol

Slide by Cassandra Pattanayak 46

Best Practices for Observational Studies

1. Determine timing of treatment assignment relative to measured variables

2. Hide outcome data until design phase is complete

3. Identify key covariates likely related to outcomes and/or treatment assignment. If key covariates not observed or very noisy, usually better to give up and find a better data source.

4. Remove units not similar to any units in opposite treatment group

5. Estimate propensity scores, as a way to…

6. Find subgroups (subclasses or pairs) in which the treatment groups are balanced on covariates (not always possible; inferences limited to subgroups where balance is achieved)

Analysis often exploratory rather than defined by protocol

Slide by Cassandra Pattanayak 47

Best Practices for Observational Studies

1. Determine timing of treatment assignment relative to measured variables

2. Hide outcome data until design phase is complete

3. Identify key covariates likely related to outcomes and/or treatment assignment. If key covariates not observed or very noisy, usually better to give up and find a better data source.

4. Remove units not similar to any units in opposite treatment group

5. Estimate propensity scores, as a way to…

6. Find subgroups (subclasses or pairs) in which the active treatment and control groups are balanced on covariates (not always possible; inferences limited to subgroups where balance is achieved)

7. Analyze according to pre-specified protocol

Slide by Cassandra Pattanayak 48

Best Practices for Observational Studies

1. Determine timing of treatment assignment relative to measured variables

Design Observational Study to Approximate Hypothetical, Parallel Randomized Experiment

5. Estimate propensity scores, as a way to…

6. Find subgroups (subclasses or pairs) in which the active treatment and control groups are balanced on covariates (not always possible; inferences limited to subgroups where balance is achieved)

7. Analyze according to pre-specified protocol

Slide by Cassandra

49

Propensity Score Estimation

- Fit logistic regression, regressing W on covariates and any interactions or transformations of covariates that may be important
- Force all primary covariates to be in the model; choose others via variable selection (likelihood ratio, stepwise…)
- Trim units with no comparable counterpart in opposite group, and iterate between trimming and fitting propensity score model

Subclassification

- Divide units into subclasses within which the propensity scores are relatively similar
- Estimate causal effects within each subclass
- Average these estimates across subclasses (weighted by subclass size)
- (analyze as a stratified experiment)

Matching

- Matching: Find control units to "match" the units in the treatment group
- Restrict the sample to matched units
- Analyze the difference for each match (analyze as matched pair experiment)
- Useful when one group (usually the control) is much larger than the other

Decisions

- Estimating propensity score:
  - What variables to include?
  - How many units to trim, if any?
- Subclassification:
  - How many subclasses and where to break?
- Matching:
  - with or without replacement?
  - 1:1, 2:1, … ?
  - how to weight variables / distance measure?
  - exact matching for any variable(s)?
  - calipers for when a match is "acceptable"?
  - …

Weighting

- Weighting provides a convenient and flexible way to do causal inference
- However, conventional weighting methods that place the propensity score in the denominator can go horribly wrong
- The overlap weighting method is a new alternative developed to avoid this problem

To Do

- Read Ch 19
- Midterm on Monday, 3/17!