Determining Cause and Effect from Observational Data:
An Introduction to Causal Inference

Amy Nail, Ph.D.

Honestat Statistics & Analytics

amynail@honestat.com

Analytics>Forward

March 10, 2018

Example: Plantar fasciitis

- Pain on the bottom of the foot and on the heel
- Hurts when getting out of bed in the morning
- Hurts when standing up after sitting for a while
- Hurts a few hours after walking, running, hiking, rock climbing ☹
- When severe, hurts during these activities ☹

- The pain can keep you from participating in fun things ☹

Goal: compare new treatment to old

Old treatment (control)

- Stretching & strengthening exercises

New treatment (treatment)

- Same as old, PLUS
- Trigger-point dry-needling of calf muscles

Response variable:

Change in range of motion (CROM) of ankle joint in dorsiflexion after 4 wks of treatment

CROM = 4wk ROM – baseline ROM

Bigger is better

Other variables that might affect recovery from PF

- Gender
- Age

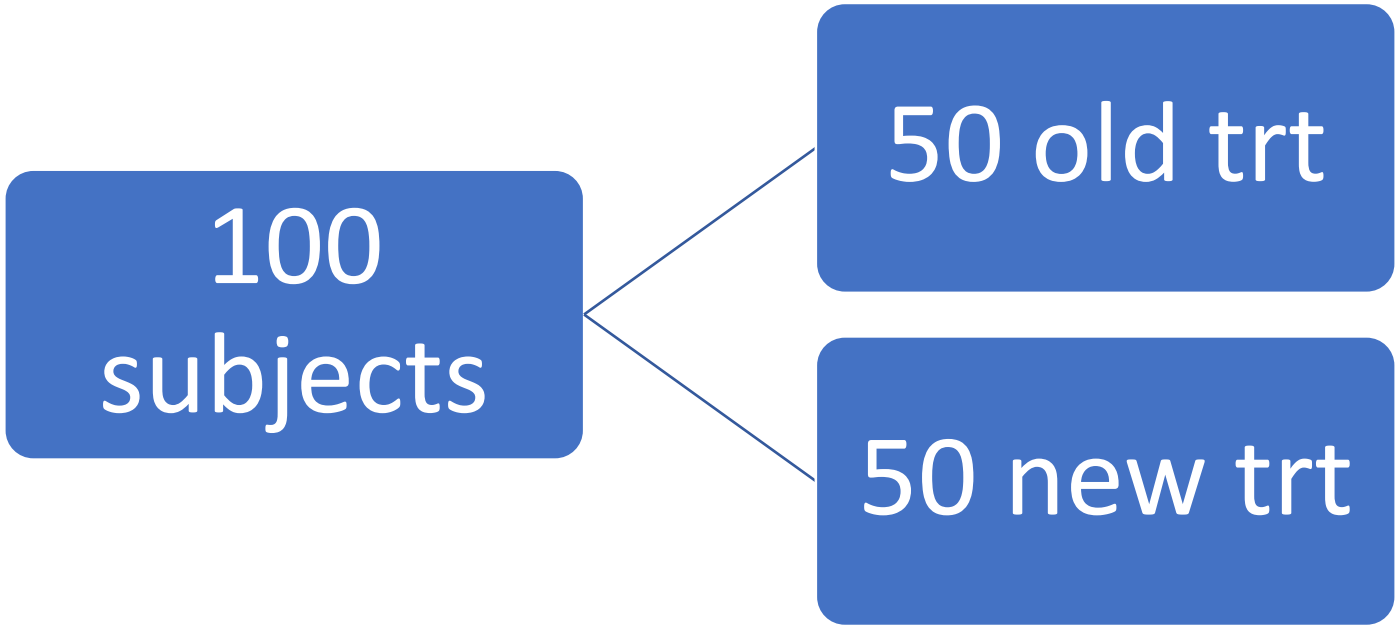
Definition of **confounding variable**

Both must apply

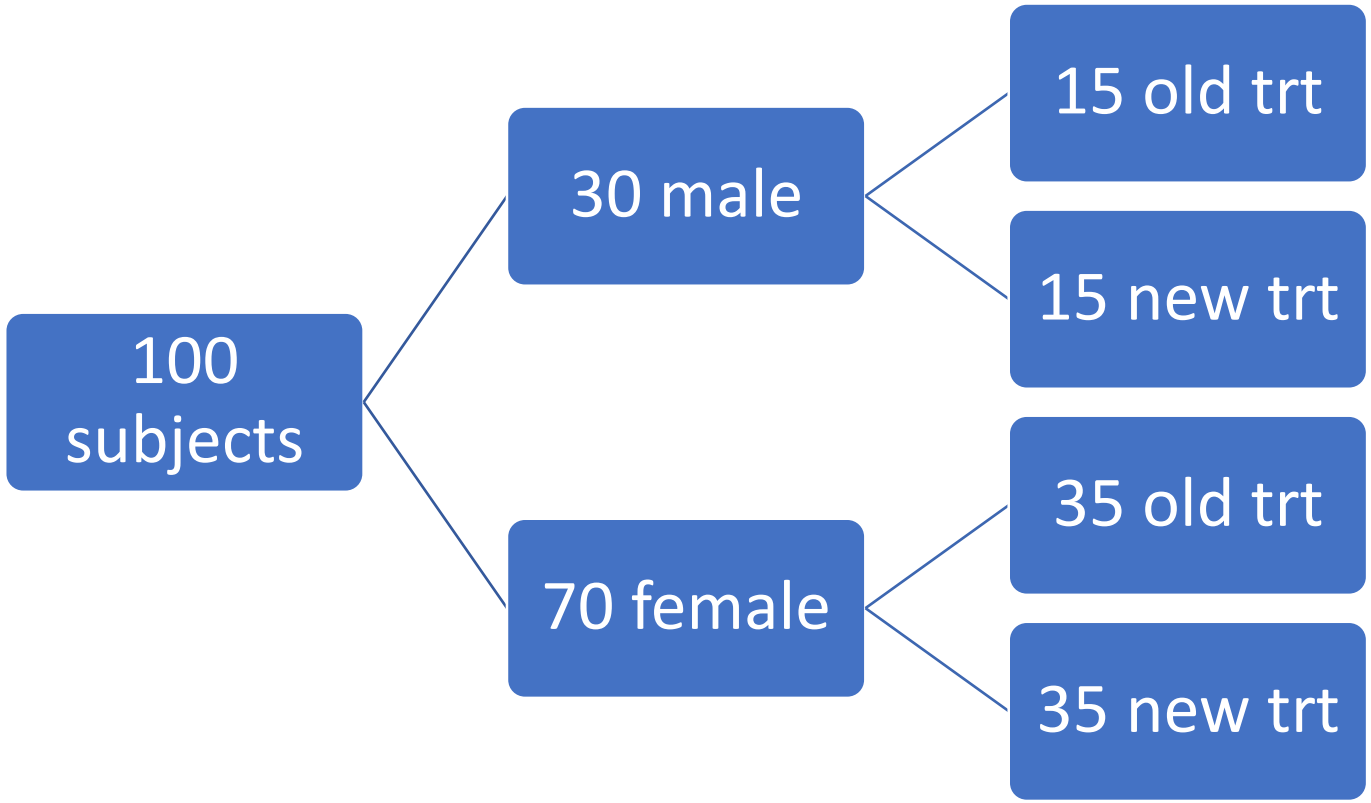
1. The variable affects the response variable, AND

2. The variable is not balanced between treatment and control groups

Controlled experiment/ Randomized controlled trial (**RCT**) Could use **Simple Random Sample (SRS)**



If sample isn't balanced over confounders, could use **Stratified Random Sample**



Instead of RCT, use historical controls (observational data)

Dr. Jones has a database of 50 patients treated with the old treatment

- Use those patients as controls

Give new treatment to next 50 patients

- Use them as treatment group

Groups may not be balanced

(Historical) Control group

- 15 male
- 35 female

- Average age: 32 years

Treatment group

- 28 male
- 22 female

- Average age: 25 years

How can we use probability theory to create balance?

Counterfactual framework

Y $\equiv$ What the subject's CROM *is observed to be* after 4 wks

$Y1$ $\equiv$ What the subject's CROM *would have been* after 4 wks if that subject were in the treatment group

If subject in treatment group, $\mathrm{Y} = Y1$

$Y0 \equiv$ What the subject's CROM *would be* after 4 wks if that subject were to be in the control group

If subject in control group, $\mathrm{Y} = Y0$

Definition:
Average Treatment Effect (ATE)

$$\Delta \equiv \mathrm{E}(Y1) - \mathrm{E}(Y0)$$

In English:

ATE ≡ Mean Y1 – Mean Y0

Estimators of ATE (Δ)

$\hat{\Delta}_1$	Difference of means
$\hat{\Delta}_{1R}$	$\hat{\Delta}_1$ with regression-adjustment
$\hat{\Delta}_S$	Stratification estimator
$\hat{\Delta}_{SR}$	$\hat{\Delta}_S$ with regression-adjustment
$\hat{\Delta}_{IPW}$	Inverse-probability-weighting
$\hat{\Delta}_{DR}$	$\hat{\Delta}_{IPW}$ with regression-adjustment (Doubly-robust)

Things we need for each estimator

1. Formula or method to calculate estimate
2. Formula or method to get confidence intervals (CI)

How to compare estimators

1. Want to minimize bias
2. Want to minimize variance of estimator
   (e.g., width of CI)
3. Want a 95% CI to have 95% coverage probability

Why I chose these estimators

- There is a head-to-head comparison of these in Lunceford & Davidian 2004

- Extensive section with theoretical results
- Extensive simulation study to back up the theory

$\widehat{\Delta}_1$ Difference of means

$$\widehat{\Delta}_1 \equiv \frac{1}{50}\sum_{\substack{trt \\ group}} Y_i - \frac{1}{50}\sum_{\substack{ctrl \\ group}} Y_i$$

More notation

Z = 1 if subject in treatment group

0 otherwise

X = The vector of potential confounders

(X_1, X_2)

Gender, age

Another way to calculate $\widehat{\Delta}_1$

Notice if we fit the model

$$Y = \beta_0 + \beta_Z Z$$

Then

$$\widehat{\Delta}_1 = \widehat{\beta_Z}$$

$\widehat{\Delta}_{1R}$ Regression-adjustment estimator

Fit the model

$$Y = \beta_0 + \beta_Z Z + \beta_1 X_1 + \beta_2 X_2$$

Use

$$\widehat{\Delta}_{1R} \equiv \widehat{\beta_Z}$$

What we know so far

$$Bias(\hat{\Delta}_{1R}) < Bias(\hat{\Delta}_1)$$

Conditions (both required):

1. No missing confounders
2. Regression model is correctly specified

Assumption of no missing confounders

In English (using our example):

The vector $X = (X_1, X_2)$ = (gender, age) contains all variables that are related to ***both*** the outcome Y and the treatment assignment Z

In Statistics jargon:

The potential outcome vector is independent of treatment assignment given X

In symbols: $(Y_0, Y_1) \perp Z|X$

Why is this assumption important?

- For two (or more) people who have the ***exact same*** values of X, we can calculate ATE!

  Female 24 years old

- Calculate mean CROM for all such people in treatment group, $\overline{Y_1}$
- Calculate mean CROM for all such people in control group, $\overline{Y_0}$
- For this group,
- $\hat{\Delta} \equiv \overline{Y_1} - \overline{Y_0}$

Problem

- We will not get exact matches

Solution

- Use **groups** that are **close** to being alike
- Use multi-dimensional distance metrics to create groups

Even better solution—introducing the…

Propensity score, $\boldsymbol{e(X)}$:

The probability that a given person is in the treatment group

Use a ***propensity model*** to calculate $e(X)$

- Use logistic regression
- Response variable: treatment status, Z
- Input variables: gender, age, X

Powerful result (for statisticians)

Rosenbaum & Rubin 1983 (using probability theory) showed that

$X \perp Z|e(X)$

In words: the confounders are independent of treatment status given the propensity score

What does that mean?!

- The propensity score, $e(X)$, is a single number that can be used to represent the entire covariate vector $X = (X_1, X_2)$

- It has probabilistic properties that distance metrics don't have.

$\widehat{\Delta}_S$ Stratification estimator

1. Divide the dataset into K evenly-spaced strata based on the propensity score. Here, $K = 5$

   [0,0.2), [0.2,0.4), [0.4, 0.6), [0.6, 0.8), [0.8, 1]

2. Within each stratum, calculate the treatment effect, $\widehat{\Delta}_j, j = 1, \ldots K$

   $$\widehat{\Delta}_j = \frac{1}{nt_j}\sum_{\substack{trt \\ group}} Y_i \quad - \quad \frac{1}{nc_j}\sum_{\substack{ctrl \\ group}} Y_i$$

3. The average of the stratum-specific treatment effects is the overall treatment effect

   $$\widehat{\Delta}_S \equiv \sum_{j=1}^{K} \widehat{\Delta}_j$$

Problem with stratification estimator

- The strata can be too large, and the degree of homogeneity may be low within a stratum.

Solution

- Do regression adjustment within each stratum

$\widehat{\Delta}_{SR}$ Stratification with regression adjustment

1. Divide the dataset into K evenly-spaced strata based on $e(X)$.
2. Within each stratum, fit a regression model

$$Y = \beta_0 + \beta_Z Z + \beta_1 X_1 + \beta_2 X_2$$

3. The stratum-specific treatment effect will be $\widehat{\Delta}_j = \widehat{\beta_Z}$
4. The overall treatment effect is the average

$$\widehat{\Delta}_{SR} \equiv \sum_{j=1}^{K} \widehat{\Delta}_j$$

What we know so far

$$Bias(\hat{\Delta}_{SR}) < Bias(\hat{\Delta}_{s})$$

Conditions (all required):

1. No missing confounders
2. Propensity model is correctly specified
3. Regression model is correctly specified

- Done with
  - Stratification estimators
- Move on to
  - Weighting estimators

$\widehat{\Delta}_{IPW}$ Inverse-probability-weighting

Recall the definition of ATE: $\Delta \equiv \mathrm{E}(Y1) - \mathrm{E}(Y0)$

$$\widehat{\Delta}_{IPW} = \frac{1}{50} \sum_{\substack{trt \\ group}} \frac{1}{\widehat{e_i(X)}} Y_i - \frac{1}{50} \sum_{\substack{ctrl \\ group}}^{50} \frac{1}{1 - \widehat{e_i(X)}} Y_i$$

Why does IPW make sense? Simpler example.

	Total	Treatment	Control
Male	10	6	4
Female	20	5	15

Why does IPW make sense? Simple example.

	Total	Treatment	Control
Male	10	6	4
Female	20	5	15

$$\hat{e}(male) = \frac{6}{10}$$

Why does IPW make sense? Simple example.

	Total	Treatment	Control
Male	10	6 * 10/6 = 10	4
Female	20	5	15

$$\hat{e}(male) = \frac{6}{10}$$

Why does IPW make sense? Simple example.

	Total	Treatment	Control
Male	10	6 * 10/6 = 10	4
Female	20	5	15

$$\hat{e}(male) = \frac{6}{10}$$

$$1 - \hat{e}(male) = \frac{4}{10}$$

Why does IPW make sense? Simple example.

	Total	Treatment	Control
Male	10	6 * 10/6 = 10	4 * 10/4 = 10
Female	20	5	15

$$\hat{e}(male) = \frac{6}{10}$$

$$1 - \hat{e}(male) = \frac{4}{10}$$

Why does IPW make sense? Simple example.

	Total	Treatment	Control
Male	10	6 * 10/6 = 10	4 * 10/4 = 10
Female	20	5	15

$$\hat{e}(male) = \frac{6}{10}$$

$$1 - \hat{e}(male) = \frac{4}{10}$$

$$\hat{e}(female) = \frac{5}{20}$$

Why does IPW make sense? Simple example.

	Total	Treatment	Control
Male	10	6 * 10/6 = 10	4 * 10/4 = 10
Female	20	5 * 20/5 = 20	15

$$\hat{e}(male) = \frac{6}{10}$$

$$1 - \hat{e}(male) = \frac{4}{10}$$

$$\hat{e}(female) = \frac{5}{20}$$

Why does IPW make sense? Simple example.

	Total	Treatment	Control
Male	10	6 * 10/6 = 10	4 * 10/4 = 10
Female	20	5 * 20/5 = 20	15

$$\hat{e}(male) = \frac{6}{10}$$

$$1 - \hat{e}(male) = \frac{4}{10}$$

$$\hat{e}(female) = \frac{5}{20}$$

$$1 - \hat{e}(female) = \frac{15}{20}$$

Why does IPW make sense? Simple example.

	Total	Treatment	Control
Male	10	6 * 10/6 = 10	4 * 10/4 = 10
Female	20	5 * 20/5 = 20	15 * 20/15 = 20

$$\hat{e}(male) = \frac{6}{10}$$

$$1 - \hat{e}(male) = \frac{4}{10}$$

$$\hat{e}(female) = \frac{5}{20}$$

$$1 - \hat{e}(female) = \frac{15}{20}$$

$\widehat{\Delta}_{IPW}$ has some problems, however

- Suppose you get $\hat{e}(X) = 0.000000001$.
  - (That is, probability of being treated very close to 0.)
- Then $^{1}/_{\hat{e}(X)} = {}^{1}/_{0.000000001}$ =1,000,000,000
- Applying very large weights to individuals can create numeric instabilities in calculations

- The same thing happens if probability of being treated is very close to 1, because then the inverse of $1 - \hat{e}(X)$ gets very large.

Progression from $\widehat{\Delta}_{IPW}$ to $\widehat{\Delta}_{DR}$ (IPW with regression-adjustment)

- Researchers realized that $\widehat{\Delta}_{IPW}$ belonged to a larger class of estimators called ***ratio estimators***
- Ratio estimators had been thoroughly examined by missing data researchers

Is there another ratio estimator with better statistical properties than $\widehat{\Delta}_{IPW}$?

- Get rid of numerical instability when $\hat{e}(X)$ close to 0 or 1
  - By using a different, and better ratio for weighting

- Lower (or same) bias
- Lower (or same) variance

YES! There is!

$\widehat{\Delta}_{DR}$ IPW with regression-adjustment Aka Doubly-robust estimator

$$\hat{\Delta}_{DR} =$$

$$\frac{1}{50}\sum_{\substack{trt\\group}}\frac{Y_i-(Z_i-\hat{e}_i)m_{1i}}{\hat{e}_i}-\frac{1}{50}\sum_{\substack{ctrl\\group}}\frac{Y_i+(Z_i-\hat{e}_i)m_{0i}}{1-\hat{e}_i}$$

Remarkable properties of $\widehat{\Delta}_{DR}$

Out of all possible ratio estimators,

$\widehat{\Delta}_{DR}$ is the minimum variance unbiased estimator

(read: statistical holy grail)

Assumptions required

- No missing confounders
- ***One*** of the two models is correct
  - Propensity model OR
  - Regression-adjustment model

What we know

$$Bias(\hat{\Delta}_{DR}) < Bias(\hat{\Delta}_{IPW})$$

Conditions:

1. No missing confounders
2. Propensity model **OR** regression model is correctly specified

More comparisons are made in Lunceford & Davidian 2004

- When all assumptions are satisfied, comparing between all 3 groups
- When propensity model is not correctly specified (in different ways)
- When regression model is not correctly specified (in different ways)
- When extra variables are added

- Other comparisons elsewhere:
  - What if both models are not correctly specified?

- Hope you learned from what we had time for!

References

Lunceford, Jared K.; Davidian, Marie. 2004 "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study." *Statistics in Medicine* 23(19): 2937-2960.

Luo, Zhehui; Gardiner, Joseph C.; Bradley, Cathy J. 2010 "Applying propensity score methods in medical research: Pitfalls and prospects." *Medical Care Research and Review* 67(5): 528-554. Doi: 10.1177/1077558710361486.

Stuart, Elizabeth A. 2010 "Matching methods for causal inference: A review and a look forward." *Statistical Science* 25(1): 1-21. Doi: 10.1214/09-STS313.

Thank you

Amy Nail, Ph.D.

Honestat Statistics & Analytics

amynail@honestat.com

Bonus slides

(with additional reference list at the end)

Caution: look for nonlinear effects in models

- Did you look for nonlinear effects in propensity and regression-adjustment models? Do you know what graphs to produce to examine the need for nonlinear effects?
- Does your software allow nonlinear effects in propensity and regression-adjustment models?
- Do you know how to model nonlinear effects in these models when your software does allow it OR does not allow it (you can always code it yourself if your software doesn't do it)?

For confidence intervals…

- Getting uncertainty quantification (confidence intervals) right is important for decision-making. Getting their width right, and getting their shape right. Not all are symmetric.
- Can I assume normality, or should I use a bootstrap?
- Quick answer: if overall sample size is less than 500, use bootstrap CI
- Caution: This does not mean calculate bootstrap SE and use +/- 1.96*SE
- Instead, it means use the 2.5^{th} and 97.5^{th} percentiles of bootstrap estimators, because the CI may not be symmetric
- Devil is in details: boostrapping can be problematic if lots of categorical variables with small numbers in each category
- If your software doesn't do bootstrapping and your sample size is small, code it yourself and/or hire someone like myself to guide you in doing this.

Time-to-event data (aka Survival analysis) 1 of 2

- Several papers have been written by Peter Austin, but
- Though he cites Lunceford & Davidian, he doesn't seem to have read the entire paper. If he had, he would have realized that one of the methods he tests for time-to-event data—stratification--has already been shown not to work well for continuous and binary response variables, but he tests the time-to-event analog anyway. Furthermore, Lunceford & Davidian show that a slight modification—stratification with regression adjustment—works much better than stratification alone for continuous and binary response variables, but he never tests the analog to stratification with regression adjustment for time-to-event data.
- Austin's papers do not use doubly robust methods.

Time-to-event data (aka Survival analysis) 2 of 2

- Doubly robust methods are given in
  - Bai, Tsiatis, and O'Brien 2013, Section 2
  - Bai 2014 (her dissertation), Chapter 1

- How to program these?
  - No SAS procedure or macro yet
  - No R package of function yet
  - No Python package yet
  - I've done it for a client. I may write packages so that others can easily implement these methods. I'd love to find volunteers to help with this effort.

How to implement methods: SAS 1 of 2

- SAS/STAT ® Version 14.2 (SAS 9.4 Maintenance level TS1M4)
  - Proc CAUSALTRT
    - For continuous and binary response, not for time-to-event response
  - Proc PSMATCH

- Doubly robust methods macro (If you don't have SAS 9.4)
  - For continuous and binary response variables only
  - Funk, et al. 2007 paper below describes the macro, and the link to the macro is in the DOWNLOADING & INSTALLING section of the paper
  - www2.sas.com/proceedings/forum2007/189-2007.pdf

How to implement methods: SAS 2 of 2

- For survival analysis, you would have to code any of these methods yourself.
- I have coded these methods for survival analysis in SAS for a previous client. (SR and DR analogs for time-to-event data)

How to implement methods: R

- Packages
  - Drgee
  - Drtmle
- Note
  - Doubly robust methods are also used in the missing data context, so you may find

How to implement methods: Python

- CausalInference 0.1.2
  - Laurence Wong (I know nothing about this person.)
  - Appears to have doubly robust methods
  - Doesn’t appear to get CI via bootstrapping
- Adam Kelleher is working on a package
- You can hire someone like myself to help your Python programmers implement the methods using existing regression and logistic regression capabilities.

References for Bonus slides

- Austin_2013_PerformanceOfDiffPropensityScoreMethodsForEstMarginalHazardRatios_StatsInMed
- Austin_2013_UseOfPropensityScoreMethodsWithSurvivalOutcomes_StatsInMed
- Bai, Xiaofei. 2014. *Doubly-robust Estimators in Observational Studies with and without a Stratified Sub-sample.* Dissertation. https://catalog.lib.ncsu.edu/record/NCSU3088536
- Bai, Xiaofe; Tsiatis, Anastasios A.; O'Brien, Sean M. 2013 "Doubly-robust estimators of treatment-specific survival distributions in observational studies with stratified sampling." *Biometrics* 69: 830-839.