Matching to Produce Causal Estimates in Non-experimental Settings

Kaitlyn Cook, Tom Grodzicki, Harrison Reeder, Lauren Yoo

Carleton College

February 26, 2015

Motivating Example

- Examine the effect of job training on unemployment
- Let z be treatment status
  - Job training ($z = 1$)
  - No job training ($z = 0$)
- Let r^{obs} be the observed outcome
  - Unemployed ($r^{obs} = 1$)
  - Employed ($r^{obs} = 0$).
- Let x be the vector of observed covariates

Potential Outcomes

An individual's possible outcomes given different treatments

A *causal effect* compares these two potential outcomes

Potential Outcomes

Only one potential outcome is ever observed

- "Missing data" problem
- Solution: look at causal effects on a population level
- Compare outcomes between *groups* of treated and non-treated units

The Estimand: the Odds Ratio

- Compares odds of unemployment in the job training group to that in the no job training group:

$$\gamma = \frac{\frac{\overline{r_1^{obs}}}{1-\overline{r_1^{obs}}}}{\frac{\overline{r_0^{obs}}}{1-\overline{r_0^{obs}}}}$$

The Estimand: the Odds Ratio

- Compares odds of unemployment in the job training group to that in the no job training group:

$$\gamma = \frac{\frac{\overline{r_1^{obs}}}{1-\overline{r_1^{obs}}}}{\frac{\overline{r_0^{obs}}}{1-\overline{r_0^{obs}}}}$$

- 40 individuals: 20 treated, 20 non-treated
  - Among treated: 5 cases
  - Among non-treated: 10 cases

The Estimand: the Odds Ratio

- Compares odds of unemployment in the job training group to that in the no job training group:

$$\gamma = \frac{\frac{\overline{r_1^{obs}}}{1-\overline{r_1^{obs}}}}{\frac{\overline{r_0^{obs}}}{1-\overline{r_0^{obs}}}}$$

- 40 individuals: 20 treated, 20 non-treated
  - Among treated: 5 cases
  - Among non-treated: 10 cases

$$\gamma = \frac{\frac{0.25}{0.75}}{\frac{0.5}{0.5}} = \frac{1}{3}$$

The Assumption: Covariate Balance

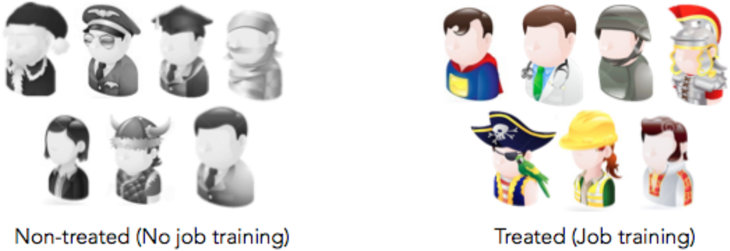
- Treated and non-treated groups must "look alike"
- Assumption not met $\Rightarrow$ covariate imbalance
  - Covariate distributions differ
- Depends on treatment assignment mechanism

Study Designs: Randomized Experiment

- Random treatment assignment
- Covariate balance by design

Study Designs: Randomized Experiment

- Random treatment assignment
- Covariate balance by design



Study Designs: Prospective Observational Study

- Treatment assignment is not random
- Treated and non-treated groups may substantially differ

Study Designs: Prospective Observational Study

- Treatment assignment is not random
- Treated and non-treated groups may substantially differ

Non-treated (No job training)

Treated (Job training)

Study Designs: Retrospective Study

- Subjects sampled based on outcomes
- Oversample "case" individuals, then sample "non-case" individuals
- Treatment assignment is not random
- Treated and non-treated groups may substantially differ

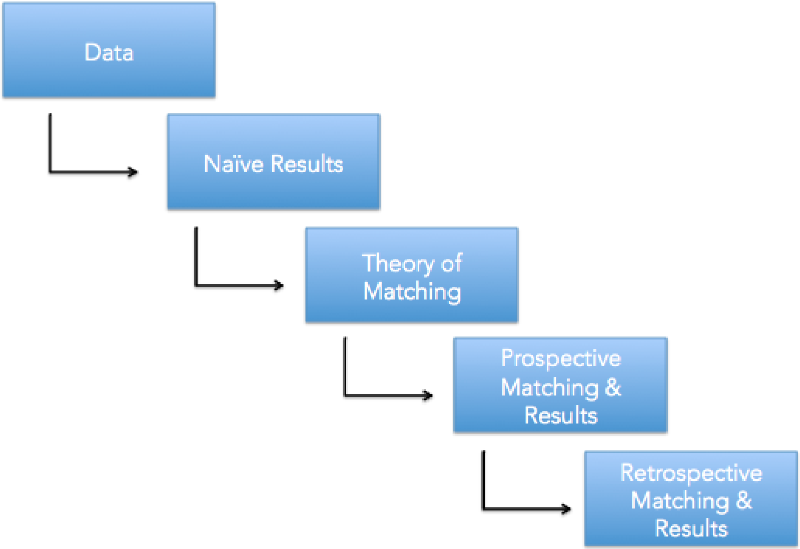
Study Designs: Retrospective Study

- Subjects sampled based on outcomes
- Oversample "case" individuals, then sample "non-case" individuals
- Treatment assignment is not random
- Treated and non-treated groups may substantially differ

Non-case (Employed)

Case (Unemployed)

Outline


Data
Naïve Results
Theory of Matching
Prospective Matching & Results
Retrospective Matching & Results

Our Data, Part I

The National Supported Work (NSW) Experiment

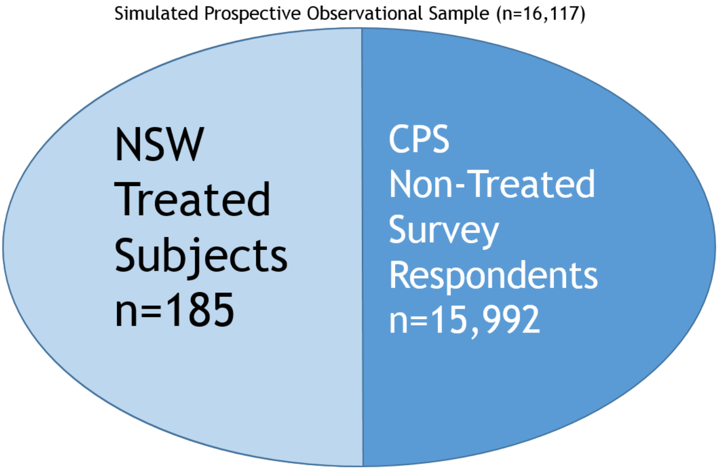
- Examined the effect of job training programs on employment
- 445 participants, 185 treated
- Outcome of interest: unemployment in 1978
- Covariates measured:
  - age
  - years of education
  - race (white, black, hispanic)
  - college degree
  - marital status
  - 1974 earnings
  - 1975 earnings

Our Data, Part II

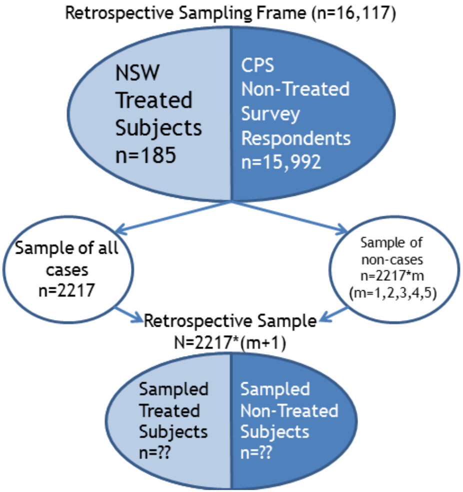
Current Population Survey (CPS)

- Administered by the Census Bureau and the Bureau of Labor Statistics
- 15,992 respondents, all non-treated
- Recorded same covariates as NSW experiment

Simulating Prospective Study



Simulating Retrospective Study



Results from Randomized Experiment

Study	Naive Plug-in
Experimental	0.59

Table: Randomized experiment estimation of odds ratio γ for unemployment

$$\gamma = \frac{\frac{\overline{r_1^{obs}}}{1-\overline{r_1^{obs}}}}{\frac{\overline{r_0^{obs}}}{1-\overline{r_0^{obs}}}}$$

- Odds ratio of 1.00 indicates a treated individual is just as likely to be unemployed as an untreated individual
- This is a value we will hope to recover with alternative methods

Naive Plug-in Results

Study	Naive Plug-in
Experimental	0.59
Prospective	2.04
Retrospective 1:1	2.04
Retrospective 1:5	2.04

Table: Prospective Study Treated v. Non-Treated, unmatched estimation of odds ratio γ for unemployment indicator

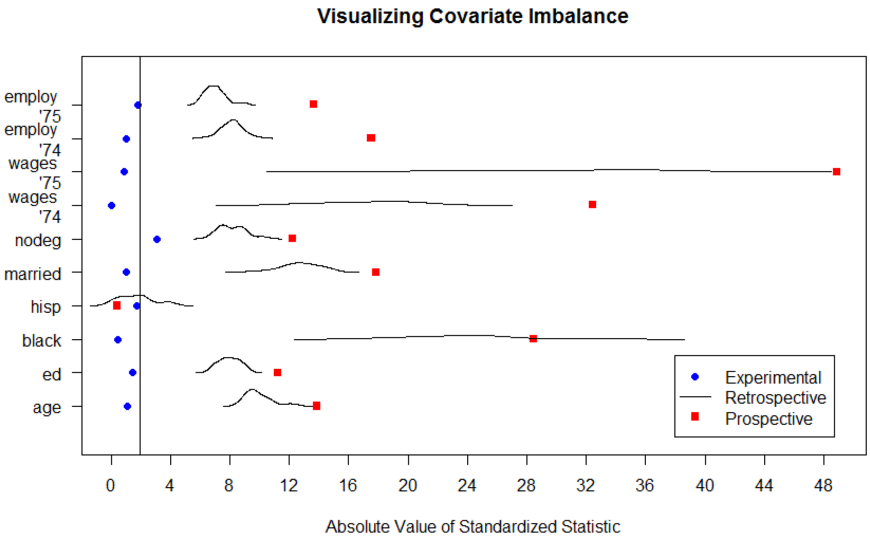


Figure: Love Plot depicting covariate imbalance

Logistic Regression

Let $p_i = P(r_i^{obs} = 1|x, z)$

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta^{treat} z_i + \beta x_i, i = 1, \ldots, n \tag{1}$$

- Controls for covariates
- Binary outcome variable $\rightarrow$ Binary logistic Regression
- Produces Log odds ratio

Results of Logistic Regression

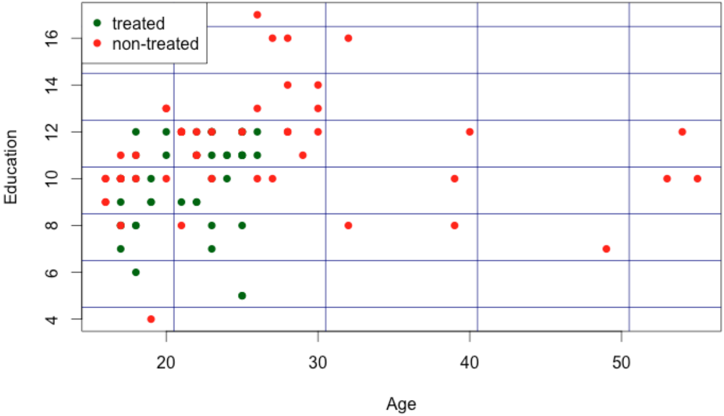
Study	Naive Plug-in	Logistic Regression
Experimental	0.59	0.58
Prospective	2.04	0.87
Retrospective 1:1	2.04	1.15
Retrospective 1:5	2.04	0.90

Table: Unmatched estimation of odds ratio γ for unemployment indicator

- Controlling for covariates means holding other covariates constant
- Covariates constant $\neq$ Covariates equal

Matching

- Match treated unit to one or more non-treated units that have similar covariates
- "Recreate" randomized experiment



Difficulty in Matching

Multi-dimensional Matching Problem

In our example, we have 8 covariates to match on: Age, education, black, hispanic, marriage status, earnings in 1974, earnings in 1975

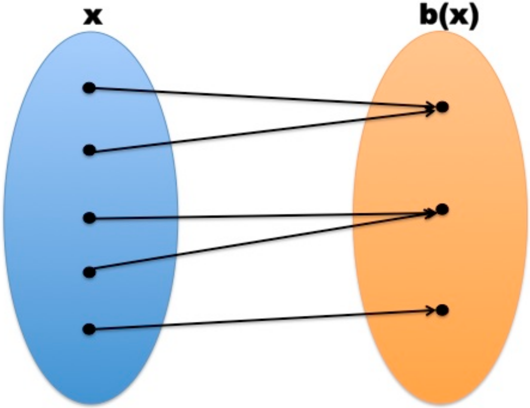
Solution

Balancing Score

Balancing Score

Definition

A balancing score $b(x)$ is a function of the observed covariates x such that $x \perp\!\!\!\perp z | b(x)$.



Propensity Score $e(x)$

Definition

Propensity score $e(x) = P(z = 1|x)$.

- The conditional probability of being assigned the treatment given the observed covariates.
- One-dimensional balancing score

Theorem 1

Theorem

Let $b(x)$ be a function of x. Then $b(x)$ is a balancing score if and only if propensity score $e(x)$ is a function of $b(x)$.

The proof uses:

- Definition of function
- Conditional Independence: $P(Z|X, Y) = P(Z|Y)$
- Law of Iterated Expectation: $E[E[Z|X]] = E[Z]$

Theorem 2

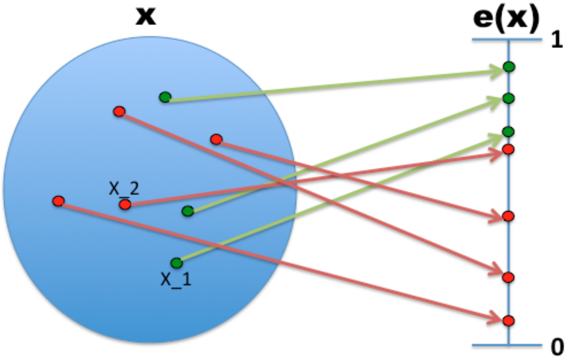
Theorem

Treatment assignment and the observed covariates are conditionally independent given the propensity score $e(x)$, that is,

$$x \perp\!\!\!\perp z | e(x),$$

- We can generate random assignment of treatment by conditioning on $e(x)$
- Matching on $e(x)$, we get treated and non-treated groups that look alike

Propensity Score Matching Example



- Ex:
  - x_1 = (25 years old, Hispanic, Not-married, Earned \$5,000 earnings in 1974,etc) $\rightarrow$ e(x_1) = .89
  - x_2 = (26 years old, White, Not-married, Earned \$4,900 in 1974,etc) $\rightarrow$ e(x_2) = .87

Effects of Matching on a Balancing Score

- Covariate balance
- Obtain unbiased average treatment effect conditioned on $b(x)$

Prospective Matched Results

- Matched treated units to non-treated units using the propensity score

Study	Conditional Logistic Regression
Experimental	0.58
Prospective	0.67

Table: Estimation of odds ratio for unemployment

- "Conditional Logistic Regression" is a form of logistic regression that incorporates the matching information.

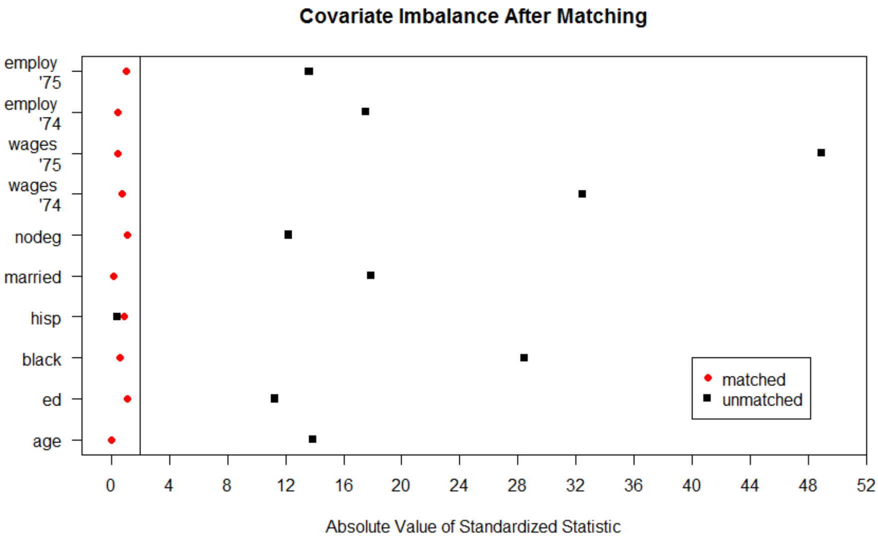
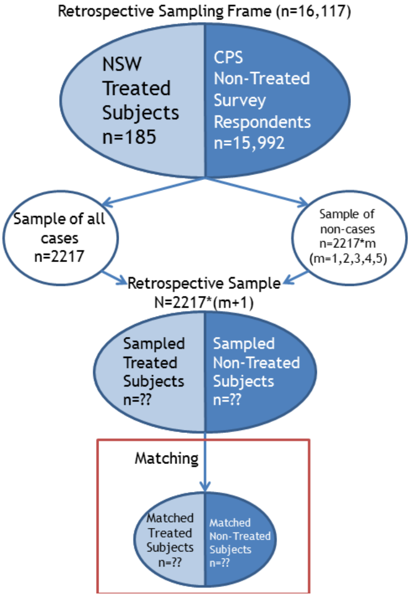


Figure: Love plot depicting covariate imbalance in matched prospective design using propensity score matching

Matching in the Retrospective Setting

- In prospective setting, matching and sampling are intertwined
- Retrospective challenge: we've already sampled subjects based on their case/non-case status
  - Observed relationship between treatment status and covariates conditioned on being sampled

Matching in the Retrospective Setting



Difficulties with Retrospective Matching

- Let s be a sampling indicator with

$$s = \begin{cases} 1 & \text{if retrospectively sampled} \\ 0 & \text{otherwise.} \end{cases}$$

- Regressing covariates x against treatment z in a retrospective sample estimates

$$P(z = 1|x, s = 1)$$

  - The propensity score $P(z = 1|x) \neq P(z = 1|x, s = 1)$
  - nor a function of it $\Rightarrow$ not a balancing score

The Prospective Exposure Case (PEC) Score

PEC score

$b_{pro}^{ty}(x) = P(z = t, r^{obs} = y|x)$ for $t, y \in \{0, 1\}$

For a subject with covariates x, PEC score is three dimensional summary showing joint conditional probabilities of treatment and outcome:

	Non-Case (y=0)	Case (y=1)
Non-Treated (t=0)		$b_{pro}^{01}(x)$
Treated (t=1)	$b_{pro}^{10}(x)$	$b_{pro}^{11}(x)$

The Prospective Exposure Case (PEC) Score

PEC score

$b^{ty}_{pro}(x) = P(z = t, r^{obs} = y|x)$ for $t, y \in \{0, 1\}$.

	Non-Case (y=0)	Case (y=1)
Non-Treated (t=0)		$b^{01}_{pro}(x)$
Treated (t=1)	$b^{10}_{pro}(x)$	$b^{11}_{pro}(x)$

Theorem

$b^{ty}_{pro}(x)$ is a balancing score, i.e., $e(x) = f(b^{ty}_{pro}(x))$ for some function f.

Proof.

$e(x) = P(z = 1|x) = b^{10}_{pro}(x) + b^{11}_{pro}(x)$ $\square$

The Retrospective Exposure Case (REC) Score

- Analogous to PEC score, but conditions on retrospective sampling.

REC Score

$b_{ret}^{ty}(x) = P(z = t, r^{obs} = y | x, s = 1)$ for $t, y \in \{0, 1\}$

	Non-Case (y=0)	Case (y=1)
Non-Treated (t=0)		$b_{ret}^{01}(x)$
Treated (t=1)	$b_{ret}^{10}(x)$	$b_{ret}^{11}(x)$

The Retrospective Exposure Case (REC) Score

Theorem

The REC score $b_{ret}^{ty}(x)$ is a balancing score, i.e., $e(x) = f(b_{ret}^{ty}(x))$ for some function f.

We prove this by writing the PEC score as a function of the REC score.

The Retrospective Exposure Case (REC) Score

Theorem

Let $p_1 = P(s = 1 | r^{obs} = 1)$, *and* $p_0 = P(s = 1 | r^{obs} = 0)$. *Assume sampling is independent of treatment and covariates, given case status. Then the REC score* $b^{ty}_{ret}(x)$ *is a balancing score, i.e.,* $e(x) = f(b^{ty}_{ret}(x))$ *for some function* f.

Proof.

$$b^{ty}_{ret}(x) = \frac{p_y b^{ty}_{pro}(x)}{p_0 b^{00}_{pro}(x) + p_0 b^{01}_{pro}(x) + p_1 b^{10}_{pro}(x) + p_1 b^{11}_{pro}(x)}$$

Inverting this, we can show $b^{ty}_{pro}(x) \propto p_{1-y} b^{ty}_{ret}(x)$. Thus, $e(x)$ is a function of $b^{ty}_{ret}(x)$. □

Estimating the REC Score

To estimate the REC score from a retrospective sample we use *multinomial regression*

- similar to logistic regression, but for modeling a multinomial variable like REC score from covariates.
- predicts each dimension of REC Score for each subject

Matching on the REC Score

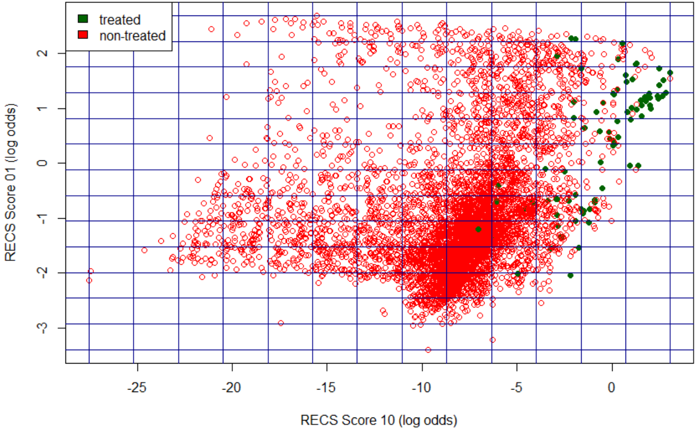


Figure: Plot showing two dimensions of RECS with strata

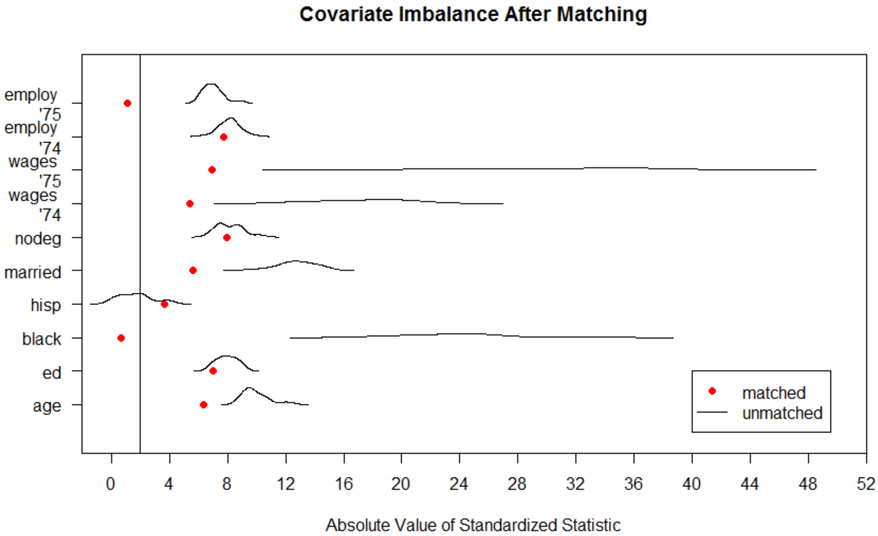


Figure: Love plot depicting covariate imbalance in matched retrospective dataset using RECS matching

Retrospective Matched Estimates

Study	Conditional Logistic Regression
Experimental	0.58
Retrospective 1:1	0.94
Retrospective 1:5	0.60

Table: Retrospective Study Treatment v. Control

- On average, 1:5 Retrospective analysis shows that receiving the job training decreases the odds of being unemployed by 40%.
  - Almost exactly recovers experimental value.

Primary Findings

- Matching is an effective way to fix covariate imbalance in prospective studies
- Results indicate REC score matching works
- Retrospectively matched sample still contains covariate imbalance

Limitations

- Matching
  - Only observed variables are matched on
  - Large sample needed to find good matches

Looking Ahead

- Put standard errors on estimates
- Implement REC score matching on another dataset
- Figure out why the RECS matching did so well
- Find a better balancing score to match on
- Lots of observational data and causal inference

Thank you!

- The inestimable Dave Watson
- Carleton Math & Stats faculty and mentors
- Our many friends and colleagues
- Donald Rubin & Paul Rosenbaum, inventors of e(x)