Contemporary Methods in Causal Inference for Program Evaluation

Noah Greifer
UNC Chapel Hill
Department of Psychology & Neuroscience
11/1/18

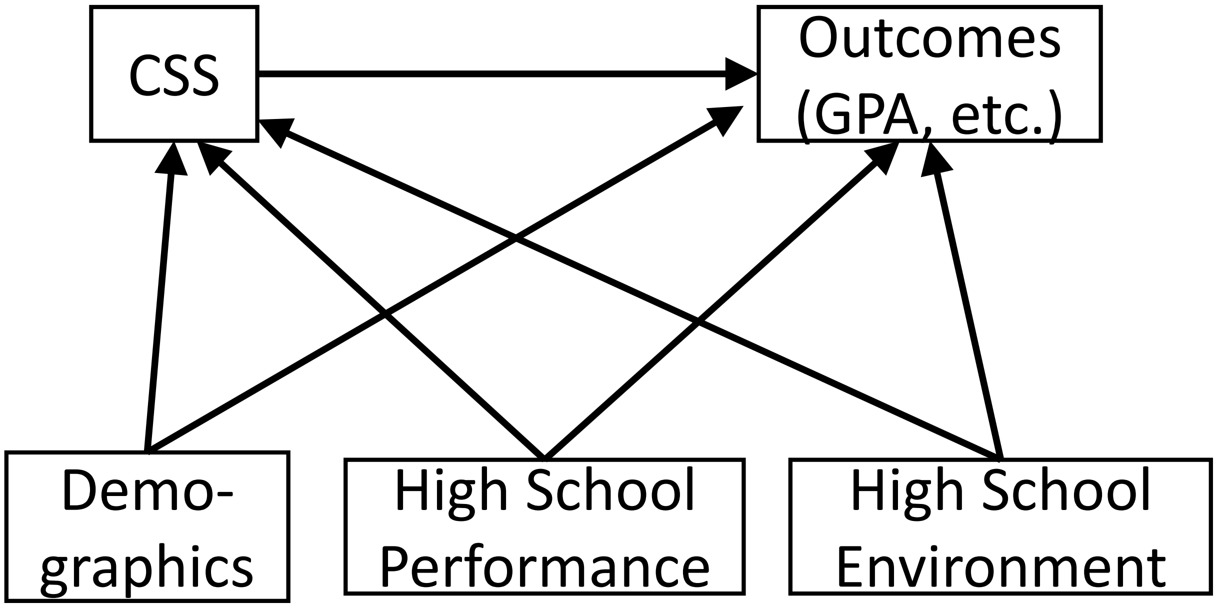
Agenda

- Causal inference basics
- Causal inference methods
  - Classic methods: Regression, PS matching, PS weighting
  - New methods
  - Additional challenges: clusters and missing data
- Application: evaluating CSS
- Demonstration of methods in R

Evaluating CSS

- Interested in the causal effect of Chancellor's Science Scholars (CSS) program on student outcomes
- Why can’t we just compare CSS and non-CSS students?
  - Intense selection: CSS only accepts the best
  - Even if CSS were not effective, CSS students would have better outcomes
- Confounding:
  - Selection into treatment is caused by variables that also cause variation in the outcome

Evaluating CSS



Causal Inference Basics

- Potential Outcomes
  - Y^1 and Y^0
  - Outcome *were* a student to be in CSS or not in CSS
  - Only one is observed
  - Individual treatment effect: $\tau_i = Y_i^1 - Y_i^0$

CSS
CHANCELLOR'S
Science Scholars

GPA:
3.6

GPA:
3.0

CSS
CHANCELLOR'S
Science Scholars

GPA: 3.2

CSS
CHANCELLOR'S
Science Scholars

GPA: 3.2

Causal Effects

- Average Treatment Effect in the population (ATE)
  - $E[\tau] = E[Y^1 - Y^0] = E[Y^1] - E[Y^0]$
  - Difference between giving everyone treatment versus giving everyone control
- Average Treatment Effect in the treated (ATT)
  - $E[\tau|T = 1] = E[Y^1|T = 1] - E[Y^0|T = 1]$
  - Difference between giving treatment to those who received treatment versus giving control to those who received treatment
- Average Treatment Effect in the control (ATC)
  - $E[\tau|T = 0] = E[Y^1|T = 0] - E[Y^0|T = 0]$
  - Difference between giving treatment to those who received control versus giving control to those who received control

Causal Effects

- $E[Y^1|T = 1]$ is known; Y^1 for those with $T = 1$ is just Y
- We need to estimate $E[Y^0|T = 1]$
  - What if those who received treatment had instead received control?
- Simulate $E[Y^0|T = 1]$ using control group
  - In control group, we know Y^0
  - Need subset of the control group that is *exchangeable* with treated group

Assumptions

- Conditional Exchangeability (CE):
  - We have measured a sufficient set of variables required to remove confounding
- Positivity:
  - All units have a nonzero probability of being in treatment or control
- Stable Unit Treatment Value Assumption (SUTVA):
  - Outcome do not depend on treatment status of others
- Consistency:
  - No unmeasured versions of treatment

Methods

Regression/ANCOVA

- ANCOVA:
  - $Y = \mu_t + \boldsymbol{\beta X} + \epsilon$
  - Comparison between μ_1 and μ_0 is the treatment effect estimate
  - Biased if any treatment effect heterogeneity
- Regression:
  - $Y = \beta_0 + \tau T + \boldsymbol{\beta_1 X} + \boldsymbol{\beta_2 X} T + \epsilon$
  - If all X centered, τ is the treatment effect estimate
  - Generally a good strategy, but has weaknesses

Regression/ANCOVA

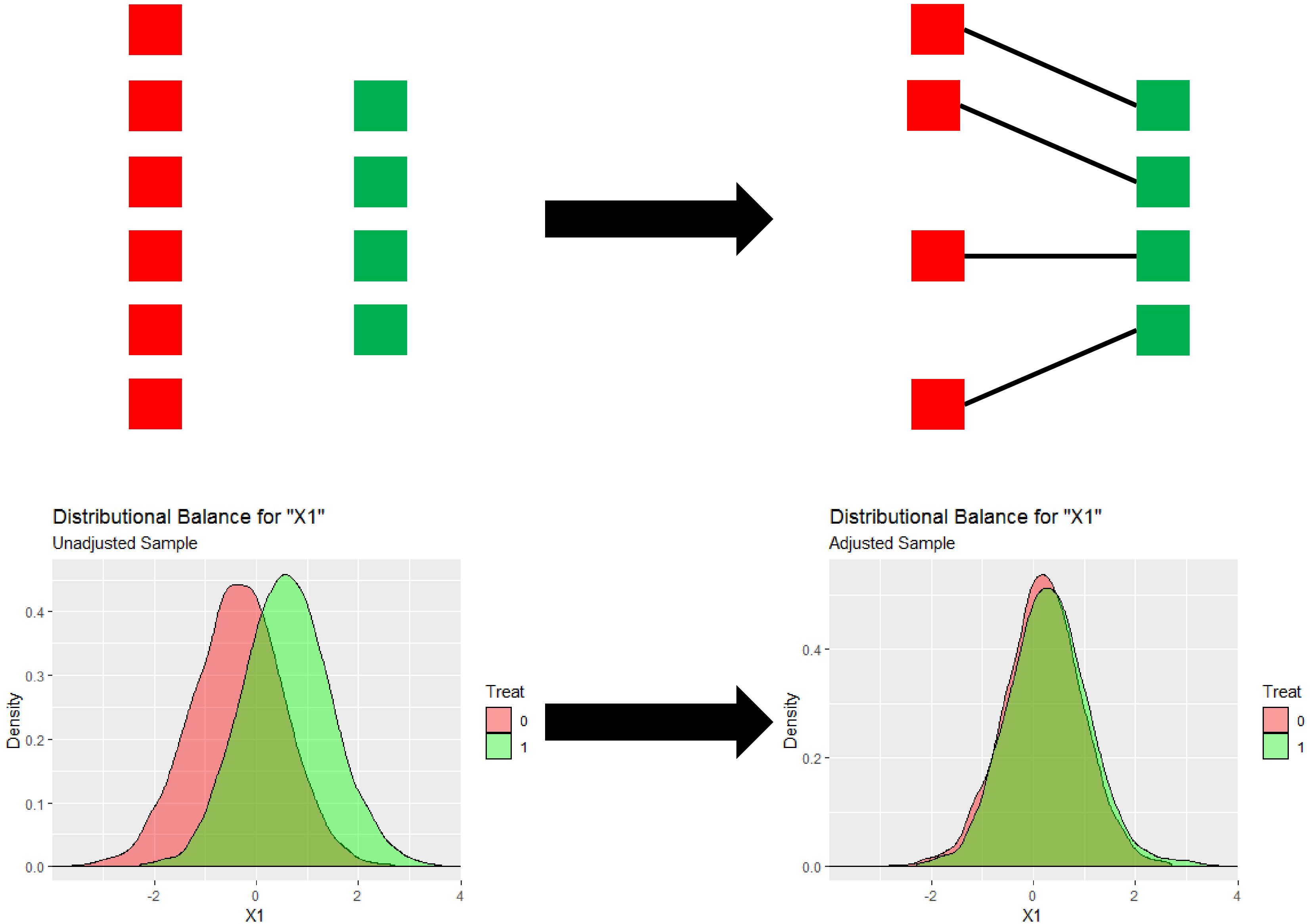
- Strengths:
  - Low standard errors → high power if there is an effect
  - Easy to implement (not if you want to get fancy, which you should)
  - Does a good job at eliminating bias (even better if you want to get fancy)
- Weaknesses
  - Requires correct parametric form
  - Model may not be stable with many variables
  - Less straightforward to estimate ATT
  - Doesn’t separate design form analysis: easy to cheat (even by accident)

Design/Preprocessing

- Matching, Weighting, Stratification
- Done without looking at outcome variable until final estimation step
  - Can perform many exploratory analyses without biasing inference
- Less sensitive to functional form

Matching

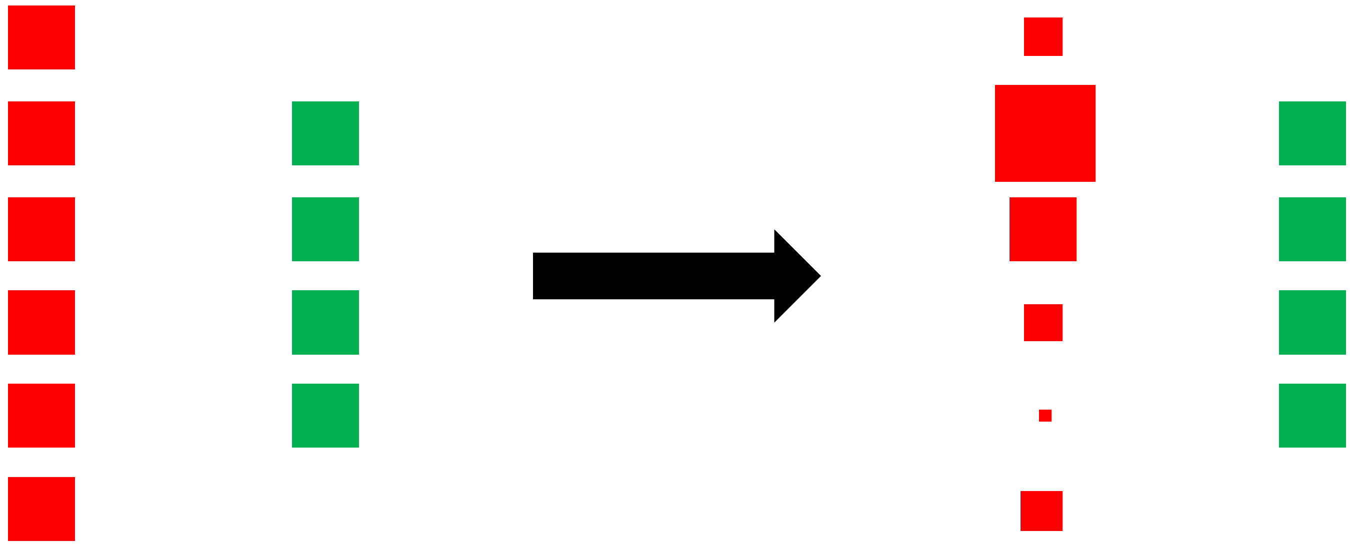
- Finding a set of controls with a similar covariate distribution to the treated
  - Discarding the rest
  - Distinct from pairing: finding pairs/small subsets of treated and controls
- Historically, matching is done by finding pairs; paired units form matched set
  - Debate over whether to retain pairing if done this way
- Effect estimate is comparison of means in matched sample
  - Independent samples t-test, paired sample t-test, other paired/non-paired methods

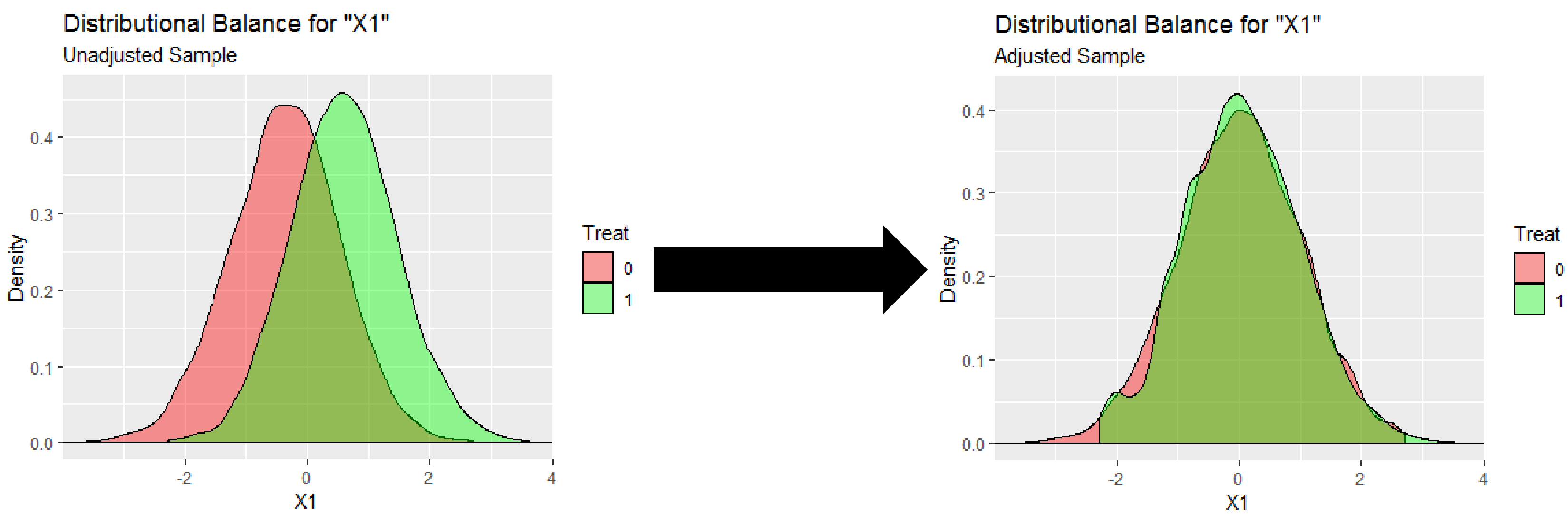


Distributional Balance for "X1"
Unadjusted Sample
Density
X1
Treat
0
1
Distributional Balance for "X1"
Adjusted Sample
Density
X1
Treat
0
1

Weighting

- Estimate weights (like sampling weights) that when applied to the sample yield balanced samples
- Effect estimate is comparison of weighted means
  - Weighted t-test, weighted regression
  - Standard errors by bootstrap, sandwich standard error, or, in some cases, analytical formula





Distributional Balance for "X1"
Unadjusted Sample
Density
X1
Treat
0
1
Distributional Balance for "X1"
Adjusted Sample
Density
X1
Treat
0
1

Matching vs. Weighting

Matching

- Exact pairing eliminates functional form assumptions
- Easy to explain and interpret
- Can only estimate ATT or ATC
- Can discard many units, decreasing power
- Relies on good matches

Weighting

- Flexible and smooth
- Generally more power and less bias
- Can estimate any estimand

Propensity Scores

- One-dimensional summary of all covariates
- $\widehat{ps} = \hat{P}(T = 1|\boldsymbol{X})$
- Rosenbaum & Rubin (1983):
  - If conditioning on X is sufficient to eliminate bias, conditioning on propensity score is sufficient to eliminate bias
- Propensity score matching
- Propensity score weighting
- Propensity score stratification
- Propensity score ANCOVA

Preprocessing Process

1. Decide on estimand, variables, assumptions
2. Preprocess: matching/weighting
3. Assess balance and sample size
   - If poor, redo steps 2 and 3
4. Estimate treatment effect

Historical Approaches: Matching

- Mahalanobis distance matching
- Propensity score matching
  - PS estimate with logistic regression or other parametric model using maximum likelihood
- Many options:
  - With or without replacement
  - With or without a caliper
  - 1:1, 1:k, or variable 1:k
  - Top to bottom, bottom to top, random

Historical Approaches: Matching

- Problems:
  - Balance not guaranteed
  - Need to manually search through PS specifications and matching options to find good balance while retaining power
  - Problems with bias and efficiency
  - Some options affect inference

Historical Approaches: Weighting

- Propensity score weighting
  - PS estimated with logistic regression or other parametric model using maximum likelihood
  - $w_i = T_i + (1 - T_i)\frac{\widehat{ps}_i}{1-\widehat{ps}_i}$
- Options:
  - Trimmed/truncated weights
  - Stabilized weights

Historical Approaches: Weighting

- Problems:
  - Balance not guaranteed
  - Need to manually search through PS specifications and to find good balance while retaining power
  - Often extreme weights cause instability in effect estimates → low power, high variability
  - Trimming methods *ad hoc* and can change inference

Historical Approaches

- In general, parametric approaches
  - Maximize likelihood, not related to causal inference goals of balance
  - Require manual respecification
  - Require knowledge or testing of nonlinearities and interactions
  - Are sensitive to model choices

Machine Learning: Matching

- Genetic matching
  - Use “evolutionary algorithm” to find best matches by trying out many matches and optimizing toward best ones
  - Automatically seeks out balance, no need for iterative balance checking
  - No need to estimate PS

Machine Learning: Matching

- Problems
  - Computationally intensive
  - Doesn’t always find a good solution; now what?
  - Solution may not be truly optimal

Machine Learning: Weighting

- Generalized boosted modeling (GBM):
  - Estimates PS using boosted logistic regression
  - Doesn’t require functional form assumptions; automatically includes interactions and nonlinearities
  - Automatically seeks out balance, no need for iterative balance checking
  - Decent performance in simulations and empirical studies
- SuperLearner:
  - Combines many machine learning methods to find the best
  - Focus on prediction rather than balance

Machine Learning: Weighting

- Problems:
  - Computationally intensive
  - Doesn’t always find a good solution; now what?
  - Solution may not be truly optimal
  - Some methods focus on good prediction rather than balance, which is not useful for causal inference

Machine Learning

- In general, machine learning methods
  - Automate the process of finding good balance, but treat it as a mostly random search
  - Can have an objective function related to balance
  - Can account for nonlinearities and interactions
  - Are not fully optimal
  - Computationally intensive
  - Frequently fail, but can be effective

Optimization Hybrids: Matching

- Optimal PS Matching
  - Finds matched pairs that minimize overall propensity score distance
  - Similar options to nearest neighbor
  - Similar performance to nearest neighbor
- Optimal Full PS Matching
  - Finds strata containing at least one treated unit that minimize overall propensity score distance across strata
  - Uses every unit, so none are discarded
  - Usually analyzed using weights generated from matching process rather than as matched data
  - Tends to have very good performance, a balance between matching and weighting

Optimization Hybrids: Weighting

- Covariate Balance Propensity Score (CBPS)
  - Estimates PS using logistic regression
  - Uses generalized method of moments to include balance as optimization criterion along with prediction
  - Just-identified version focuses only on balance and not prediction
  - Often yields close to exact balance, especially just-identified version

Optimization Hybrids

- In general, optimization hybrids
  - Combine optimization with some parametric formulation
  - Optimize some criterion using algorithm rather than random search
  - Satisfy those who desire optimization but respect classic approaches
  - Often yield excellent performance
  - Less sensitive to modeling choices
  - Can fail to converge with little advice on how to fix

Optimization Methods: Matching

- Balanced Optimal Subset Selection (BOSS)
  - Finds subset of controls that yield balance with treated
  - No pairing, only matching
  - No propensity score required
  - Users decide exactly what balance means, optimizer satisfies conditions (if possible)
- Cardinality matching/*designmatch*
  - Finds matched pairs that satisfy user-specified balance constraints
  - Exact balance, fine balance, and approximate balance are possible
  - Can optimize size of matched set for given balance constraints
  - Current implementation in R

Optimization Methods: Matching

- Problems:
  - Requires good matches
  - Computationally intensive
  - Can fail to converge or require approximate solution
  - Not well implemented in software
    - BOSS must be manually programmed
    - *designmatch* full of bugs and limitations, though promising if fixed
  - Require specification of balance constraints
    - But easy specifications are allowed

Optimization Methods: Weighting

- Entropy balancing
  - Bypasses propensity scores to estimate weights directly
  - Guarantees exact balance on all covariate if possible
  - Weights do not vary much from 1 → stability, power
  - Doubly robust and semiparametric efficient
- Minimal Approximately Balancing Weights/*optweight*
  - Uses can specify exactly how much balance is required; can manage bias/variance trade-off
  - Weights are optimized for power and stability
  - Solution more likely to be found
  - Failure to converge can be easily diagnosed
  - Not computationally intensive; sometimes faster than logistic regression!

Optimization Methods: Weighting

- Problems:
  - Exact balancing may be too strict (but can be relaxed)
  - May fail to converge (but can be diagnosed)
  - Requires good substantive knowledge to be most effective (but still very effective even without)

Optimization Methods

- In general, optimization methods
  - Create balanced sets without estimating propensity scores
  - Can achieve perfect or user-specified balance
  - Doubly robust and efficient
  - Easy to use, technology exists
  - Tend to outperform other methods when used correctly

What should I use?

- Whatever works best!
  - Best balance while maximizing effective sample size/minimizing variability of the weights
- (Optimization will almost always be best)
- Weighting will almost always get you best balance and sample size, but doesn’t have advantage of pairing
  - Weighting methods tend to be more refined and easier to use

Clustered Data and Missing Data

Clustered Data

- Clustered data: students within school/cohorts, patients within hospitals, etc.
- Need to account for unit-level confounding AND cluster-level confounding
  - E.g., if gender affects treatment and outcome, but also proportion of each gender in cluster affects treatment and outcome
- With regression, we might use a multilevel model or fixed effects model to account for cluster membership

Clustered Data

- Approaches: CAC, CWC, Hybrids
- Conditioning Within Clusters (CWC)
  - Analysis takes place separately within each cluster, then estimates are combined at the end, conditioning on cluster
  - Automatically eliminates cluster-level confounding
  - Can be challenging to find good balance and sample size, especially with matching

Clustered Data

- Conditioning Across Clusters (CAC)
  - Analysis ignores cluster membership or treats it as a covariate
  - Requires balancing on cluster-level confounders
  - Easy to do, but hard to eliminate bias due to confounding
- Hybrid Approaches
  - Preferential CWC
  - Conditioning within latent classes/similar clusters

Clustered Data

- Notes:
  - Clustering must be taken account of in preprocessing, outcome model, or both
  - If interested in moderation by cluster (i.e., cluster-specific effects), need to use CWC and include cluster in outcome model
  - If interested in overall effect and cluster is an instrumental variable (affects treatment but not outcome), conditioning on cluster can induce bias
  - If using propensity scores, let balance be your guide, not theory or simulation

Missing Data

- Approaches: deletion, missingness indicators, multiple imputation
- Deletion
  - Deleting cases with missing data
  - Always the wrong choice, especially with small sample size and missingness in many variables

Missing Data

- Missingness Indicators
  - Treat missingness as its own variable, fill in value for missing data with a constant
  - Perform modeling and balance checking with filled in variables and missingness indicator
  - Method has some success, but tends not to be as effective as multiple imputation

Missing Data

- Multiple Imputation
  - Fill in missing values with best guess; create many data sets with different best guesses
  - Best guesses generated from predictive model
  - Recommendations:
    - Use many imputations
    - Include outcome in imputation
    - Impute missing outcomes
    - Use multiple imputation with chained equations (MICE)
    - Use machine learning for prediction models
  - Two approaches: across and within

Missing Data

- Multiple Imputation
  - Across approach
    - Estimate propensity scores within each imputed dataset
    - For each unit, compute average propensity score across imputed datasets
    - Perform matching/weighting on average propensity score
    - Perform effect estimation in matched/weighted sample
    - Benefits:
      - Don’t need to estimate multiple outcome models or pool results
      - Tends to be effective when you don’t use the outcome to impute the predictors
    - Problems:
      - Balance not guaranteed and hard to assess
      - Tends to be ineffective compared to within approach when you do use the outcome to impute predictors
      - Not compatible with missingness in outcomes
      - Not compatible with approaches that don’t use propensity scores
      - Unable to include missing covariates in outcome model

Missing Data

- Multiple Imputation
  - Within approach
    - Perform entire analysis within each imputed data set
    - Combine effect estimates at the end using Rubin's rules
    - Benefits:
      - Straightforward to achieve and assess balance
      - Tends to work well when outcomes are used to impute predictors
      - Statistically validated and understood
      - Compatible with all preprocessing and analysis methods
    - Problems:
      - Many places for analysis to fail; each imputed dataset is an opportunity for failure to converge
      - May not be straightforward to combine estimates, e.g., for model comparison
      - Increase in standard error due to imputation uncertainty

Overall Recommendations

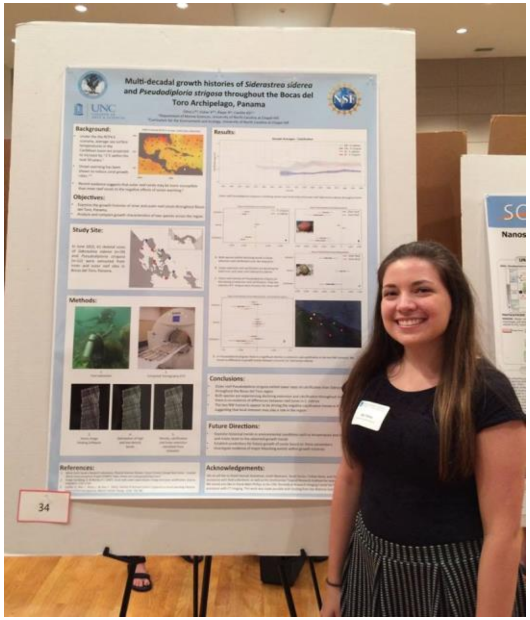
- Preprocessing
  - Use optimization-based weights
    - Entropy balancing if you have a large sample size and want a well-established method
    - Optweight if you need to maximize power and don’t mind using a new method (it will soon become more mainstream)
  - Avoid 1:1 matching unless treated and control groups are both huge and good matches exist
  - If matching, use pairing and perform paired analysis (accounting for pair membership)
- Clustered Data
  - CWC if possible; otherwise hybrid approach with careful eye on cluster-level covariates
- Missing Data
  - Multiple imputation, “within” approach
    - Use best practices in multiple imputation
    - Ensure average balance across imputations is satisfactory

Estimating the Causal Effect of CSS Participation

Chancellor's Science Scholars

- Academic enrichment program at UNC
- Funded in part by HHMI
- Intends to increase presence of underrepresented people in STEM PhD programs
  - Racial minorities, gender minorities, first-generation students, rural students, etc.
- Involves 13 pillars, including advising, mentorship, financial scholarship
- Intensive and competitive selection process based on demographics, high school scores, and interview

CSS
CHANCELLOR'S
Science Scholars



Research Goal

- Estimate the causal effect of CSS on prognostic outcomes
  - Cumulative GPA, DFWU rate, STEM major

Data

- Data on four cohorts: 2013, 2014, 2015, 2016
  - CSS and non-CSS (Science Interested)
- Followed for (up to) 4 years

Cohort	Admission	Year 1	Year 2	Year 3	Year 4	n_{CSS}	$n_{non\text{-}CSS}$
2013						23	1207
2014						33	1348
2015						35	1501
2016						36	1785

Data

- Variables:
  - Treatment: Participation in CSS (Y/N)
  - Outcomes:
    - Cumulative GPA
    - DFW unit rate per unit attempted
    - STEM major (Y/N)

Data

- Covariates:
  - Age
  - Sex
  - Race/ethnicity
  - URM status
  - State residency
  - Citizenship
  - FGC student
  - Financial need
  - ACT Math, Science, English
  - SAT Math, Reading & Writing
  - High school type
  - High school sector
  - Any APs taken
  - No. APs
  - No. APs ≥ 3
  - No. APs ≥ 4
  - No. APs = 5
  - Any Science APs taken
  - No. Science APs
  - No. Science APs ≥ 3
  - No. Science APs ≥ 4
  - No. Science APs = 5
  - AP Calculus score
  - STEM major at application

Methods

- Causal Estimand:
  - Marginal effect of CSS on outcomes for CSS participants
  - Target population: students like CSS students
    - Average treatment effect on the treated; ATT
  - Integrating over cohorts, allowing for effect moderation by cohort
    - Involves estimating effects in each cohort and then combining results
- Problems:
  1) Confounding
  2) Incomplete cohort design
  3) Missing Data

Methods

- Problem: Confounding
  - CSS students very different from non-CSS due to competitive selection process
  - If unaddressed, effect estimate will be biased

Missing Data

Cohort	Admission	Year 1	Year 2	Year 3	Year 4	n_{CSS}	$n_{non\text{-}CSS}$
2013						23	1207
2014						33	1348
2015						35	1501
2016						36	1785

Missing Data

Cohort	Admission	Year 1	Year 2	Year 3	Year 4	n_{CSS}	$n_{non\text{-}CSS}$
2013						23	1207
2014						33	1348
2015						35	1501
2016						36	1785

- Missing predictors: race/ethnicity, ACT scores, SAT scores, high school type, high school sector, high school GPA
- Missing outcomes: Cumulative GPA, STEM major

Missing Data

- Multiple Imputation
  - Create 20 data sets
  - Generate predictions of missing values within each data set
    - Multiple imputation with chained equations (MICE)
    - Random forests for imputation of all variables
    - Impute outcome and use outcomes to impute predictors
  - Perform analysis separately within each data set
  - Combine results using Rubin’s rules for estimates and standard errors

Methods

- Analysis by outcome year

Cohort	Admission	Year 1	Year 2	Year 3	Year 4	n_{CSS}	$n_{non\text{-}CSS}$
2013						23	1207
2014						33	1348
2015						35	1501
2016						36	1785

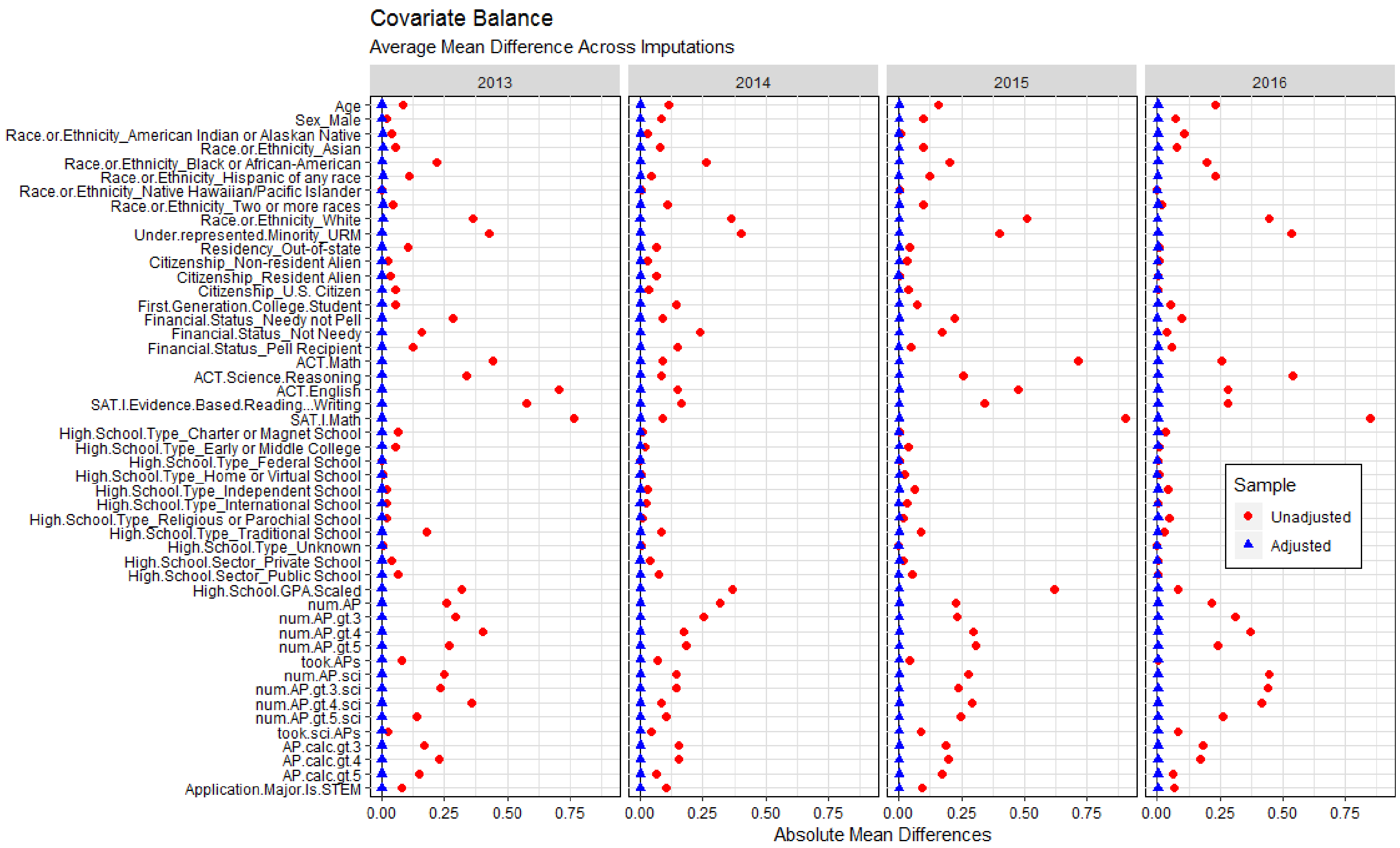
Overall Analysis Plan

1. Multiply impute data
2. Estimate balancing weights in each cohort in each imputed data set
3. Estimate treatment effect in each cohort in each imputed data set
4. Combine effect estimates across cohorts and imputed data sets for final estimate

Balancing Weights

- Using optweights
  - Minimize variance of the weights while satisfying approximate balance constraints
  - Balance constraints:
    - 0.001 for differences in proportion for binary variables
    - 0.001 for standardized mean differences ($\frac{\bar{x}_1 - \bar{x}_0}{s_1}$) for continuous variables
    - 0.1 for standardized mean differences for square of ACT Math, SAT Math, and HS GPA
  - Convergence:
    - In several cohort-imputations, optimization failed to converge due to strict tolerances
    - In some subsets, need to relax constraints of race/ethnicity and square of High School GPA

Balance



Covariate Balance
Average Mean Difference Across Imputations
2013
2014
2015
2016
Age
Sex_Male
Race.or.Ethnicity_American Indian or Alaskan Native
Race.or.Ethnicity_Asian
Race.or.Ethnicity_Black or African-American
Race.or.Ethnicity_Hispanic of any race
Race.or.Ethnicity_Native Hawaiian/Pacific Islander
Race.or.Ethnicity_Two or more races
Race.or.Ethnicity_White
Under.represented.Minority_URM
Residency_Out-of-state
Citizenship_Non-resident Alien
Citizenship_Resident Alien
Citizenship_U.S. Citizen
First.Generation.College.Student
Financial.Status_Needy not Pell
Financial.Status_Not Needy
Financial.Status_Pell Recipient
ACT.Math
ACT.Science.Reasoning
ACT.English
SAT.I.Evidence.Based.Reading...Writing
SAT.I.Math
High.School.Type_Charter or Magnet School
High.School.Type_Early or Middle College
High.School.Type_Federal School
High.School.Type_Home or Virtual School
High.School.Type_Independent School
High.School.Type_International School
High.School.Type_Religious or Parochial School
High.School.Type_Traditional School
High.School.Type_Unknown
High.School.Sector_Private School
High.School.Sector_Public School
High.School.GPA.Scaled
num.AP
num.AP.gt.3
num.AP.gt.4
num.AP.gt.5
took.APs
num.AP.sci
num.AP.gt.3.sci
num.AP.gt.4.sci
num.AP.gt.5.sci
took.sci.APs
AP.calc.gt.3
AP.calc.gt.4
AP.calc.gt.5
Application.Major.Is.STEM
0.00
0.25
0.50
0.75
Absolute Mean Differences
Sample
Unadjusted
Adjusted

Effect Estimation: GPA

- $E[GPA] = \beta_0 + \beta_1 Y_{2014} + \beta_2 Y_{2015} + \beta_3 Y_{2016} + \beta_4 CSS + \beta_5 Y_{2014} \times CSS + \beta_6 Y_{2015} \times CSS + \beta_7 Y_{2016} \times CSS$
- Effect estimate = $\frac{\beta_4+\beta_5+\beta_6+\beta_7}{4}$
- Weighted regression with robust standard error
- Combining results across imputed data sets

Effect Estimation: STEM Major

- Binomial regression with identity link
  - Actually doesn't matter because of saturated model
- Estimand: difference in probability of being a STEM major at end of year

Effect Estimation: DFWU/AH

- Zero-inflated negative binomial regression of DFWU on cohort and CSS
- Log Attempted Hours (AH) as offset
- Mixture of students who never fail a class (fail ineligible) and students who might fail one or more classes (fail eligible)

Effect Estimation

	GPA	STEM Major	DFWU: Ineligible	DFWU: Rate
Non-CSS	3.287	.963	.706	.170
CSS	3.537	.899	.795	.092
ATT	**0.250**	**-.064**	**.089**	**.542**
P-value	**.000**	**.026**	**.122**	**.000**

Conclusions

- CSS has positive impacts on students after the first year
  - Negative STEM major effects are possibly an artifact and disappear in later years
- Effects are meager due to the pre-existing condition of excellence in CSS and weighted comparison group, but still meaningful

R Example

https://github.com/ngreifer/Causal-Webinar-11-1-18/