Berkeley Public Health

Estimation approaches for causal inference: parametric and semi-parametric estimators

Jennifer Ahern, PhD MPH
Professor of Epidemiology
Associate Dean for Research
University of California, Berkeley, School of Public Health
Laura Balzer, PhD Mphil
Assistant Professor
University of Massachusetts, Amherst, School of Public Health & Health Sciences

Acknowledgments

SERtalks

Jennifer Ahern
University of California, Berkeley

Laura Balzer
University of Massachusetts-Amherst

- NICHD/NIH Office of the Director DP2HD080350
- NIAID R01AI125000, U01AI099959, UM1AI068636

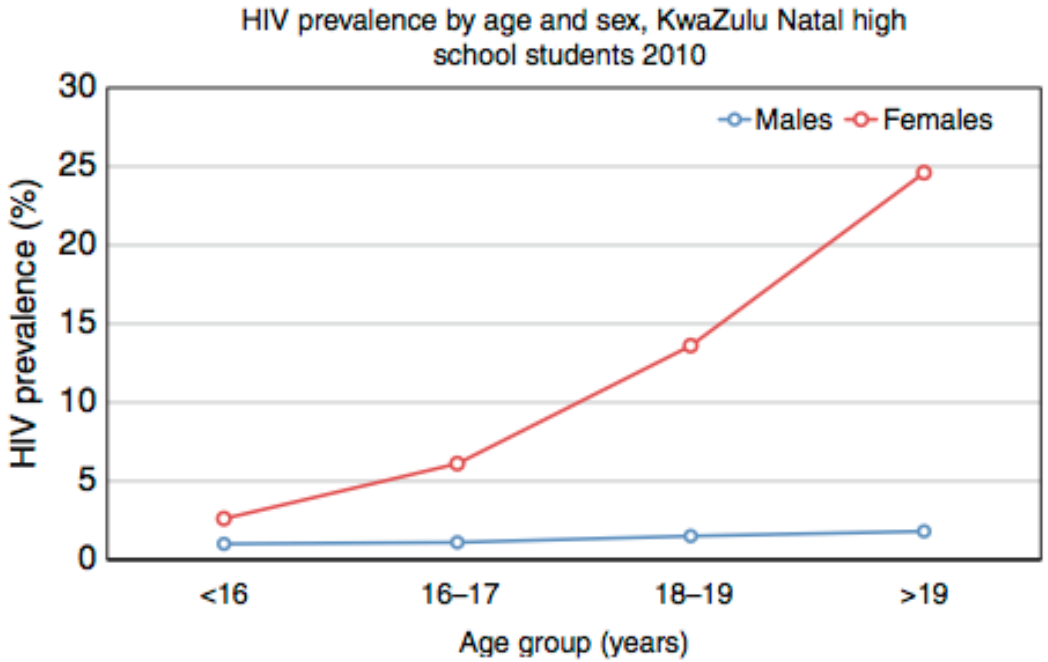
- Alan Hubbard
- Maya Petersen
- Mark van der Laan

Overview

- Causal Roadmap
  - Applied example of HIV prevention intervention
  - More in depth look at estimation methods

Causal roadmap

- Adolescent girls and young women in Southern and Eastern Africa are at high risk of HIV infection
- Interested in the impact of a school-based prevention package and cumulative HIV incidence



Dellar *et al.*, 2015

Causal roadmap

- Collect or gain access to data on school-based prevention packages, HIV incidence, and covariates

- Usual approach:
  - Note that the outcome is binary, and thus use logistic regression for analysis
  - Estimate conditional odds ratios by exponentiating the regression coefficient on the prevention package

- Problem:
  - This allows the tool (i.e., logistic regression) to define the question you answer, rather than starting with the question and picking amongst tools that allow you to answer that question

- Solution:
  - ?

Causal roadmap

- Framework for estimating causal parameters from real data

1. State the scientific question and hypothetical experiment
2. Define the causal model and parameter of interest
3. Link the causal model to the observed data and define the statistical model

Causal roadmap

4. Assess identifiability: link the causal effect to a parameter estimable from the *observed* data
5. Choose and apply the estimator
6. Derive an estimate of the sampling distribution (statistical uncertainty)
7. Make inferences (interpret findings)

Notation

- Notation/terminology
  - Exposure = treatment
    - Term treatment often used in causal inference even with exposures that are not medical treatments
  - *A* = exposure
    - A=1 for exposed (treated)
    - A=0 for unexposed (control)
  - *Y* = outcome (as usual)
  - *W* = set of measured confounding variables

Notation

- $E[Y|A=a]$: expected outcome Y among those who experience exposure $A=a$ in our population
  - Descriptive/association
- $E[Y_a]$: expected counterfactual outcome Y_a when all experience exposure $A=a$ in our population
  - Causal
- Generally $E[Y|A=a]$ does *not* equal $E[Y_a]$
  - A central problem in causal inference

Notation

- $E[Y|A=a,W=w]$: expected outcome Y among those who experience exposure $A=a$ and have covariates $W=w$ in our population
  - e.g. the mean outcome among exposed men
  - Multivariable regression models estimate conditional expectations – expected value of the outcome conditioned on ("among those") with the exposure and covariate level
    - E.g. logistic regression provides estimates of the conditional log odds ratio
- $E\{E[Y|A=a,W]\}$: expected outcome Y among those who experience exposure $A=a$ and have covariates $W=w$, averaged across the covariate strata in our population
  - No longer conditional on the confounders, instead averaged across the confounders – a marginal expectation

1. State the question

1. State the scientific question and hypothetical experiment
   - A helpful way to be clear about the scientific question is to explicitly state the experiment that would unambiguously yield estimates of the causal effect of interest

1. State the question

- Example: what is the effect of a school-based prevention package on the cumulative HIV incidence among young women in East Africa?
- We consider one hypothetical experiment today
    - What would be the difference in HIV incidence if all schools received the prevention package compared to no schools?
    - To sharply frame our question, we want to be more specific
        - The target population (what age group? where?)
        - The exposure (what are the package components?)
        - The outcome (over what time frame?)
        - Ways to change the exposure and their plausibility

1. State the question

- Example: what is the effect of a school-based prevention package on the cumulative HIV incidence among young women in East Africa?
  - Other interesting hypothetical experiments:
    - What would be the difference in HIV incidence if all schools received the package, compared to if package implementation remained as observed?
    - What would be the difference in HIV incidence if an additional 10% of schools received the package, compared to if package implementation remained as observed?
    - Note flexibility in experiments

2. Define the causal parameter

- Causal model used to formalize what we know
  - Which variables affect each other
  - The role of unmeasured/background factors
  - Functional form of relationships
- Causal models
  - Potential outcomes framework (Rubin 1986)
  - Structural causal model and causal graphs (Pearl 2000)
- For our example, we are interested in the ATE
  - $E[Y_1] - E[Y_0]$
  - Difference in the expected counterfactual cumulative HIV incidence if all schools received the prevention package vs. no schools received the package

3. Link causal to observed

- Observed data are *O=(W,A,Y)*
  - *W* as measured covariates, *A* as the exposure, and *Y* as the outcome
- We assume the causal model provides a description of our study under
  - Existing conditions (i.e. the real world)
  - Interventions (i.e. the counterfactual world)
- This provides a link between the causal world and the real (observed data) world

4. Assess identifiability

4. Assess identifiability: link the causal effect to a parameter estimable from the *observed* data
   - Temporality: exposure precedes outcome
   - Consistency*:* $Y_a = Y$ when $A=a$
   - Stability: no interference between units
   - Randomization: no unmeasured confounding
   - Positivity: sufficient variability in the exposure within confounder strata

4. Assess identifiability

- The G-computation identifiability result (Robins 1986):
  - Under the above assumptions

$$E[Y_1 - Y_0] = E\{E(Y|A=1, W) - E(Y|A=0,W)\}$$

- The right hand side is our parameter of interest (a.k.a. our statistical estimand)
- We have departed from the causal world and entered the observed data and statistical world

5. Choose and apply the estimator

- An estimator is an algorithm that when applied to the data generates an estimate of the parameter of interest
- Several estimators available for the statistical parameter - $E\{E(Y|A=1, W) - E(Y|A=0,W)\}$
  - Substitution estimators (e.g., parametric G-computation)
  - Propensity score based (e.g., IPTW, matching)
  - Double robust (e.g., TMLE, A-IPTW)

5. Choose and apply the estimator

- Pause to recall the usual approach
- Run logistic regression of the outcome (HIV incidence) *Y* on the exposure (prevention package) *A* and the baseline confounders *W:*

$$\text{Logit}[E(Y|A,W)] = \beta_0 + \beta_1 A + \beta_2 W1 + \ldots + \beta_{19} W18$$

- Exponentiate the coefficient on the exposure and interpret the association in terms of an odds ratio
  - Conditional OR: "while holding all other factors constant"
- Potential problem: our target parameter does not equal $e^{\beta 1}$
  - Letting the estimation approach drive the question
- Potential problem: relies on the main terms logistic regression being correct

5. Estimator: Parametric

- Assume we know the relation between the covariates *W*, the exposure *A* and the outcome *Y*, and have correctly specified this relation with a set of constants called "parameters"
  - e.g., specify a regression with main terms for covariates and a few interactions or squared terms that we think are reasonable
  - If we had this knowledge, we would encode in our causal model
  - We want to avoid introducing new assumptions during estimation

With parametric models we are likely assuming we know more than we actually know… like Brainy Smurf

5. Estimator: Non-parametric

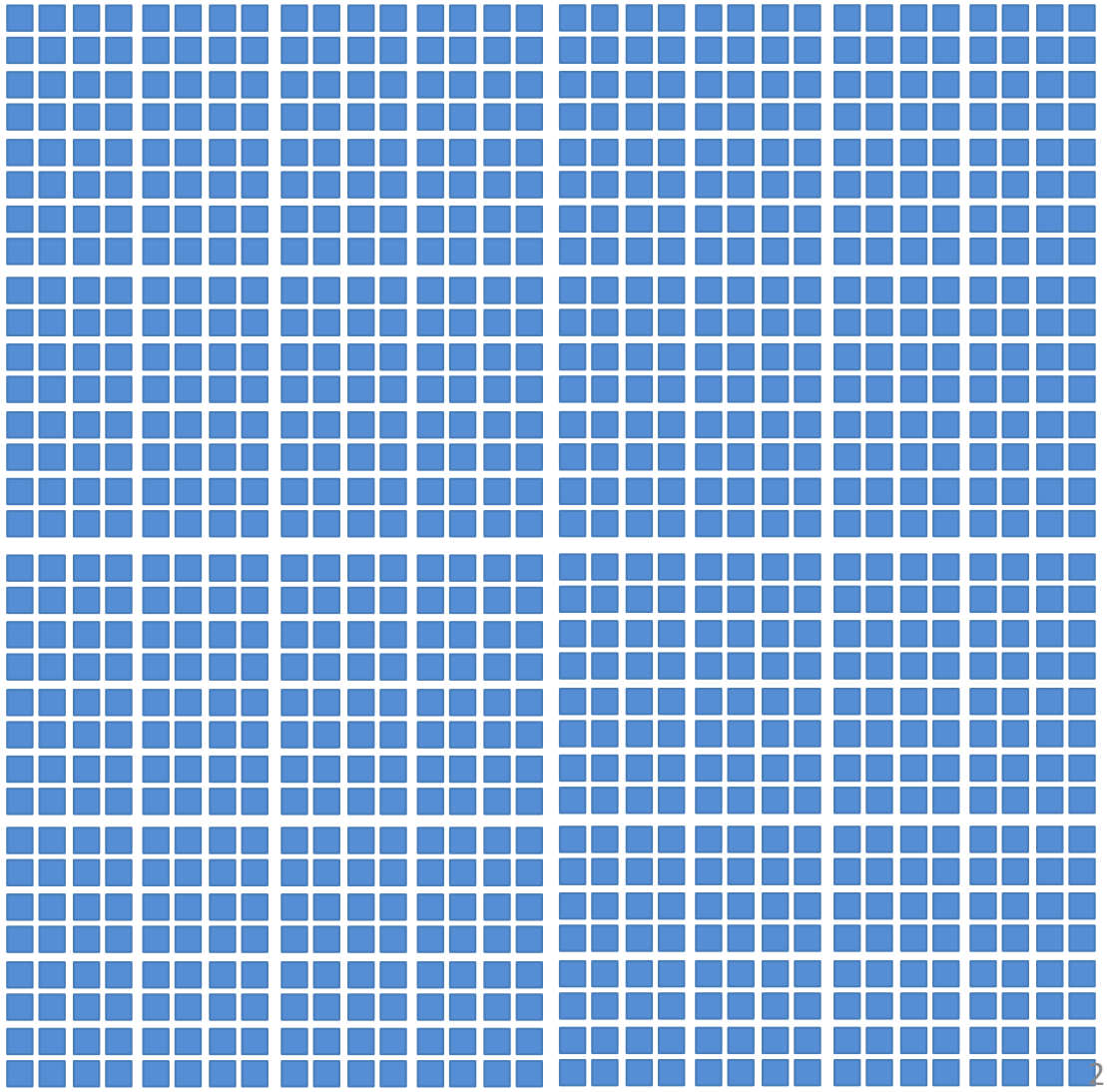
- Acknowledge that we do not know the form of the relations between the covariates *W*, the exposure *A* and the outcome *Y*
  - Divide the data into all combinations of *(A,W)*; calculate and average stratum specific *A* and *Y* relations
  - Unfortunately, we typically have too many covariates and/or continuous covariates, resulting in empty cells

Jon Snow ("Game of Thrones" - not father of modern epidemiology) doesn't know enough to specify a parametric regression

Curse of dimensionality:

of strata increases exponentially with dimension of W!

# of binary RV's in W	# of strata of W
0	1
1	2
2	4
3	8
4	16
5	32
6	64
7	128
8	256
9	512
10	1024



5. Estimator: Semi-parametric

- We often “know nothing” about the functional form, but also need to smooth over data with weak support during estimation
  - Utilize “data-adaptive estimation” or “machine learning”
  - Could choose one algorithm (e.g., stepwise regression, loess or polynomial splines), but no basis for choosing one over the other
  - Instead allow a large class of algorithms to compete
    - Select the best algorithm with cross-validation
    - We will focus on SuperLearner

Elastigirl says “You need to be more flexible”

5. Estimator

- Statistical parameter:

$$E\{E(Y|A=1,W) - E(Y|A=0,W)\}$$

  - Equals the ATE if the identifiability assumptions hold
- Estimation approaches we will cover today
  - Simple substitution estimator, a.k.a. parametric G-computation, parametric g-formula
  - Inverse probability of treatment weighting (IPTW)
  - Targeted maximum likelihood estimation (TMLE) with Super Learner

- Simulated data on 200 schools, including A (prevention package), Y (cumulative incidence of HIV) and 4 confounders (W1-W4)

Simple substitution estimator

Intuition:

- Can think of causal inference as a problem of missing information
  - Know the outcome of each unit under the observed exposure
  - Missing the outcome under the other exposure condition
- Use parametric regression to estimate outcomes for all units under both exposed and unexposed conditions after controlling for the measured confounders
- Average and compare predicted outcomes

Simple substitution estimator

1. Estimate the mean outcome *Y* as a function of the exposure (treatment) *A* and measured confounders *W*
   - E.g. run main terms logistic regression
2. Use estimates from 1. to predict outcomes for each unit while "setting" the exposure to different values
3. Average predictions to estimate the marginal risks in the population under the exposure and no exposure
   - Compare the estimates – i.e., take the difference in the means

Substitution estimator

Observed data

…	HIV incidence| A = 1	HIV incidence| A = 0
School 5		10.5%
School 6		2.3%
School 7	8.8%	
School 8		14.8%
School 9	1.3%	
School 10		0.0%
…		

Observed cumulative incidence of HIV - schools in school 6 did not have the intervention

Substitution estimator

Estimated values

…	HIV incidence| A = 1	HIV incidence| A = 0
School 5	2.8%	5.4%
School 6	2.2%	4.2%
School 7	5.1%	9.5%
School 8	3.4%	6.5%
School 9	2.1%	4.0%
School 10	3.6%	6.9%
…		
All Schools	3.9%	7.3%

Estimated cumulative incidence of HIV if schools in school 8 had the intervention

Substitution estimator

Estimated values

…	HIV incidence| A = 1	HIV incidence| A = 0
School 5	2.8%	5.4%
School 6	2.2%	4.2%
School 7	5.1%	9.5%
School 8	3.4%	6.5%
School 9	2.1%	4.0%
School 10	3.6%	6.9%
…		
All Schools	3.9%	7.3%

Average across all schools estimates cumulative incidence of HIV if all schools had the intervention

Substitution estimator

Estimated values

…	HIV incidence| A = 1	HIV incidence| A = 0
School 5	2.8%	5.4%
School 6	2.2%	4.2%
School 7	5.1%	9.5%
School 8	3.4%	6.5%
School 9	2.1%	4.0%
School 10	3.6%	6.9%
…		
All Schools	3.9%	7.3%

$E\{E(Y|A=1,W)\} - E\{E(Y|A=0,W)\} = 3.9\% - 7.3\% = -3.4\%$

IPTW- Inverse Probability of Treatment Weighting

Intuition:

- Can think of confounding as a problem of biased sampling
  - Certain exposure-covariate subgroups are over represented relative to what you would see in a randomized trial
  - Other exposure-covariate subgroups are under represented
- Apply weights to up-weight under-represented units and down-weight over-represented units
- Average and compare weighted outcomes

IPTW

1. Estimate the probability of being exposed/treated *A* as a function of measured confounders *W*: *P(A=1|W)*
   - Propensity score
   - E.g. run main terms logistic regression
2. Use estimates from 1. to calculate weights:
   - For exposed/treated: *1/P(A=1|W)*
   - For unexposed/untreated: *1/P(A=0|W*)
   - Equivalent to:

$$wt_i = \frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A = 1|W_i)} + \frac{\mathbb{I}(A_i = 0)}{\hat{\mathbb{P}}(A = 0|W_i)}$$

IPTW

3. Apply weights – average outcome in weighted exposed and weighted unexposed to estimate the marginal risks in the population under the exposure and no exposure, respectively
   - Compare the estimates – i.e., take the difference in the weighted means for the exposed and unexposed

IPTW

- IPTW illustration
  - Simple example of one confounder – W1

	A=1	A=0	Total		P-score
W1 = 1	37	67	104		
W1 = 0	25	71	96		

  - What are the propensity scores for W1=1 and W1=0?

IPTW

- IPTW illustration
  - Simple example of one confounder – W1

	A=1	A=0	Total		P-score
W1 = 1	37	67	104	37/104=	0.36
W1 = 0	25	71	96	25/96=	0.26

  - What happens when you apply these weights to the Ns from the original table?

	A=1	A=0	Total
W1 = 1	37	67	104
W1 = 0	25	71	96

IPTW

- IPTW illustration
  - Simple example of one confounder – W1

	A=1	A=0		P-score
W1 = 1	37 x 1/0.36 = 104	67 x 1/0.64 = 104		0.36
W1 = 0	25 x 1/0.26 = 96	71 x 1/0.74 = 96		0.26

  - After weighting, the treated and untreated populations have the same distribution of the confounder as the total population
  - The populations are exchangeable with respect to that confounder

IPTW

- IPTW actual propensity scores and weights

…	A	IPTW Weight	Y	Y*Wt for A=1	Y*Wt for A=0
School 5	0	1.49	10.5%	0	15.6%
School 6	0	1.89	2.3%	0	4.3%
School 7	1	3.60	8.8%	31.6%	0
School 8	0	1.96	14.8%	0	29.1%
School 9	1	2.39	1.3%	3.1%	0
School 10	1	3.10	0.0%	0.1%	0
…					
All Schools				4.0%	7.2%

IPTW

- IPTW actual propensity scores and weights

…	A	IPTW Weight	Y	Y*Wt for A=1	Y*Wt for A=0
School 5	0	1.49	10.5%	0	15.6%
School 6	0	1.89	2.3%	0	4.3%
School 7	1	3.60	8.8%	31.6%	0
School 8	0	1.96	14.8%	0	29.1%
School 9	1	2.39	1.3%	3.1%	0
School 10	1	3.10	0.0%	0.1%	0
…					
All Schools				4.0%	7.2%

Weighted average of outcomes across all schools estimates cumulative incidence of HIV if all schools had the intervention

IPTW

- IPTW actual propensity scores and weights

…	A	IPTW Weight	Y	Y*Wt for A=1	Y*Wt for A=0
School 5	0	1.49	10.5%	0	15.6%
School 6	0	1.89	2.3%	0	4.3%
School 7	1	3.60	8.8%	31.6%	0
School 8	0	1.96	14.8%	0	29.1%
School 9	1	2.39	1.3%	3.1%	0
School 10	1	3.10	0.0%	0.1%	0
…					
All Schools				4.0%	7.2%

$$E\{E(Y|A=1,W)\} - E\{E(Y|A=0,W)\} = 4.0\% - 7.2\% = -3.2\%$$

IPTW

- Formulae for IPTW in two steps

$$\Psi_{IPTW}(\hat{\mathbb{P}}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=1)}{\hat{\mathbb{P}}(A=1|W_i)}Y_i - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=0)}{\hat{\mathbb{P}}(A=0|W_i)}Y_i$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i=1)wt_iY_i}_{\text{Est. of } \mathbb{E}[\mathbb{E}(Y|A=1,W)]} - \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(A_i=0)wt_iY_i}_{\text{Est. of } \mathbb{E}[\mathbb{E}(Y|A=0,W)]}$$

- Formula for IPTW in one step

$$\Psi_{IPTW}(\hat{\mathbb{P}}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}(A_i=1)}{\hat{\mathbb{P}}(A=1|W_i)} - \frac{\mathbb{I}(A_i=0)}{\hat{\mathbb{P}}(A=0|W_i)}\right)Y_i$$

TMLE – targeted maximum likelihood estimation

Intuition:

- Again, think of causal inference as a problem of missing information
- Predict the outcomes for all units under both exposed and unexposed conditions
  - Use a flexible estimation approach to avoid assuming that a parametric regression is correct
  - Or use parametric knowledge if have it
- Incorporate information on the covariate-exposure relation to **improve** the initial estimator
  - Again use a flexible estimation approach or parametric knowledge if available
  - Second chance to control confounding (double-robust)
  - Hone our estimator towards the parameter of interest
  - Gives normal-curve inference
- Average and compare targeted predictions

TMLE

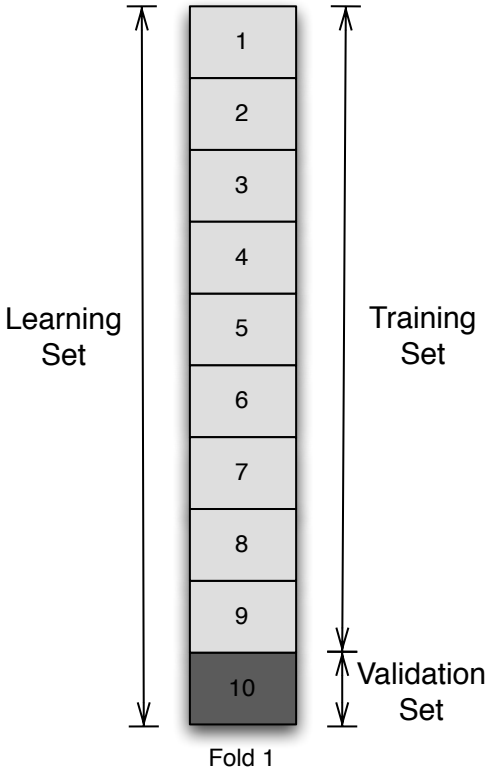
1. Flexibly estimate the mean outcome *Y* as a function of the exposure (treatment) *A* and measured confounders *W*
   - e.g. use the semi-parametric algorithm Super Learner
2. Flexibly estimate the propensity score *P(A=1|W)*
   - e.g. use the semi-parametric algorithm Super Learner
3. Target the initial estimator *E(Y|A,W)* with information in the estimated propensity score
4. Average targeted predictions of the outcome under the exposure and under no exposure, and compare the estimates

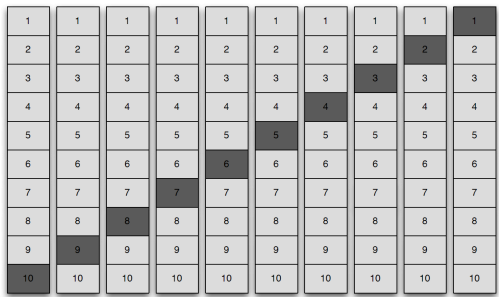
What is SuperLearner?

- Machine learning algorithm: automated approach to learn complex relationships from real data
- Uses cross-validation (data-splitting) to evaluate the performance of a library of candidate estimators
  - Library can consist of simple (e.g. main terms regression models), semi-parametric (e.g. stepwise regression, loess), and more aggressive algorithms (e.g. random forest)
  - Performance is measured by a loss function
    - e.g. Mean squared error (MSE)

What is SuperLearner?

- Cross-validation: allows us to compare algorithms based on how they perform on *independent* data
  - Partition the data in "folds"
  - Fit each algorithm on the training set
  - Evaluate its performance (called "risk") on the validation set
    - e.g. calculate the MSE for observations in the validation set
  - Rotate through the folds
  - Average the cross-validated risk estimates across the folds to obtain one measure of performance for each algorithm





What is SuperLearner?

- We could choose the algorithm with the best performance (i.e. smallest cross-validated MSE)
- SuperLearner can build the best combination of algorithm-specific estimates

Who do Captain Planet and SuperLearner need to succeed?
Our estimators combined!

Why do I need to target?

- We could use Super Learner to predict the outcomes for each unit while "setting" the exposure to different values
- We could then average the predictions to estimate outcomes under different "set" exposures, and contrast these estimates
- But SuperLearner is focused on estimating *E(Y|A,W)*
  - This is not our parameter of interest
  - Wrong bias-variance trade-off
- Also not good for inference (e.g. 95% CI)

What is targeting?

- Use information in the estimated propensity score *P(A=1|W)* to update the initial (Super Learner) estimator of *E(Y|A,W)*
- Involves running a univariate regression
- Use the estimated coefficient to update our initial predictions of the outcome under the exposure and under no exposure

Like Robin Hood, we target to hit the bullseye!

How do I target?

1. Use SuperLearner to estimate the propensity score *P(A=1|W)*
2. Calculate the "clever covariate"

$$H(A,W) = \frac{\mathbb{I}(A=1)}{P(A=1|W)} - \frac{\mathbb{I}(A=0)}{P(A=0|W)}$$

3. Run logistic regression of the outcome *Y* on this covariate using *logit* of the initial estimator *E(Y|A,W)* as offset
   - Where *logit(x)=log(x/1-x)*
4. Plug in the estimated coefficient ε to yield our targeted estimator *E*(Y|A,W)*

How do I target?

5. Use the targeted estimator *E*(Y|A,W)* to predict outcomes for each unit while "setting" the exposure to different values
6. Average the predictions to estimate the marginal risks in the population under the exposure and no exposure
   – Compare the estimates – i.e., take the difference in the means

TMLE

Targeted estimates

…	HIV incidence| A = 1	HIV incidence| A = 0
Village 5	5.8%	8.9%
Village 6	0.7%	1.2%
Village 7	9.7%	14.6%
Village 8	4.6%	7.1%
Village 9	1.0%	1.6%
Village 10	1.6%	2.5%
…		
All Villages	4.5%	6.9%

$$E\{E(Y|A=1,W)\} - E\{E(Y|A=0,W)\} = 4.5\% - 6.9\% = -2.4\%$$

Summary by Estimator

- Estimates of $E\{E(Y|A=1,W) - E(Y|A=0,W)\}$ based on
  - Simple Subs -3.4%
  - IPTW -3.2%
  - TMLE -2.4%

- Truth -2.1%
- Simulation setting so truth is known
- Estimator performance (bias, variance, MSE) has to be assessed over repeated (~500) draws from the data generating mechanism

Estimator properties

Simple substitution estimator

- Relies on consistently estimating the mean outcome $E(Y|A,W)$
  - Sometimes we have a lot of knowledge about how the relationship between the outcome *Y* and the exposure-covariates *(A,W)*
  - Other times, our knowledge is limited, and assuming a parametric regression model can result in bias and misleading inferences

Estimator properties

IPTW

- Relies on consistently estimating the propensity score *P(A=1|W)*
  - Sometimes we have a lot of knowledge about how the exposure was assigned
  - Other times, our knowledge is limited and assuming a parametric regression model can result in bias and misleading inferences
- Unstable estimator under positivity violations
  - When covariate groups only have a few exposed or unexposed observations, weights can blow up
  - When there are covariate groups with 0 exposed or unexposed observations, weights will not blow up
    - BUT – the estimator will likely be biased, variance will be underestimated

Estimator properties

TMLE

- Double robust - will yield a consistent estimate if *either* the conditional mean $E(Y|A,W)$ or the propensity score $P(A=1|W)$ is consistently estimated
  - Two chances!
- Semi-parametric efficient
  - Lowest asymptotic variance (most precision) among a large class if both consistently estimated
- Substitution estimator
  - Robustness under positivity violations, strong confounding, and rare outcomes
- Software
  - Among others, *ltmle* package in *R*

6. Statistical uncertainty

- Derive an estimate of the sampling distribution
  - Non-parametric bootstrap
    - Re-sample the observed data with replacement; apply the entire estimation algorithm (including full Super Learner algorithm) to the re-sampled data; repeat X times; estimate the variance using the resulting point estimates
  - Influence-curve based inference
    - Will not discuss here, but available in the *R* packages

7. Interpret findings

7. Make inferences (interpret findings)
   - Final step – consider whether and to what degree identifiability assumptions have been met to inform the strength of interpretation

7. Interpret findings

- If there were major concerns about identifiability (e.g., temporal ordering unclear, key confounder not measured) the parameter can be interpreted as an **association**

$$E\{E(Y|A=1,W) - E(Y|A=0,W)\}$$

- In words (1): the difference in the outcome associated with everyone exposed, compared to everyone unexposed in the population, accounting for the measured confounders
- In words (2): the marginal difference in the expected outcome with the exposure and the expected outcome with no exposure, after controlling for measured confounding

7. Interpret findings

- Example: The difference in the risk of HIV associated with all schools exposed to prevention package compared with no schools exposed is X, accounting for region, baseline prevalence, SES measures…
- Alternative: as close as we can get to the causal effect of *A* on *Y* given the limitations of the data – detail all limitations!!

7. Interpret findings

- If the authors believe identifiability assumptions are met, the parameter can be interpreted as the average treatment effect

$$E[Y_1-Y_0]$$

- In words: the difference in the expected outcome would be X if everyone were exposed compared with if everyone were unexposed
- Example: There would be a X difference in the risk of HIV if all schools in the population received the prevention package vs. no schools

Summary and Discussion

- Introduced the causal road map approach (briefly) and examined three estimators
  - Simple substitution estimator
  - IPTW
  - TMLE
  - Hopefully, you have increased your intuitive and technical understanding of these estimators

Summary and Discussion

- TMLE with SuperLearner
- TMLE as Robin Hood
  - Stealing from the rich
    - Combining the best of simple substitution and IPTW
  - And giving to the poor
    - Giving us estimators that reduce bias and are efficient

Summary and Discussion

- TMLE with Super Learner as part of a toolbox (i.e. the Roadmap)
- Be careful when specifying your machine learning library
  - e.g Schomaker et al. (2018); Tran et al. (2010, 2016)

Summary and Discussion

- Be wary of overfitting, especially in high dimensional settings (e.g. longitudinal exposures)
  - If using parametric regressions, specify carefully and consider *a priori* reduction in adjustment variables
  - Internal sample-splitting helps
    - e.g. Zheng & van der Laan (2011)

Summary and Discussion

- Double robust estimators (e.g. TMLE or A-IPW) make it possible to use machine learning during estimation
  - Basis for valid statistical inference (i.e. 95% CI)
- But does not solve all our problems
  - Still need our estimates of the conditional mean outcome and propensity score to converge to the truth fast enough
    - Area of ongoing work
    - Progress: Highly adaptive LASSO (van der Laan, 2017)

Summary and Discussion

- Practical positivity violations happen
  - Poor support for the exposures of interest
  - Can lead to bias and underestimates of variance
    - e.g. Petersen et al. (2012,2014); Tran et al. (2010)

Summary and Discussion

- Introduction to the roadmap
  - Necessitates clearly defined scientific questions, and assures the parameters estimated will match the questions posed
  - Elaborates what assumptions are necessary to interpret a parameter estimated from observed data as a causal effect
  - When assumptions are not met, the unmet assumptions provide clear guidance on how future research must be improved to increase the potential for causal interpretation

Summary and Discussion

- Working in this framework can improve the interpretability and relevance of epidemiologic research
- Applicable to other causal questions and data structures
  - Effects among treated/untreated, mediation, longitudinal interventions, stochastic interventions, dynamic regimes…
  - We just focused on the ATE for pedagogic purposes

Interested in More?

- SER workshop – Ahern and Balzer SER Boston 2020
- Coverage here similar conceptually to chapter (Ahern and Hubbard) in the Methods in Social Epidemiology text (2nd ed., Kaufman and Oakes editors)
- Analytically, follows Introduction to Causal Inference course (http://www.ucbbiostat.com/) (Petersen and Balzer)
  - Winner of 2014 Causality in Statistics Education Award

A ***few*** references – *not* complete

- Ahern, Hubbard. A roadmap for estimating and interpreting population intervention parameters. In Oakes and Kaufman, editors, *Methods in Social Epidemiology*. Jossey-Bass, San Francisco, 2nd edition, 2017.
- Balzer, Petersen, van der Laan. Tutorial for causal inference. In Buhlmann, Drineas, Kane, and van der Laan, editors, *Handbook of Big Data*. Chapman & Hall/CRC, 2016.
- Hernan, Robins. Estimating causal effects from epidemiological data. *J Epidem and Community Health*, 2006.
- Lendle, Schwab, Petersen, van der Laan. *ltmle*: An R package implementing targeted minimum loss-based estimation for longitudinal data. *J Stat Software*, 81(1):1-21, 2017.
- Naimi, Balzer. Stacked generalization: An introduction to Super Learning. *Euro J Epi*, 2018.
- Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.
- Petersen, Balzer. Introduction to Causal Inference. *UC Berkeley*, 2014. www.ucbbiostat.com
- Petersen, van der Laan. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology, 2014.*
- Polley, LeDell, Kennedy, van der Laan. *SuperLearner*: Super Learner Prediction, 2017. http://CRAN.R-project.org/package=SuperLearner.
- Robins. A new approach to causal inference in mortality studies with sustained exposure periods– application to control of the healthy worker survivor effect. *Mathematical Modelling*, 1986.
- Rosenbaum, Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika,* 1983
- van der Laan, Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

A ***few*** references – *not* complete

- Petersen, Porter, Gruber, Wang, van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21:31-54, 2012.
- Saddiki, Balzer. A Primer on Causality in Data Science, *arXiv*, 2019: https://arxiv.org/abs/1809.02408.
- Tran, Yiannoutsos, Wools-Kaloustian, Siika, van der Laan, Petersen. Double robust efficient estimators of longitudinal treatment effects: Comparative performance in simulations and a case study. *International Journal of Biostatistics*, 2019
- Tran, Yiannoutsos, Musick, Wools-Kaloustian, Siika, Kimaiyo, van der Laan, Petersen. Evaluating the impact of a HIV low-risk express care task-shifting program: a case study of the Targeted Learning roadmap. *Epidemiologic Methods*, 5(1):69-91, 2016.
- Tran, Petersen, Schwab, van der Laan. Robust variance estimation and inference for causal effect estimation. *arXiv* preprint arXiv:1810.03030, 2018.
- Schomaker, Luque-Fernandez, Leroy, Davies. Using longitudinal targeted maximum likelihood estimation in complex settings with dynamic interventions. *arXiv* preprint arXiv:1802.05005, 2018.
- van der Laan. A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *The International Journal of Biostatistics*, 13(2), 2017.
- Zheng and van der Laan. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning: Causal Inference for Observational and Experimental Data*, Chapter 27, pages 459-474. Springer, New York, 2011.