ISES-ISEE 2018 Pre-Conference Course PC01: Introduction to Double Robust Estimation for Causal Inference

Laura B. Balzer, PhD MPhil
Materials with Drs. Maya Petersen & Jennifer Ahern

Department of Biostatistics & Epidemiology
School of Public Health & Health Sciences
University of Massachusetts, Amherst

UMASS AMHERST

2018 ISES-ISEE JOINT ANNUAL MEETING OTTAWA-CANADA 26-30 AUGUST

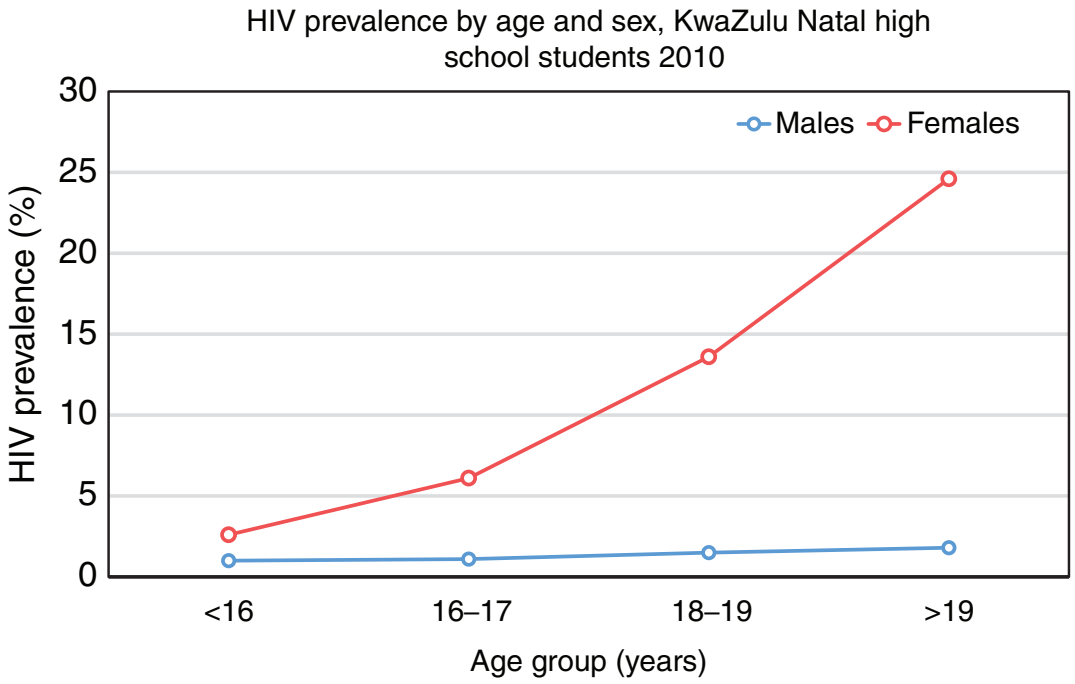
Outline

- Morning: Introduce the "Roadmap" for causal inference
- Afternoon: Implementation & comparison of estimators in `R`
- Focusing on estimation with a binary exposure at single time point

- Coverage follows UMass Biostat690B
  - Old version available at `www.ucbbiostat.com`
    - Winner of 2014 Causality in Statistics Education award

Causal Roadmap - Running example

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



Dellar et al., 2015

Causal Roadmap - Running example

One approach:

- Collect or gain access to data on school-based prevention packages, HIV incidence, and covariates
- Note that the outcome is binary or a proportion, and thus use logistic regression for the analysis
- Estimate the conditional odds ratio by exponentiating the coefficient on the exposure
- Interpret as the change in the relative odds of the outcome associated with a unit increase in the exposure, while holding all other factors constant

Causal Roadmap - Running example

Potential problems:

- Allows the tool (logistic regression) to drive the question being answered
  - What is an odds ratio anyway?
  - Interpretation changes if include/exclude additional terms
- Assumes the logistic regression is correctly specified
  - If this model is wrong, can have biased point estimates and misleading inference

One Solution:

- Causal roadmap provides a way forward!

Causal Roadmap to the Rescue

1. Scientific question
2. Causal model
3. Counterfactuals & causal parameter
4. Observed data & statistical model
5. Identifiability & statistical parameter
6. Estimation
7. Interpretation

(van der Laan & Rose, 2011; Petersen & van der Laan, 2014; Balzer *et al.*, 2016)

1. Specify the scientific question

- What is the effect of a school-based prevention package on the cumulative HIV incidence among young women in East Africa?
- Consider one hypothetical experiment:
  - What would be the difference in HIV incidence if all schools in the target population received the prevention package compared to if no schools received the package?
  - Inference about the proportion of seroconversions under different conditions
  - Many other hypothetical experiments possible

1. Specify the scientific question

- Difference in HIV incidence if all schools in the target population received the prevention package compared to if no schools?
- To sharply frame our question, we want to specify
  - The target population (what age group? where?)
  - The exposure (what package?)
  - The outcome (over what time frame?)
  - Ways to change the exposure and their plausibility

Where are we?

1. Scientific question ✓
2. Causal model
3. Counterfactuals & causal parameter
4. Observed data & statistical model
5. Identifiability & statistical parameter
6. Estimation
7. Interpretation

2. Define the Causal Model

- Causal modeling formalizes our knowledge - however limited
  - Which variables affect each other?
  - The role of unmeasured/background factors?
  - The functional form of the relationships?
- Focus on the structural causal model and corresponding causal graph (Pearl,2000)
  - Many other causal frameworks

2. Specify the causal model

Notation/Terminology

- U: set of unmeasured background factors
- W: set of measured confounders
- A: the exposure
  - $A = 1$ for the exposure = treatment = intervention
  - $A = 0$ for unexposed = untreated = control
  - A is an indicator of the prevention package
- Y: the outcome
  - Cumulative incidence of HIV
  - Proportion of seroconversions
- These are cluster-level (school) variables

2. Specify the causal model

- The structural causal model (SCM) translates our knowledge of the study design into a set of equations or causal graph
- A possible study:
  1. Randomly sample a school
  2. Measure its baseline covariates
     - Region, age-sex distribution, SES measures, baseline HIV prevalence
  3. Observe whether the school receives the combination prevention package
  4. Measure the cumulative incidence of HIV

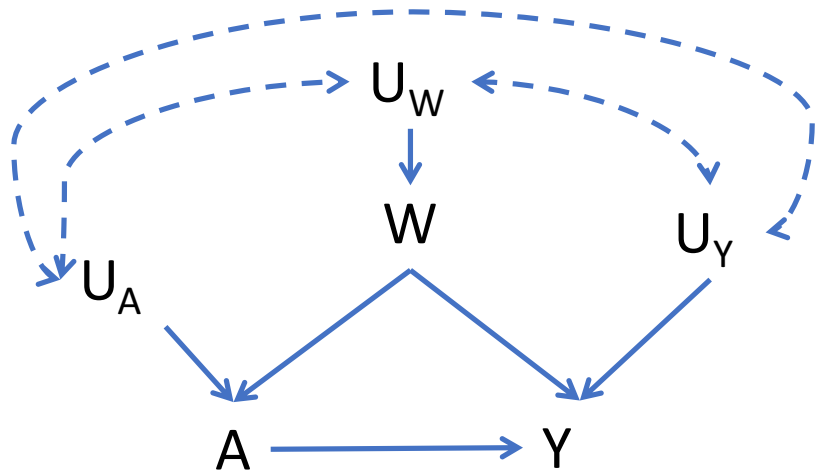
2. Specify the causal model

Structural Causal Model:

$$W \leftarrow f_W(U_W)$$
$$A \leftarrow f_A(W, U_A)$$
$$Y \leftarrow f_Y(W, A, U_Y)$$

- No assumptions
  - On the background factors (U_W, U_A, U_Y)
  - On the functions (f_W, f_A, f_Y)

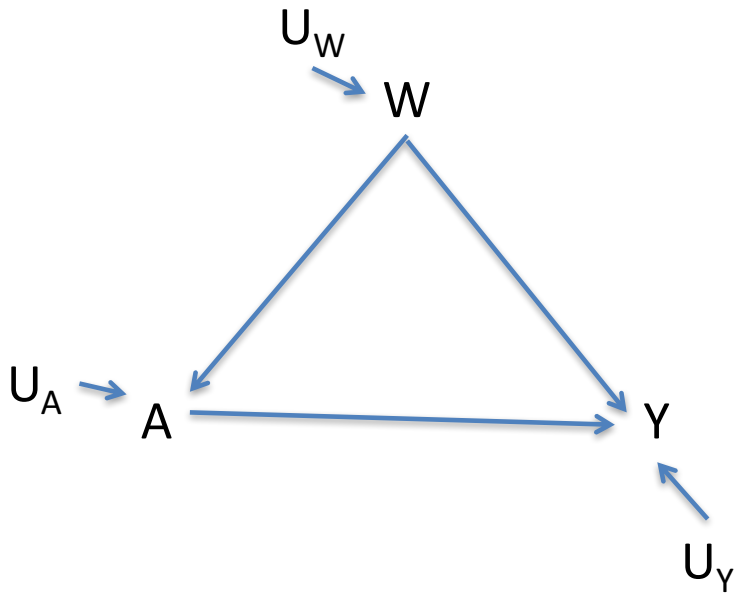
Causal graph:



- The potential correlations between the unmeasured factors are represented with double-headed arrows

2. Specify the causal model

If we believed the no unmeasured confounders assumption, one possible causal graph



- Background factors are all independent
- Still no functional form assumptions
- Wishing for something does not make it true

Where are we?

1. Scientific question ✓
2. Causal model ✓
3. Counterfactuals & causal parameter
4. Observed data & statistical model
5. Identifiability & statistical parameter
6. Estimation
7. Interpretation

3a. Specify the counterfactual (potential) outcomes

- Counterfactual (potential) outcomes are defined by modifications to the data generating process described by the causal model

$$\begin{aligned} W &\leftarrow f_W(U_W) \\ A &\leftarrow 1 \\ Y(1) &\leftarrow f_Y(W, 1, U_Y) \end{aligned} \qquad \begin{aligned} W &\leftarrow f_W(U_W) \\ A &\leftarrow 0 \\ Y(0) &\leftarrow f_Y(W, 0, U_Y) \end{aligned}$$

- $Y(a)$: the counterfactual cumulative incidence of HIV, if possibly contrary to fact, the school was assigned exposure level $A = a$
  - $Y(1)$: the counterfactual cumulative HIV incidence if, possibly contrary to fact, the school received the prevention package ($A = 1$)
  - $Y(0)$: the counterfactual cumulative HIV incidence if, possibly contrary to fact, the school continued with the standard of care ($A = 0$)

3b. Specify the causal parameter

Use counterfactuals to define the target causal parameter

- e.g. the difference in the expected counterfactual cumulative incidence of HIV if all schools received the package vs. continued with the standard of care:

$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

  - Known as the average treatment effect (ATE)
  - For a binary outcome, the causal risk difference: $\mathbb{P}(Y(1) = 1) - \mathbb{P}(Y(0) = 1)$

- Many other causal parameters possible

3. Specify counterfactuals & the causal parameter

Why is causal inference easy for Hermione?

3. Specify counterfactuals & the causal parameter

With her time turner, she can time travel! Hermione can obtain the counterfactual outcomes for all units under the levels of the intervention of interest.

Where are we?

1. Scientific question ✓
2. Causal model ✓
3. Counterfactuals & causal parameter ✓
4. Observed data & statistical model
5. Identifiability & statistical parameter
6. Estimation
7. Interpretation

4a. Specify the observed data

- For one school, the observed data are

$$O = (W, A, Y) \sim \mathbb{P}$$

  - W as measured confounders
  - A as the exposure (prevention package)
  - Y as the outcome (HIV incidence)
  - $\mathbb{P}$ as the true but unknown distribution
- In our `R` exercise, we have $n = 100$ schools
  - We have n i.i.d. copies of O

4b. Link causal to observed

- We assume the causal model provides a description of our study under
  - Existing conditions (i.e. the real world)
  - Specific interventions (i.e. the counterfactual world)
- The observed data were generated by sampling n independent times from a system compatible with the causal model
- Links the causal world and the real (observed data) world

4c. Specify the statistical model

- Our causal model implies the statistical model
  - Formally, the statistical model is the set of possible distributions of the observed data

4c. Specify the statistical model

- Our causal model implies the statistical model
  - Formally, the statistical model is the set of possible distributions of the observed data
- Causal model may, but often does not, place any restrictions on the statistical model
  - Often, no assumptions on the distribution of unmeasured factors or on the functional form of the structural equations
  - e.g. only says the exposure A is some function of baseline covariates W and unmeasured factors U_A, but we do not know the exact form:

$$A \leftarrow f_A(W, U_A)$$

  - If we had this knowledge, we should specify it in the causal model (Step 2)

4c. Specify the statistical model

- Causal frameworks help us to choose a statistical model reflecting our uncertainty
- All statistical models are not wrong
- Our statistical model should represent real knowledge (however limited)

4c. Continuum of statistical models (informally)

- Non-parametric: no restrictions

4c. Continuum of statistical models (informally)

- Non-parametric: no restrictions

- Semi-parametric: some restrictions

4c. Continuum of statistical models (informally)

- Non-parametric: no restrictions

- Semi-parametric: some restrictions

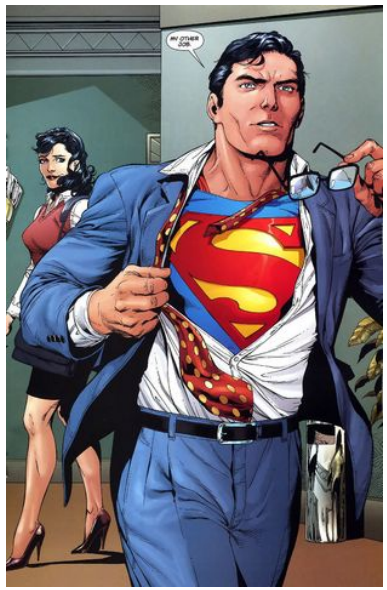
- Parametric: assumes $\mathbb{P}$ is known up to a finite number of unknown parameters

Where are we?

1. Scientific question ✓
2. Causal model ✓
3. Counterfactuals & causal parameter ✓
4. Observed data & statistical model ✓
5. Identifiability & statistical parameter
6. Estimation
7. Interpretation

5. Assess Identifiability

- Currently the parameter of interest is expressed in terms of counterfactuals: $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$
- Identifiability: what assumptions are needed to write the causal parameter as some function of the observed data distribution?



We link our day-job (estimation based on the observed data)
to our superhero-job (answering causal questions)

5. Assess Identifiability

Some intuition:

- $\mathbb{E}[Y|A=1]$: average HIV incidence Y among schools which received the prevention package $A=1$
  - Descriptive/associative
- $\mathbb{E}[Y(1)]$: average counterfactual HIV incidence $Y(1)$ when all schools receive the prevention package $A=1$
  - Causal
- Generally $\mathbb{E}[Y|A=a]$ does *not* equal $\mathbb{E}[Y(a)]$
  - Central problem in causal inference

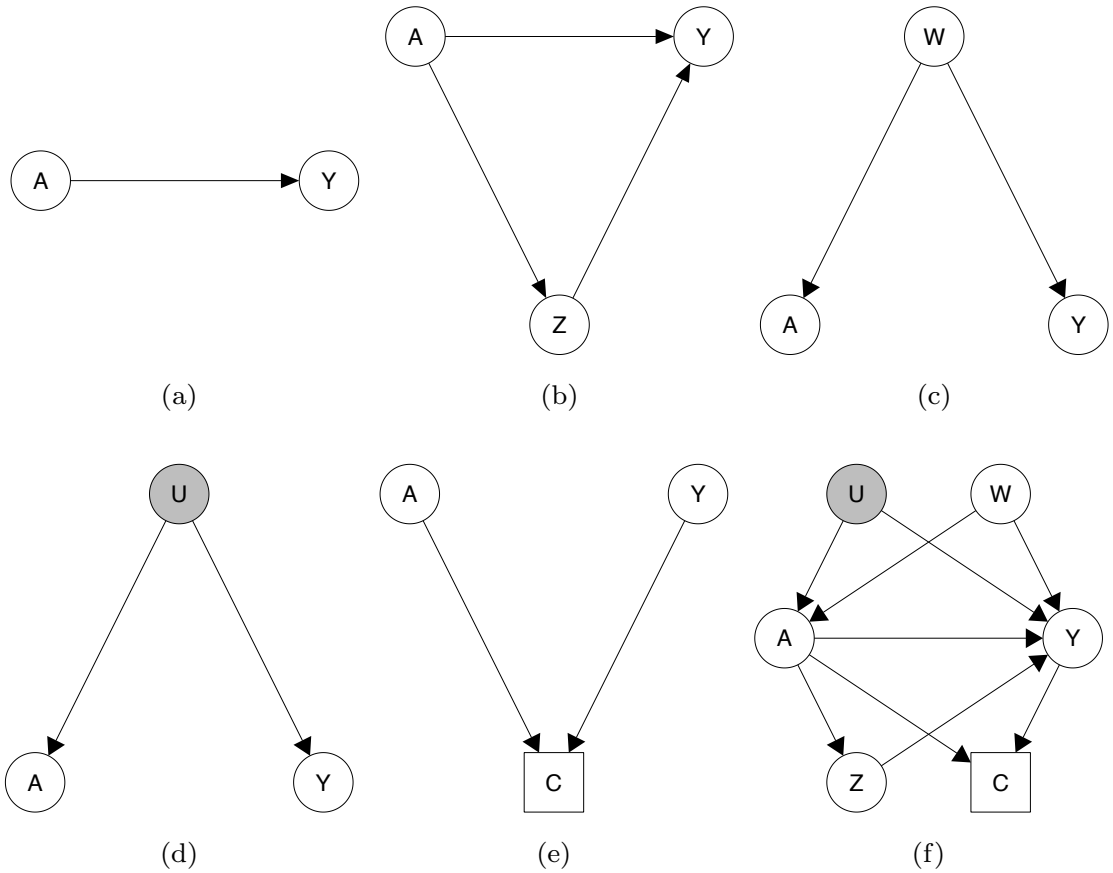
5. Assess Identifiability

- Want: all of the observed association between the exposure A and outcome Y to be due to the causal effect of interest

5. Assess Identifiability

- Want: all of the observed association between the exposure A and outcome Y to be due to the causal effect of interest
- But: many sources of association

(a) direct effects
(b) indirect effects
(c) measured confounding
(d) unmeasured confounding
(e) selection bias
(f) all
  - Many others not listed



5. Assess Identifiability

- Need:
  1. No unmeasured confounding
     - a.k.a. randomization assumption: $Y(a) \perp\!\!\!\perp A \mid W$
  2. Positivity: sufficient variability in the exposure status within strata of confounders (a.k.a. our adjustment set)
     - Ensures the statistical parameter is well-defined
- Then:

$$\begin{aligned}
\mathbb{E}[Y(a)] &= \mathbb{E}\big[\mathbb{E}[Y(a)|W]\big] \\
&= \mathbb{E}\big[\mathbb{E}[Y(a)|A=a,W]\big] && \text{under randomization} \\
&= \mathbb{E}\big[\mathbb{E}(Y|A=a,W)\big] && \text{under positivity}
\end{aligned}$$

5. Assess Identifiability

- Under the above assumptions:

$$\begin{aligned}\mathbb{E}[Y(1) - Y(0)] &= \mathbb{E}\big[\mathbb{E}(Y|A=1,W) - \mathbb{E}(Y|A=0,W)\big] \\ &= \sum_w \big[\mathbb{E}(Y|A=1,W=w) - \mathbb{E}(Y|A=0,W=w)\big]\mathbb{P}(W=w)\end{aligned}$$

- The G-computation identifiability result (Robins, 1986)
- The right hand side is our "statistical estimand"
  - Difference in the expected outcome given the exposure and confounders, and the expected outcome given no exposure and confounders, and then averaged (standardized) with respect to the covariate distribution

5. Assess Identifiability

- What if the assumptions do not hold?
  - What if we do not believe the no unmeasured confounders assumption?
  - What if we do not have time-ordering?

5. Assess Identifiability

Possible options:

- Give up
- Collect or find different or more data
- Proceed to do the best possible job estimating the target parameter
  - Still have a well-defined and interpretable target parameter
  - Coming as close to the wished-for causal parameter given the limitations in the data
- Can use the lack of identifiability to inform future data collection and studies

Where are we?

1. Scientific question ✓
2. Causal model ✓
3. Counterfactuals & causal parameter ✓
4. Observed data & statistical model ✓
5. Identifiability & statistical parameter ✓
6. Estimation
7. Interpretation

6. Estimation

- We have written the causal parameter as a function of the observed data distribution (Robins, 1986):

$$\Psi(\mathbb{P}) = \mathbb{E}\Big[\mathbb{E}(Y|A=1,W) - \mathbb{E}(Y|A=0,W)\Big]$$

- Many estimators available:
  - Parametric G-computation (a.k.a. simple substitution estimator)
  - Inverse probability of treatment weighting (IPTW)
  - Targeted maximum likelihood estimation (TMLE)
- Nothing more-or-less causal about these estimators

6. Estimation - "Standard" approach

Pause and consider the "standard" approach

1. Run logistic regression of the outcome (HIV incidence) Y on the exposure (prevention package) A and the baseline confounders W

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

2. Exponentiate the coefficient in front of the exposure ($e^{\hat{\beta}_1}$)
3. Interpret as the conditional odds ratio associated with the prevention package, while holding all the other risk factors constant

6. Estimation - "Standard" approach

Some problems:

- Our target parameter $\Psi(\mathbb{P})$ is not equal to e^{β_1}
  - Letting the estimation approach drive the question asked
  - Throwing away all our hard work!
- Relies on the main terms logistic regression being correct
  - May measure the relevant variables but do not know their exact functional relationship
  - If we had this knowledge, then we should encode it in our causal model (Step2)
  - If this parametric regression is wrong, can have biased point estimates and misleading inference

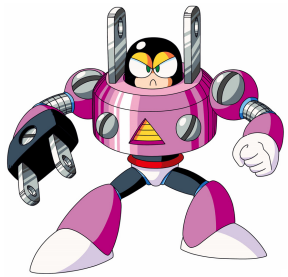
A substitution (plug-in) estimator

Consider again our target parameter:

$$\Psi(\mathbb{P}) = \mathbb{E}\big[\mathbb{E}(Y|A=1,W) - \mathbb{E}(Y|A=0,W)\big]$$
$$= \sum_w \big[\mathbb{E}(Y|A=1,W=w) - \mathbb{E}(Y|A=0,W=w)\big]\mathbb{P}(W=w)$$

- We need estimators of
  - The conditional mean outcome, given the exposure and baseline covariates $\mathbb{E}(Y|A,W)$
  - The covariate distribution $\mathbb{P}(W)$
- Substitute in (plug-in) these estimates:

$$\Psi(\hat{\mathbb{P}}) = \frac{1}{n}\sum_{i=1}^{n}\big[\hat{\mathbb{E}}(Y_i|A_i=1,W_i) - \hat{\mathbb{E}}(Y_i|A_i=0,W_i)\big]$$



  - Sample average as estimator of the covariate distribution

Simple substitution estimator (a.k.a. Parametric G-computation)

1. Estimate the mean outcome Y as a function of the exposure A and measured covariates W
   - e.g. run main terms logistic regression
2. Use the estimates from 1. to predict the outcomes for each unit while "setting" the exposure to different values
3. Point estimate of $\Psi(\mathbb{P})$ by averaging the difference in the predicted outcomes

$$\Psi(\hat{\mathbb{P}}) = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mathbb{E}}(Y_i|A_i = 1, W_i) - \hat{\mathbb{E}}(Y_i|A_i = 0, W_i)\right]$$

- If using parametric regression, also called "parametric G-computation"

Parametric G-Computation

Some 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 measured confounders
- Average and compare predicted outcomes

Parametric G-Computation

- Relies on consistently estimating the mean outcome $\mathbb{E}(Y|A, W)$
- Sometimes we have a lot of knowledge about the relationship between the outcome Y and the exposure-covariates (A, W)
  - If we had this knowledge, encode in our causal model (Step 2) & use it!
- More often, our knowledge is limited
  - Avoid introducing new assumptions during estimation
  - Assuming a parametric regression model can result in bias and misleading inferences

Brainy smurf:
Pretending to know more than we actually do

Non-parametric Estimation

- Parametric G-Computation introduced new assumptions
- Our estimation algorithm should respect our statistical model
  - Often non-parametric
- To avoid these assumptions, we could estimate $\mathbb{E}(Y|A, W)$ by getting the average outcome within all strata of exposure-covariates
- But typically have too many covariates and/or continuous covariates $\rightarrow$ empty/sparse cells

Non-parametric Estimation

- Non-parametric approach breaks down
  - Ex: with only one categorical confounder

	$W = 0$	$W = 1$	...	W=100	...
$A = 1$	310 ($n = 1$)	66 ($n = 12$)		40 (n=30)	
$A = 0$	10 ($n = 60$)	5 ($n = 4$)		? (n=0)	

Table: **Empirical mean of Y within strata of (A, W)**

"Curse of dimensionality"

Semi-parametric Estimation

- Want to avoid assumptions, but also need to smooth over data with weak support
- Relax parametric assumptions with data-adaptive algorithms
  - e.g. stepwise regression with interactions
- However, treating the final regression as if it were pre-specified ignores the model building process
  - No reliable way to obtain inference
- Algorithm tailored to maximize/minimize some criteria and is not necessarily the best algorithm for estimating $\Psi(\mathbb{P})$

Be more flexible!

Super Learner + TMLE

We need Super Learner!

- Flexible estimation approach to avoid unwarranted assumptions
- Uses cross-validation (sample splitting) to evaluate the performance of a library of candidate estimators

We need TMLE!

- Updates the initial estimator of $\mathbb{E}(Y|A,W)$ with information in the exposure mechanism $\mathbb{P}(A=1|W)$
  - Second chance to control for confounding
  - Hone our estimator to the parameter of interest
  - Central limit theorem for inference

Some More Notation

- $\mathbb{E}(Y|A,W)$ - the true conditional mean outcome, given the exposure and baseline covariates
- $\hat{\mathbb{E}}(Y|A,W)$ - an initial estimator based on n observations
- $\hat{\mathbb{E}}^*(Y|A,W)$ - the targeted estimator based on n observations

TMLE

Some intuition:

- Can think of causal inference as a problem of missing information
- Predict the outcome for all units under both exposed and unexposed conditions
  - Flexible estimation approach to avoid unwarranted assumptions
- Incorporate information in the estimated propensity score to improve the initial estimator
  - Second chance to control for confounding
  - Hone our estimator to the parameter of interest
  - Central limit theorem for inference
- Average and compare targeted predictions

Overview - TMLE

1. Estimate $\mathbb{E}(Y|A,W)$ with Super Learner
2. Estimate the propensity score $\mathbb{P}(A=1|W)$ with Super Learner
3. Target the initial estimator $\hat{\mathbb{E}}(Y|A,W)$
4. Plug-in the updated estimates into the target parameter mapping

$$\hat{\psi}^* = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mathbb{E}}^*(Y_i|A_i=1,W_i) - \hat{\mathbb{E}}^*(Y_i|A_i=0,W_i)\right]$$

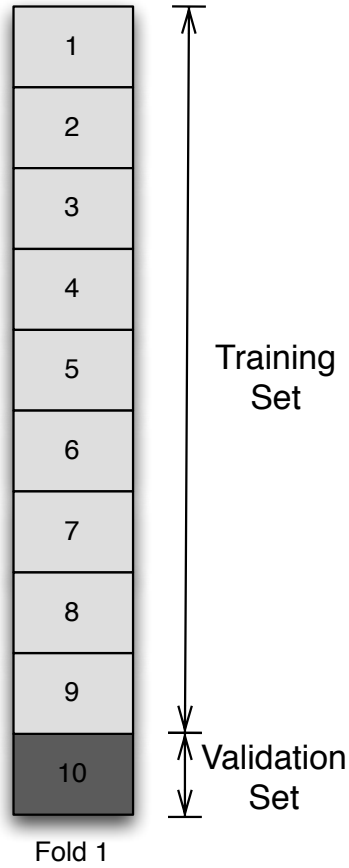
What is Super Learner?

- Machine learning algorithm: automated approach to learn complex relationship from real data
- Uses cross-validation (data-splitting) to evaluate the performance of a library of candidate estimators
- Library can consist of a 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 Super Learner?

Cross-validation: allows us to compare algorithms based on how they perform on independent data

- Partition the data into "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 Super Learner?

- We could choose the algorithm with the best performance (i.e. smallest cross-validated risk estimate)
- Instead, Super Learner builds the best weighted combination of algorithm-specific estimates
  - "Ensemble"
  - "Stacking"
  - Get weights by running a constrained regression of observed outcome on algorithm-specific cross-validated predictions

Worked examples: `https://www.biorxiv.org/content/early/2017/08/18/172395`

Why do we need to target?

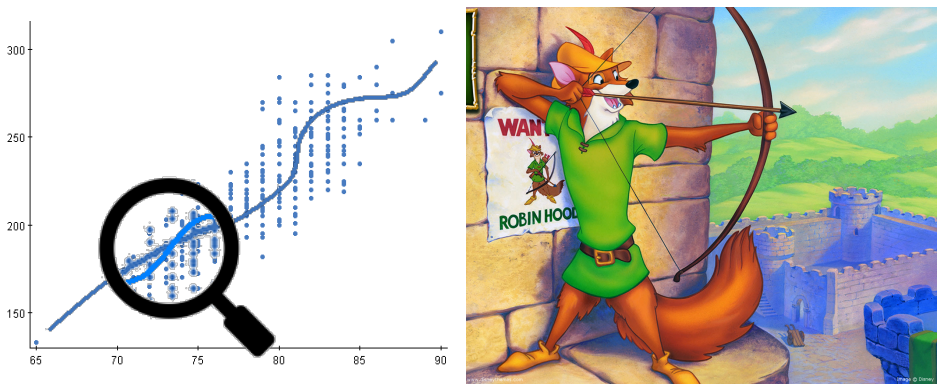
- We could use Super Learner to predict the outcomes for each unit while "setting" the exposure to different values
- Then we could plug these estimates into the target parameter mapping (i.e. average the difference in the predictions):

$$\hat{\psi} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mathbb{E}}(Y_i|A_i = 1, W_i) - \hat{\mathbb{E}}(Y_i|A_i = 0, W_i)\right]$$

- But Super Learner is focused on $\mathbb{E}(Y|A, W)$
  - Very different goal than causal effect estimation
  - Wrong bias-variance trade-off
- Also no reliable way to obtain inference

What is targeting?

- Use information in the estimated propensity score $\hat{\mathbb{P}}(A = 1|W)$ to update the initial (Super Learner) estimator $\hat{\mathbb{E}}(Y|A, W)$
- 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 we target?

1. Estimate the propensity score $\hat{\mathbb{P}}(A = 1|W)$
   - Again, use a flexible approach or parametric knowledge if available
2. Create the "clever" covariate:

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

3. Run logistic regression of the outcome Y on the clever covariate $\hat{H}$ with offset as the logit of the initial estimates.
4. Plug in the estimated coefficient $\hat{\epsilon}$ to yield our targeted estimator:

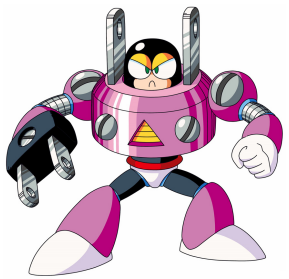
$$\text{logit}\left[\hat{\mathbb{E}}^*(Y|A, W)\right] = \text{logit}\left[\hat{\mathbb{E}}(Y|A, W)\right] + \hat{\epsilon}\hat{H}(A, W)$$

   - where $\text{logit}(x) = log(x/1 - x)$

TMLE - Point Estimate

4. Use the updated estimator $\hat{\mathbb{E}}^*(Y|A,W)$ to predict the outcomes for each unit while "setting" the exposure to different values
5. Substitute into the target parameter mapping:

$$\hat{\psi}^* = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mathbb{E}}^*(Y_i|A_i=1,W_i) - \hat{\mathbb{E}}^*(Y_i|A_i=0,W_i)\right]$$



Some nice things about TMLE

- Double robust
  - Consistent if either conditional mean $\mathbb{E}(Y|A,W)$ or the propensity score $\mathbb{P}(A=1|W)$ is consistently estimated
  - Two chances!
    - Even better with Collaborative-TMLE
- Semi-parametric efficient
  - Lowest asymptotic variance (most precision) among a large class if both consistently estimated
- Asymptotically linear
  - Normal curve for inference
- Substitution estimator
  - Robustness under positivity violations, strong confounding and rare outcomes
- Software
  - *ltmle* and *drtmle* packages in *R*

6. Statistical inference

- For all estimation approaches, we need are not satisfied with a point estimate
- To construct confidence intervals and test the null hypothesis, we need an estimate of uncertainty
- Two options (not covered here):
  - Influence-curve based inference (available in *R* packages)
  - Non-parametric bootstrap

Where are we?

1. Scientific question ✓
2. Causal model ✓
3. Counterfactuals & causal parameter ✓
4. Observed data & statistical model ✓
5. Identifiability & statistical parameter ✓
6. Estimation ✓
7. Interpretation

7. Interpretation

- Final step - consider whether and to what degree the identifiability assumptions have been met
- Statistical:
  - Estimate of the marginal difference in the cumulative incidence of HIV in treated and control schools, after adjusting for measured confounders
  - As close as we can get to causal effect given the limitations in the data
  - "Variable importance measure"
- Causal:
  - If the necessary causal assumptions hold: Estimate of the causal risk difference or the average treatment effect

Summary & Discussion

The causal roadmap to the rescue!

1. Scientific question ✓
2. Causal model ✓
3. Counterfactuals & causal parameter ✓
4. Observed data & statistical model ✓
5. Identifiability & statistical parameter ✓
6. Estimation ✓
7. Interpretation ✓

Causal Frameworks as a Tool

1. Make uncertainty and limits of knowledge explicit
2. Frame better questions
3. Understand assumptions, and when assumptions are not met, provide guidance on how future research can be improved
4. Helps ensure that parameters estimated come as close as possible to answering the causal question posed
5. Interpret results appropriately
6. Widely applicable
   - Effects among the treated/untreated, mediation, longitudinal interventions, dynamic regimes, missing data...

Summary & Discussion of Estimation

We can all be Super Learners!!

"Super Learner . . .
It's our hero . . .
Going to take bias down to zero"
(To the tune of the
"Captain Planet" theme song)

"The Power is Yours"

Summary & Discussion of Estimation

We can all be Super Learners!!

"Super Learner . . .
It's our hero . . .
Going to take bias down to zero"
(To the tune of the
"Captain Planet" theme song)

"The Power is Yours"

TMLE as Robin Hood

- Stealing from the rich
  - Combining the best of IPTW and GComp
- and giving to the poor
  - and giving us unbiased and maximally efficient estimators

Bullseye!

A few references - not a complete bibliography

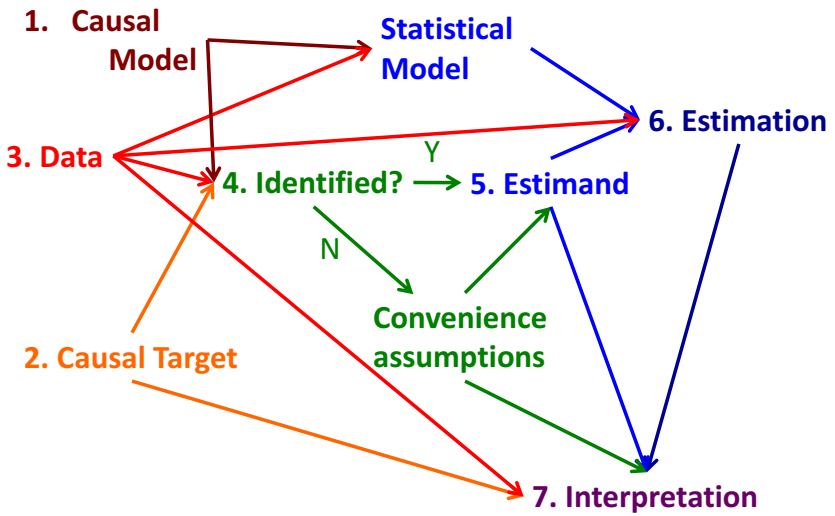
- Ahern and Balzer. Estimation and interpretation: Introduction to parametric and semi-parametric estimators for causal inference. SER Workshop, 2018.
- Balzer, Petersen, van der Laan. Tutorial for causal inference. In Buhlmann, Drineas, Kane, and van der Laan, editors, *Handbook of Big Data*. Chapman & Hall/CRC, 2016.
- Hernan and Robins. Estimating causal effects from epidemiological data. *J Epidem and Community Health*, 2006.
- Naimi and Balzer. Stacked generalization: An introduction to Super Learning. *Euro J Epi*, 2018.
- Pearl. *Causality: Models, Reasoning and Inference.* Cambridge University Press, 2000.
- Petersen and Balzer. Introduction to Causal Inference. UC Berkeley. 2014 `www.ucbbiostat.com`
- Petersen and van der Laan. Causal Models and Learning from Data: Integrating Causal Modeling and Statistical Estimation.*Epidemiology*, 2014.
- 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 and Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 1983.
- van der Laan and Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data.* Springer, 2011.

Thank you & Questions

- You!
- Jennifer Ahern
- Maya Petersen
- Mark van der Laan
- Organizers of ISES-ISEE 2018

- More info: www.ucbbiostat.com

A roadmap....



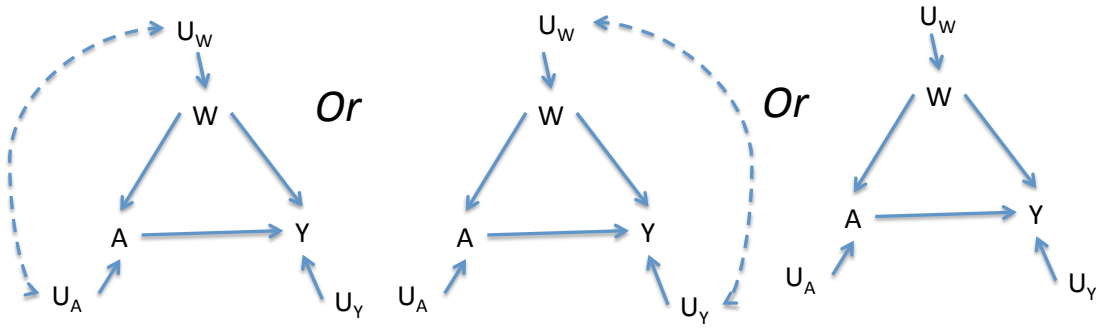
Bonus Slides!!

5. Assess Identifiability

- Back-door criterion provides a graphical approach for identifying the needed adjustment set (Pearl, 2000)
  - What variables to include in W?
  - Are they sufficient?
- For the effect of a single intervention ("point treatment"), we need the adjustment set W to
  1. Block any association between A and Y that arises from measured or unmeasured common causes
  2. Not create any new non-causal associations between A and Y
  3. Not block any of the effect of A on Y

5. Assess Identifiability

- The covariates W will satisfy the back-door criterion if



  - The covariates W blocks all back-door paths from the outcome Y into the exposure A
    - Without creating new spurious sources of association
    - Without blocking the effect of interest
  - No unmeasured common causes of the exposure A and outcome Y

- Equivalently stated by the randomization assumption: $Y(a) \perp\!\!\!\perp A \,||\, W$

5. Assess Identifiability

- Also need a positive probability of receiving each level of the exposure within strata of baseline covariates
- Positivity: sufficient variability in the exposure within confounder strata

$$\mathbb{P}(A = a|W = w) > 0 \text{ for all } w \text{ with } \mathbb{P}(W = w) > 0$$

  - a.k.a. "overlap" or the "experimental treatment assignment assumption"
  - Ensures the statistical parameter is well-defined

5. Assess Identifiability

Structural causal model describes the data generating process

1. Draw background factors $U = (U_W, U_A, U_Y)$ from $\mathbb{P}_U$

2. Generate the covariates $W = f_W(U_W)$
3. Generate the exposure $A = f_A(W, U_A)$
4. Generate the outcome $Y = f_Y(W, A, U_Y)$

- Observed data: $O = (W, A, Y)$

2. Generate the covariates $W = f_W(U_W)$
3. Set the exposure $A = 1$
4. Generate the outcome $Y = f_Y(W, A = 1, U_Y)$

- Counterfactual outcomes: $Y(1)$

- Observed outcome Y = counterfactual outcome $Y(1)$ when observed exposure A = exposure of interest (here, 1)
- Then $\mathbb{P}(Y = 1 \mid A = 1) = \mathbb{P}^*(Y(1) = 1 \mid A = 1)$

5. Assess Identifiability

- Temporality: exposure precedes the outcome ✓
  - Indicated by an arrow on the causal graph from the A to Y
  - Equivalently, Y as a function of A in the causal model
- Consistency: $Y(a) = Y|A = a$ ✓
  - Recall our causal model provides a description of the study under existing conditions (i.e. observed exposure) and interventions (i.e. set exposure)
- Stability: no interference between units ✓
  - Indicated by the outcome Y being only a function of each individual's exposure A in the causal model and graph

5. Assess Identifiability

For a binary outcome

- Under the above assumptions, we have equivalence between the causal risk difference and the marginal risk difference

$$\mathbb{P}[Y(1) = 1] - \mathbb{P}[Y(0) = 1] = \mathbb{E}\big[\mathbb{P}(Y = 1|A = 1, W) - \mathbb{P}(Y = 1|A = 0, W)\big]$$

  - The conditional risk given the exposure and confounders minus the conditional risk given no exposure and confounders, and then averaged/standardized/marginalized with respect to the covariate distribution

Equivalence between the IPTW & G-Comp. estimands

$$\mathbb{E}\left[\frac{\mathbb{I}(A=1)}{\mathbb{P}(A=1|W)}Y\right]$$

$$= \sum_{w,a,y} \frac{\mathbb{I}(A=1)}{\mathbb{P}(A=1|W=w)} y\mathbb{P}(Y=y, A=a, W=w)$$

$$= \sum_{w,a,y} \frac{\mathbb{I}(A=1)}{\mathbb{P}(A=1|W=w)} y\mathbb{P}(Y=y|A=a, W=w)\mathbb{P}(A=a|W=w)\mathbb{P}(W=w)$$

$$= \sum_{w,y} y\mathbb{P}(Y=y|A=1, W=w)\mathbb{P}(W=w)$$

[cancellation by evaluating at $A=1$]

$$= \mathbb{E}[\mathbb{E}(Y|A=1, W)]$$

Inverse Probability of Treatment Weighting (IPTW)

Some intuition:

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

IPTW

More formally:

- We can re-write our target parameter as

$$\begin{aligned}\Psi(\mathbb{P}) &= \mathbb{E}\Big[\mathbb{E}(Y|A=1,W) - \mathbb{E}(Y|A=0,W)\Big] \\ &= \mathbb{E}\left[\left(\frac{\mathbb{I}(A=1)}{\mathbb{P}(A=1|W)} - \frac{\mathbb{I}(A=0)}{\mathbb{P}(A=0|W)}\right)Y\right]\end{aligned}$$

  - where $\mathbb{I}(A=a)$ is an indicator function, equalling 1 if $A=a$ and 0 otherwise

- Suggests an alternate estimator:

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

IPTW

1. Estimate the "propensity score" (Rosenbaum & Rubin, 1983): the probability of being exposed/treated given the measured confounders $\mathbb{P}(A = 1|W)$
   - e.g. run main terms logistic regression
2. Use the estimates from 1. to calculate weights:
   - For exposed: $1/\hat{\mathbb{P}}(A = 1|W)$
   - For unexposed: $1/\hat{\mathbb{P}}(A = 0|W)$
3. Point estimate of $\Psi(\mathbb{P})$ with the average difference in weighted outcomes

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

Estimation with IPTW

- Relies on consistently estimating the propensity score $\mathbb{P}(A = 1|W)$
- Sometimes we have a lot of knowledge about how the exposure was assigned
  - If we had this knowledge, encode in our causal model (Step 2) & use it!
- More often, our knowledge is limited
  - Avoid introducing new assumptions during estimation
  - Assuming a parametric regression model can result in bias and misleading inferences

Estimation with IPTW

- Tends to be an unstable estimator under positivity violations (i.e. strong confounding)
  - 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 and variance underestimated
- Can result in unreasonable estimates
  - e.g. yield probabilities less than 0 and greater than 1
- Note: this is just one flavor of IPTW

IPTW

What do IPTW estimators and Wonder Woman have in common?

IPTW

Weight in all the right places

Step 1: Estimation with Super Learner

Requires

- Data: $O_1, \ldots, O_n \sim \mathbb{P}$
- Loss function: Measure of the dissimilarity between estimate and target.
- Candidate estimators: Throw in any parametric procedure, non-parametric algorithm, histogram estimator...

Step 1: Estimation with Super Learner

Requires

- Data: $O_1, \ldots, O_n \sim \mathbb{P}$
- Loss function: Measure of the dissimilarity between estimate and target.
- Candidate estimators: Throw in any parametric procedure, non-parametric algorithm, histogram estimator...

Uses Cross-Validation

- Evaluate estimator performance and prevent over-fitting

Step 1: Estimation with Super Learner

Requires

- Data: $O_1, \ldots, O_n \sim \mathbb{P}$
- Loss function: Measure of the dissimilarity between estimate and target.
- Candidate estimators: Throw in any parametric procedure, non-parametric algorithm, histogram estimator...

Uses Cross-Validation

- Evaluate estimator performance and prevent over-fitting

Returns the optimal prediction function as a weighted combination of candidate estimators.

- Optimal: minimizes the expected loss, called the "risk"

How does Super Learner work?

- Discrete super learner selects the algorithm with the smallest cross-validated risk.
- Super learner uses the predicted outcomes to create the best weighted combination of algorithms.

How does Super Learner work?

1. Define a loss function:

$$L\big(O, \mathbb{E}(Y|A,W)\big) = \big(Y - \mathbb{E}(Y|A,W)\big)^2$$

2. Define a library of candidate estimators:

$$\begin{aligned}
\mathbb{E}_{n,1}(Y|A,W) &= \beta_0 + \beta_1 A + \beta_2 W_1 + \beta_3 W_2 + \beta_4 W_3 \\
\mathbb{E}_{n,2}(Y|A,W) &= \beta_0 + \beta_2 A + \beta_2 W_1 + \beta_3 sin(W_2) + \beta_4 A\ xW_1^2 \\
\mathbb{E}_{n,3}(Y|A,W) &= \text{Stepwise} \\
\mathbb{E}_{n,4}(Y|A,W) &= \text{Loess} \\
&\vdots \\
\mathbb{E}_{n,k}(Y|A,W) &= \text{your advisor's favorite algorithm}
\end{aligned}$$

3. Split the data $O_1, \ldots O_n$ into $V = 10$ "folds".
   - Divide the data into ten blocks of size $n/10$.

How does Super Learner work?

4. Define nine blocks (90% of the data) to be the training set and the remaining block (10% of the data) to be the validation set.
5. Fit each estimator on the training set.
   - e.g. Use maximum likelihood estimation to fit $\mathbb{E}_{n,1}(Y|A, W)$ on 90% of the data.
6. Predict the outcomes for the validation set.
   - e.g. Plug in the observed treatment A_i and covariates W_i for validation set (the remaining 10% of the data).

How does Super Learner work?

7. Evaluate the empirical risk for each estimator.

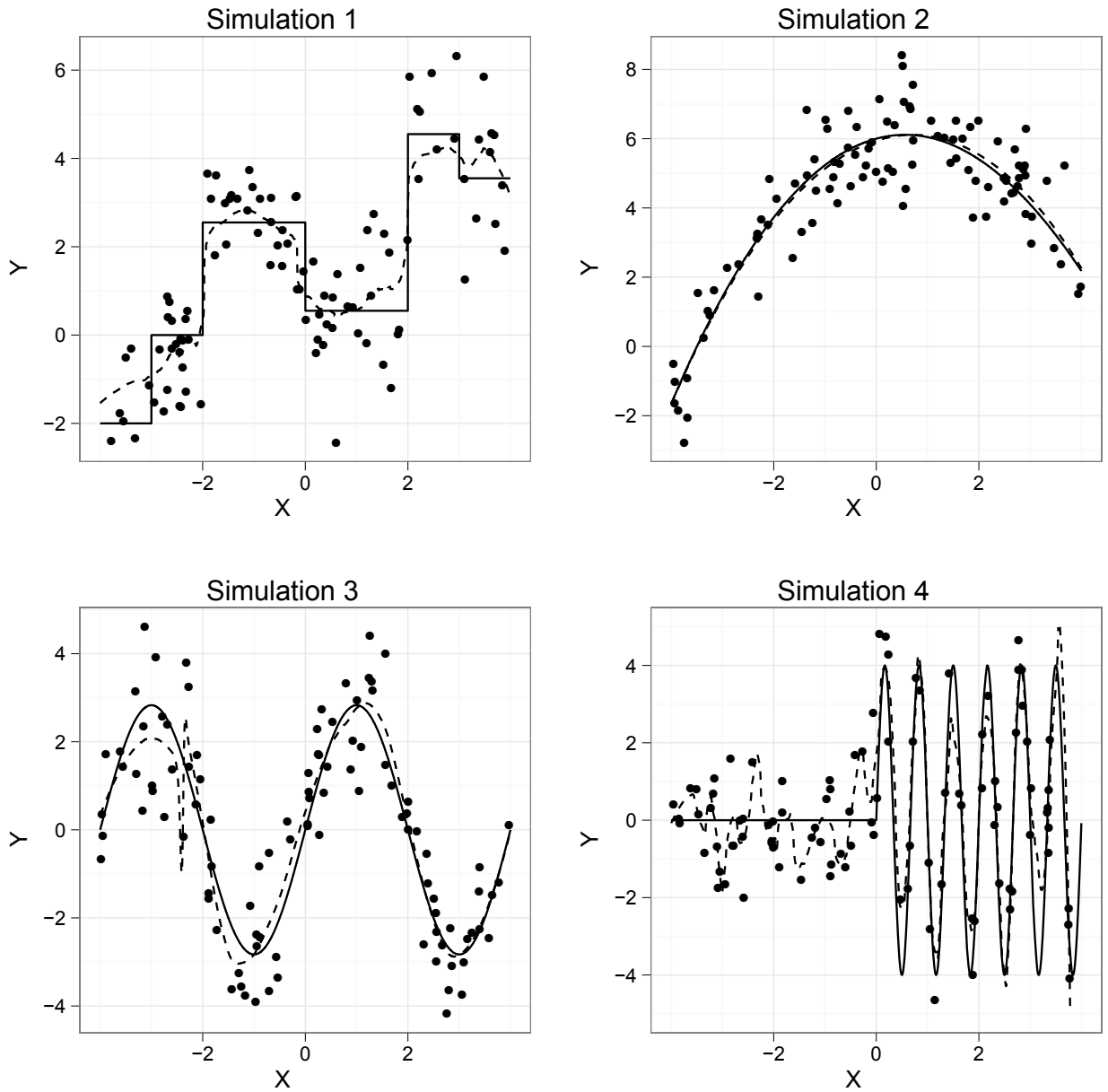
$$\text{Risk}_{n,1}(v=1) = \frac{1}{n^*}\sum_{i=1}^{n^*}\left(Y_i - \mathbb{E}_{n,1}(Y_i|A_i, W_i)\right)^2$$

with n^* as the number of observations in the validation set

8. Repeat steps 4-7 so that each block gets to serve as the validation set.
9. Calculate the cross-validated risk for each algorithm.

$$\text{CV-Risk}_1 = \frac{1}{10}\sum_{v=1}^{10}\text{Risk}_{n,1}(v)$$

Example of Super Learner's Power



TMLE - Inference with the Influence Curve

- Normal limit distribution*
- Construct 95% confidence interval as $\hat{\psi}^* \pm 1.96\sigma_n/\sqrt{n}$
- Standard error estimated with the sample variance of the estimated influence curve divided by sample size n
- Available in the *ltmle* and *drtmle* packages

Misunderstood Hero