M-estimation

Paul Zivich

Institute of Global Health and Infectious Diseases
Causal Inference Research Laboratory
University of North Carolina at Chapel Hill

October 5, 2022

Acknowledgements

Supported by NIH T32-AI007001.

Thanks to Bonnie Shook-Sa, Stephen Cole, Jessie Edwards, and others at the UNC Causal Lab (causal.unc.edu).[1]

pzivich@unc.edu @PausalZ pzivich

[1] Footnotes are reserved asides for possible discussion or questions

Overview

Introduce M-estimation

Computational M-estimation

Applications

Conclusion

Introduction to M-estimation

M-estimation: a short history

- M(aximum likelihood)-estimation
  - More general framework[2]
  - Defined as a zero of an estimating function
- Developed to study robust statistics[3,4]
  - Mean robust to outliers
- Operate under frequentist superpopulation model

[2] Stefanski LA & Boos DD (2002) *The American Statistician*, 56(1), 29-38.
[3] Huber PJ (1964) *Annals of Mathematical Statistics*, 35, 73–101.
[4] Huber PJ (1973) *Annals of Statistics*, 1, 799–821.

M-estimation: the basics

M-estimator: solution for θ in

$$\sum_{i=1}^{n} \psi(O_i; \hat{\theta}) = 0$$

where

- $O_1, O_2, ..., O_n$ are independent observations
- $\theta = (\theta_1, ..., \theta_k)$
- $\psi(.)$ is a known $k \times 1$ estimating function
  - Does not depend on i
  - Proof of CAN follows from unbiased estimating functions[5]

[5] See pages 327-329 of 'Essential Statistical Inference' by Boos & Stefanski

By-hand example

Task: estimate the mean (μ) of $\{1, 5, 3, 7, 24\}$

Using $\hat{\mu} = n^{-1} \sum_{i=1}^{n} Y_i$

$$\hat{\mu} = \frac{1 + 5 + 3 + 7 + 24}{5} = \frac{40}{5} = 8$$

The equivalent estimating function is

$$\sum_{i=1}^{n} (Y_i - \hat{\mu}) = 0$$

By-hand example

To find $\hat{\mu}$, we use a root-finding algorithm[6]

- Select a grid of values
  - $0, 5, ..., 25$
- Plug in guess for $\hat{\mu}$ into $\sum_{i=1}^{n}(Y_i - \hat{\mu})$
- Select values that straddle zero
  - $5, 10$
- Select new grid and repeat process
  - $5, 6, 7, 8, 9, 10$
- Terminate procedure when $\hat{\mu}$ that returns zero is found

End up with $\hat{\mu} = 8$

[6] This procedure is a simple example of the bisection algorithm.

M-estimation: the basics

Asymptotic sandwich variance

$$V(\theta) = B(\theta)^{-1} F(\theta) \left(B(\theta)^{-1} \right)^T$$

Empirical sandwich variance estimator

$$V_n(O_i; \hat{\theta}) = B_n(O_i; \hat{\theta})^{-1} F_n(O_i; \hat{\theta}) \left(B_n(O_i; \hat{\theta})^{-1} \right)^T$$

where

$$B_n(O_i; \hat{\theta}) = n^{-1} \sum_{i=1}^{n} -\psi'(O_i; \hat{\theta})$$

$$F_n(O_i; \hat{\theta}) = n^{-1} \sum_{i=1}^{n} \psi(O_i; \hat{\theta}) \psi(O_i; \hat{\theta})^T$$

Connections to maximum likelihood estimation

When the correct parametric family is assumed

$$B(\theta) = F(\theta) = I(\theta)$$

Therefore

$$V(\theta) = I(\theta)^{-1}$$

When the parametric family is incorrect

$$B(\theta) \neq F(\theta)$$

and the correct limiting variance is $V(\theta)$

Advantages of the sandwich estimator

Key advantages

- Robust to secondary assumptions
- Automation of the delta method
- Captures uncertainty of parameters that depend on other estimated parameters
- Less computationally intensive
  - Relative to bootstrap, Monte Carlo

By-hand example

Bread matrix

$$B_n(Y_i; \hat{\mu}) = 5^{-1} \sum_{i=1}^{5} -\psi'(Y_i; \hat{\mu})$$

Here

$$\psi'(Y_i; \hat{\mu}) = \frac{d}{d\hat{\mu}} (Y_i - \hat{\mu}) = -1$$

Therefore

$$B_n(Y_i; \hat{\mu}) = 5^{-1} \sum_{i=1}^{5} -(-1) = \frac{5}{5} = 1$$

By-hand example

Filling matrix

$$F_n(Y_i;\hat{\mu}) = 5^{-1}\sum_{i=1}^{5}\psi(Y_i;\hat{\mu})\psi(Y_i;\hat{\mu})^T$$

Here

$$\psi(Y_i;\hat{\mu})\psi(Y_i;\hat{\mu})^T = (Y_i-\hat{\mu})(Y_i-\hat{\mu}) = (Y_i-\hat{\mu})^2$$

Therefore

$$F_n(Y_i;\hat{\mu}) = 5^{-1}\sum_{i=1}^{5}(Y_i-8)^2 = 68$$

By-hand example

Sandwich matrix

$$V_n(O_i;\hat{\theta}) = B_n(O_i;\hat{\theta})^{-1} F_n(O_i;\hat{\theta}) \left(B_n(O_i;\hat{\theta})^{-1}\right)^T$$

$$V_n(O_i;\hat{\theta}) = 1^{-1} \times 68 \times 1^{-1} = 68$$

Scale by n for finite-sample variance estimate

$$n^{-1} V_n(O_i;\hat{\theta}) = 68/5 = 13.6$$

Computational M-estimation

Implementation of M-estimators

Solving 'by-hand' has issues

- More than one parameter
- May introduce math errors

However, can all be done by the computer

Procedure

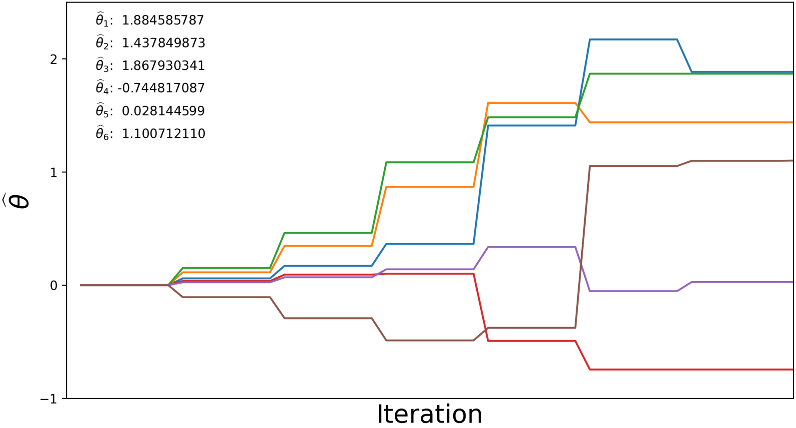
- Root-finding procedure for $\hat{\theta}$
- Numerically approximate derivatives in $B_n(O_i; \hat{\theta})$
- Matrix algebra for sandwich

Software[7]

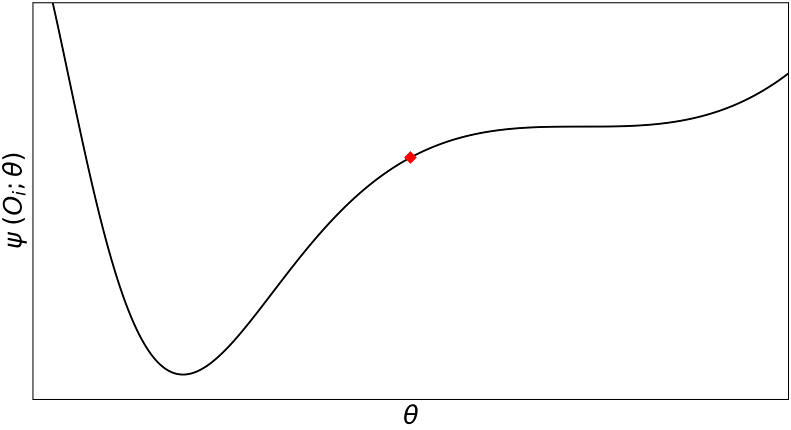
`PROC IML;`

[7] `delicatessen`: Zivich et al. *arXiv*:2203.11300, `geex`: Saul & Hudgens (2020) *J Stat Soft*

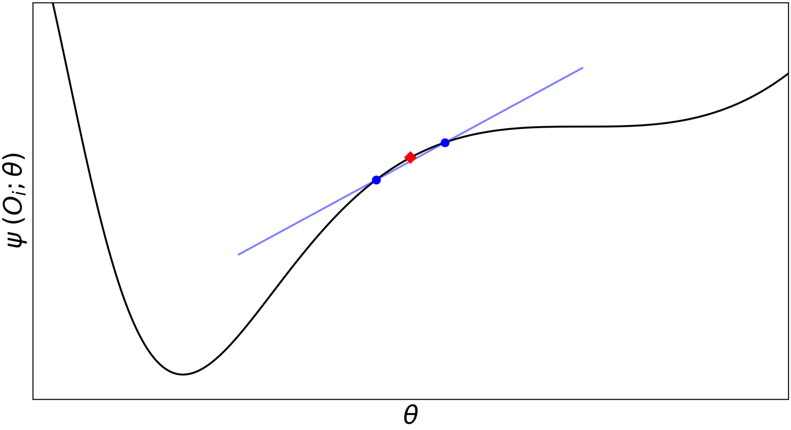
Root-finding



Numerical approximation of derivative



Numerical approximation of derivative



Application of M-estimators

Outline

Robust mean

Regression

- Simple
- Robust

Causal estimation methods

- Inverse probability weighting
- G-computation

Fusion designs

- Bridged treatment comparisons

Robust Mean

Problem with the mean

Sensitivity to outliers

- For $\{1, 5, 3, 7, 24\}$
- Observation of 24 has large impact on $\hat{\mu}$
- Mean ($\hat{\mu} = 8$) is larger than the other 4 observations

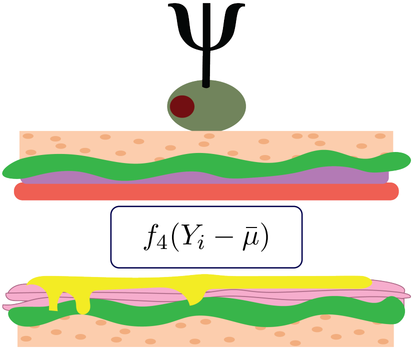
Robust mean[8]

$$\sum_{i=1}^{n} f_k(Y_i - \bar{\mu}) = 0$$

$$f_k(x) = \begin{cases} x, & \text{if } -k < x < k \\ k, & \text{if } x \geq k \\ -k, & \text{if } x \leq -k \end{cases}$$

[8] Mean and median are special cases where $k \to \infty$ and $k \to 0$, respectively

With $k = 4$



$\bar{\mu} = 5$ and $\bar{Var}(\bar{\mu}) = 3.3$

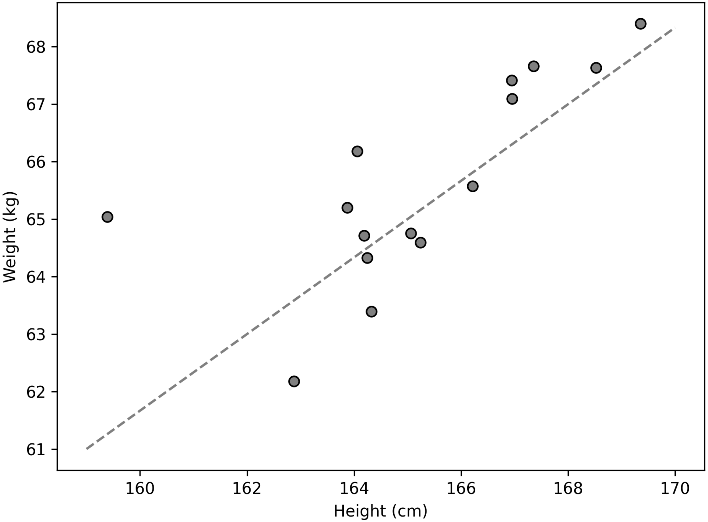
Regression

Notation

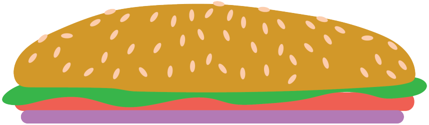
Y_i: independent variable
X_i: dependent variable

$g(X_i) = (1, X_i)$
$\beta = (\beta_0, \beta_1)$

Example



Weight (kg)
68
67
66
65
64
63
62
61
160
162
164
166
168
170
Height (cm)



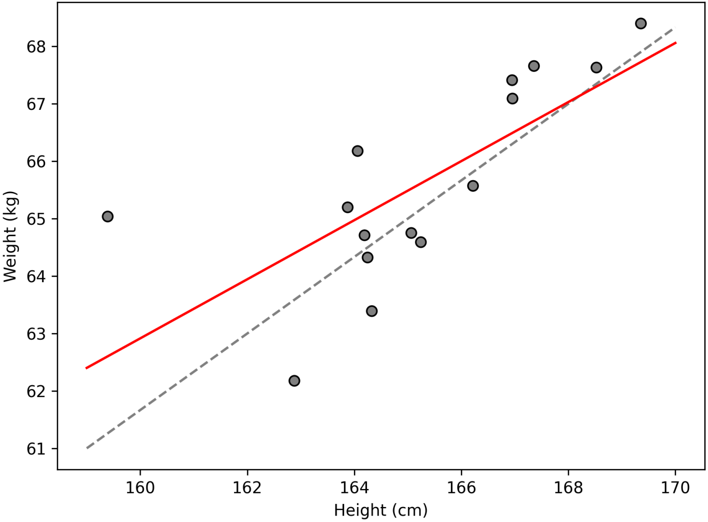
$$\left(Y_i - g(X_i)^T \hat{\beta}\right) 1$$

$$\left(Y_i - g(X_i)^T \hat{\beta}\right) X_i$$

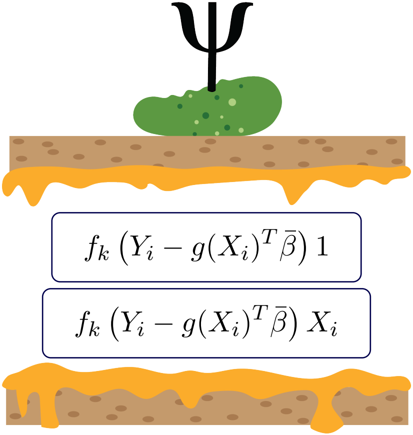
Notice: the estimating function is the score equation

- Easy to develop as M-estimators

Simple Linear Regression



Robust Linear Regression

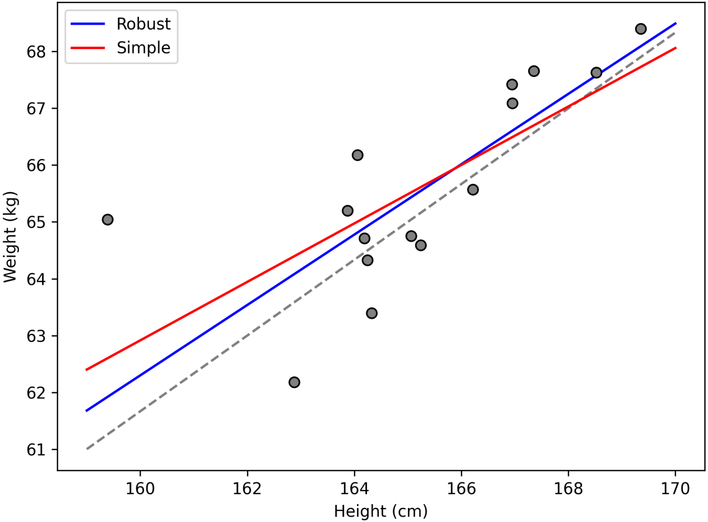


$$f_k\left(Y_i - g(X_i)^T \bar{\beta}\right) 1$$

$$f_k\left(Y_i - g(X_i)^T \bar{\beta}\right) X_i$$

Outliers can only impact up to k

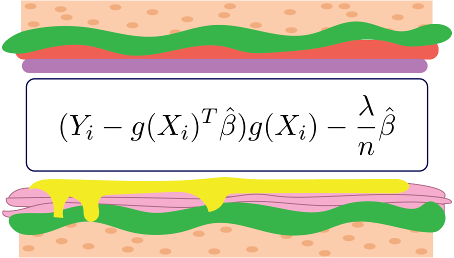
Robust Linear Regression



Other regression models

Penalized regression[9]

- Ridge or L_2 penalty



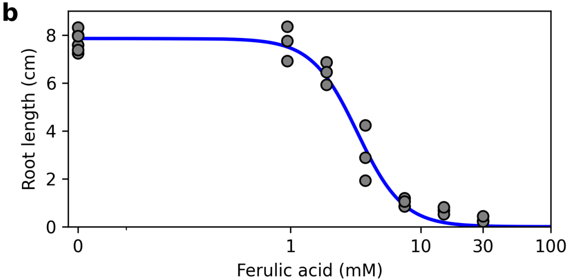
$$(Y_i - g(X_i)^T \hat{\beta}) g(X_i) - \frac{\lambda}{n} \hat{\beta}$$

[9] Fu WJ. (2003) *Biometrics*, 59, 126-132

Other regression models

Dose-response regression[10]

- 3-parameter log-logistic models[11]



[10] An H et al. (2019) *R Journal*, 11(2), 171.
[11] Example provided in Zivich et al. *arXiv*:2203.11300

Causal Effect Estimation

Notation

Y_i: outcome of interest
A_i: action of interest
Y_i^a: potential outcome under action a
W_i: vector of covariates

$g(W_i) = (1, W_i)$
$g(A_i, W_i) = (1, A_i, W_i)$

Aside: Identification vs Estimation

Following all relies on identification assumptions: causal consistency, exchangeability, positivity[12]

- Identification: writing interest parameter in terms of observable data
- Estimation: how the parameter in terms of observable data is estimated

[12] Identification should always precede estimation (see Maclaren OJ & Nicholson R (2019) *arXiv:1904.02826*, Aronow PM et al. (2021) *arXiv:2108.11342* for why)

Aside: Nuisance Parameters

Causal inference (and related) problems can be set up as

$$\theta = (\mu, \eta)$$

μ is the *interest* parameter
η is the *nuisance* parameter

- To estimate μ, need to estimate η
- But η is not of any immediate interest
- Example: causal mean and propensity scores

Motivating Example

Example from Morris et al. (2022)[13]

- Comparison of covariate adjustment methods
  - Gain power in randomized trials
  - Account for systematic error in observational studies

- Data from the *GetTested* trial[14]
  - Efficacy of e-STI testing on STI testing uptake
  - W_i: gender, age, number of sexual partners, sexual orientation, ethnicity
  - Will ignore missing data here[15]

[13] Morris TP et al. (2022) *Trials* 23(1), 1-17.
[14] Wilson E et al. (2017) *PLOS Medicine* 14(12), e1002479
[15] Don't do this. Will be a later slide on extending the M-estimators

Inverse Probability Weighting

The IPW estimator is

$$\frac{1}{n}\sum_{i}^{n}\frac{Y_i A_i}{\Pr(A=1|W_i;\hat{\alpha})} - \frac{1}{n}\sum_{i}^{n}\frac{Y_i(1-A_i)}{\Pr(A=0|W_i;\hat{\alpha})}$$

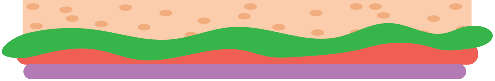
Estimate $\hat{\alpha}$ using a logistic model, $\eta = \alpha$

Estimating the variance for the RD

- Bootstrap
  - Computationally expensive
- The "GEE trick"
  - Treats $\hat{\alpha}$ as known
  - Conservative estimate of the variance[16]
- Sandwich

[16] Only true for some parameters, see Reifeis & Hudgens (2022) *Am J Epidemiol* for an exception

Inverse Probability Weighting

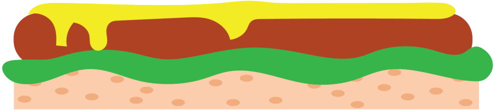


$$\left(A_i - \text{expit}(g(W_i)^T \hat{\alpha})\right) g(W_i)$$

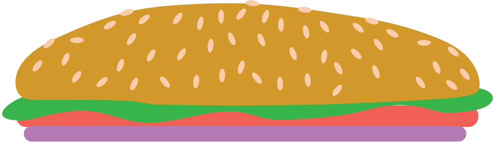
$$\frac{Y_i A_i}{\text{expit}(g(W_i)^T \hat{\alpha})} - \hat{\mu}_1$$

$$\frac{Y_i (1 - A_i)}{1 - \text{expit}(g(W_i)^T \hat{\alpha})} - \hat{\mu}_0$$

$$(\hat{\mu}_1 - \hat{\mu}_0) - \hat{\mu}_2$$



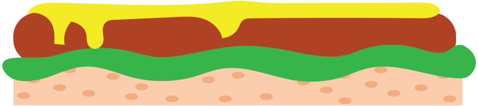
Inverse Probability Weighting



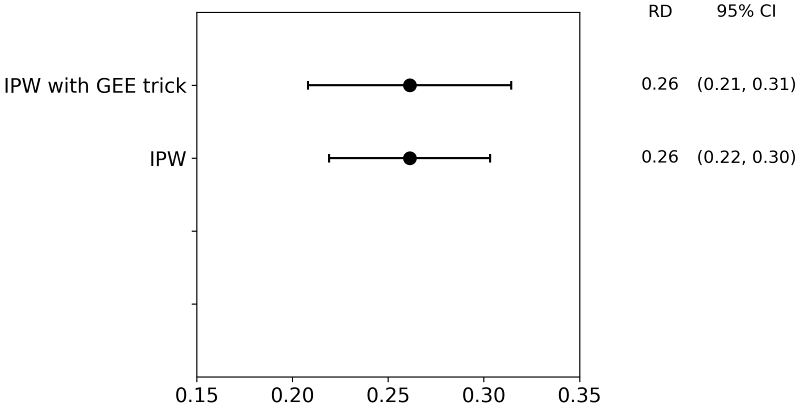
$$\frac{Y_i A_i}{\text{expit}(g(W_i)^T \alpha)} - \hat{\mu}_1$$

$$\frac{Y_i (1 - A_i)}{1 - \text{expit}(g(W_i)^T \alpha)} - \hat{\mu}_0$$

$$(\hat{\mu}_1 - \hat{\mu}_0) - \hat{\mu}_2$$



Results


RD
95% CI
IPW with GEE trick
0.26
(0.21, 0.31)
IPW
0.26
(0.22, 0.30)
0.15
0.20
0.25
0.30
0.35

G-computation

G-computation[17]

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

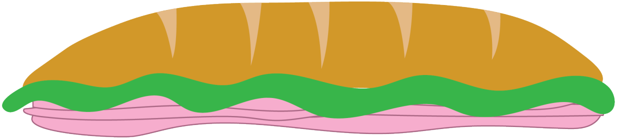
Estimate $\hat{\beta}$ using a logistic model for binary Y_i, $\eta = \beta$

Estimating the variance for the RD

- Bootstrap
- Sandwich

[17] See Snowden et al. (2011) *Am J Epidemiol* for details on this 'trick'

G-computation



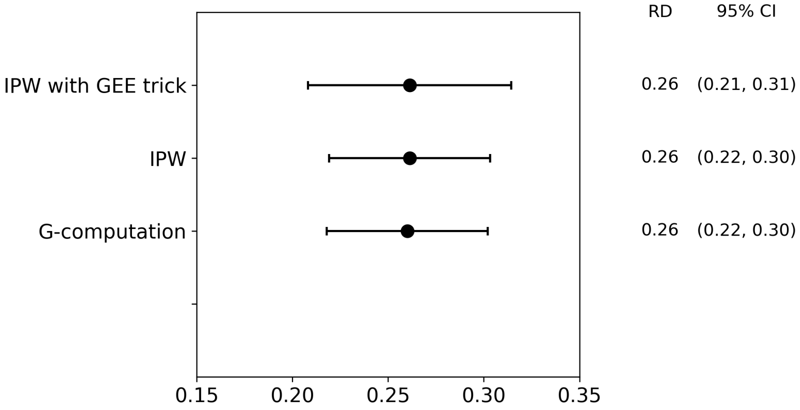
$$\Big(Y_i - \text{expit}(g(A_i, W_i)^T \hat{\beta})\Big)\, g(A_i, W_i)$$

$$\text{expit}(g(1, W_i)^T \hat{\beta}) - \hat{\mu}_1$$

$$\text{expit}(g(0, W_i)^T \hat{\beta}) - \hat{\mu}_0$$

$$(\hat{\mu}_1 - \hat{\mu}_0) - \hat{\mu}_2$$

Results


RD
95% CI
IPW with GEE trick
0.26 (0.21, 0.31)
IPW
0.26 (0.22, 0.30)
G-computation
0.26 (0.22, 0.30)
0.15
0.20
0.25
0.30
0.35

Missing Data

Do not ignore

- If MCAR, may lose efficiency
- If MAR, may be biased

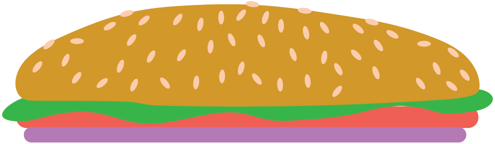
M-estimation makes extending the estimators simple

R_i: observed Y_i $(R_i = 1)$ or missing Y_i $(R_i = 0)$

$$\frac{1}{n}\sum_{i}^{n} \frac{Y_i R_i \, I(A_i = a)}{\Pr(A_i = a | W_i; \hat{\alpha}) \Pr(R_i = 1 | A_i, W_i; \hat{\gamma})}$$

$\eta = (\alpha, \gamma)$

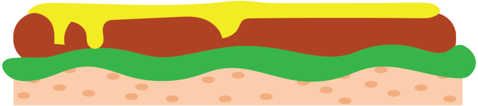
Inverse Probability Weighting with Missing Y



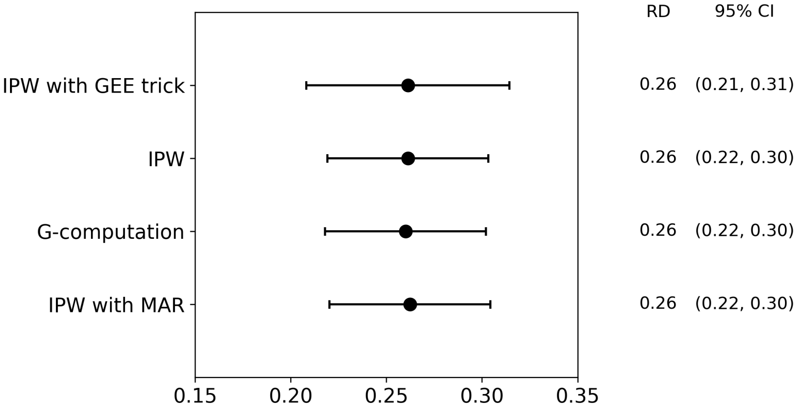
$$\left(A_i - \text{expit}(g(W_i)^T \hat{\alpha})\right) g(W_i)$$

$$\left(R_i - \text{expit}(g(A_i, W_i)^T \hat{\gamma})\right) g(A_i, W_i)$$

$$\frac{Y_i A_i R_i}{\text{expit}(g(W_i)^T \hat{\alpha}) \; \text{expit}(g(A_i, W_i)^T \hat{\gamma})} - \hat{\mu}_1$$



Results


RD
95% CI
IPW with GEE trick
0.26
(0.21, 0.31)
IPW
0.26
(0.22, 0.30)
G-computation
0.26
(0.22, 0.30)
IPW with MAR
0.26
(0.22, 0.30)
0.15
0.20
0.25
0.30
0.35

Fusion Designs

What is a fusion design?

Combine data across sources in a principled way to address a question none of the constituent data sets could address as well alone[18]

Examples

- Transporting the average causal effect
- Measurement error corrections
- Two-stage studies
- Bridged treatment comparisons

[18] See Cole et al. (2022) *Am J Epidemiol* for examples

Notation

T_i: time of event
C_i: time of censoring
$T_i^* = \min(T_i, C_i)$
$\Delta_i = I(T_i = T_i^*)$
$F(t)$: risk at time t

A_i: action of interest, $\{1, 2, 3\}$
W_i: vector of covariates

Bridged Treatment Comparisons

Bridged treatment comparisons[19]

$$(\overbrace{\Pr(Y^3|S=1) - \underbrace{\Pr(Y^2|S=1)) + (\Pr(Y^2|S=1)}_{\text{Bridge}} - \Pr(Y^1|S=1)}^{\text{Parameter of Interest}})$$

- Target population ($S_i = 1$): 3 vs 2
- Secondary population ($S_i = 0$): 2 vs 1

[19] See Breskin et al. (2021) *Stats in Med* and Zivich et al. (2022) *arXiv:2206.04445* for details on identification

Motivating Example

What is the one-year risk difference function comparing triple versus mono antiretroviral therapy (ART) on a composite outcome for the ACTG 320 trial?

- Outcome: AIDS, death, or a large decline in CD4 ($>$50%)
- ACTG 320
  - Randomized to triple ART ($a = 3$) versus dual ART ($a = 2$)
- ACTG 175
  - Randomized to dual ART ($a = 2$) versus mono ART ($a = 1$)

Bridged Treatment Comparisons

Estimator

$$\hat{\mu}_t = \left(\hat{F}^3_{320}(t) - \hat{F}^2_{320}(t)\right) + \left(\hat{F}^2_{175}(t) - \hat{F}^1_{175}(t)\right)$$

Tasks

- Incorporate treatment assignment
- Account for informative loss to follow-up
- Transport ACTG 175 results to ACTG 320 population[20]

[20] Westreich et al. (2017) *Am J Epidemiol*, 186(8), 1010-1014

Bridged Treatment Comparisons

Estimator for ACTG 320 pieces:

$$\hat{F}^{a}_{320}(t) = n_{320}^{-1} \sum_{i=1}^{n} \frac{I(A_i = a) I(S_i = 1) I(T_i^* \leq t) \Delta_i}{\pi_A(S_i; \hat{\eta}) \pi_C(W_i, A_i, S_i; \hat{\eta})}$$

where $a \in \{2, 3\}$,

$$n_{320} = \sum_{i=1}^{n} I(S_i = 1)$$

$$\pi_A(S_i) = \Pr(A_i = a | S_i; \hat{\eta})$$

$$\pi_C(W_i, A_i, S_i; \hat{\eta}) = \Pr(C_i > t | W_i, A_i, S_i; \hat{\eta})$$

Bridged Treatment Comparisons

Estimator for ACTG 175 pieces:

$$\hat{F}^{a}_{175}(t) = \hat{n}^{-1}_{175} \sum_{i=1}^{n} \frac{I(A_i = a) I(S_i = 1) I(T_i^* \le t) \Delta_i}{\pi_A(S_i; \hat{\eta}) \pi_C(W_i, A_i, S_i; \hat{\eta})} \times \frac{1 - \pi_S(W_i; \hat{\eta})}{\pi_S(W_i; \hat{\eta})}$$

where $a \in \{1, 2\}$

$$\hat{n}_{175} = \sum_{i=1}^{n} I(S_i = 0) \frac{1 - \pi_S(W_i; \hat{\eta})}{\pi_S(W_i; \hat{\eta})}$$

$$\pi_S(V_i; \hat{\eta}) = \Pr(S_i = 1 | W_i; \hat{\eta})$$

Bridged Treatment Comparisons: Diagnostic

Notice that[21]

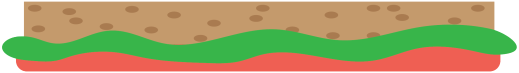
$$E\left[\hat{F}^2_{320}(t) - \hat{F}^2_{175}(t)\right] = 0$$

Offers a testable implication

- Compare difference in data
- Difference from zero indicates ≥ 1 assumption is violated

[21] Zivich et al. (2022) *arXiv:2206.04445* proposed this diagnostic and a permutation test for the whole risk difference curve

Bridged Treatment Comparisons

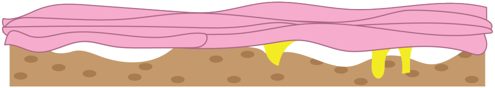


$$I(S_i = 0)\left(I(A_i = 1) - \hat{\gamma}_{0,1}\right)$$
$$I(S_i = 0)\left(I(A_i = 2) - \hat{\gamma}_{0,2}\right)$$
$$I(S_i = 1)\left(I(A_i = 2) - \hat{\gamma}_{1,2}\right)$$
$$I(S_i = 1)\left(I(A_i = 3) - \hat{\gamma}_{1,3}\right)$$

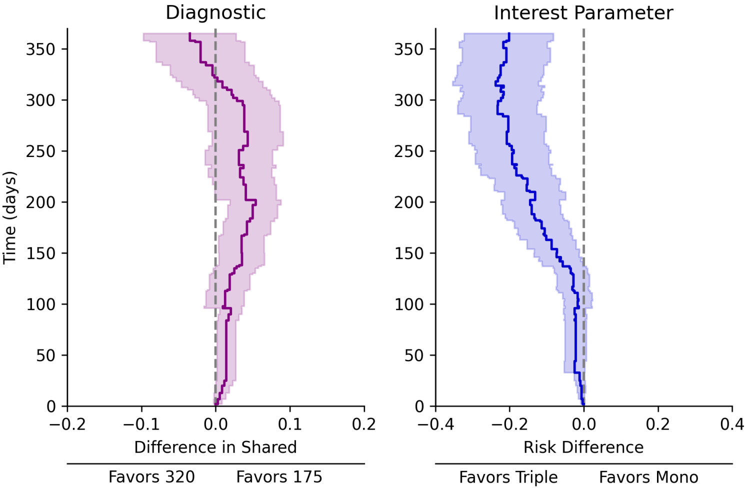
$$\Big(I(S_i = 1) - \text{expit}(W_i^T \hat{\delta})\Big) W_i$$

$$\psi_{AFT}(O_i; \hat{\lambda}, \hat{\beta}, \hat{\alpha})$$

$$\psi_{RD(t)}(O_i; \hat{\mu}_t, \hat{\gamma}_{a,s}, \hat{\delta}, \hat{\lambda}, \hat{\beta}, \hat{\alpha})$$



Bridged Treatment Comparisons[22]



[22] Results presented using twister plots (Zivich et al. (2021) *Am J Epidemiol*)

Conclusions

Key Advantages

Stacking estimating functions together

- Natural way to build an estimator
- Connects to interest versus nuisance parameters
- Sandwich variance
  - Percolates uncertainty of nuisance parameters
  - Automation of the delta-method
  - Computationally efficient

Existing estimators

- Many can be expressed as M-estimators
- Score function

Flexible software to implement M-estimators

Limitations

Valid estimating functions

- $\psi(O_i; \theta)$ must not depend on i
  - Excludes models like Cox PH model
- Non-smooth estimating functions
  - Bread estimator may not be valid

Finite dimensional nuisance model

- Nuisance parameters assumed to be finite dimension
- Unclear how (and if) data-adaptive algorithms could be used

Further Reading

Introductory papers

- Stefanski LA & Boos DD. (2002). The calculus of M-estimation. *The American Statistician*, 56(1), 29-38.
- Cole SR, Edwards JK, Breskin A, et al. (2022). Illustration of Two Fusion Designs and Estimators. *American Journal of Epidemiology*.
- Jesus J & Chandler RE. (2011). Estimating functions and the generalized method of moments. *Interface Focus*, 1(6), 871-885.

Software

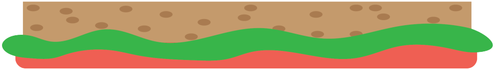
- deli.readthedocs.io
- bsaul.github.io/geex/

Thanks

Slides & code available at: github.com/pzivich/Presentations

pzivich@unc.edu @PausalZ pzivich

Appendix

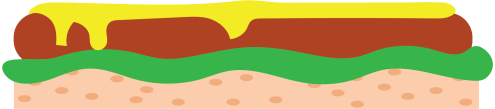


$$\left(A_i - \text{expit}(g(W_i)^T \alpha)\right) g(W_i)$$

$$(Y_i - \mu_1) \frac{A_i}{\text{expit}(g(W_i)^T \alpha)}$$

$$(Y_i - \mu_0) \frac{(1 - A_i)}{1 - \text{expit}(g(W_i)^T \alpha)}$$

$$(\mu_1 - \mu_0) - \mu_2$$



Augmented Inverse Probability Weighting

The AIPW estimator is

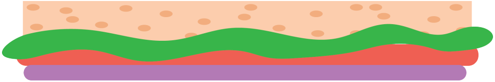
$$\frac{1}{n}\sum_{i=1}^{n} \tilde{Y}_i^1 - \tilde{Y}_i^0$$

$$\tilde{Y}_i^a = \frac{Y_i I(A_i = a)}{\Pr(A_i = a | W_i; \hat{\alpha})} + \frac{E[Y_i | A_i = a, W_i; \hat{\beta}](...)}{\Pr(A_i = a | W_i; \hat{\alpha})}$$

Estimating the variance for the RD

- Bootstrap
- Outer product of influence functions
- Sandwich

Augmented Inverse Probability Weighting



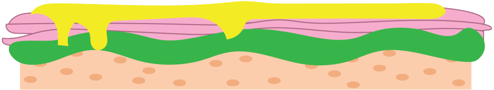
$$\left(A_i - \text{expit}(g(W_i)^T \hat{\alpha})\right) g(W_i)$$

$$\left(Y_i - \text{expit}(g(A_i, W_i)^T \hat{\beta})\right) g(A_i, W_i)$$

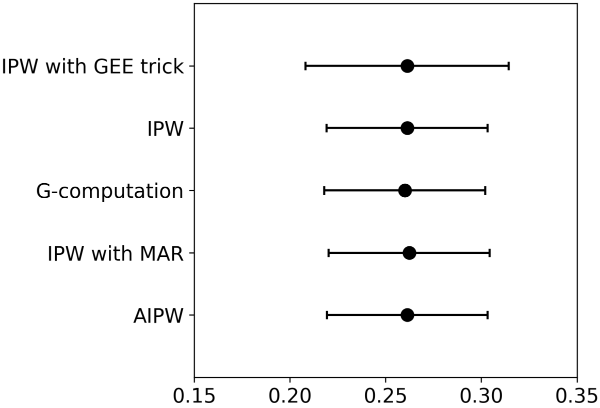
$$\tilde{Y}_i^1 - \hat{\mu}_1$$

$$\tilde{Y}_i^0 - \hat{\mu}_0$$

$$(\hat{\mu}_1 - \hat{\mu}_0) - \hat{\mu}_2$$



Results



	RD	95% CI
IPW with GEE trick	0.26	(0.21, 0.31)
IPW	0.26	(0.22, 0.30)
G-computation	0.26	(0.22, 0.30)
IPW with MAR	0.26	(0.22, 0.30)
AIPW	0.26	(0.22, 0.30)