An Application of Targeted Maximum Likelihood Estimation to Health Economic Evaluation

Susan Gruber
Department of Epidemiology, Harvard School of Public Health

TMLE work in collaboration with Mark van der Laan

Causal Inference

Goal: *Quantify the effect of treatment A on outcome Y*

Road Map

1. Specify a scientific research question
2. Formulate question that can be answered from data, i.e., target parameter of probability distribution P_0
3. Define a mapping from data to target parameter
$$\psi_0 = \Psi(P_0)$$
4. Develop a procedure to estimate P_0
5. Apply mapping
$$\psi_n = \Psi(P_n)$$

Example: Estimating additive effect of binary point treatment

- Naive approach can be biased due to confounding
  selection bias, informative dropout, etc.
- Traditional parametric modeling approach is inadequate
  - Parametric models are seldom correct, and break down for high-dimensional data
  - Maximum likelihood estimation typically involves a global bias/variance tradeoff
- **Targeted** maximum likelihood instead

Fundamental Issues: Bias and Variance

- Parametric model misspecification can bias estimates
- Advantages of a semi-parametric approach
  - weakens modeling assumptions
  - feasible with high dimensional data
- Global bias/variance tradeoff of maximum likelihood estimation not optimal when parameter of interest is low-dimensional.
  Can we make a better tradeoff for the parameter we care about?
- Ideal estimator
  - double robust
  - achieves semi-parametric efficiency bound

Demystifying Double Robustness

Kang and Schafer (2007)

- comparison of DR estimators, inverse probability weighting estimators (IPCW), OLS on simulated data
- extreme propensity scores (near positivity violations)
- correct and misspecified outcome regression and propensity score models
- demonstrated that DR estimators can perform worse than non-DR parametric regression based estimators

Rejoinders to K&S

- Robins et al, emphasize *boundedness*: estimates fall within the parameter space with probability 1
- Tsiatis and Davidian propose strategies to address positivity violations (no IP weighting for observations with pscore close to 0, use regression predictions)
- other respondents addressed the dual misspecification problem

Advances in TMLE address many of these issues

Relative Performance of Targeted Maximum Likelihood Estimators

Porter, et. al. (2011)

- Boundedness

  substitution estimator that remains within the parameter space

- Near Positivity Violations

  C-TMLE approach to estimating the treatment assignment/censoring mechanism

- Model Misspecification

  data-adaptive Super Learning can replace parametric model specification

Simulation Study in Health Economics

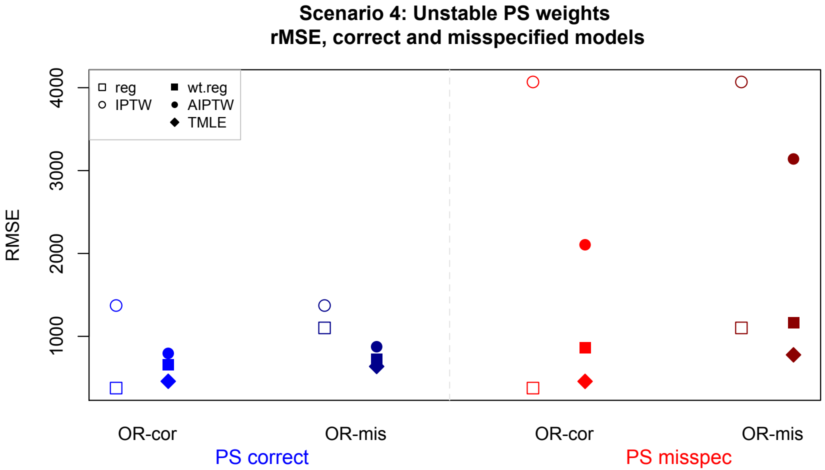
joint work with Noémi Krief, Rosalba Radice, Richard Grieve, Jasjeet S. Sekhon

- Estimating ATE on Incremental Net Benefit (INB)
- INB is a composite outcome (Cost and QALY)
- Analogous to Kang and Schafer
  - extreme propensity scores
  - correct and misspecified outcome regression and propensity score models
  - comparison of estimators

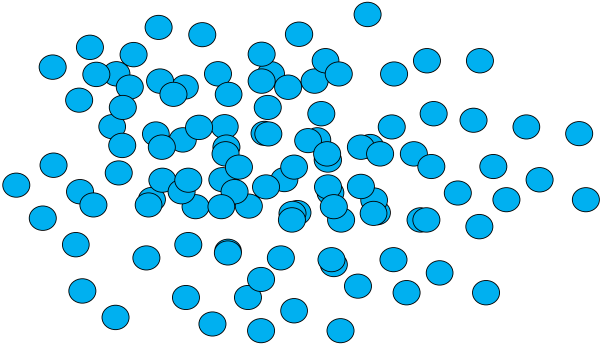
Non-DR	DR
regression	weighted regression
IPTW*	augmented IPTW
	TMLE

*inverse probability of treatment weighting

Hernán et al, 2000, Robins and Rotnitzky 2001

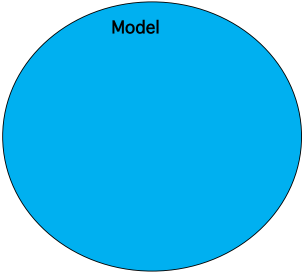

Scenario 4: Unstable PS weights
rMSE, correct and misspecified models
RMSE
1000
2000
3000
4000
reg
IPTW
wt.reg
AIPTW
TMLE
OR-cor
OR-mis
OR-cor
OR-mis
PS correct
PS misspec

Motivation for TMLE



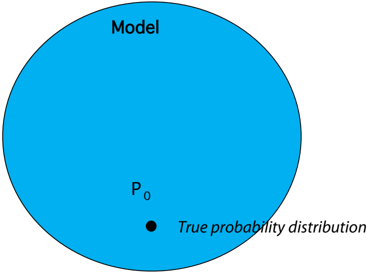
Observations (O_1, ... O_n)

Motivation for TMLE

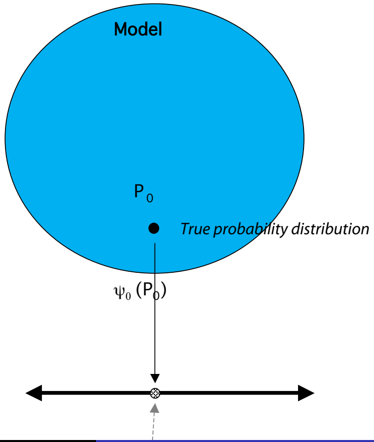


Model: set of possible probability distributions of the data

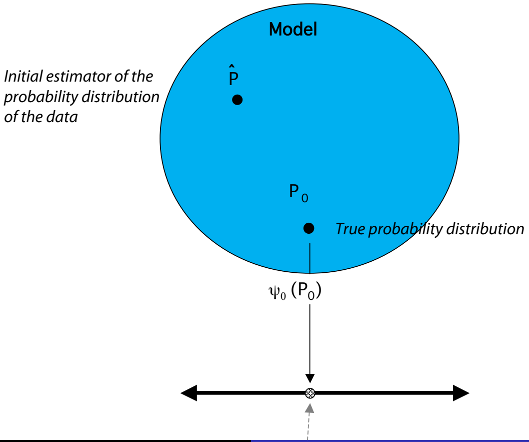
Motivation for TMLE



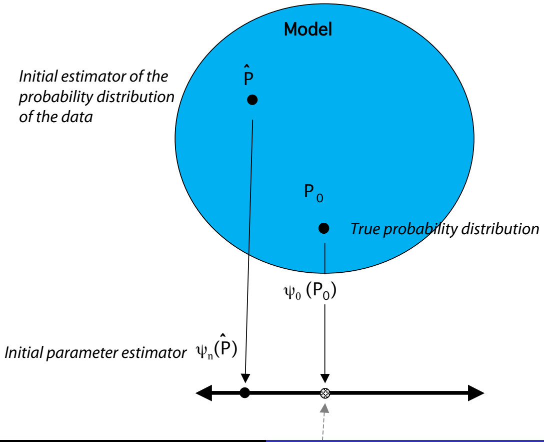
Motivation for TMLE



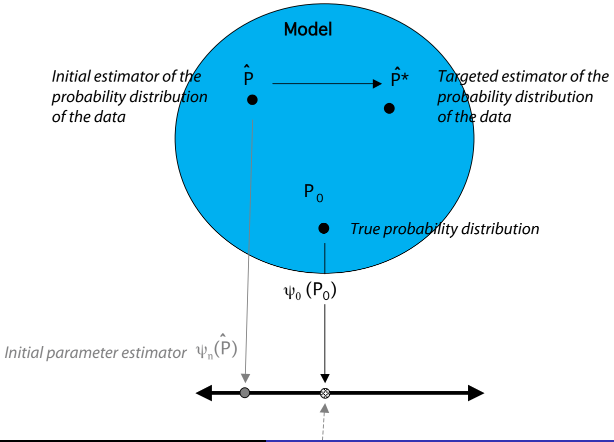
Motivation for TMLE



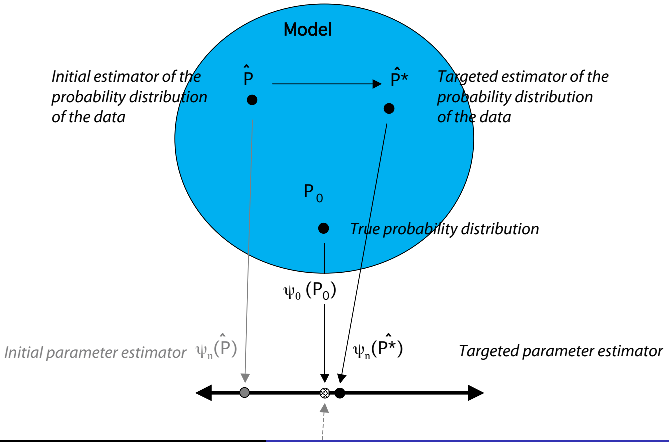
Motivation for TMLE



Motivation for TMLE



Motivation for TMLE



Example: TMLE to estimating additive effect of binary point treatment (ATE)

$$O = (Y, A, W) \sim P_0$$

Y: outcome, A: binary treatment indicator, W: covariate vector

Likelihood factorizes:

$$\mathcal{L}(O) = \underbrace{P(Y \mid A, W)}_{Q_Y} P(A \mid W) \underbrace{P(W)}_{Q_W}$$

Define

$$Q_0 = (Q_{0_Y}, Q_{0_W})$$
$$g_0 = P_0(A \mid W)$$

Double Robustness: TMLE consistent if either Q_0 or g_0 is estimated consistently, and asymptotically efficient when both are correct.

TMLE Algorithm

- Step 1: Obtain initial estimate

$$\bar{Q}_n^0(A, W) = \hat{E}(Y \mid A, W)$$

TMLE Algorithm

- Step 1: Obtain initial estimate

$$\bar{Q}_n^0(A, W) = \hat{E}(Y \mid A, W)$$

- Step 2: Target initial estimate (logit scale)

$$\bar{Q}_n^*(A, W) = \bar{Q}_n^0(A, W) + \epsilon H_{g_n}^*(A, W)$$

  - Estimate g_0 (propensity score)
  - construct $H_{g_n}^*(A, W)$, parameter-specific fluctuation covariate
  - maximum likelihood to fit ϵ

TMLE Algorithm

- Step 1: Obtain initial estimate

$$\bar{Q}_n^0(A, W) = \hat{E}(Y \mid A, W)$$

- Step 2: Target initial estimate (logit scale)

$$\bar{Q}_n^*(A, W) = \bar{Q}_n^0(A, W) + \epsilon H_{g_n}^*(A, W)$$

  - Estimate g_0 (propensity score)
  - construct $H_{g_n}^*(A, W)$, parameter-specific fluctuation covariate
  - maximum likelihood to fit ϵ

- Step 3: Evaluate parameter: $\psi_n^{TMLE} = \Psi(\bar{Q}_n^*)$

Key idea: $\Psi(\bar{Q}_n^*)$ less biased than $\Psi(\bar{Q}_n^0)$

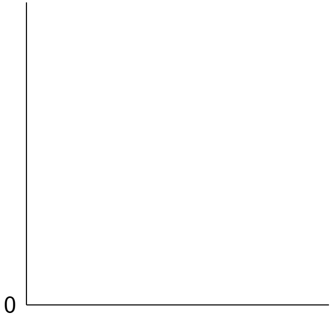
Influence Curve (IC)

IC is a function that describes estimator behavior under slight perturbations of the empirical distribution

Hampel, 1974

Influence Curve (IC)

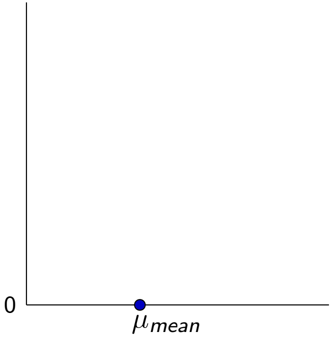
IC is a function that describes estimator behavior under slight perturbations of the empirical distribution



Hampel, 1974

Influence Curve (IC)

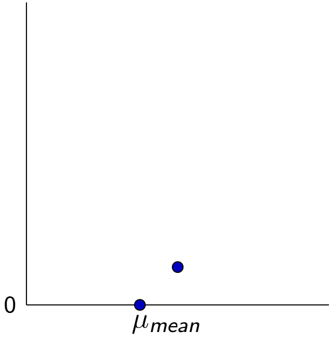
IC is a function that describes estimator behavior under slight perturbations of the empirical distribution



Hampel, 1974

Influence Curve (IC)

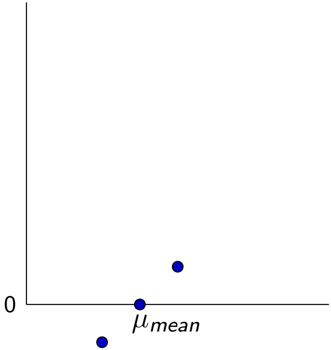
IC is a function that describes estimator behavior under slight perturbations of the empirical distribution



Hampel, 1974

Influence Curve (IC)

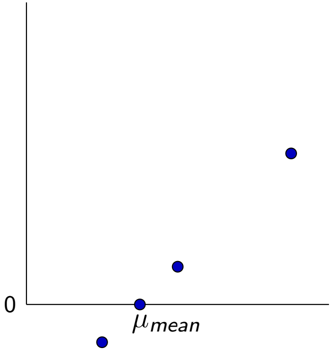
IC is a function that describes estimator behavior under slight perturbations of the empirical distribution



Hampel, 1974

Influence Curve (IC)

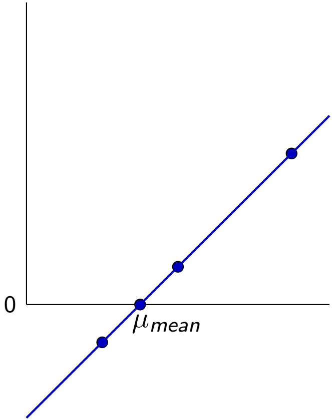
IC is a function that describes estimator behavior under slight perturbations of the empirical distribution



Hampel, 1974

Influence Curve (IC)

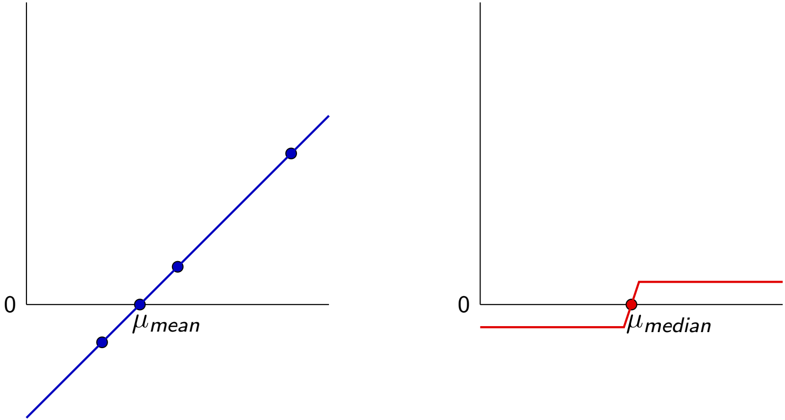
IC is a function that describes estimator behavior under slight perturbations of the empirical distribution



Hampel, 1974

Influence Curve (IC)

IC is a function that describes estimator behavior under slight perturbations of the empirical distribution



Hampel, 1974

Influence Curve

- Influence Curve has mean 0 at the true parameter value, so can be used as an estimating equation
- Consider the mean

$$
\begin{aligned}
IC^{mean} &= x - \mu \\
0 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)
\end{aligned}
$$

- Every estimator for a given parameter has its own influence curve (not necessarily unique)
- Among all these influence curves, one has minimal variance. This is the Efficient Influence Curve (canonical gradient)
- An examination of an estimator's influence curve gives insight into its behavior

ATE parameter

$$\psi_0^{ATE} = E_0(E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W))$$

$$IC_{eff}^{ATE} = \left(\frac{I(A=1)}{g_0(1 \mid W)} - \frac{I(A=0)}{g_0(0 \mid W)}\right)[Y - \bar{Q}(A, W)] + \bar{Q}(1, W) - \bar{Q}(0, W) - \psi$$

$$H_{g_0}^{*ATE}(A, W) = \frac{I(A=1)}{g_0(1 \mid W)} - \frac{I(A=0)}{g_0(0 \mid W)}$$

$H_{g_0}^{*}(A, W)$ is derived such that the maximum likelihood procedure that fits ϵ also solves score equations that span the efficient influence curve for the target parameter.

Influence curve (IC)

- empirical mean of IC for regular asymptotically linear (RAL) estimator provides linear approximation of estimator
- thus, VAR(IC) provides asymptotic variance of estimator
- IC-based inference: p-value, test statistic, confidence interval
- Solving the efficient IC equation confers double robustness

Example: Estimating an Additive Treatment Effect

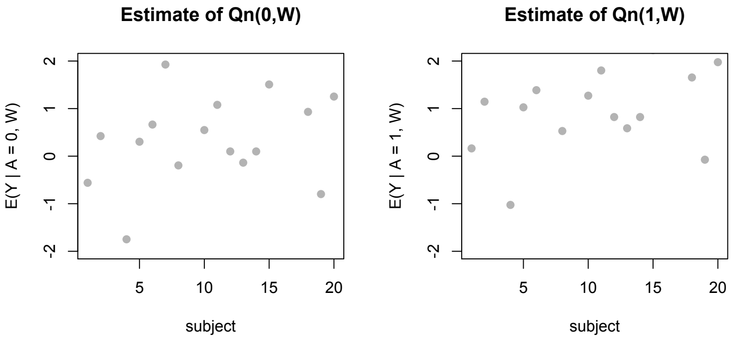
$$\psi_0^{ATE} = E_0(E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W))$$

1. Generate Data, P_0 : $(\psi_0 = 1)$

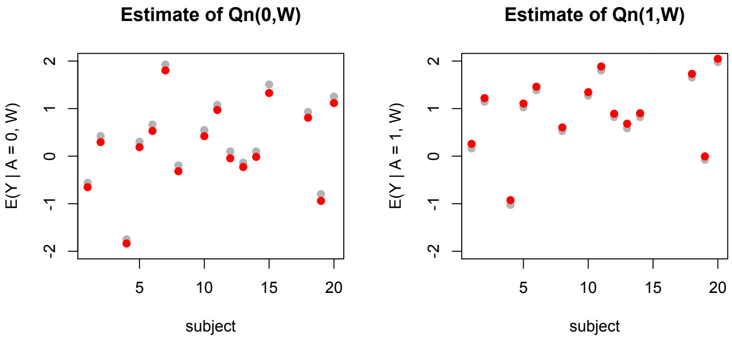
$$W_1, W_2, W_3 \sim N(0, 1)$$
$$g_0(A = 1 \mid W) = expit(0.2 + 0.2W_1 + 0.3W_2)$$
$$\bar{Q}_0(A, W) = A - 3AW_1 - 2W_1W_3$$

2. Illustrate TMLE performance, $\bar{Q}_0$ is misspecified, g_0 correct

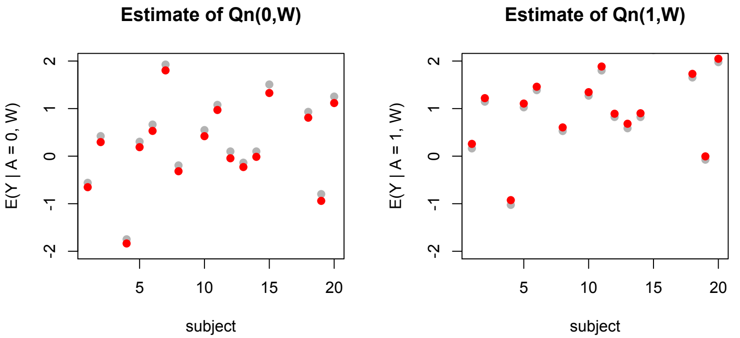
```
Qmis <- glm(Y ~ A + W1 + W2 + W3 + W1*W3)
gcor <- glm(A ~ W1 + W2, family = binomial)
```



Initial estimate of counterfactual predicted outcomes based on misspecified model for $\bar{Q}_0$ when A is set to 0 (l) and A is set to 1 (r).



Targeted estimate of counterfactual predicted outcomes based on misspecified model for $\bar{Q}_0$ when A is set to 0 (l) and A is set to 1 (r).



Targeted estimate of counterfactual predicted outcomes based on misspecified model for $\bar{Q}_0$ when A is set to 0 (l) and A is set to 1 (r).

$$\Psi(\bar{Q}_n^0) = 0.72,\ \Psi(\bar{Q}_n^*) = 0.93$$

one sample, $n = 500$, $\psi_0 = 1$

Collaborative Targeted Minimum Loss-Based Estimation (C-TMLE)

Key Insight: Maximal bias reduction achieved by adjusting for residual confounding in $(Q_0 - Q_n^0)$ only

- standard TMLE: external estimate of g_0
- C-TMLE: estimate only the required portion of g_0
  - same target parameter
  - goal is is reduce MSE
  - recommended when there are near positivity violations

C-TMLE algorithm

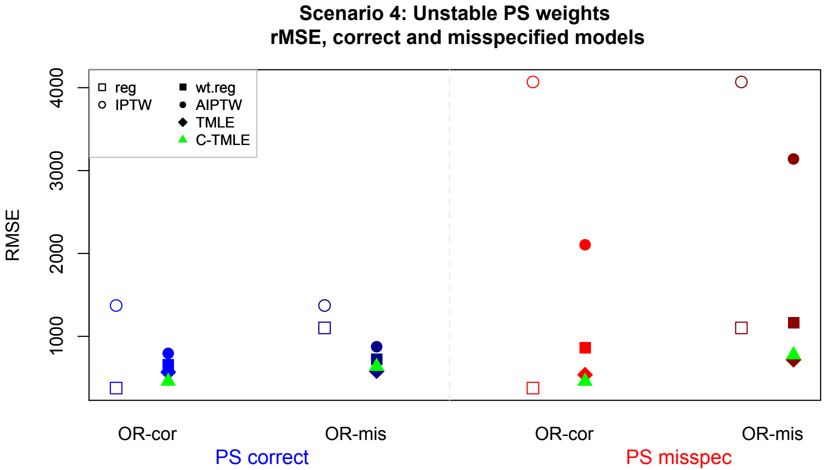
Step 1. Estimate $\bar{Q}_n^0 = \hat{E}(Y \mid A, W)$

Step 2. Collaborative targeted construction of candidate treatment mechanism estimates, $\hat{g}_1, \ldots, \hat{g}_K$

Step 3. Create candidate TMLE estimators $\bar{Q}_1^*(\hat{g}_1), \ldots \bar{Q}_K^*(\hat{g}_K)$

Step 4. Select the best candidate, $\bar{Q}_n^* = \bar{Q}_k^*(\hat{g}_k)$, using cross validation *(loss function for Q)*

Step 5. Evaluate parameter: $\psi_n^* = \Psi(Q_n^*)$


Scenario 4: Unstable PS weights
rMSE, correct and misspecified models
RMSE
1000
2000
3000
4000
reg
IPTW
wt.reg
AIPTW
TMLE
C-TMLE
OR-cor
OR-mis
OR-cor
OR-mis
PS correct
PS misspec

Summary

TMLE family of estimators

- locally efficient
- substitution estimator (respects known bounds on $\mathcal{M}$)
- listening to data in principled way trumps intuition
- C-TMLE, Super Learning

Acknowledgements

Richard Grieve, Noémi Krief, Rosalba Radice, *LSHTM*

Nick Black, Jan vanderMeulen, *LSHTM*

Jas Sekhon, Kristin Porter, *UC Berkeley*

Jamie Robins, Miguel Hernán *Harvard School of Public Health*

References

- Software
  - S. Gruber. tmle. R package version 1.2.0-1, URL http://CRAN.R-project.org/package=tmle, 2012
  - S. Gruber and M.J. van der Laan, tmle: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software* (in press).
- Papers
  - N. Kreif, R. Radice, R. Grieve, J.S. Sekhon. Regression-adjusted matching and double-robust methods for estimating average treatment effects in health economic evaluation. (2011)
  - S. Gruber and M.J. van der Laan. An Application of Collaborative Targeted Maximum Likelihood Estimation in Causal Inference and Genomics. *The International Journal of Biostatistics*, 6(1), 2010.
  - S. Gruber and M.J. van der Laan. A Targeted Maximum Likelihood Estimator of a Causal Effect on a Bounded Continuous Outcome. *The International Journal of Biostatistics*, 6(1), 2010.
  - F. Hampel. The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association*, 69(346), 38393, 1974.

- M. A. Hernan, B. Brumback, and J. M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5):561570, 2000.
- J. Kang and J. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22:52339, 2007.
- K.E. Porter, S. Gruber, M.J. van der Laan and J.S. Sekhon. The Relative Performance of Targeted Maximum Likelihood Estimators. *The International Journal of Biostatistics*, 7(1), 2011.
- J.M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, 2000.
- J.M. Robins, A. Rotnitzky. Comment on the Bickel and Kwon article, Inference for Semiparametric Models: Some Questions and an Answer. *Statistica Sinica*, 11(4), 920 936, 2001.
- M.J. van der Laan and D.E. Rubin. Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, 2(1), 2006.
- M.J. van der Laan, E. Polley, and A. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(25), 2007. ISSN 1.

- M.J. van der Laan and S. Gruber. Collaborative double robust penalized targeted maximum likelihood estimation. The International Journal of Biostatistics, 2010.

- Books

- M.J. van der Laan and S. Rose. *Targeted Learning: Prediction and Causal Inference for Observational and Experimental Data*. Springer, New York, 2011.

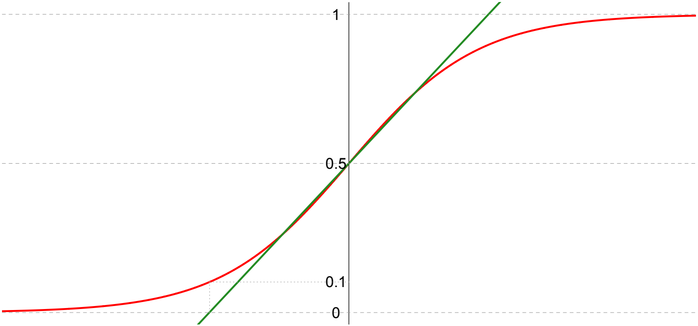
Additional Slides

TMLE for Bounded Continuous Outcomes

- Logistic Fluctuation
    - Enforces bounds on the problem
    - Reduces bias and variance (w.r.t. linear fluctuation) when there is sparsity in the data
    - Robustifies estimator when applied to sparse data

Logistic fluctuation ensures TMLE estimate respects bounds on semi-parametric model

- $Y \in [a, b]$ maps to $Y^* \in [0, 1]$ $\qquad Y^* = \frac{(Y-a)}{(b-a)}$



Logistic fluctuation ensures TMLE estimate respects bounds on semi-parametric model

- $Y \in [a, b]$ maps to $Y^* \in [0, 1]$ $\quad Y^* = \frac{(Y-a)}{(b-a)}$
- Causal Additive Treatment Effect for Y^*

$$\psi_0^* = \Psi^*(P_0) = E_0[E_0(Y^* \mid A = 1, W) - E_0(Y^* \mid A = 0, W)]$$

- Fluctuate on logit scale to target initial estimate

$$Q^*_{n,Y^*}(A, W) = expit[logit(Q^0_{n,Y^*}) + \epsilon h(A, W)]$$

- logistic regression to fit ϵ, $h(A, W) = \frac{I(A=1)}{g(1|W)} - \frac{I(A=0)}{1-g(1|W)}$

(same as for linear fluctuation)