Evaluating treatment effectiveness under model misspecification: a comparison of targeted maximum likelihood estimation with bias-corrected matching

Noemi Kreif[1], Susan Gruber[2], Rosalba Radice[1], Richard Grieve[1], Jasjeet S. Sekhon[3]

[1]Department of Health Services Research and Policy, LSHTM,
[2]Department of Epidemiology, Harvard School of Public Health,
[3]Travers Department of Political Science, UC Berkeley

CSM Seminar, 31th January 2013

Background

PhD thesis "Statistical methods to address selection bias in economic evaluations that use patient-level observational data"

- Funding by ESRC
- Supervisor Richard Grieve, advisory committee James Carpenter, John Cairns, Rhian Daniel

- Applying and extending methods from the causal inference literature for cost-effectiveness analysis

Challenges:

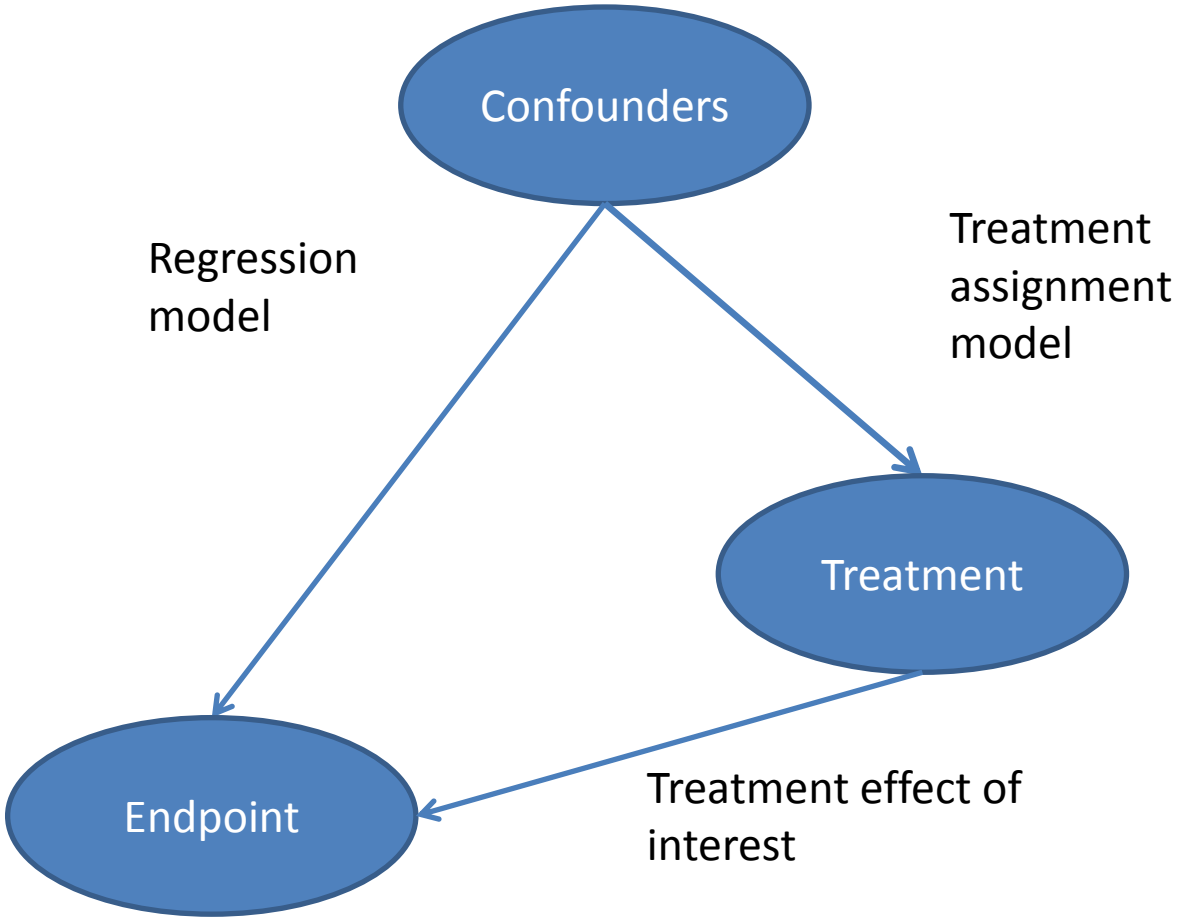
- Estimating treatment effects
- Correlated costs and outcomes
- Endpoints with skewed distributions
- Nonlinear relationships between the covariates and the endpoints
- Subgroup-effects of interest

Context

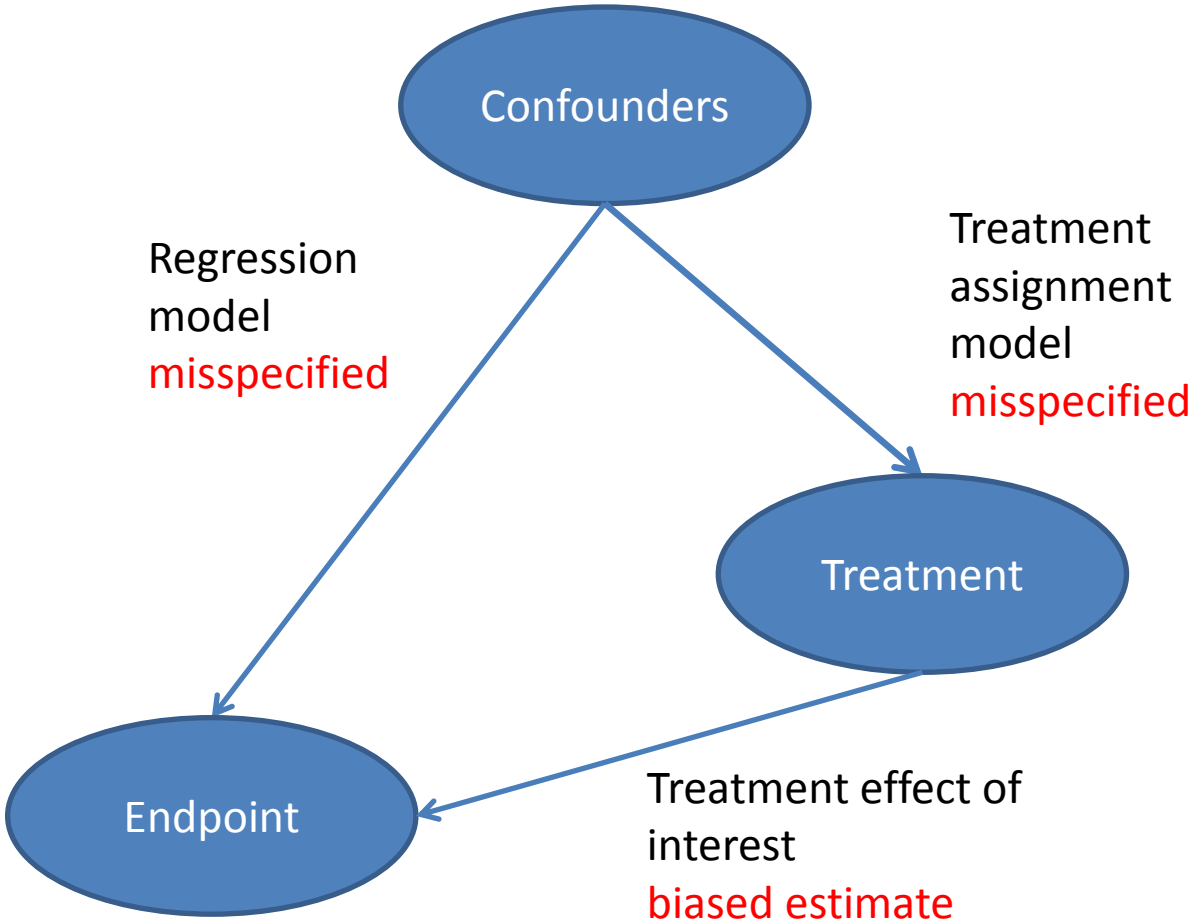
- Health policy makers require estimates of clinical and cost-effectiveness
- Observational data often used
- Motivating example: comparative effectiveness of alternative types of hip replacement on patients' post-operative quality of life

- Main challenge: control for confounding
  - Assume no unmeasured confounding
- Statistical methods to adjust for observed confounding:
  - Modelling endpoint: regression methods
  - Modelling assignment mechanism: propensity score (PS) methods

Context


Confounders
Regression model
Treatment assignment model
Treatment
Endpoint
Treatment effect of interest

Context


Confounders
Regression model misspecified
Treatment assignment model misspecified
Treatment
Endpoint
Treatment effect of interest biased estimate

Context

- "Combined" approaches recommended (Rubin, 1973):
  - Regression + PS weights: double-robust methods (Robins and colleagues)
  - PS matching + regression: regression-adjusted matching, bias-corrected matching

- Interest in finite sample performance when both models are misspecified (Kang and Schafer, 2007)

- Machine-learning methods to reduce bias due to misspecification (Porter et al, 2011; Lee et al., 2010; Austin, 2010)

- Little work on comparing matching and DR methods (Waernbaum, 2010; Busso et al., 2011)

Objective: to compare two approaches that combine the PS with regression: targeted maximum likelihood estimation (van der Laan and Rubin, 2006) and bias-corrected matching (Abadie and Imbens, 2010)

Outline

1. Potential outcomes framework and notation
2. Limitations of “standard” methods
   - Regression
   - PS methods: matching, IPTW
3. Double-robust methods
   - Targeted maximum likelihood estimation
4. Bias-corrected matching
5. Motivating case study
6. Simulation design
7. Simulation results
8. Discussion

Potential outcomes framework

- $Y(1)$ is the potential outcome under treatment and $Y(0)$ under control
- W : vector of confounders
- A : binary, time-constant treatment
- Parameter of interest: average treatment effect (ATE): $\psi = E\left[Y(1) - Y(0)\right]$

- Under no unmeasured confounding:

$$\psi = E\left[\left[\left(Y \mid A=1, W\right) - \left(Y \mid A=0, W\right)\right] \mid W\right]$$

- Further assumptions: consistency and positivity

Positivity assumption

(Westreich and Cole, 2010)

- Model for the treatment assignment: $g(A,W) = P(A|W)$

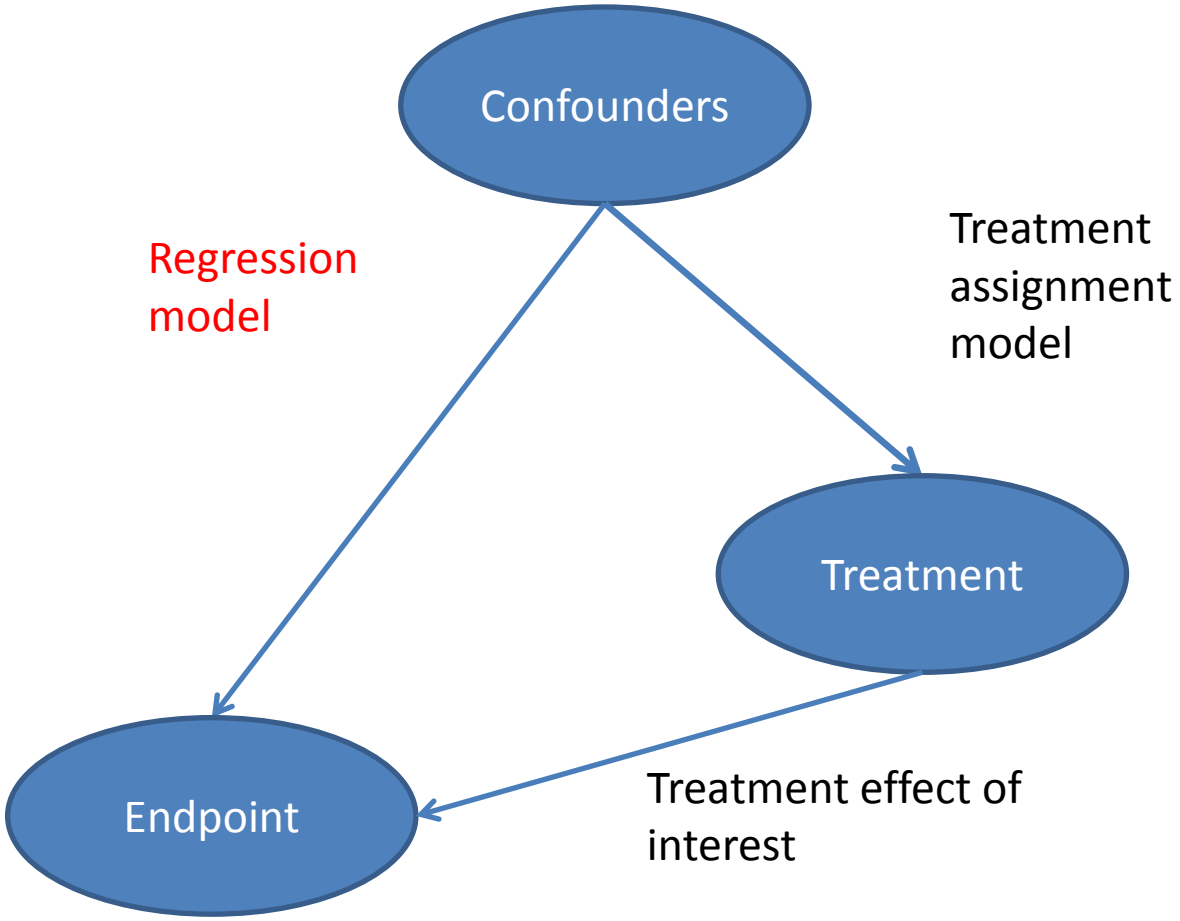
- Positivity assumption holds if $0 < g(A,W) < 1$ for any W
  - possible to have units at every combination of the observed confounders
  - experimental treatment assignment, common support, good overlap

- Structural positivity violations -> ATE cannot be identified

- "Practical positivity violations" – finite sample issue, small numbers or no individuals for some W

- Problem for methods that use $\hat{g}(A,W)$
- Here, referred to as poor overlap

Controlling for observed confounding



Regression

- To estimate expected potential outcomes, controlling for observed confounders

- Regression function: $Q(A,W) = E[Y \mid A,W]$

- Regression estimator: $\hat{\psi}^{reg} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{Q}(1,W_i) - \hat{Q}(0,W_i)\right\}$

- In practice, $\hat{Q}(A,W_i)$ obtained using fixed, parametric models e.g. GLMs

- If large imbalance in the baseline distribution of confounders, regression estimates sensitive to misspecification (Ho et al, 2007)

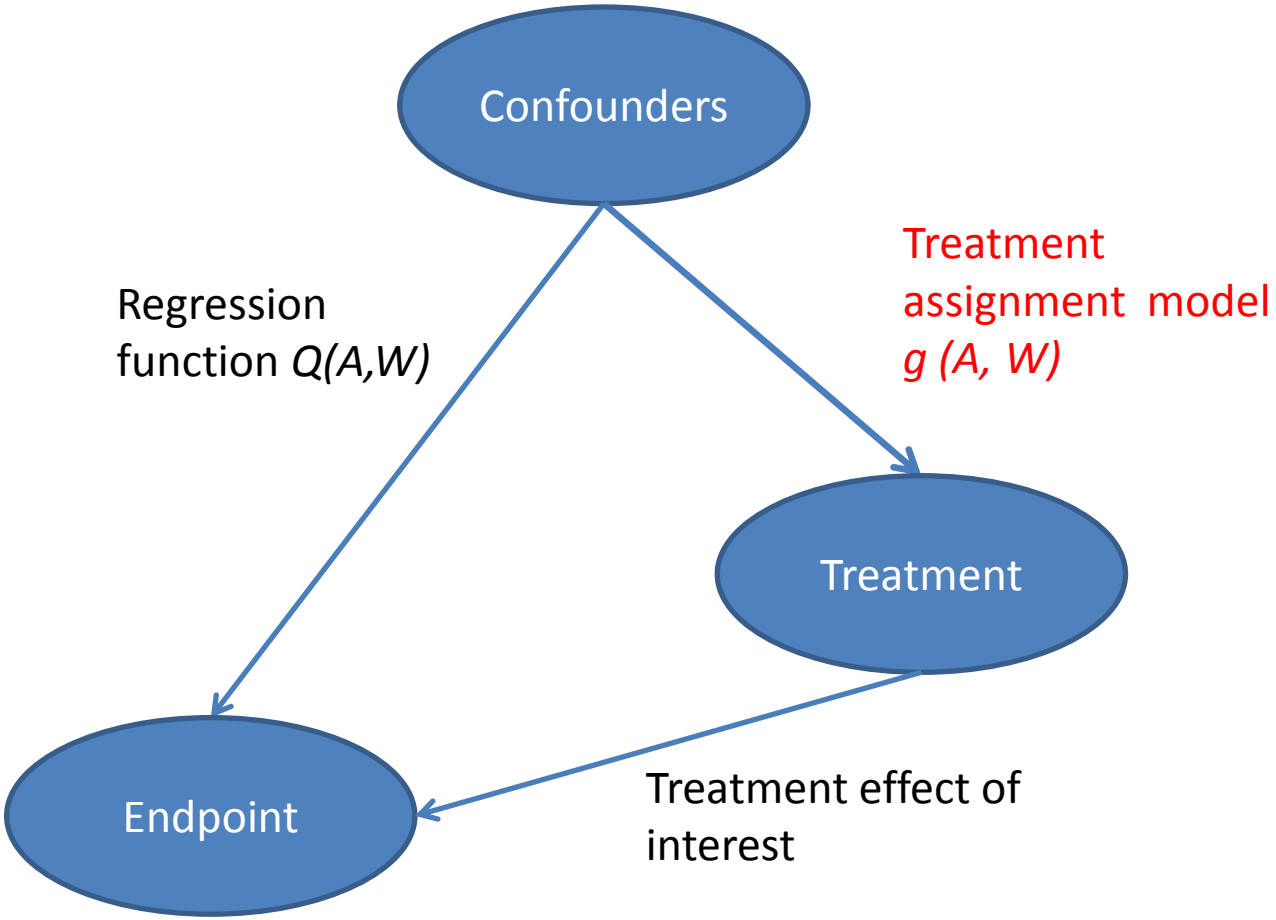
- Objective: reduce functional form misspecification

Machine learning estimation for the endpoint regression function

- Instead of fixed, parametric models, more flexible approaches to estimate the regression function
  - Regression trees, random forests (Austin, 2012)

- "Super learner" (van der Laan and Polley, 2007)
  - Prediction algorithm
    - For causal inference: predictions for both potential outcomes
  - User defines the range of algorithms (from fixed parametric models, to nonparametric)
  - Selects convex combination based on cross validation
  - Software: http://cran.r-project.org/web/packages/SuperLearner/index.html

Controlling for observed confounding



Propensity score methods

- Propensity score: $g(1,W) = \Pr(A = 1 | W)$

- Matching: impute missing counterfactual, using information from the closest matches (measured by $\hat{g}(1,W)$)
- To create balance in matched sample

- Inverse probability of treatment weighting (IPTW): reweighting treated and control sample to create balance

- Weights: $\frac{A_i}{\hat{g}(1,W_i)}$ for treated and $\frac{1-A_i}{1-\hat{g}(1,W_i)}$ for control

- Poor overlap -> $\hat{g}(.)$ close to 0 or 1 -> extreme weights –> bias, inefficiency for IPTW (weight truncation)

- IPTW more sensitive to misspecification than matching (Waernbaum, 2011)

Machine learning estimation of the propensity score

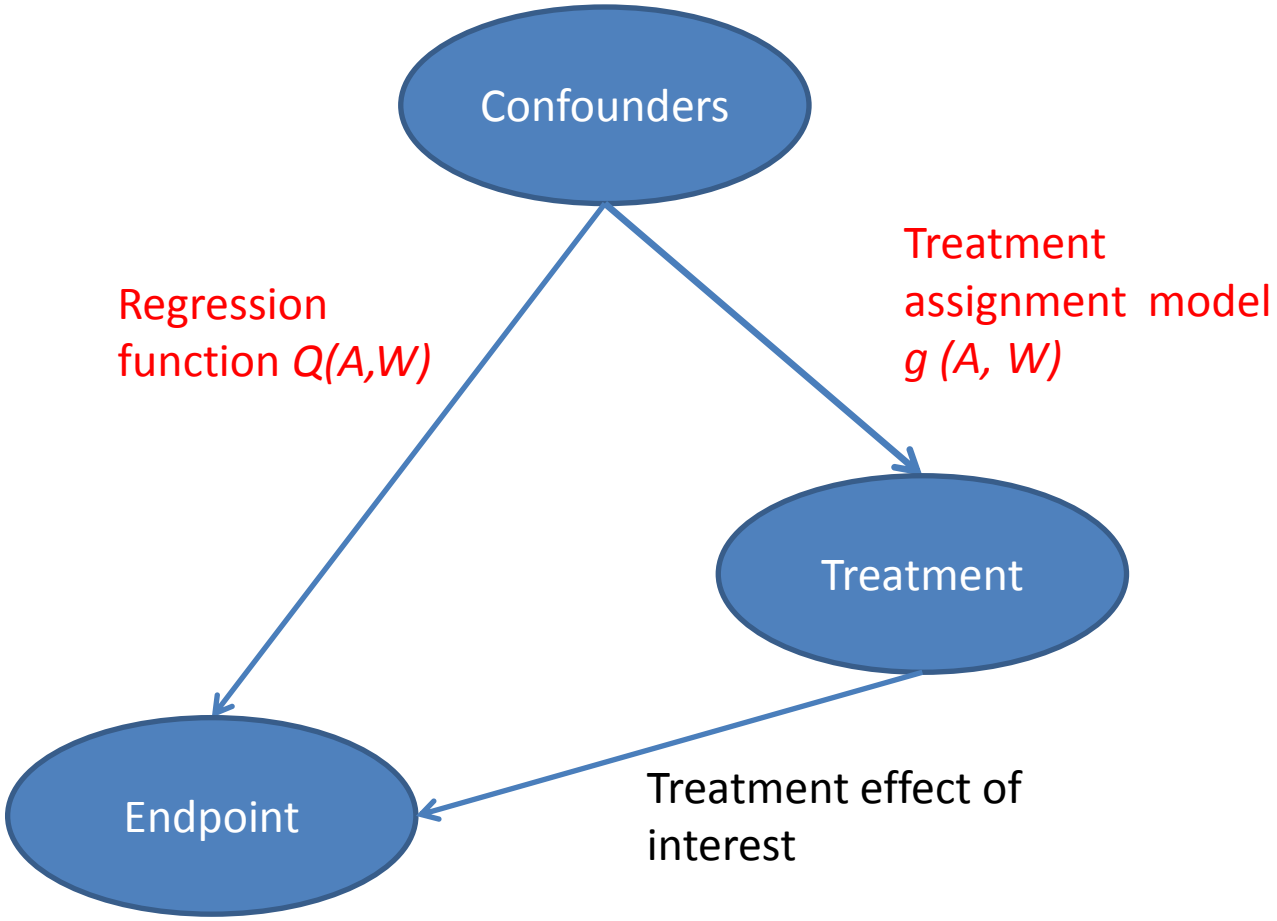
- PS balancing score if correctly specified (Rosenbaum and Rubin, 1983)

- Standard in applied work: main terms logistic regression
  - Untested assumptions: linearity in the log odds, no interactions, higher order terms

- Alternatives: regression trees, neural networks (Westreich et al., 2010)

- Boosted classification and regression trees (CARTs) (Lee et al., 2010)
  - Uses generalised boosted (logistic) regression
  - Demonstrated to protect against extreme PS weights -> bias reduction, efficiency improvement
  - Selects PS to maximise balance
  - Software: R package “twang” http://cran.r-project.org/web/packages/twang/index.html

Controlling for observed confounding



Double-robust methods

- Combine $Q(A,W)$ and $g(A,W)$
- DR property: consistent if either $Q(A,W)$ or $g(A,W)$ is correctly specified (Robins et al., 1994)

- Common implementations: weighted regression, augmented IPTW

  - When PS extreme, DR methods can perform worse than OLS regression (Kang and Schafer, 2007)
  - Debate, new methods proposed (Robins et al., 2007; Tan 2010; Porter et al., 2010; Robins et al., 2012)

- Boundedness is a desirable property: $\hat{Q}(A,W)$ always lies within the range of the possible values of the endpoint

- Targeted maximum likelihood estimation (TMLE)

Targeted maximum likelihood estimation

Van der Laan and Rubin, 2006

- Traditional ML: aims to maximise the likelihood function for the whole distribution of the data

- TMLE: designed to reduce bias in the parameter of interest (e.g. ATE), by "targeting" an initial estimate of the endpoint regression function

- Solves the efficient influence curve estimating equation
  - double-robust.
  - semi-parametric efficient if both the initial $Q(A,W)$, and $g(A,W)$ are correctly specified

Targeted maximum likelihood estimation

Van der Laan and Rubin, 2006

2 stage estimator:

1. Obtain initial estimate of $E\left[Y(A,W)\right]=Q^0(A,W)$

2. Update (fluctuate) initial estimate, using information from the treatment assignment mechanism

$$\hat{Q}^1(A,W)=\hat{Q}^0(A,W)+\hat{\varepsilon}\hat{h}(A,W)$$

- fluctuation on the logistic scale (continuous endpoints rescaled between 0 and 1), to ensure boundedness (Gruber and van der Laan, 2010)

- for the ATE, $h(A,W)=\frac{A}{\hat{g}(1,W)}-\frac{1-A}{1-\hat{g}(1,W)}$

- Estimator for the ATE $\hat{\psi}^{TMLE}=\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{Q}^1(1,W_i)-\hat{Q}^1(0,W_i)\right\}$

Targeted maximum likelihood estimation

```
m <- glm(Y ~ A + W1*W3 + W2)
Q <- cbind(QAW = predict(m),
    Q1W = predict(m, newdata = data.frame(A = 1, W1, W2, W3)),
    Q0W = predict(m, newdata = data.frame(A = 0, W1, W2, W3)))

g <- glm(A ~ W1 + W2, family = binomial)
g1W <- predict(g, type = "response")
h <- cbind(A/g1W - (1-A)/(1-g1W), 1/g1W, -1/(1-g1W))
epsilon <- coef(glm(Y ~-1 + h[,1] + offset(Q[,"QAW"])))
Qstar <- Q + epsilon*h
psi <- mean(Qstar[,"Q1W"] - Qstar[,"Q0W"])
```

(Further R code available in Gruber and van der Laan, 2012)

- Demonstrated to outperform other DR methods, when there is poor overlap (Porter et al. 2011)
- Recommended to be applied with machine learning estimation for *Q(.)* and *g(.),* to reduce bias due to misspecification
- Software implementation: R package "TMLE" (Gruber and van der Laan, 2012)

http://cran.r-project.org/web/packages/tmle/index.html

Bias-corrected matching

(Abadie and Imbens, 2010)

- Reminder: matching estimators impute missing potential outcome
- Idea: to "clean up" residual imbalances between treatment groups after matching
- Subtract an estimate of the bias, from each individual-level estimate of treatment effect

1. Obtain matched data
2. Obtain regression predictions for the potential outcomes
3. Construct missing counterfactual using
   - Observed values of the matches
   - Adjustment term

$$\hat{Y}(0,W_i) = \begin{cases} Y_i & \text{if} \quad A_i = 0 \\ \frac{1}{M} \sum_{j \in \zeta_{M(i)}} Y_j + \hat{Q}(0,W_i) - \hat{Q}(0,W_j) & \text{if} \quad A_i = 1 \end{cases}$$

$$\hat{Y}(1,W_i) = \begin{cases} \frac{1}{M} \sum_{j \in \zeta_{M(i)}} Y_j + \hat{Q}(1,W_i) - \hat{Q}(1,W_j) & \text{if} \quad A_i = 0 \\ Y_i & \text{if} \quad A_i = 1 \end{cases}$$

Bias-corrected matching

(Abadie and Imbens, 2010)

- Estimator: $$\hat{\psi}^{BCM} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{Y}(1,W_i) - \hat{Y}(0,W_i)\right\}$$

- Double-robust properties suggested: "BCM estimators have the advantage of an additional layer of robustness, because matching ensures consistency for any given value of the smoothing parameters, without accurate approximations to either the regression function or the PS "(AI,2010)

- Implementations use OLS for adjustment: if non-linearities not severe, bias reduced (Busso et al., 2011)

- Software implementation: "NNMATCH" package in Stata, implements BCM with OLS

ideas.repec.org/c/boc/bocode/s439701.html

Motivating case study

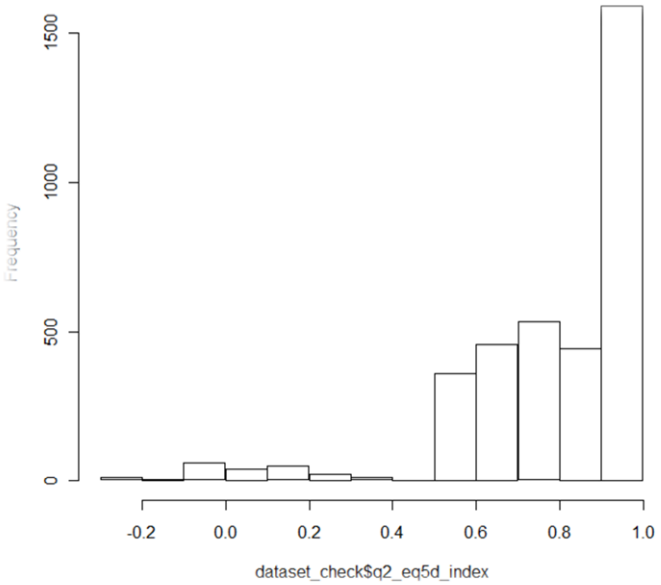
- Effect of alternative types of hip replacement on patients' post-operative quality of life (QoL)
  - EQ5D-3L

- Interest for policy makers: clinical and cost-effectiveness

- PROMs: large observational dataset

- Bounded endpoint with spike at 1

- Regression modelling challenging (Basu and Manca, 2010)



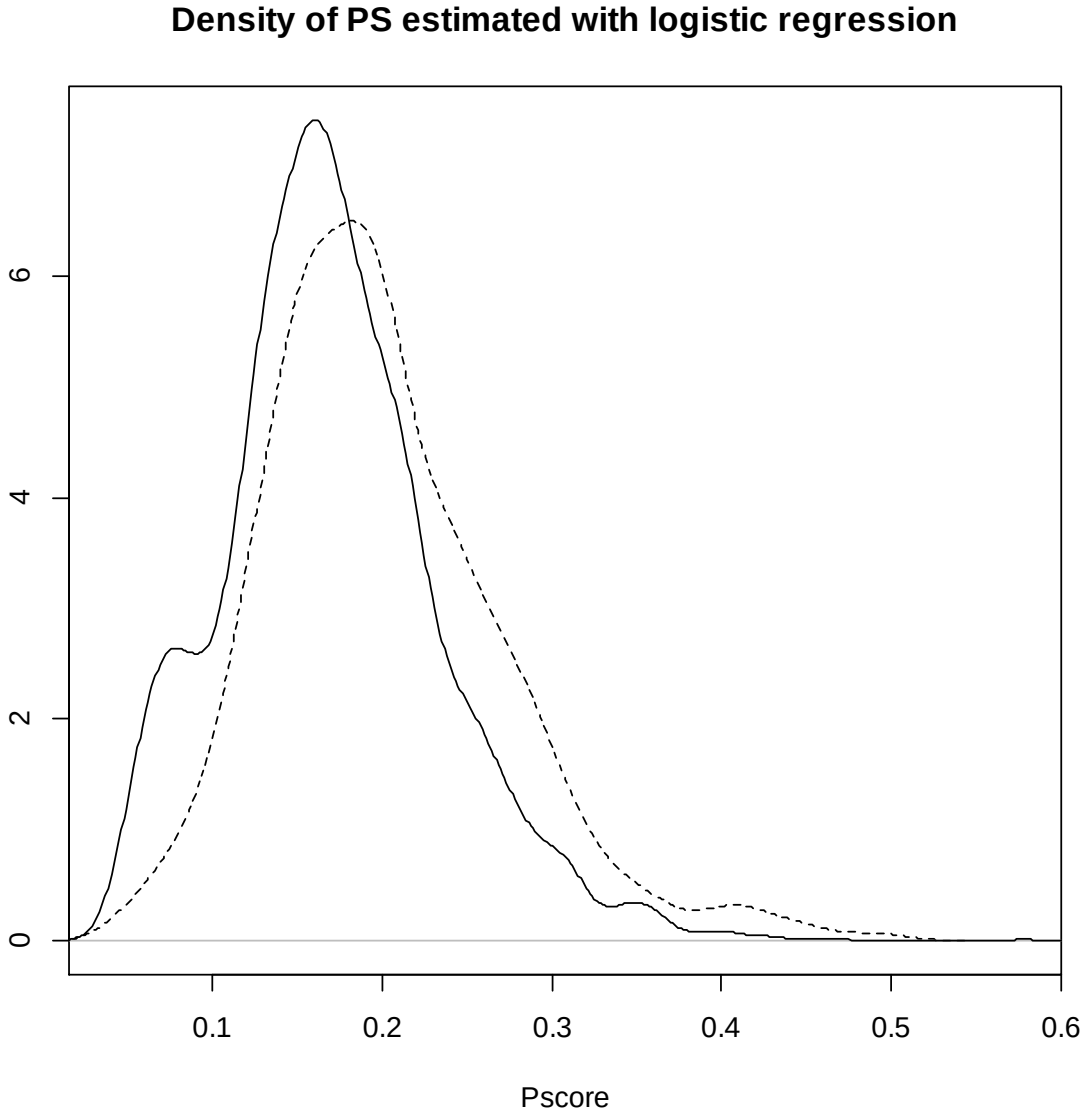
- Causal question: QoL difference between patients with hybrid vs. uncemented hip replacement (male, aged 65-74 (n = 3,583))

Methods in the case study

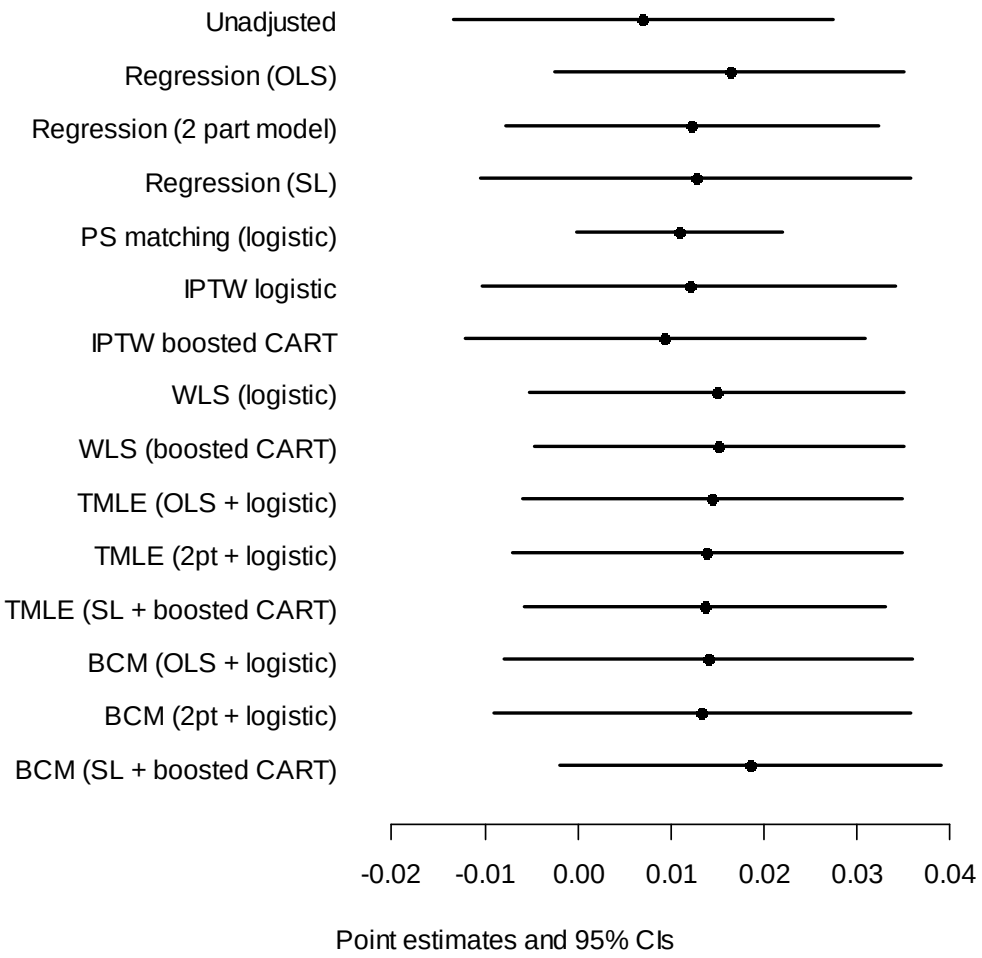
	Fixed parametric model	**Machine learning**
TMLE	OLS + logistic regression Two part model + logistic regression	Super learner + boosted CARTs
Bias-corrected PS matching	OLS + logistic regression Two part model + logistic regression	Super learner + boosted CARTs

Also implemented: regression, IPTW, matching, weighted least squares (both fixed parametric and machine learning methods)

Case study: estimated PS



Case study: estimated ATEs



Simulation design

Focus: to compare finite sample performance of TMLE and BCM for range of realistic scenarios, when using fixed parametric models and machine learning, under misspecification

Hypotheses:

- When overlap good: reweighting methods (weighted regression, TMLE) outperform BCM (more efficient, when correctly specified)

- When overlap poor, BCM expected to outperform TMLE (matching more robust, when misspecified)

- Using appropriate machine learning methods is anticipated to reduce bias compared to using misspecified, fixed parametric models

Simulation design

	Overlap	Confounder-endpoint association	Endpoint distribution
DGP 1	Good	Moderate	Normal
DGP 2	Good	Strong	Normal
DGP 3	Poor	Strong	Normal
DGP 4	Poor	Strong	Gamma
DGP 5	Poor	Strong	Semi-continuous

Simulation design

- Confounders W_1-W_8, correlated, binary and normal

- **PS** : $\log\left(\frac{PS}{1-PS}\right) = \alpha_0 + k_1(\boldsymbol{\alpha}\mathbf{W} + \alpha_6 W_6^{\,2} + \alpha_7 W_7^{\,2} + \alpha_8 W_8^{\,2} + \alpha_9 W_1 W_2 + \alpha_{10} W_1 W_3)$

- **Endpoint**:

DGP1-3 normal

$$Y \sim N(\mu_N, 1)$$

$$\mu_N = \beta_0 + \psi A + k_2(\boldsymbol{\beta}\mathbf{W} + \beta_6 W_6^{\,2} + \beta_7 W_7^{\,2} + \beta_8 W_8^{\,2} + \beta_{10} W_6^{\,3} + \beta_{11} W_7^{\,3} + \beta_{12} W_8^{\,3} + \beta_{13} W_1 W_2 + \beta_{10} W_1 W_7)$$

DGP4 gamma

$$Y \sim Gamma\left(100, \frac{\mu_G}{100}\right)$$

$$\log(\mu_G) = \beta_0 + \psi A + \boldsymbol{\beta}\mathbf{W} + \beta_6 W_6^{\,2} + \beta_7 W_7^{\,2} + \beta_8 W_8^{\,2} + \beta_{10} W_6^{\,3} + \beta_{11} W_7^{\,3} + \beta_{12} W_8^{\,3} + \beta_{13} W_1 W_2 + \beta_{10} W_6 W_7)$$

DGP5 Semi-continuous (mixture of beta distributed 1-Y` and 1)

Correct specification (model follows DGP)

Misspecification (main terms logistic regression, OLS)

Simulation design

- Scenarios:
  - (a) g and Q correctly specified
  - (b) Q correct, g misspecified
  - (c) Q misspecified, g correct
  - (d) Both g and Q misspecified
    - (d1) Misspecified fixed, parametric models
    - (d2) Machine learning estimation

Fixed, parametric models

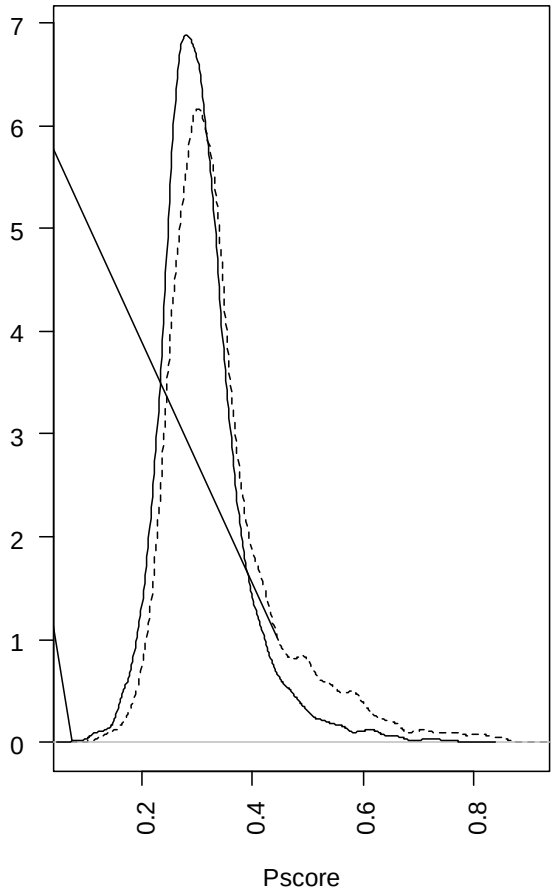
- 1000 datasets of n = 1000

- Measures of performance: relative bias, variance, root mean squared error (RMSE), 95% CI coverage

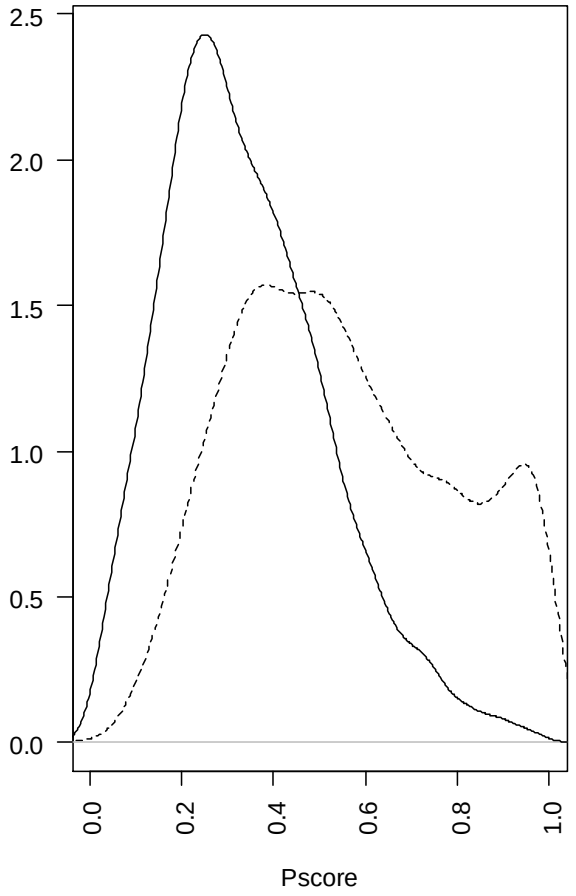
Simulation design

Distributions of true PS

Good overlap (DGP 1 & 2)



Poor overlap (DGP 3 to 5)



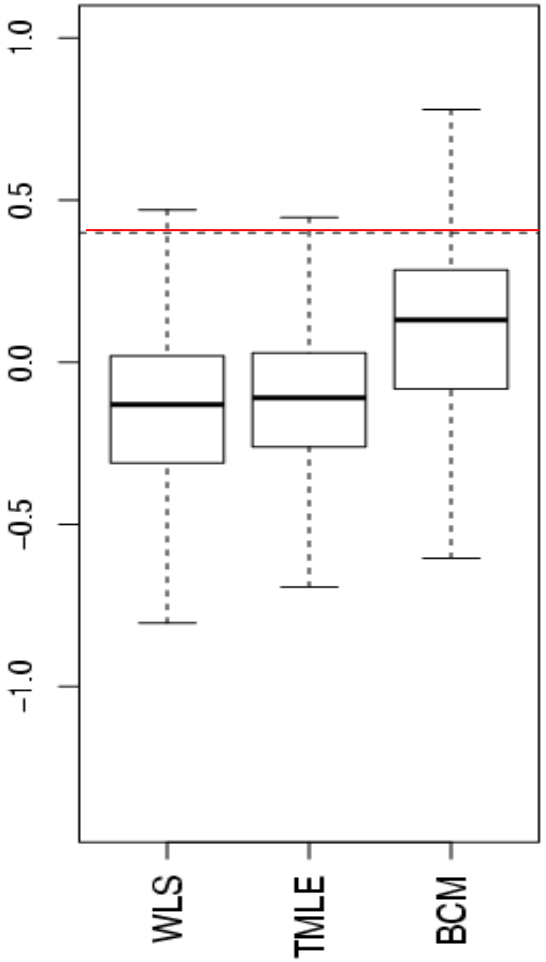
Simulation results

- DGP 1 – “Ideal scenario”: Normal endpoint, good overlap, moderate confounder-endpoint relationship
  - Scen (a) Close to zero bias
  - Scen (b) and (c) Double robustness for WLS, TMLE and BCM
  - Scen (d1) Bias across methods similar, between 7%- 15%
  - Scen (d2) Bias decreases to 1%-2% for WLS, TMLE and BCM
- DGP 2 – Stronger confounding leads to bias of 30%-50% in (d), machine learning reduces it to 3%-10%

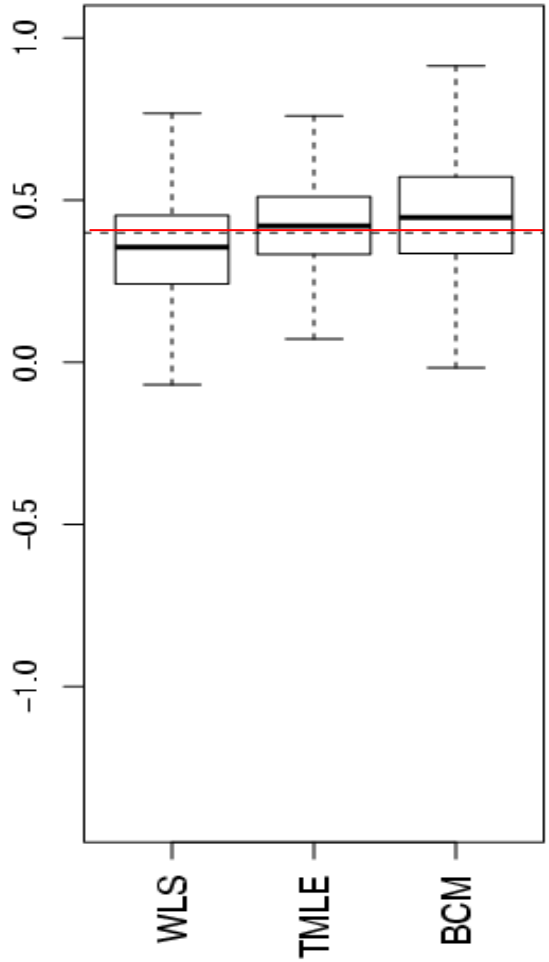
Simulation results, DGP 3

Poor overlap, normal endpoint, Q and g misspecified

(d1) Fixed, parametric models for Q and g



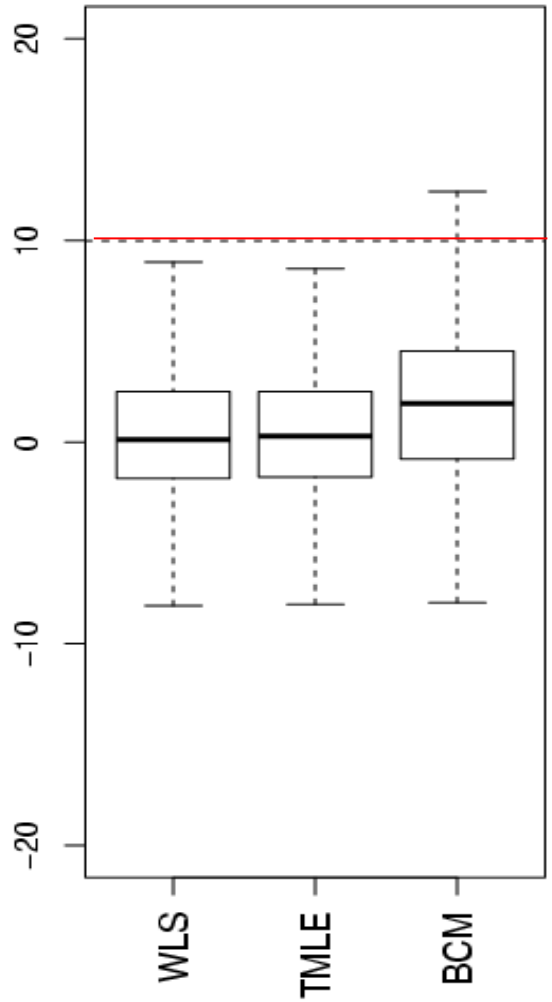
(d2) Machine learning for Q and g



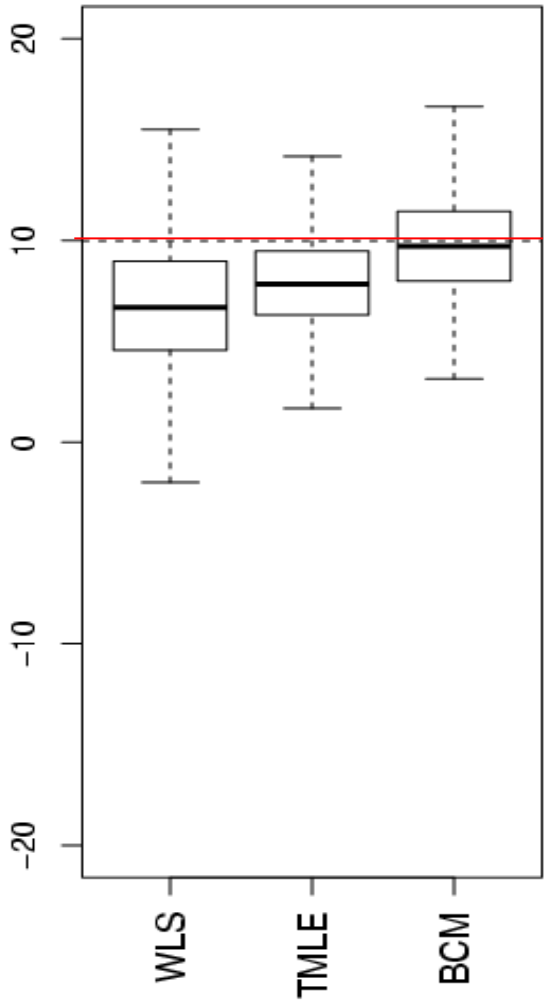
Simulation results, DGP 4

Poor overlap, gamma-distributed endpoint, Q and g misspecified

(d1) Fixed, parametric models for Q and g

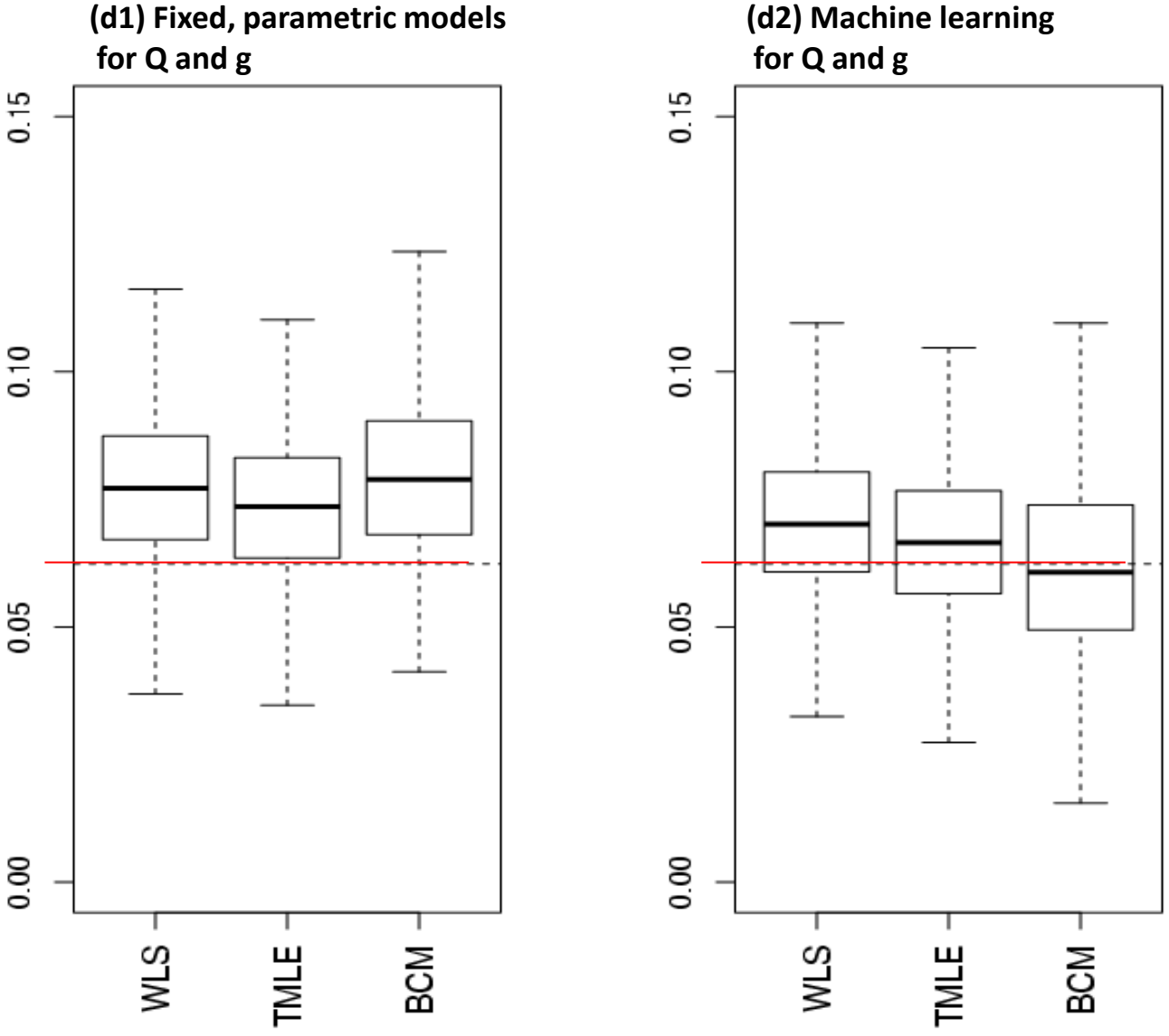


(d2) Machine learning for Q and g



Simulation results, DGP 5

Poor overlap, semi-continuous endpoint, Q and g misspecified



Discussion

- Combining information from the conditional distribution of the endpoint and the PS can reduce bias due to functional form misspecification, when machine learning approaches used
- Bias-variance trade off between BCM and TMLE
- BCM seems more robust in challenging settings of poor overlap and non-normal endpoints
- R code to implement BCM with machine learning part of this work, TMLE user friendly software available
- Main limitation: no unmeasured confounders
- Further research:
  - Alternative machine learning methods for the PS
  - Collaborative maximum likelihood estimation (van der Laan and Gruber, 2010)
    - Optimal choice of variables in the PS, improves TMLE for poor overlap

References

Rubin, D. B. (1973). "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies." Biometrics 29(185-203).

Robins, J., A. Rotnitzky, et al. (1994). " Estimation of regression coefficients when some regressors are not always observed." Journal of the American Statistical Association **89: 846-866.**

Westreich, D. and S. R. Cole (2010). "Invited Commentary: Positivity in Practice." American Journal of Epidemiology 171(6): 674-677.

Petersen, M.L., Porter, K., Gruber, S., Wang, Y., Laan., M.J.v.d.: Diagnosing and Responding to Violations in the Positivity Assumption. U.C. Berkeley Division of Biostatistics Working Paper Series (2010)

Austin, P. C. (2012). "Using Ensemble-Based Methods for Directly Estimating Causal Effects: An Investigation of Tree-Based G-Computation." Multivariate Behavioral Research(47.1): 115-135.

van der Laan, M. J., E. C. Polley, et al. (2007). "Super Learner." Statistical Applications in Genetics and Molecular Biology 6.

Waernbaum, I. (2011). "Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation." Stat Med 31(15): 1572-1581.

Westreich, D., J. Lessler, et al. (2010). "Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression." Journal of Clinical Epidemiology 63(8): 826-833.

References

Lee, B. K., J. Lessler, et al. (2010). "Improving propensity score weighting using machine learning." Statistics in Medicine 29(3): 337-346.

Kang, J. D. Y. and J. L. Schafer (2007). "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data." Statistical Science **22(4): 523-539.**

van der Laan, M. J. and D. Rubin (2006). "Targeted maximum likelihood learning." The International Journal of Biostatistics **2(1).**

Gruber, S. and M. van der Laan (2010). "A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome." Int J Biostat. **6(1).**

Gruber, S. and M. J. van der Laan (2012). "tmle: An R Package for Targeted Maximum Likelihood Estimation." Journal of Statistical Software **51(13): 1-35.**

Abadie, A. and G. W. Imbens (2011). "Bias-Corrected Matching Estimators for Average Treatment Effects." Journal of Business & Economic Statistics **29(1): 1-11.**

Busso, M., J. DiNardo, et al. (2011). New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators. Working paper.