Evaluating baseline variable selection procedures for adjusted marginal treatment effect estimators: Application to Alzheimer's Disease trials

Melody Dehghan, Elizabeth Colantuoni, Michael Rosenblum

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Objectives

At the end of this session, you should be able to

- Explain counterfactual outcomes within the setting of a randomized trial
- Define the marginal or average treatment effect
- Describe how baseline covariate adjustment may improve precision in estimated marginal treatment effects
- Describe what a variable selection procedure is
- Explain the utility of lasso regression

©2021 Johns Hopkins University. All rights reserved.

Randomized trial

Subject	Outcome under Treatment	Outcome under Control	Random assignment A	Observed outcome
1	Y1(1)	Y1(0)	0	Y1 = Y1(0)
2	Y2(1)	Y2(0)	1	Y2 = Y2(1)
…	…	…	…	
n	Yn(1)	Yn(0)	1	Yn = Yn(1)

Individual causal effect: Yi(1) – Yi(0)

Average or marginal treatment effect: E[Yi(1) – Yi(0)]

Unbiased estimate marginal treatment effect: $\theta = \frac{\sum A_i Y_i}{\sum A_i} - \frac{\sum(1-A_i)Y_i}{\sum(1-A_i)}$

©2021 Johns Hopkins University. All rights reserved.

Baseline covariate adjustment

- The goal is to reduce the $Var(\hat{\theta})$
  - Reduce width of confidence intervals
  - Improve power for a fixed sample size
  - Reduce required sample size for a fixed power

- ANCOVA approach for linear outcomes
  - Takes advantage of chance imbalance in variables that are correlated with Y
  - Estimator:

$$Y_i = \beta_0 + \beta_A A_i + \beta_1 X_{1i} + \ \ldots + \beta_p X_{pi} + \varepsilon_i$$

  - Estimated marginal treatment effect: $\hat{\theta} = \hat{\beta}_A$

©2021 Johns Hopkins University. All rights reserved.

Properties of adjusted estimator

- ANCOVA estimator is consistent[1]
    - i.e. unbiased for the marginal treatment effect
    - This holds even if ANCOVA model is incorrectly specified
- We will use this estimator in this talk

- Several alternative estimators with enhanced properties[2]
    - Augmented doubly robust estimator: guaranteed to be as precise or more precise than the unadjusted estimator

1 Yang L, Tsiatis AA. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. The American Statistician. 2001; 55:314–321.
2 Rotnitzky A., Lei Q., Sued M, Robins JM. Improved double-robust estimation in missing data and causal inference models. Biometrika. 2012; 99(2):439–456.

©2021 Johns Hopkins University. All rights reserved.

Statistical Motivation

- Large body of work developing baseline covariate adjusted estimators for the marginal treatment effect

- Several systematic reviews reporting the use of baseline covariate adjustment, or lack there of

- Less guidance about when and how to select baseline variables

©2021 Johns Hopkins University. All rights reserved.

Practical Motivation:

Our work is motivated by the ongoing HOPE4MCI trial

- Treatment: ABG101 vs. placebo
- Outcome: 18-month change in Clinical Dementia Rating-Sum of Boxes score
  - CDR-SB score: higher scores indicate worse cognition
- Designed to detect a 30% reduction in mean outcome
  - Sample size: 160
- Pre-planned analysis: estimate the treatment effect adjusting for 8 baseline variables
  - Baseline variables selected by correlating 18-month change in CDR-SB and variables within the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort

©2021 Johns Hopkins University. All rights reserved.

Objectives:

Compare different procedures for selecting the baseline variables to include in the ANCOVA estimator of the marginal treatment effect in a randomized trial.

In all cases: use a pre-specified candidate variable list and pre-planned variable selection procedure.

Comparison 1: pre-selection (before trial starts) vs. post-selection (using trial data).

Comparison 2: selection procedures: stepwise inclusion, lasso, random forest.

©2021 Johns Hopkins University. All rights reserved.

Notations and Definitions

- We assume a randomized controlled trial where we observe n independent participants, each with data vector (W_i, A_i, Y_i) from an unknown probability distribution P
  - W_i is a m x 1 column vector of baseline variables
  - A_i is the treatment arm indicator (where 1= treatment and 0= placebo)
  - Y_i is a continuous valued outcome

- We assume no missing values and 1:1 randomization
- Target is marginal treatment effect -> defined above
  - ANCOVA estimator

©2021 Johns Hopkins University. All rights reserved.

Baseline variable selection procedures

- Stepwise selection using cross-validated (CV) R^2
- Lasso regression
- Random forest (VSURF)

Default settings were used for lasso and random forest
Lasso R package glmnet
Random forest R package VSURF

©2021 Johns Hopkins University. All rights reserved.

Baseline Variable Selection Procedure: CV-R^2

- Compute the sum of squared residuals for Y based on estimates of the study arm specific mean of Y, sm_1 and sm_0 for the treatment and control arms, respectively

- Fit the adjusted regression for each study arm, ($a \in 0,1$): $Q^a(W,B^{(a)})$ for $E(Y=1|A=a, W)$ where $B^{(a)}$ are the set of association parameters from the study arm specific regression model

- Compute the relative efficiency and approximate reduction in the required sample size

$$\widehat{RE} = \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{sm_{A_i}}\right)^2}{\sum_{i=1}^{n}\left(Y_i - Q^a\left(W, \widehat{B^{A_i}}\right)\right)^2}, \quad \widehat{RR_n} = 1 - \frac{1}{\widehat{RE}}$$

- To avoid being overly optimistic, the estimate of the $\widehat{RE}$ are derived using a leave-one-out cross-validation (CV) procedure

Note: RR_n= (equivalent) relative reduction in sample size

©2021 Johns Hopkins University. All rights reserved.

Baseline Variable Selection Procedure: CV-R^2

The CV-R^2 variable selection procedure is:

- Step 1: Compute RR_n for each of the M baseline variables one at a time & rank from highest to lowest RR_n.
- Step 2: If highest ranked variable has $RR_n <= 1/n$, then use unadjusted
  - Else, do 1 pass over each variable from highest to lowest rank and include each variable that adds at least 1/n to current variable set's RR_n

©2021 Johns Hopkins University. All rights reserved.

Linear Regression Review

- Simple linear regression
  - $Y = X\beta + \epsilon; \epsilon \sim N(0, \sigma^2)$

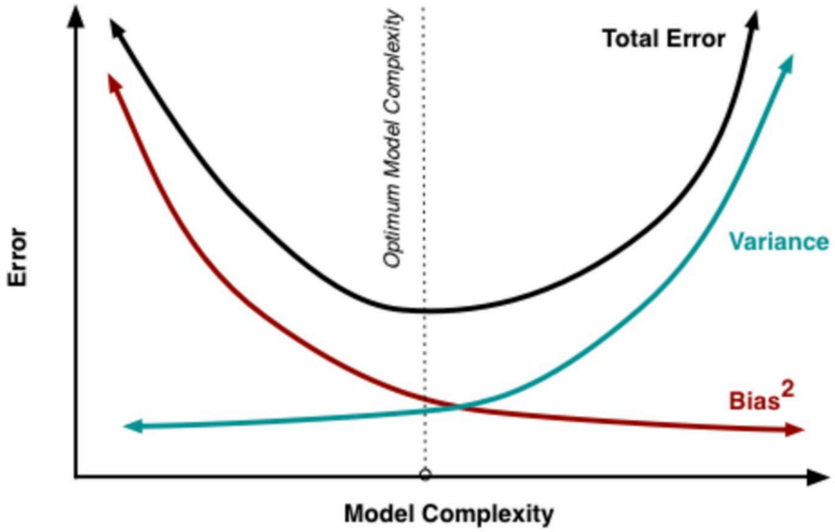
- Ordinary Least Squares (OLS): minimize the loss function and obtaining the OLS estimates, $\widehat{\beta_{OLS}} = (X'X)^{-1}(X'Y)$
  - $L_{OLS}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 = \left|\left|y - X\hat{\beta}\right|\right|^2$

- OLS procedure yields unbiased estimates of β with variance given by
  - $Variance(\widehat{\beta_{OLS}}) = \sigma^2(X'X)^{-1}$

©2021 Johns Hopkins University. All rights reserved.

Regularization: Ridge, Lasso, Elastic Net

- Sometimes due to OLS estimator's unbiased property, there can be large variances
    - Highly correlated predictor variables
    - Many predictors, e.g. more predictors than sample size



- Solution: reduce variance at the cost of introducing bias
    - Unbiased OLS: right side of the picture
    - Regularization: move left towards the optimum (lowest MSE)

©2021 Johns Hopkins University. All rights reserved.

Regularization: Lasso

- Lasso penalizes the sum of the absolute values of coefficients (L1 penalty)
  - $L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(yi - x_i'\hat{\beta})^2 + \lambda \sum_{j=1}^{m}|\widehat{\beta_j}|$
  - λ = regularization penalty
    - At λ = 0, $\hat{\beta}_{lasso} = \hat{\beta}_{OLS}$
    - As λ → ∞, $\hat{\beta}_{lasso}$ → 0, bias increases, variance decreases
    - As λ → -∞, variance increases

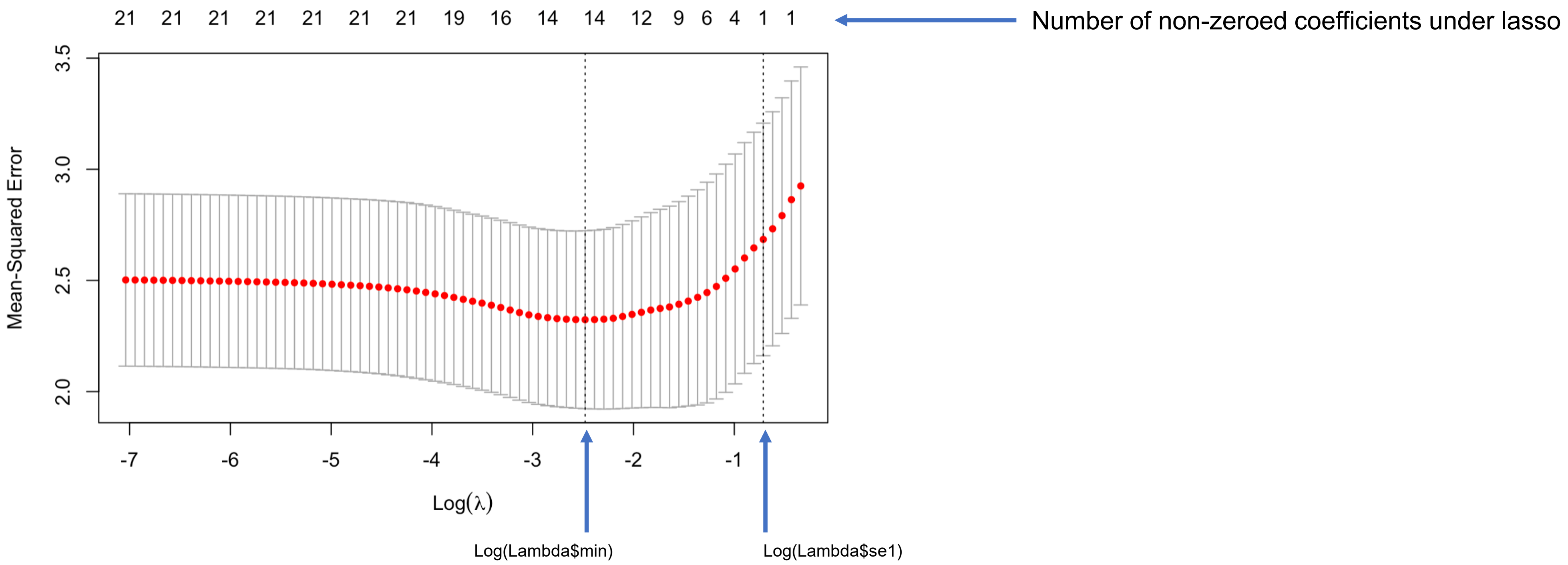
- Can choose an optimal value of λ based on AIC, BIC, cross-validation…
- For lasso, high values of λ → coefficients are zeroed under lasso

©2021 Johns Hopkins University. All rights reserved.

Lasso Regression with ADNI

- For selecting variables within ADNI, λ = 10-fold cross-validation

```r
lasso_cv <- cv.glmnet(x_subset, y_subset, alpha = 1, standardize = TRUE, nfolds = 10)
plot(lasso_cv)
```



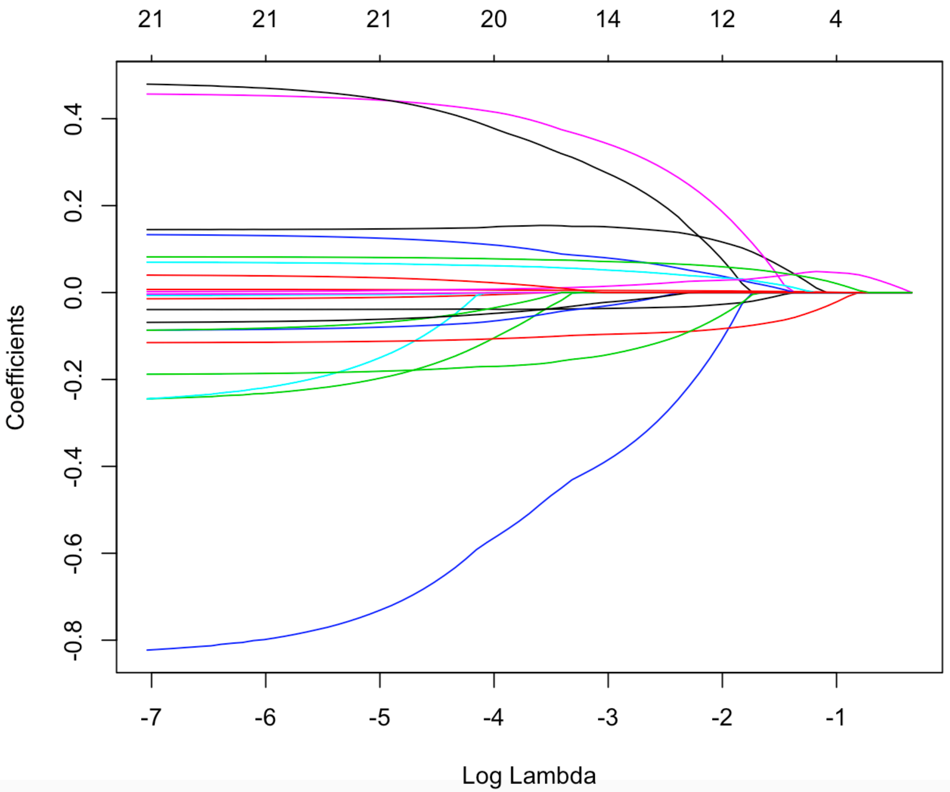
- Lambda$min= minimum MSE of the cross validation; lambda$se1= adding one standard error to the minimum MSE
  - Choosing lambda$se1= a more regularized model

©2021 Johns Hopkins University. All rights reserved.

Lasso Regression with ADNI

- Each line = coefficient for one variable for different λ

```
res <- glmnet(x_subset, y_subset, alpha = 1, standardize = TRUE)
plot(res, xvar = "lambda")
```



- Higher the λ, more coefficients shrink towards 0

©2021 Johns Hopkins University. All rights reserved.

Lasso Regression with ADNI

Lasso Regression

```r
model_cv <- glmnet(x_subset, y_subset, alpha = 1, lambda = lambda_cv, standardize = TRUE)
coef(model_cv)
```

```
22 x 1 sparse Matrix of class "dgCMatrix"
                         s0
(Intercept)    -0.993496382
CDRSB_Baseline  0.140912435
AGE             .
female          .
married        -0.272692875
divorced        .
APOE4           0.279358639
MMSE           -0.033885148
LDELTOTAL      -0.091582197
HMSCORE        -0.108866457
GDTOTAL        -0.006089670
CATANIMSC       .
TRAASCOR        .
TRAAERRCOM      0.192696345
TRABSCOR        0.004391774
TRABERRCOM      .
TRABERROM       0.060564748
ADAS11          0.044939544
ADAS13          0.020888725
AVDEL30MIN     -0.007570606
AVDELTOT        .
FAQ             0.066750085
```

. = coefficient zeroed under lasso regularization → variables not chosen by lasso

Linear Regression

```
Call:
lm(formula = y_subset ~ x_subset)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3145 -0.9793 -0.0817  0.9942  4.5712 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)   
(Intercept)             0.428367   2.559277   0.167  0.86729   
x_subsetCDRSB_Baseline  0.144369   0.204035   0.708  0.48025   
x_subsetAGE            -0.014871   0.019175  -0.776  0.43920   
x_subsetfemale         -0.252555   0.270310  -0.934  0.35157   
x_subsetmarried        -0.837803   0.390902  -2.143  0.03362 * 
x_subsetdivorced       -0.259520   0.544787  -0.476  0.63447   
x_subsetAPOE4           0.459072   0.166838   2.752  0.00662 **
x_subsetMMSE           -0.039050   0.068477  -0.570  0.56931   
x_subsetLDELTOTAL      -0.115610   0.048418  -2.388  0.01813 * 
x_subsetHMSCORE        -0.189024   0.172678  -1.095  0.27533   
x_subsetGDTOTAL        -0.087584   0.091038  -0.962  0.33749   
x_subsetCATANIMSC      -0.006609   0.026475  -0.250  0.80319   
x_subsetTRAASCOR       -0.003625   0.006355  -0.570  0.56918   
x_subsetTRAAERRCOM      0.485209   0.269680   1.799  0.07390 . 
x_subsetTRABSCOR        0.007270   0.002328   3.123  0.00213 **
x_subsetTRABERRCOM     -0.089421   0.087110  -1.027  0.30621   
x_subsetTRABERROM       0.134526   0.080138   1.679  0.09519 . 
x_subsetADAS11          0.070451   0.073167   0.963  0.33708   
x_subsetADAS13          0.001489   0.057829   0.026  0.97949   
x_subsetAVDEL30MIN     -0.069791   0.050836  -1.373  0.17173   
x_subsetAVDELTOT        0.041019   0.037116   1.105  0.27077   
x_subsetFAQ             0.082430   0.036522   2.257  0.02538 * 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.419 on 158 degrees of freedom
Multiple R-squared:  0.3866,	Adjusted R-squared:  0.305 
F-statistic: 4.741 on 21 and 158 DF,  p-value: 4.334e-09
```


©2021 Johns Hopkins University. All rights reserved.

VSURF

- VSURF (random forest) is a two-step algorithm- first step involves ranking the variables according to variable importance (VI) and eliminating the unimportant ones
- Step 1- Algorithm first ranks the variables by an averaged VI over about 50 forests. Variable elimination is decided by variable importance of X^j
  - $VI(X^j) = \frac{1}{ntree} \sum_t \left(err\widehat{OOB}_t^j{}_t - errOOB_t \right)$
  - $errOOB_t^j$ = error for a single tree t in the *OOBt* sample
  - Error= MSE for the regression and misclassification rate for classification
  - Next, randomly permute the values of X^j in the OOB_t and get a perturbed sample, OOB_t
  - $err\widehat{OOB}_t$= error of the predictor t on this perturbed sample

©2021 Johns Hopkins University. All rights reserved.

VSURF

- When comparing a prognostic variable to a useless variable, the prognostic variable will have a larger variability of VI across repetitions of random forests
- Threshold value in step 1 is therefore decided by the standard deviation of the VI of the useless variables
    - Specifically, the minimum prediction value by a CART model in which the X are the variable ranks and the Y are the standard deviations of the Vis
    - The algorithm selects variables with an averaged VI greater than the threshold

©2021 Johns Hopkins University. All rights reserved.

VSURF

- Step 2- method selects two variable subsets

1. Interpretation- includes all variables highly correlated with the outcome
    - Algorithm selects models with the smallest OOB error from nested collections of random forests

2. Prediction- more limited subset, only includes the smallest subset of variables with very low redundancy
    - Algorithm takes the variables chosen in the interpretation step and tests the variables in a stepwise sequence of random forest models
    - Ideally, a predictive variable's OOB error decrease should much greater than the average variation of noisy variables

- We use the subset of variables for prediction for variable selection in each simulated dataset

©2021 Johns Hopkins University. All rights reserved.

Simulation studies

- Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal multicenter cohort study with observational data from cognitively normal older adults, patients with MCI and patients with mild to moderate AD dementia

- Two curated datasets selected to reflect the inclusion/exclusion criteria for the HOPE4MCI trial
  - Dataset 1: Y and W weakly correlated
  - Dataset 2: Y and W strongly correlated

©2021 Johns Hopkins University. All rights reserved.

Simulation studies: pre-select variables

- Basic idea:
  - Pre-select variables using an ADNI dataset
  - Generate hypothetical trials
    - Resampled with replacement n = 160 participants from ADNI dataset
    - Estimate marginal treatment effect after adjusting for pre-selected variables

  - Considered several scenarios
    - Prognostic baseline variables with positive treatment effect
    - No prognostic baseline variables (scramble Y) with positive treatment effect

©2021 Johns Hopkins University. All rights reserved.

Simulation studies: post-select variables

- Basic idea:
  - Pre-specify M baseline variables
  - Generate hypothetical trials
    - Resampled with replacement n = 160 participants from ADNI dataset
  - Select m < M variables using hypothetical trial data

- Two scenarios:
  - Resample data from cohort 1: Y and W are weakly correlated
    Resample data from cohort 2: Y and W are strongly correlated

©2021 Johns Hopkins University. All rights reserved.

Marginal Treatment Effect Estimators

- We estimated the marginal treatment effect using the unadjusted estimator and 5 ANCOVA estimators adjusting for:
  1. Baseline CDR-SB score only
  2. All M candidate prognostic baseline variables

  Adjust for the $m \leq M$ baseline variables pre- or post-selected via:
  1. CV-R^2 procedure
  2. Lasso regression procedure
  3. RF (VSURF) procedure

©2021 Johns Hopkins University. All rights reserved.

Simulation metrics: in one scenario

- Bias: mean of the ANCOVA estimators over all hypothetical trials – true treatment effect

- Variance: variance of all ANCOVA estimators over all hypothetical trials

- Mean squared error: variance + bias^2

- Estimated reduction in required sample size when using the adjusted estimator compared to the unadjusted estimator with fixed power
  - Relative efficiency: MSE(unadjusted) / MSE(adjusted)
  - Reduction in required sample size: 1 -1/relative efficiency

©2021 Johns Hopkins University. All rights reserved.

Results: pre-selecting baseline variables

Scenario		Prognostic Variables Bias	Var[1]	MSE[2]	100*RR_n[3]	No Prognostic Variables Bias	Var[1]	MSE[2]	100*RR_n[3]
Cohort 1	Unadj	-0.001	0.062	0.062	0.0	-0.006	0.061	0.061	0.0
	Y_0	-0.001	0.062	0.062	-0.2	-0.006	0.061	0.061	-0.7
	All Cov	0.005	0.054	0.054	13.3	-0.006	0.071	0.071	-16.6
	CV-R^2	0.002	0.052	0.052	16.0	-0.006	0.061	0.061	-0.5
	Lasso	0.004	0.052	0.052	17.0	-0.006	0.061	0.061	0.0
	RF	0.003	0.055	0.055	12.3	-0.006	0.063	0.063	-4.3
Cohort 2	Unadj	0.005	0.055	0.055	0.0	-0.001	0.055	0.055	0.0
	Y_0	0.005	0.051	0.051	7.1	-0.001	0.055	0.055	-0.2
	All Cov	0.011	0.040	0.040	27.4	-0.001	0.064	0.064	-16.1
	CV-R^2	0.006	0.039	0.039	28.8	-0.001	0.056	0.056	-1.5
	Lasso	0.010	0.038	0.038	29.7	-0.001	0.055	0.055	0.0
	RF	0.007	0.046	0.046	15.5	-0.001	0.056	0.056	-2.4

[1] Var corresponds to variance

[2] MSE corresponds to mean squared error= $variance + bias^2$

[3] RR_n corresponds to the relative reduction in required sample size comparing the adjusted estimator to the unadjusted estimator, 1 - MSE(adjusted) / MSE(unadjusted)

©2021 Johns Hopkins University. All rights reserved.

Results: post-selecting baseline variables

Scenario		Prognostic Variables				No Prognostic Variables			
		Bias	Var[1]	MSE[2]	100*RR_n[3]	Bias	Var[1]	MSE[2]	100*RR_n[3]
Cohort 1	Unadj	-0.001	0.062	0.062	0.0	-0.006	0.061	0.061	0.0
	Y_0	-0.001	0.062	0.062	-0.2	-0.006	0.061	0.061	-0.7
	All Cov	0.005	0.054	0.054	13.3	-0.006	0.071	0.071	-16.6
	CV-R^2	0.004	0.055	0.055	12.5	-0.006	0.065	0.065	-6.7
	Lasso	0.022	0.050	0.050	19.4	0.002	0.060	0.060	1.1
	RF	0.005	0.054	0.055	12.5	-0.006	0.062	0.062	-1.8
Cohort 2	Unadj	0.005	0.055	0.055	0.0	-0.001	0.055	0.055	0.0
	Y_0	0.005	0.051	0.051	7.1	-0.001	0.055	0.055	-0.2
	All Cov	0.011	0.040	0.040	27.4	-0.001	0.064	0.064	-16.1
	CV-R^2	0.008	0.040	0.040	26.5	-0.001	0.059	0.059	-7.7
	Lasso	0.028	0.037	0.038	30.8	0.011	0.055	0.055	-0.2
	RF	0.010	0.044	0.045	18.6	-0.000	0.056	0.056	-1.5

[1] Var corresponds to variance

[2] MSE corresponds to mean squared error= $variance + bias^2$

[3] RR_n corresponds to the relative reduction in required sample size comparing the adjusted estimator to the unadjusted estimator, 1 - MSE(adjusted) / MSE(unadjusted)

©2021 Johns Hopkins University. All rights reserved.

Simulation Study Results

- Regardless of when and which baseline variables are selected, all estimators have similar and small bias and produced roughly 95% coverage of the marginal treatment effect, with coverage ranging from 92.8% to 95.3%
- Similar performance of pre- and post-selecting when data generating distributions are identical

©2021 Johns Hopkins University. All rights reserved.

Simulation Study Results

- Adjusting for all candidate prognostic baseline variables resulted in the largest precision loss when baseline variables are not prognostic (-17.3% to -16.1%) and performed similarly to the three variable selection procedures when baseline variables are prognostic
- The lasso procedure resulted in
  - The largest precision gains (15.5% to 34.8%) under prognostic baseline variables
  - The lasso procedure resulted in the smallest precision loss (-0.8% to 1.1%) under no prognostic baseline variables

©2021 Johns Hopkins University. All rights reserved.

Conclusions and future work

- Post-selecting baseline variables using the lasso procedure resulted in the largest precision gains in all scenarios with prognostic baseline variables, and no loss when baseline variables are not prognostic

- The baseline variable selection procedure should be pre-planned, including when and how baseline variables are selected

- Trialists must also decide whether to assume efficiency gains from covariate adjustment and set sample size accordingly
  - It is not guaranteed that the strength of the correlation between the W and Y will be similar in the trial planning data vs. the actual trial

- Future work includes-
  - To prove theoretical results on the asymptotic of such a procedure
  - A similar evaluation using different outcome types, e.g. binary, survival endpoints

©2021 Johns Hopkins University. All rights reserved.