Synthetic Data Simulation Results

Alan Hubbard

2024-04-08

Simulation Code

```
generate_data_simple <- function(N,Xranges=c(-1,1,-1,1),betaA=c(0,0.1,-0.4),
                                 betaY0=c(0,1,2,-1),betaC=c(1,7/5,5,3),sdy=1){
  # A simple data generating process
  X1 <- runif(N,Xranges[1],Xranges[2])
  X2 <- runif(N,Xranges[3],Xranges[4])
  pi0 <- plogis(betaA[1]+betaA[2]*X1*X2+betaA[3]*X1)
  A <- rbinom(N,1,prob=pi0)
  muY0 <- betaY0[1]+betaY0[2]*X1*X2 + betaY0[3]*X2^2 +betaY0[4]*X1
  CATE <- betaC[1]*X1^2*(X1+betaC[2]) + (betaC[3]*X2/betaC[4])^2
  muY = muY0+A*CATE
  Y <- rnorm(N,sd=sdy,mean= muY)
  return(tibble(X1=X1,X2=X2,A=A,Y=Y))
}
```

Synthetic data simulation details

We have created a structure to

__1. Simulate complex data from known distributions

```
_a. Can be cross-sectional,

_b. longitudinal,

_c. have missing data, etc.
```

__2. Define a set of parameters we wish to estimate and derive inference from the synthetic data

```
_a. Marginal parameters, like means

_b. Regression estimates for working models

_c. Causal parameters based on known causal model
```

Simulation Structure

Investigate the performance of competing methods for synthetic data by:

_1. Simulate the data.

_2. Estimate the data-generating distribution using different methods (HAL, SuperLearner, etc.): looking at additional methods.

_3. Synthethesize the data from the estimated DGDs based on competing methods.

_4. Repeat 1-3 1000 times.

_5. Evaluate the comparison of the distribution of synthetic data-based results to those based on the actual data.

X-Sectional Data Example

- Data: $O = (W, A, Y)$
- Causal Model: $W \rightarrow A \rightarrow Y$
- $W_1, W_2 \sim Uniform$
- Complex DGD
  - $logit(P(A = 1|W)) = \alpha_0 + \alpha_1 * A + \alpha_2 * A * W_1 * W_2 + \alpha_3 * W_1$
  - $E(Y_0|W) = \beta_0 + \beta_1 * W_1 * W_2 + \beta_2 * W_2^2 + \beta_3 * W_1$
  - $E(Y_1|W) - E(Y_0|W) = \gamma_0 + \gamma_1 * W_1^2 * (W_1 + \gamma_2) + \gamma_3 * W_2^2$
  - $Y = E(Y_0|W) + A * (E(Y_1|W) - E(Y_0|W)) + e, e \sim N(0, \sigma)$

Parameters of Interest

- Simple ones
  - $P(A = 1)$
  - EY
- Standard Regression parameters
  - Coefficients in working model: $b_0 + b_1 * W_1 + b_2 * W_2 + b_3 * A$
- Causal parmaters estimated with (targeted) machine learning
  - $ATE = E(E(Y|A = 1, W) - E(Y|A = 0, W))$

Methods for generating Synthetic Data

- Undersmoothed Highly adaptive lasso (HAL) - undersmoothing based on ATE
- SuperLearner
- Compare estimates using synthetic data to those based on the data behind synthetic data

Estimators used to get parameters from synthetic (and actual) data

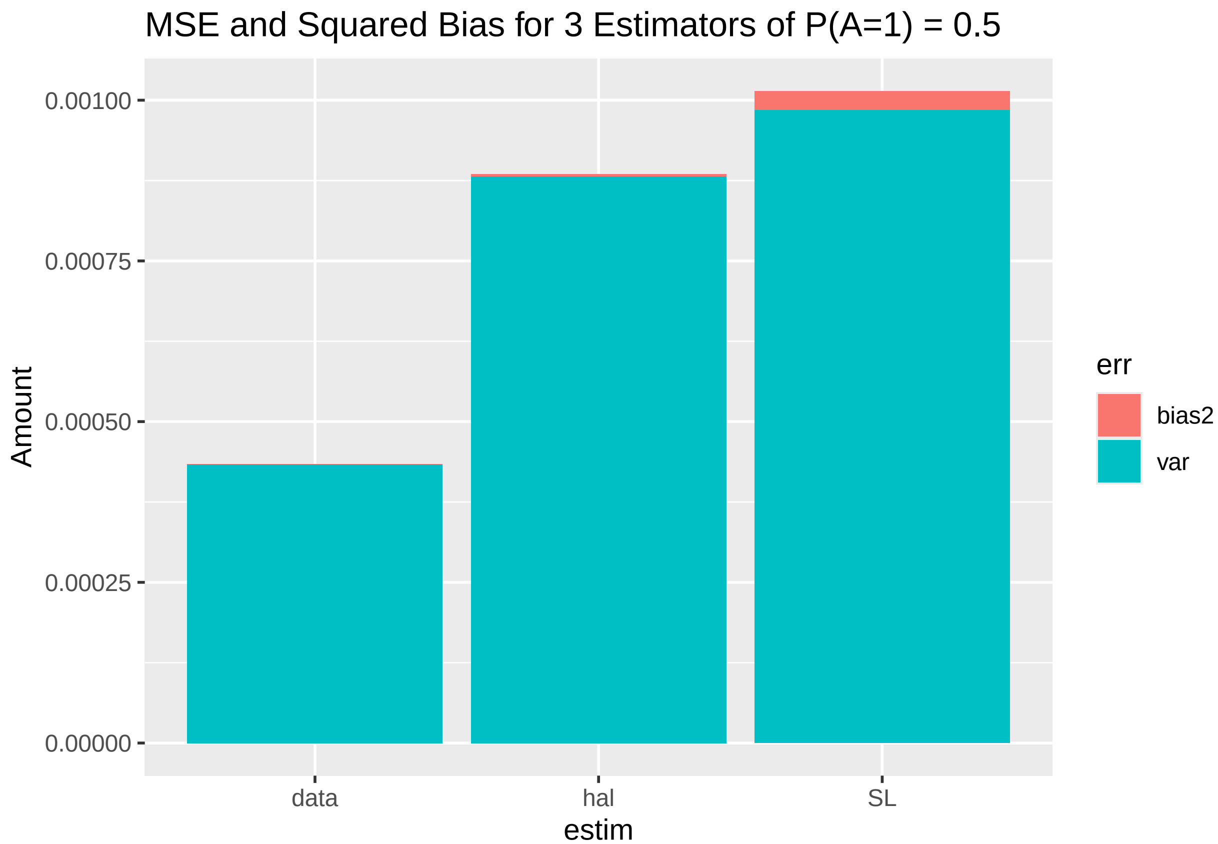
- Simple parameters - simple averages
- Coefficients - ordinary least squares
- ATE - TMLE with SL
- Evaluations: bias, variance, MSE and coverage

Straightforward issues with inference from Synthetic Data

- If the methods used to estimate the parameters that define the model for the DGD are unbiased, synthetic data based on them will not be.
- For example, if my model is $X \sim N(\mu, \sigma)$
- Want to synthesize data based on a sample $X_i, i = 1, ..., n$
- Generate synthetic data from $N(\bar{X}, \sigma) : X_i^*, i = 1, ...n$
- Estimate μ with $\bar{X}^*$
- MSE of estimate based on synthetic data is: $MSE(\bar{X}^*) = E(\bar{X}^* - \bar{X})^2 + E(\bar{X} - \mu)^2$
- To get good coverage in this simple conrext, would increase the margin of error in confidence intervals by a factor of $\sqrt{2} : \hat{\theta} \pm \sqrt{2} * 1.96 * SE(\hat{\theta})$.
- Could also generate a synthetic data sample that is very large relative to the size of the original data to get estimates, but based inference (calculate standard erros) on a synthetic sample the same size as the original data.

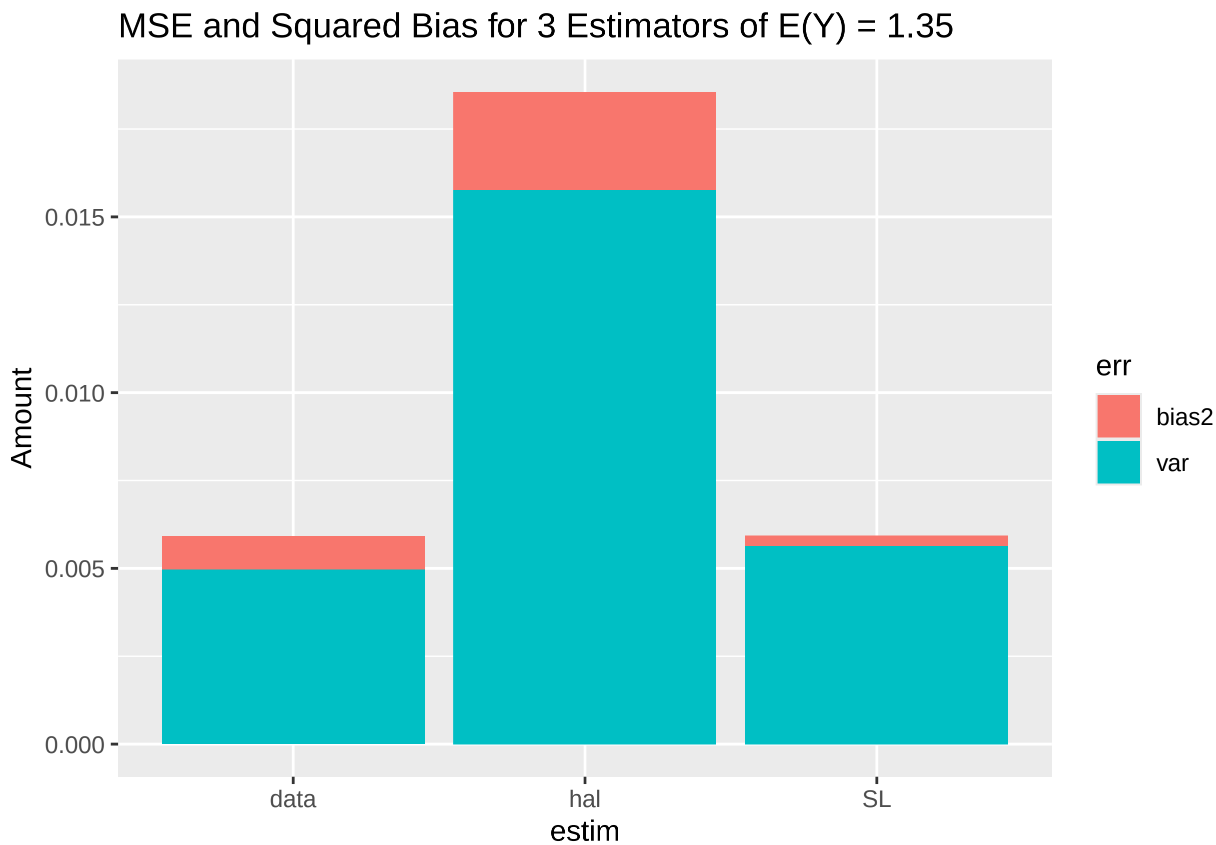
Results

$P(A = 1)$

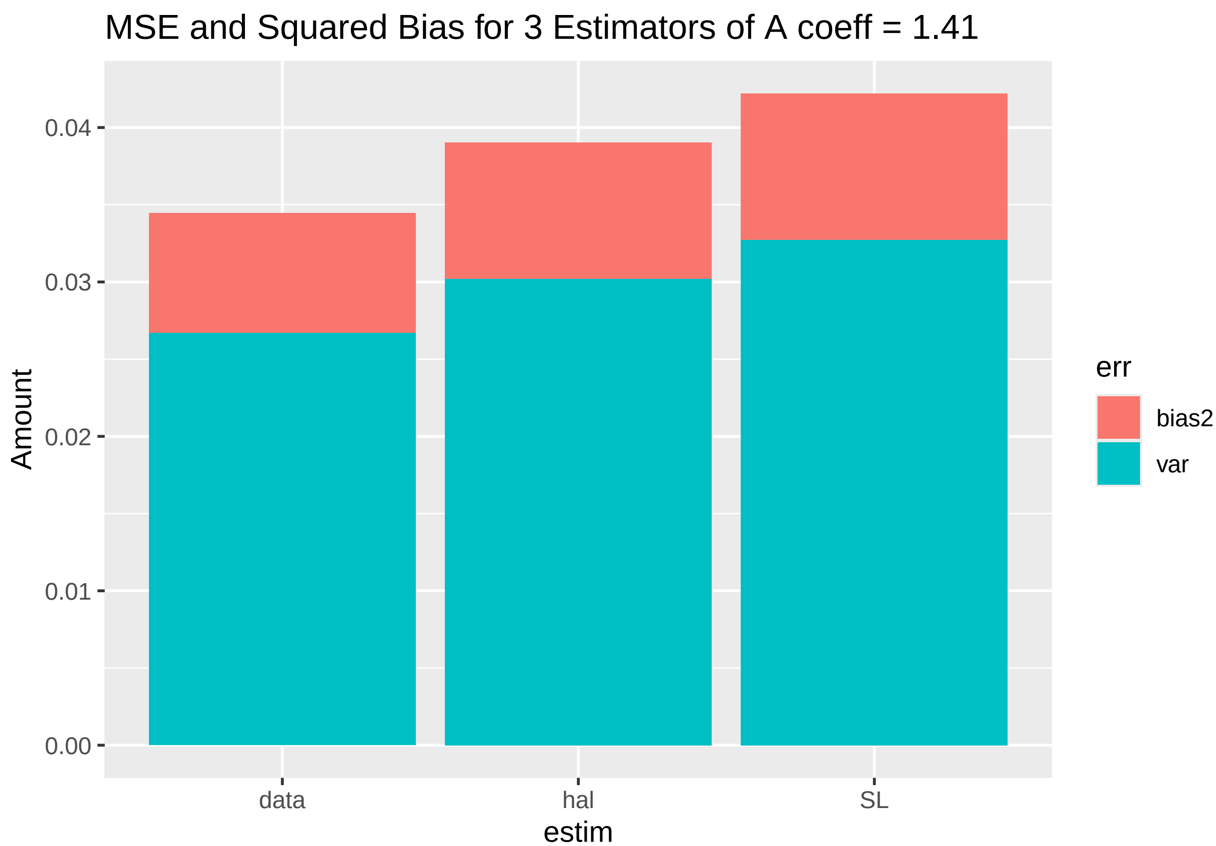


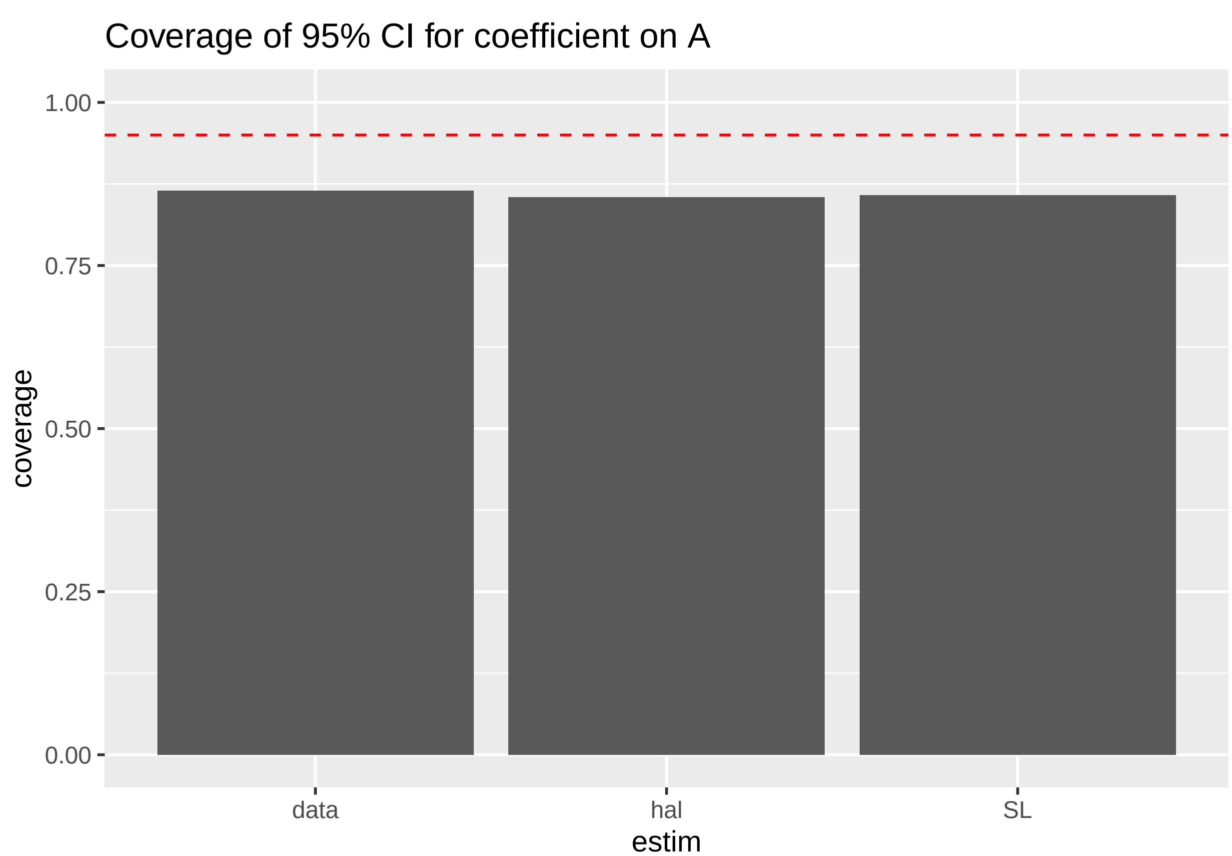
MSE and Squared Bias for 3 Estimators of P(A=1) = 0.5
Amount
0.00100
0.00075
0.00050
0.00025
0.00000
data
hal
SL
estim
err
bias2
var

EY

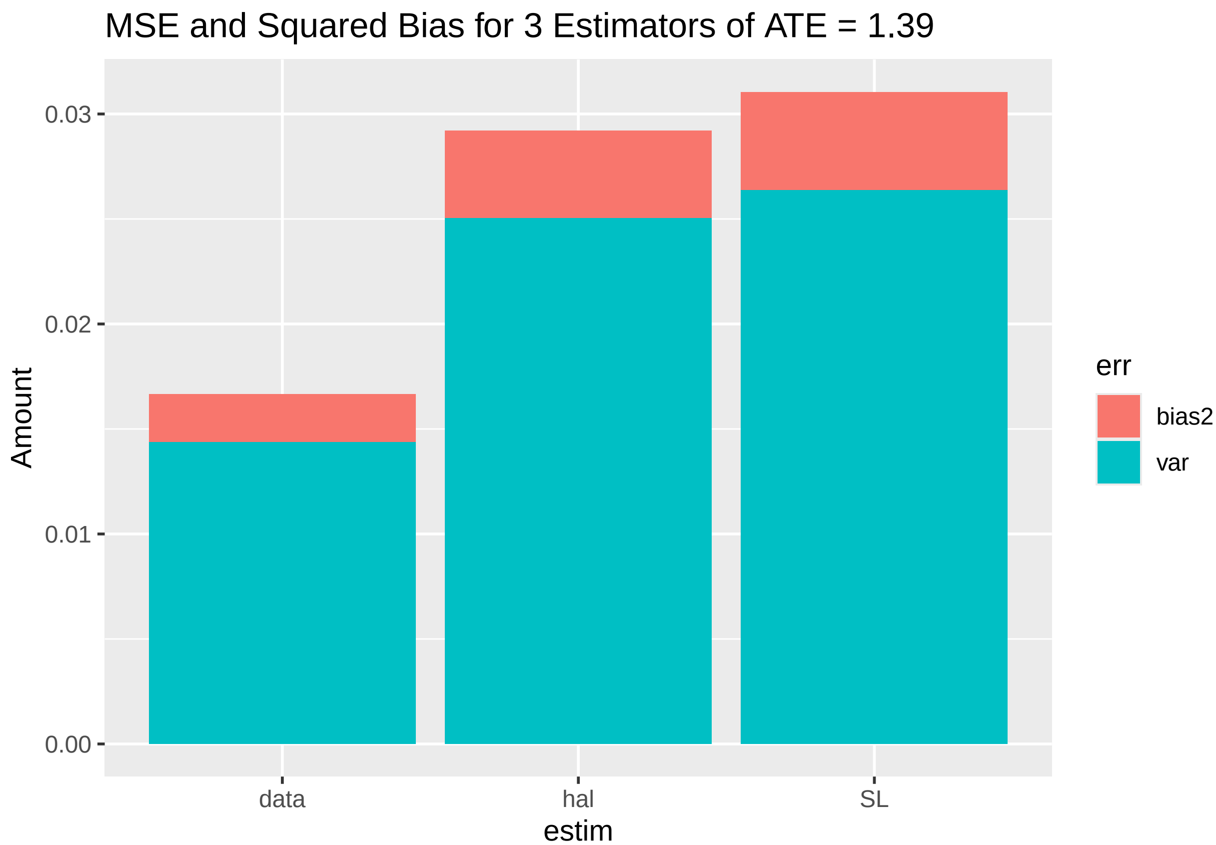

MSE and Squared Bias for 3 Estimators of E(Y) = 1.35
Amount
0.015
0.010
0.005
0.000
data
hal
SL
estim
err
bias2
var

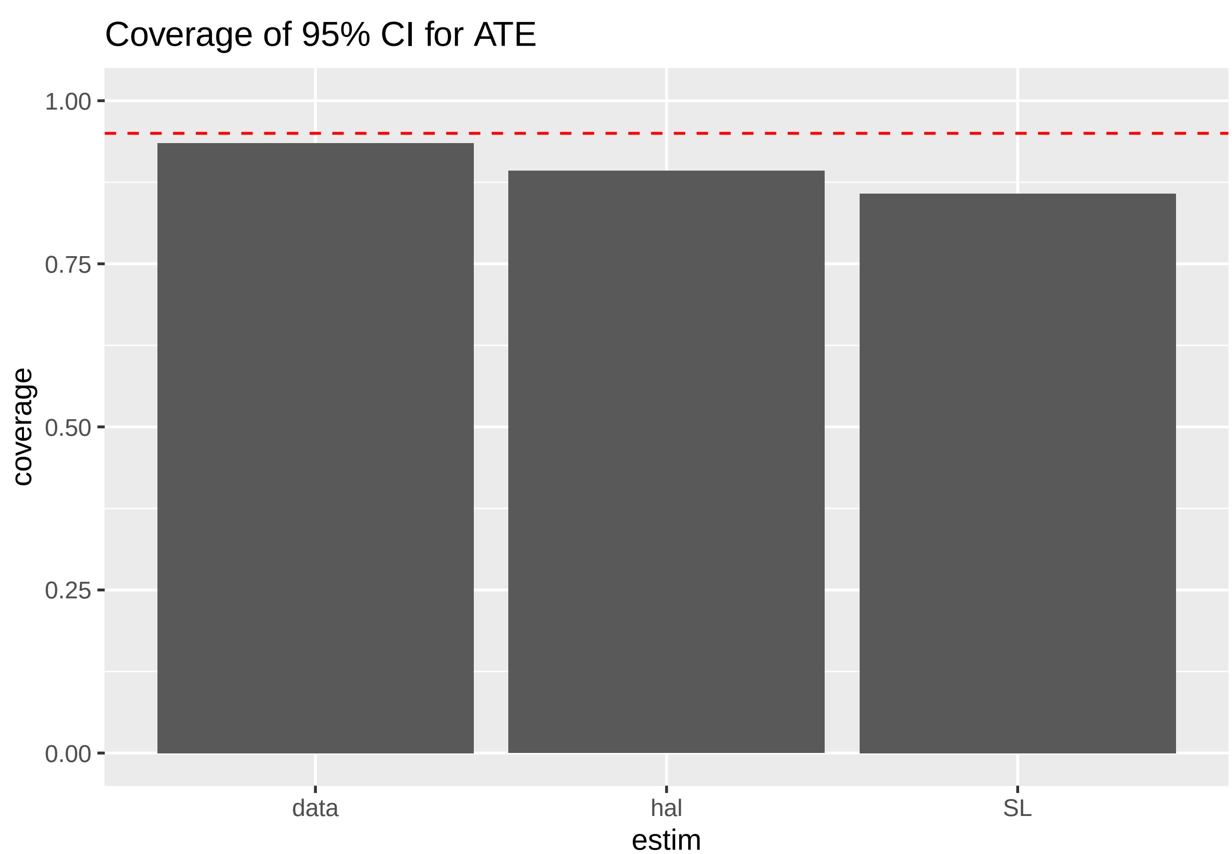
Regression coefficients from working model





ATE


MSE and Squared Bias for 3 Estimators of ATE = 1.39
Amount
0.03
0.02
0.01
0.00
data
hal
SL
estim
err
bias2
var


Coverage of 95% CI for ATE
coverage
1.00
0.75
0.50
0.25
0.00
data
hal
SL
estim

Other work in progress

- More complex simulations that mimic N3C data (e.g., missing data, more complex models based on long Covid N3C data)
- Methods for obscuring baseline covariates for privacy
  - Estimating joint distribution of discretized baseline covariates
  - Apply required coarsening of identifying variables (age, etc.)
- Comparisons to existing synthetic data algorithms available (usually very similar in framework, but use more arbitrary modeling assumptions)
- Comparisons to CV-HAL
- Develop general algorithm that works directly on HAL objects and users can querry parameters by providing functions which operate on HAL objects that contain DGD components.
- Exploring generative AI but skeptical they are applicable to sample sizes of data based on N3C without having a pre-trained model on similar data.