R Lab 4:
Super Learner!

topics:

- SL
- SL
- SL
- SL
- R HW 3 hints

IMPT REMINDERS

- Josh: remember to record this
- R HW 4 is posted soon (already?)
- Use TEMPLATE for R HW 4 so you won't miss any Qs
- Solutions to this lab will be posted soon too; R code will be very helpful

- Remember to **`install.packages("SuperLearner")`** for this lab / HW (now??)

R Lab 4 / HW 4

WARNING!!!!!

A lot less writing, a bit more code.

If you have struggled with coding in previous HWs…

START THIS ONE EARLY AND COME TO OFFICE HOURS!!!!!!11!!11!

As usual, use the example code in R Lab 4 answers to help with HW 4.

SL: Motivation

- Previous discussion of g-comp formula (R Lab 2, R HW 2) noted that for it to work well, you need a correctly-specified outcome regression (think “Regression Model 4” from R HW 2, for example - had lowest bias and MSE). Same for IPTW.

- But how do we know how to specify? What interactions to choose? What functional form? Squared terms? Link functions? etc…… Incorrect specification leads to bias.

- Trad Stats Ppl typically think in terms of GLMs, but machine learning offers many more prediction mechanisms that might be employed.

- GOAL: find a way of creating outcome predictions without testing tons of model options (garden of forking paths, etc)…

“Trying a bunch of regressions, looking at the results, fiddling with the specifications, and choosing the ‘best’ (using arbitrary criteria) leads to biased point estimates and misleading inference.”

SL: Big ideas

- Fit many competing algorithms, using CV to test predictions
- “Risk” measures MSE (or some other loss function, like cross-entropy/LL)
- Pick algorithm with lowest risk (“Discrete” SL) or…
- Ensemble SL: Create a weighted combination (ensemble) of all the algorithms
- … use this to predict Y(0) and Y(1) to estimate ATE

…(*****in a special way with TMLE, stay tuned*****)

Lab 4 Tasks

- For some data from an unknown DGP…
  - Split the data into 10 folds
  - Fit the models 9/10 of the data, excluding fold #1
  - Test the predictions from the model on the left-out 1/10 of the data
  - Record the error in these predictions, repeat for all the other folds #2 - #10
  - “Risk” is the average MSE across all 10 iterations
- Compare the different models by their risk
- Finally, protect against overfitting by repeating this process over a set of SuperLearners with CV.SuperLearner

Lab 4 Tasks

- VERY UNREALISTIC SCENARIO: SL ensemble contains a bunch of different GLMs

$$\bar{Q}^a(W) = logit^{-1}\left[\beta_0 + \beta_1 W1 + \beta_2 W3 + \beta_3 W1^* W3 + \beta_5 W4^2\right]$$
$$\bar{Q}^b(W) = logit^{-1}\left[\beta_0 + \beta_1 W1 + \beta_2 log(W2) + \beta_3 W3 + \beta_4 W4 + \beta_5 W3^* W4\right]$$
$$\bar{Q}^c(W) = logit^{-1}\left[\beta_0 + \beta_1 W1 + \beta_2 W2 + \beta_3 W4 + \beta_4 W1^* W2 + \beta_5 W1^* W4 + \beta_6 W2^* W4 + \beta_7 W1^* W2^* W4\right]$$
$$\bar{Q}^d(W) = logit^{-1}\left[\beta_0 + \beta_1 W1 + \beta_2 sin(W2^2) + \beta_3 W1^* sin(W2^2) + \beta_4 log(W4)\right]$$

- In reality, would want a diversity of algorithms and specifications

- You will be given a data set from an unknown data-generating mechanism and try to find the best predictor of $\bar{Q}_0(W) = \mathbb{E}_0(Y \mid W)$, using both discrete and ensemble SL.

SL discrete/ensemble example

$$\bar{Q}^a(W) = logit^{-1}\left[\beta_0 + \beta_1 W1 + \beta_2 W3 + \beta_3 W1^* W3 + \beta_5 W4^2\right]$$
$$\bar{Q}^b(W) = logit^{-1}\left[\beta_0 + \beta_1 W1 + \beta_2 log(W2) + \beta_3 W3 + \beta_4 W4 + \beta_5 W3^* W4\right]$$
$$\bar{Q}^c(W) = logit^{-1}\left[\beta_0 + \beta_1 W1 + \beta_2 W2 + \beta_3 W4 + \beta_4 W1^* W2 + \beta_5 W1^* W4 + \beta_6 W2^* W4 + \beta_7 W1^* W2^* W4\right]$$
$$\bar{Q}^d(W) = logit^{-1}\left[\beta_0 + \beta_1 W1 + \beta_2 sin(W2^2) + \beta_3 W1^* sin(W2^2) + \beta_4 log(W4)\right]$$

	Risk	Coef
SL.glm.EstA_All	0.1809579	0.0000000
SL.glm.EstB_All	0.1779846	0.0000000
SL.glm.EstC_All	0.1652105	0.3553639
SL.glm.EstD_All	0.1555183	0.6446361

Which is best?

Not unusual for some algorithms to get no weight.

SL details

- “In reality, would want a diversity of algorithms and specifications” … how do we do this?
- SL uses “wrappers” to standardize inputs and outputs of common ML functions. (more on this in next slides)
- Here’s an example of a library with the four GLMs:

```
SL.library <- c('SL.glm.EstA', 'SL.glm.EstB', 'SL.glm.EstC', ‘SL.glm.EstD')

X <- subset(ObsData, select= -Y)

SL.out <- SuperLearner(Y=ObsData$Y, X=X, SL.library = SL.library, family='binomial',
                                cvControl=list(V=10))
```

SL details

- In Lab 4, there are four GLMs specified in an accompanying R file:

You can make your own based on the template here:

```r
# ###############
# Estimator A
SL.glm.EstA<- function(Y, X, newX, family, ...) {
  if(family$family=='binomial') {
    fit.glm<- glm(Y~ W1*W3 + W4sq, data=X, family=family)
    pred <- predict(fit.glm, newdata=newX, type='response')
    fit<- list(object=fit.glm)
  }
  if(family$family=='gaussian'){}

  out <- list(pred=pred, fit=fit)
  class(out$fit) <- c('SL.glm.EstA')
  return(out)
}
```

SL details

- All built-in functions are visible by running `listWrappers()`. But you will often want to modify their default settings.
- Why? Force in variables you KNOW matter, make sure certain interactions are tested, etc…
- My SL libraries often have a mix of default and custom settings:

```
SL.library <- c(
'SL.glm.EstA',
'SL.glm.EstB',
'SL.glm.interaction',
'SL.glmnet', 'SL.glmnet_alpha1', 'SL.glmnet_alpha0.5',
'SL.mean',
'SL.ranger', 'SL.ranger_PRUNE3', 'SL.ranger_PRUNE5',
'SL.xgboost')
```

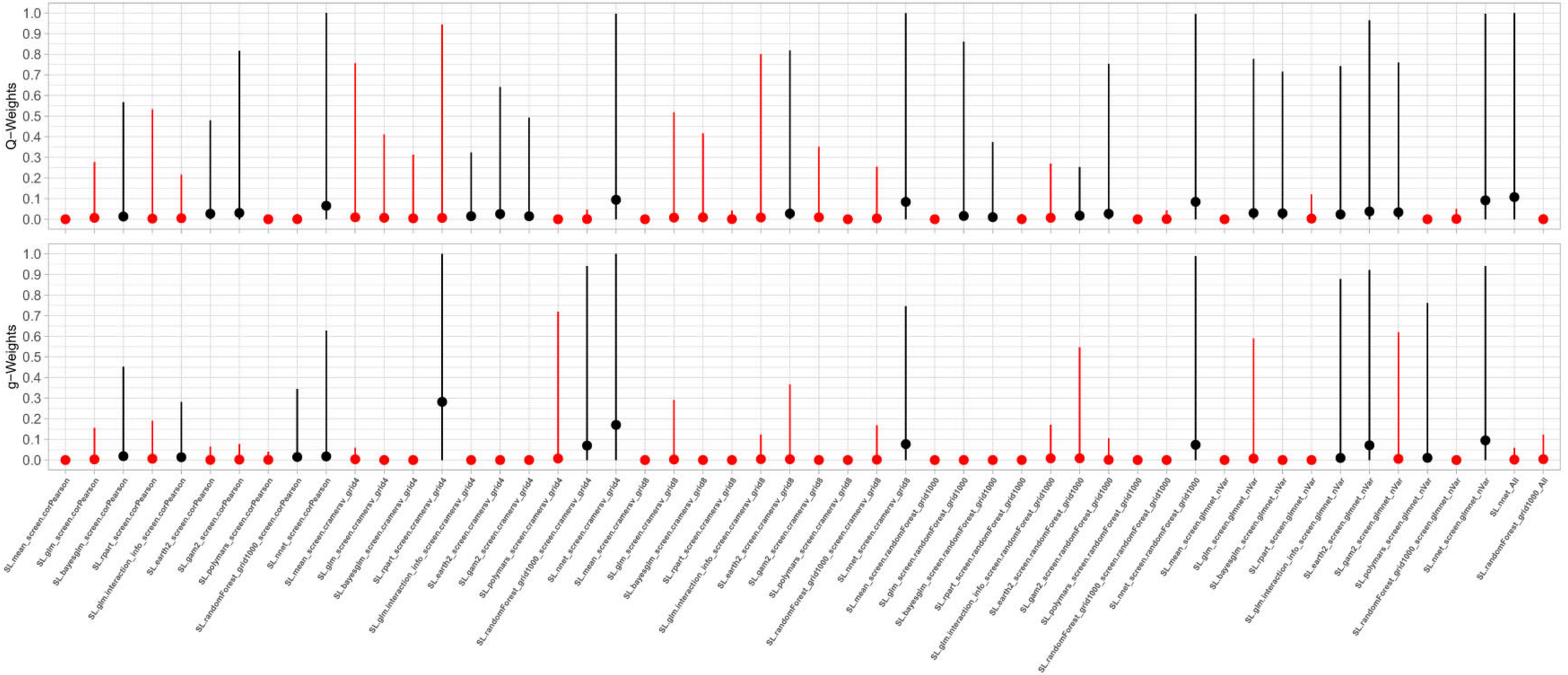
```
> listWrappers()
All prediction algorithm wrappers in SuperLearner:

 [1] "SL.bartMachine"      "SL.bayesglm"         "SL.biglasso"
 [4] "SL.caret"            "SL.caret.rpart"      "SL.cforest"
 [7] "SL.earth"            "SL.extraTrees"       "SL.gam"
[10] "SL.gbm"              "SL.glm"              "SL.glm.interaction"
[13] "SL.glmnet"           "SL.ipredbagg"        "SL.kernelKnn"
[16] "SL.knn"              "SL.ksvm"             "SL.lda"
[19] "SL.leekasso"         "SL.lm"               "SL.loess"
[22] "SL.logreg"           "SL.mean"             "SL.nnet"
[25] "SL.nnls"             "SL.polymars"         "SL.qda"
[28] "SL.randomForest"     "SL.ranger"           "SL.ridge"
[31] "SL.rpart"            "SL.rpartPrune"       "SL.speedglm"
[34] "SL.speedlm"          "SL.step"             "SL.step.forward"
[37] "SL.step.interaction" "SL.stepAIC"          "SL.svm"
[40] "SL.template"         "SL.xgboost"

All screening algorithm wrappers in SuperLearner:

[1] "All"
[1] "screen.corP"           "screen.corRank"        "screen.glmnet"
[4] "screen.randomForest"   "screen.SIS"           "screen.template"
[7] "screen.ttest"          "write.screen.template"
```

Build Your Own Spicy Version

Another example: Default k-nearest neighbors algorithm. k = 10 by default.

You could modify! There is no built-in CV here to pick k…

```
> SL.knn
function (Y, X, newX, family, k = 10, ...)
{
    .SL.require("class")
    if (family$family == "gaussian") {
        stop("SL.knn only available for family = binomial()")
    }
    fit.knn <- class::knn(train = X, test = newX, k = k, cl = Y, prob = TRUE)
    pred <- (as.numeric(fit.knn) - 1) * attr(fit.knn, "prob") +
        (1 - (as.numeric(fit.knn) - 1)) * (1 - attr(fit.knn,
            "prob"))
    fit <- list(k = k)
    out <- list(pred = pred, fit = fit)
    class(out$fit) <- c("SL.knn")
    return(out)
}
```



Baumann, Philipp F. M., Schomaker, Michael and Rossi, Enzo. "Estimating the effect of central bank independence on inflation using longitudinal targeted maximum likelihood estimation" *Journal of Causal Inference*, vol. 9, no. 1, 2021, pp. 109-146. https://doi-org.silk.library.umass.edu/10.1515/jci-2020-0016

Baumann, Philipp F. M., Schomaker, Michael and Rossi, Enzo. "Estimating the effect of central bank independence on inflation using longitudinal targeted maximum likelihood estimation" *Journal of Causal Inference*, vol. 9, no. 1, 2021, pp. 109-146. https://doi-org.silk.library.umass.edu/10.1515/jci-2020-0016

Practical issues with SL

- It's temping to use it mindlessly…Only as good as algorithms in the ensemble
  - Do you expect interactions? Make sure you have algorithms with them (glm.interaction, random forest, etc)
  - Computation time can get long if your SL library gets big and fancy
  - Diversity of algorithms is best. Always include "SL.mean". SL can still overfit!!!!!!!
  - Can include copies of algorithms with different parameters (splines with different numbers of knots, for example)
  - Work through roadmap to think about which covariates to include - screeners can be used
- "Effective" sample size is smaller with rare outcomes… overfitting still possible
- set.seed() can give different results in some cases! (though averaging possible)

Putnam Data Sciences

269 subscribers

SUBSCRIBED

HOME | VIDEOS | PLAYLISTS | CHANNELS | ABOUT

Uploads PLAY ALL

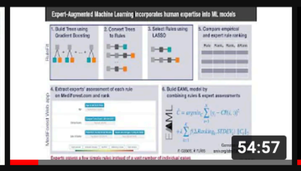
SORT BY

Targeted Learning: Towards a future informed by real worl...

114 views • 1 month ago

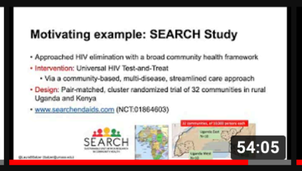
BAA Project to Advance Regulatory...

44 views • 2 months ago



Expert Augmented Machine Learning

92 views • 3 months ago

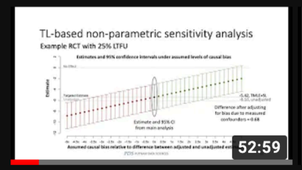


Challenges and Solutions in the Analysis of Cluster...

123 views • 4 months ago

Practical Considerations for Specifying a Super Learner

177 views • 5 months ago

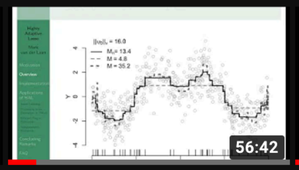


Developing a Targeted Learning-Based Statistical...

182 views • 6 months ago

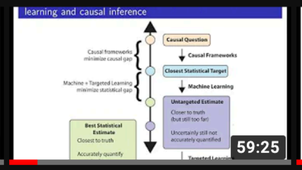
Targeted Machine Learning in Action in the ICU

78 views • 7 months ago



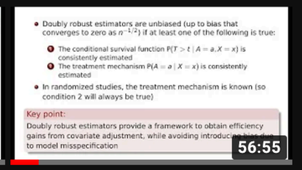
Highly Adaptive Lasso (HAL) in Causal Inference

605 views • 10 months ago



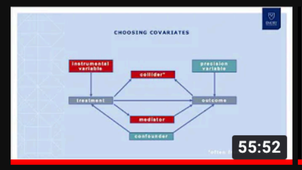
Cross-validated Targeted Maximum Likelihood...

355 views • 1 year ago



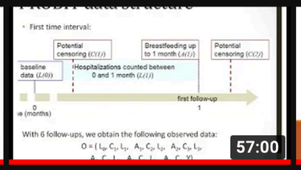
Covariate adjustment in randomized studies with...

191 views • 1 year ago



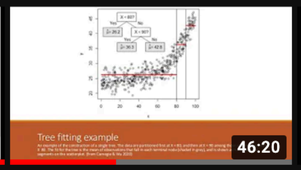
Practical Issues in Targeted Learning

281 views • 1 year ago



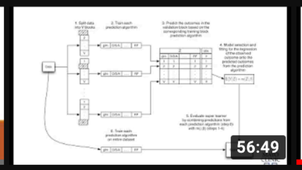
Causal Inference of Longitudinal Exposures,...

447 views • 1 year ago



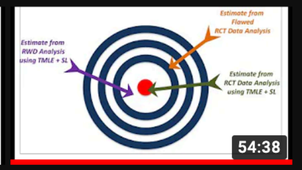
Introduction to Bayesian Additive Regression Trees...

3.8K views • 1 year ago



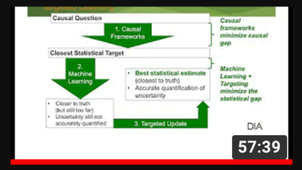
3. An introduction to Super Learning

976 views • 1 year ago



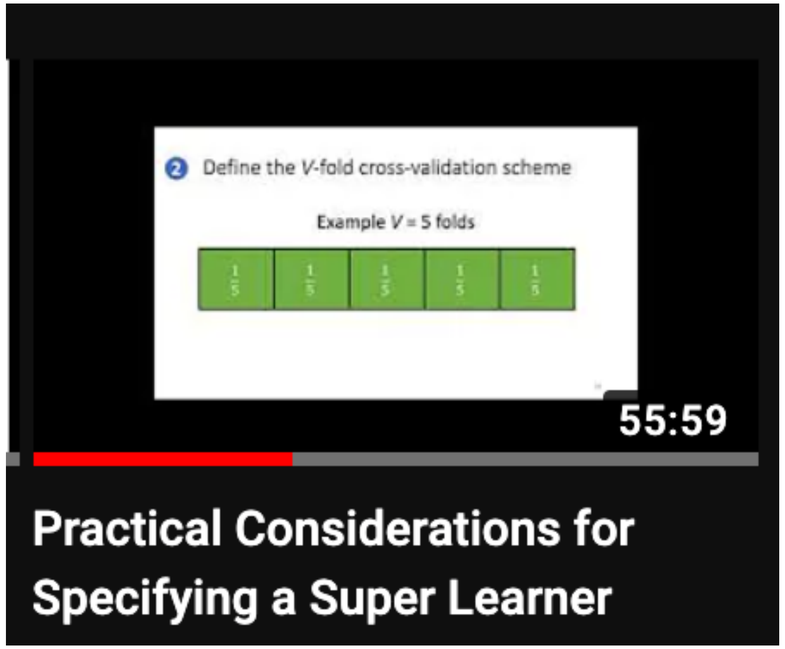
2. An Introduction toTargeted Maximum Likelihood...

1.8K views • 1 year ago



1. Targeted Machine Learning for Causal Inference based...

1K views • 1 year ago



Pro Tips

- mcSuperLearner() runs multi-core, so can be faster when algorithms are complicated or data sets large. Sometimes a little unstable though. Use at your own risk; might help on final project.

- sl3 is newest, modern version of SuperLearner. Still in development and intermittently stable/unstable. Google “tlverse handbook” to learn more about this. We won’t use it this year in the course but probably will starting next year. Same big ideas, just a more modern implementation. Better parallel processing.

- With clustered data, need to be careful

- More data -> fewer CV folds. Less data -> More CV folds (up to LOO)

HW 4 Tasks

- For some data from an unknown DGP…
  - Split the data into 20 folds
  - Fit the models 19/20 of the data, excluding fold #1
  - Test the predictions from the model on the left-out 1/20 of the data
  - Record the error in these predictions, repeat for all the other folds #2 - #20
  - “Risk” is the average MSE across all 20 iterations
- Compare the different models by their risk
- Finally, protect against overfitting by repeating this process over a set of SuperLearners with CV.SuperLearner

HW 4 Tasks

- You will start by hand-coding the process using boring GLMs.
- Second section: Using fun GLM + ML ensemble and true SuperLearner package. You should get slightly better performance than in the first round with just GLMs.
- Feel free to create a bonus question comparing the performance of a BIG super learner ensemble with a custom function!
- Finally, use CV.SuperLearner to get a more “honest” MSE (risk) assessment by adding a CV layer

R HW 3 hints

$$W \sim Bernoulli(.5)$$

$$A \mid W \sim Bernoulli(0.2 + 0.6 \times W)$$

$$Y \mid A, W \ =_d 1000 \ + \mathbb{I}(\tilde{U} < logit^{-1}(W \times A)),$$

$$\hat{\Psi}_{IPTW}(\mathbb{P}_n) = \frac{1}{n} \sum_{i=1}^{n} \frac{A_i}{\mathbb{P}_0(A = 1 | W_i)} Y_i$$

Questions???

IMPT REMINDERS

- Josh: stop the recording
- Use TEMPLATE for R HW 4 so that you won’t forget any Qs!
- I didn’t cover EVERYTHING in R Lab 4; please look over the answer key. R code will be very helpful, plus more detail and context
- Thurs (tmrw) is holiday but I still have normal office hours