PH C240B: Assignment 4

Rachael Phillips

12/1/2017

Load longdata.RData from bCourses. You may use Odat.long or Odat.wide for your analysis. They are identical except for the format. The columns are labeled according to how we normally notate them. $A1$'s are the binary treatments, $A2$'s are the censoring indicators, Y's are the binary survival outcomes and L's are the covariates, which are either continuous or binary as you will see upon looking at the dataframe. Fortunately, the data is already cleaned, as in, NA's are produced for all nodes after outcome is 1 (death) and also after censoring occurs.

1

Assuming sequential randomization and positivity as usual, estimate the difference in counterfactual mean outcomes at time 5 under treatment and not: $\mathbb{E}[Y(5)_{\bar{a}=(0,0,1,1,1)}] - \mathbb{E}[Y(5)_{\bar{a}=(0,0,0,0,0)}]$, given no one is allowed to leave the study (i.e., censoring intervened upon to be 0 at all times). Give estimates using IPTW and TMLE (debiasing each outcome regression with a targeting step, moving backward in time and otherwise known as L-TMLE) and show all code used to do so. Your superlearner library should include three algorithms, one of each of the following: glm, mean, xgboost. Provide instructions on how I may access your superlearner results in this estimation process.

```r
Odat = Odat.wide[,c(2:5,7:27)]

Anodes = grep("A1",colnames((Odat)))
Cnodes = grep("A2",colnames((Odat)))
Lnodes = grep("L",colnames((Odat)))
Ynodes = grep("Y", colnames((Odat)))

for (col in Cnodes) Odat[,col] = BinaryToCensoring(is.censored = Odat[col])

Odat[,"Y_1"][is.na(Odat$Y_1)] = 1
Odat[,"Y_2"][is.na(Odat$Y_2)] = 1
Odat[,"Y_3"][is.na(Odat$Y_3)] = 1
Odat[,"Y_4"][is.na(Odat$Y_4)] = 1
Odat[,"Y_5"][is.na(Odat$Y_5)] = 1

results.ltmle <- ltmle(Odat,
                  Anodes = Anodes,
                  Cnodes = Cnodes,
                  Lnodes = Lnodes,
                  Ynodes = Ynodes,
                  abar = list(c(0,0,1,1,1), c(0,0,0,0,0)),
                  survivalOutcome = TRUE,
                  SL.library = c("SL.mean", "SL.glm", "SL.xgboost"),
                  variance.method = "ic"
                  )

# TMLE results
summary(results.ltmle, "tmle")
```

```
## Estimator:  tmle
## Call:
## ltmle(data = Odat, Anodes = Anodes, Cnodes = Cnodes, Lnodes = Lnodes,
##     Ynodes = Ynodes, survivalOutcome = TRUE, abar = list(c(0,
##         0, 1, 1, 1), c(0, 0, 0, 0, 0)), SL.library = c("SL.mean",
##         "SL.glm", "SL.xgboost"), variance.method = "ic")
##
## Treatment Estimate:
##    Parameter Estimate:  0.58119
##     Estimated Std Err:  0.019826
##               p-value:  <2e-16
##     95% Conf Interval: (0.54233, 0.62005)
##
## Control Estimate:
##    Parameter Estimate:  0.60397
##     Estimated Std Err:  0.028616
##               p-value:  <2e-16
##     95% Conf Interval: (0.54788, 0.66005)
##
## Additive Treatment Effect:
##    Parameter Estimate:  -0.022775
##     Estimated Std Err:  0.033962
##               p-value:  0.50248
##     95% Conf Interval: (-0.08934, 0.04379)
##
## Relative Risk:
##    Parameter Estimate:  0.96229
##   Est Std Err log(RR):  0.056938
##               p-value:  0.49961
##     95% Conf Interval: (0.86068, 1.0759)
##
## Odds Ratio:
##    Parameter Estimate:  0.90996
##   Est Std Err log(OR):  0.14122
##               p-value:  0.50404
##     95% Conf Interval: (0.68995, 1.2001)
```

```r
# IPTW results
summary(results.ltmle, "iptw")
```

```
## Estimator:  iptw
## Call:
## ltmle(data = Odat, Anodes = Anodes, Cnodes = Cnodes, Lnodes = Lnodes,
##     Ynodes = Ynodes, survivalOutcome = TRUE, abar = list(c(0,
##         0, 1, 1, 1), c(0, 0, 0, 0, 0)), SL.library = c("SL.mean",
##         "SL.glm", "SL.xgboost"), variance.method = "ic")
##
## Treatment Estimate:
##    Parameter Estimate:  0.58422
##     Estimated Std Err:  0.021316
##               p-value:  <2e-16
##     95% Conf Interval: (0.54244, 0.626)
##
## Control Estimate:
##    Parameter Estimate:  0.60754
##     Estimated Std Err:  0.034266
##               p-value:  <2e-16
##     95% Conf Interval: (0.54038, 0.6747)
##
## Additive Treatment Effect:
##    Parameter Estimate:  -0.023324
##     Estimated Std Err:  0.039981
##               p-value:  0.55964
##     95% Conf Interval: (-0.10168, 0.055037)
##
## Relative Risk:
##    Parameter Estimate:  0.96161
##   Est Std Err log(RR):  0.06654
##               p-value:  0.55632
##     95% Conf Interval: (0.84403, 1.0956)
##
## Odds Ratio:
##    Parameter Estimate:  0.90767
##   Est Std Err log(OR):  0.16684
##               p-value:  0.56146
##     95% Conf Interval: (0.65451, 1.2587)
```

```
# SuperLearner results
#results.ltmle$fit$g

# SuperLearner Q results
#results.ltmle$fit$Q
```

2

Let us assume you have estimated counterfactual means for all 32 possible regimes, $\bar{a}$, over 5 time points and you have summarized the counterfactual means via the use of a marginal structural model (MSM):

$$\underset{\alpha,\beta}{argmin} \sum_{\bar{a}} (\mathbb{E}Y(5)_{\bar{a}} - (\alpha + \beta sum(\bar{a})))^2$$

where $sum(\bar{a})$ = sum of treatments over the 5 time points in the study. Notice, this is just a least squares projection of the counterfactual means onto a line. How would you interpret this coefficient to a layperson, i.e. someone not technical?

This coefficient is the estimated line that describes the change in the probability of our outcome (survival or death) caused by differing levels of treatment (or exposure) over five time points, summarizing how the treatment affects the outcome. We can think of this line as an estimated relationship between our treatment and outcome. Specifically, this is the line that best fits our estimation and we estimated this line using a marginal structural model (MSM). MSMs use a multi-step estimation strategy to separate confounding control (i.e. control from a variable that influences both the treatment and the outcome) from the estimation of the parameters of interest, allowing us to obtain unbiased estimates of the counterfactual means.

Representing this treatment-outcome relationship as a line of best fit gives us the ability to communicate to practitioners the rate, the slope, of outcome under the cumulative treatment regimes.

3

Continuing from the previous question, what if you had some regimes that were much more common than others in the data and you decided to offer estimates of the weighted projection:

$$\underset{\alpha,\beta}{argmin} \sum_{\bar{a}} P_n(\bar{A} = \bar{a})(\mathbb{E}Y(5)_{\bar{a}} - (\alpha + \beta sum(\bar{a})))^2$$

How would you interpret this coefficient to a layperson, ie someone not technical? $P_n(\bar{A} = \bar{a})$ is just the empirical probability of observing $\bar{a}$. Also, in what situation does this change the parameter defined in the previous sections? Explain briefly.

This coefficient is the estimated line that describes the change in the probability of our outcome (survival or death) caused by **observing** differing levels of treatment (or exposure) over five time points, summarizing how the **observed** treatment regimes affect the outcome. We can think of this line as an estimated relationship between our treatment and outcome where we *also allow the commonality of the treatment to affect this treatment-outcome relationship.* This line is estimated similarly to the previous section but, contrary to the previous coefficient, this line is weighted by the probability of observing a treatment regime. This type of weighting can be useful when some regimes are much more common than others.

Like the previous section, we can still interpret this coefficient/line in terms of a rate (where the rate is just the slope of this line). However, the rate here is influenced by how many people got the treatment regimes.

Collaborators & Resources

Tommy Carpenito, Yue You, Jonathan Levy