Problem Set 4

November 19, 2018

1) Run Monte carlos with series, neural nets, and random forests.

a) Install the "neuralnet" and "randomForest" packages in R.

b) Draw five separate x variables with random correlation using the following code:

```
################################################
#Draw observable explanatory variables
x1 = rgamma(n,2,1); x2 = rnorm(n,0,2);
x3 = rweibull(n,2,2); x4 = rlogis(n,2,1);
x5 = rbeta(n,2,1);
x = cbind(x1,x2,x3,x4,x5)
################################################
#transform into independent random variables
# find the current correlation matrix
c1 <- var(x)

# cholesky decomposition to get independence
chol1 <- solve(chol(c1))

x <-  x %*% chol1
################################################
#generate random correlation matrix

R <- matrix(runif(ncol(x)^2,-1,1), ncol=ncol(x))

RtR <- R %*% t(R)

corr <- cov2cor(RtR)

# check that it is positive definite
sum((eigen(corr)$values>0))==ncol(x)
################################################
#transform according to this correlation matrix
x <- x \%*\% chol(corr)
```

c) You'll estimate three sets of three models. For each, set the seed to 0, the sample size to 1,000, and allocate 50% of the data to a test sample.

Specification 1: let $y = x_1 + \frac{x_3 x_2^2}{10} + \frac{x_4 x_1 x_5}{10}$.

Specification 2: let $y = x_1 + \frac{x_3 x_2^2}{10} + \frac{x_4 x_1 x_5}{10} + u$ where $u \sim N(0,1)$.

Specification 3: let $y = log(|x_1^4/10| + |x_2| + x_3^2) + x_4 x_2 sin(x_5) + u$ where $u \sim N(0,1)$.

For each specification,

i) Estimate a neural net with 3 hidden layers, each with 64, 32, and 16 neurons respectively.

ii) Estimate a series using the poly function. Set the degree to 3.

iii) Estimate a random forest. Use 1000 trees with 4 covariates sampled each time.

iv) Calculate the MSE on the test set. Repeat i-iii as many times as you can and average the MSE to get a single set of MSE's for each specification.

d) For specification 1, which performs best? Why?

e) For specification 2, which performs best? Why?

f) For specification 3, which performs best? Why?

2) Go back to problem set 3. In addition to the five models you estimated there, estimate a random forest with the same five predictors you used in question 6. How does it perform, in terms of MSE, relative to the cross-validated flexible logit model with those predictors? Why do you think this is?

3) Group markets using kmeans and agglomerative hierarchical clustering.

a) Load the file "PS4_mkt.R".

b) For c) through e), use the average price, average distance, nonstop miles, number of carriers, and hhi as covariates.

c) Use the kmeans function to cluster the origin and destination pairs in the data into 2 clusters. Calculate summary statistics for each of them. How would you best characterize

these clusters qualitatively?

d) Now, use the kmeans function to cluster the origin and destination pairs in the data into 4 clusters.

e) Use the hclust function to perform agglomerative clustering. Plot the dendrogram. Using the cutree function with k = 4, compare the results of this clustering algorithm with those in part d.