Data Science & Strategic Pricing

Instructor: Jacob LaRiviere, Director of Economics and Data Science at Amazon

Email: [jlarivi1@uw.edu](mailto:jlarivi1@uw.edu), [lghhager@uw.edu](mailto:lghhager@uw.edu) (TA)

# Course Assignments & Reading

Course assignments should be knitted Rmarkdown files and turned in at the start of class on Canvas unless otherwise noted. Feel free to work in groups but everyone is required to turn in their own work with answers written in your own words. Keep in mind that in most cases a good answer is one precise sentence; quality is heavily favored over quantity. This will be graded on a full credit, half credit and no credit basis. All work must be typed. At the beginning of each class the professor will lead a discussion around these questions. Students will be called on, potentially at random, to add their insight. This part of class will contribute heavily to your course participation grade.

**Week 4, due October 26**

**Assignment to be turned in.** Please turn in your Rmarkdown output and answers on canvas

1. In the last assignment you calculated the MSE on a test set. Let’s expand that code to include 5-fold cross validation.
   1. Create 5 partitions of the data of equal size.
   2. Create 5 training datasets using 80% of the data for each one. This is done by “appending” the data together using [rbind](https://www.statmethods.net/management/merging.html).
   3. Estimate a complex model using OLS which includes price, featured, brand, brand\*price and lagged price, all the sociodemographic variables and interactions of EDUC and HHSIZE with price on each of the training sets then the MSE on the test sets using the predict command.
      1. Calculate the MSE for the model on the test set for each fold (e.g., there will be five sets of model parameters and five test set MSEs with 5-fold cross validation).
      2. Average across the MSEs to get the cross validated MSE for an OLS model run on that particular set of features.
2. Now lets take that same model from (1.c) and run a LASSO using [glmnet](https://glmnet.stanford.edu/articles/glmnet.html) which is a workhorse R package for LASSO, Ridge and Elastic Nets.
   1. First remember to install the glmnet package and library to your R session.
   2. Remember to estimate a LASSO you must pass glmnet a matrix of data for candidate features and a vector as candidate outcomes:

set.seed(720)

lasso\_v1 <- cv.glmnet(x, y, alpha=1)

lasso\_v1\_cv <- glmnet(x, y, alpha=1)

As an alternative to defining products in a dataframe, turning that into a matrix then passing that matrix to glmnet, you can cut out the middle man with this.

X <- model.matrix(formula, df\_RHS)

# Note that df\_RHS is a dataframe of just the features you’d like # to use as predictors of Y (this should include lagged price)

# Now pass X to glmnet; [documentation](https://www.rdocumentation.org/packages/MatrixModels/versions/0.4-1/topics/model.Matrix)

In addition to the variables in the original dataframe, try to create tons of new features that you think could plausibly be predictive of quantity sold. This could include lagged prices, interactions of several features, etc.

* 1. Investigate the coefficients of the cross validated LASSO model. Code from class is here:

x <- as.matrix(oj\_cross[ ,5:17])

y <- as.numeric(as.matrix(oj\_cross[ ,4]))

set.seed(720)

#lasso\_v1 <- cv.glmnet(x, y, alpha=1)

**lasso\_v1 <- glmnet(x, y, alpha=1)**

#Results

plot(lasso\_v1)

**coef(lasso\_v1, s=lasso\_v1$lambda.min)**

# Now ready for cross validation version of the object

cvfit <- cv.glmnet(x, y, alpha=1)

#Results

plot(cvfit)

cvfit$lambda.min

log(cvfit$lambda.min)

coef(cvfit, s = "lambda.min")

Which are the parameters the cross validated LASSO model kicks out of the model? What is the ratio of number of features to number of observations? How might that relate to overfitting from “sampling error”?

* 1. Can you look that the glmnet objects and figure out what the out of sample (e.g., test set) average MSE was with the cross validated LASSO model relative to the model in 1.c?
  2. What is the advantage of using LASSO for choosing model complexity as opposed to using your intuition as an economist?
     1. In what part of this process did you use your intuition as an economist? (*HINT: what’s in the X matrix?)*

1. Now estimate the model with only the variable selected with the LASSO procedure but with OLS to avoid attenuation bias in the coefficients (similar to this [paper](https://www.aeaweb.org/articles?id=10.1257/aer.p20151022)).
   1. Let’s return to the orange juice assignment and get very precise about how to interpret coefficients. What is the predicted elasticity in the following cases?
      1. For Dominicks when the lagged price is $1 (NOTE: did you interact lagged price with current period price?) If not, does lagged price impact the *elasticity* this period or *log move* this period.
      2. For Tropicana
      3. For Tropicana when its featured
      4. What is the 95% confidence intervals for Tropicana
   2. Which product has the most elastic demand?
      1. Should that product have the highest markup over costs or lowest markup over costs? Why?
2. Go back to using logmove and log(price).
   1. Estimate a 3x3 matrix own price and cross price elasticities for Dominicks, Minute Maid, and Tropicana using only the current week’s prices. Be sure to estimate separate models for sales of Dominicks, MM and Tropicana (e.g., you’ll run three separate regressions with the same RHS variables but different LHS variables). It doesn’t need to be overly complicated, but make sure there is an interpretable elasticity estimate. NOTE: This will require three different regressions & add in socio demographic controls for each store.
   2. Do the same but add in interactions for whether or not each brand is featured.
      1. How do the estimates change?
      2. What product’s sales suffer the most when Minute Maid is both featured and lowers its price?
   3. Which two products are the most competitive with each other?
      1. How did you infer that looking at the cross price elasticity?
      2. What do you expect that to mean about the correlation of the prices of those two products? Would they be more correlated or less correlated than the price of other pairs of products?
   4. BONUS: Estimate a 3x3 matrix own price and cross price elasticities for Dominicks, Minute Maid, and Tropicana using the current week’s prices, whether the good is featured or not and last week’s own price and the competitor’s prices.