Week 2, HW3: Interpretation and Cross Validation

Instructor: Jacob LaRiviere

Email: jlariv@microsoft.com

1) Let's return to the orange juice dataset and investigate how store demographics are related to demand.
   a. Take the "fully interacted" model from HW2 (logmove ~ log(price)*brand*feat) and add in the store demographics as linear features (e.g. + demo1 + demo2+…).
   b. What demographics are significantly (t>2) related to demand?
   c. How much did the adjusted R-squared improve with the addition of these variables?
   d. Use 5-fold cross validation to compare the MSE of the model with and without sociodemographic controls.
      i. There are two ways to do this. The first is by hand.
      ii. The second is using built in functions. I'm giving you code to do it by hand.

2) Let's focus on two variables `HVAL150` ("percent of HHs with homes >$150K") and one of your choosing.
   a. What are the means and percentiles of each of these variables?
      HINT: `summary(oj$HVAL150)`
   b. Using your coefficient estimates from the regression in 1b:
      i. If we move from the median value of `HVAL150` to the 75th percentile (3rd quartile), how much does log(quantity) change each week on average?
         HINT: using `coef(reg_output)["var_name"]` exports the coefficient on "`var_name`" from the regression model "reg_output". Similarly, `summary(df$var_name)` will output a bunch of summary statistics for the variable var_name in data frame df. Using `summary(df$var_name)["3rd Qu."]` will take the level of the 3rd quantile from the summary of `var_name`.
         Because we estimate things in logs you'll want to take the exponent of everything.
      ii. If we move from the median value of `HVAL150` to the 75th percentile (3rd quartile), how much does log(quantity) change each week on average?
      iii. Base on this analysis, which is the more important predictor of demand?
   c. Now let's see if these variables impact price sensitivity. Add two interaction terms (with logprice) to the model to test this.
      i. What are the coefficients on the interaction terms?
      ii. Recall, positive values indicate lower price sensitivity and negative values indicate greater price sensitivity. Do your estimates make sense based on your intuition?
      iii. What are the coefficient estimates on the constants `HVAL150` and your variable of choice? How do they compare to your regression from 1b?
      iv. Similar to 2b, if we move from the median value of each variable to the 3rd quartile, how much does elasticity change? Based on this, which is more important to price sensitivity?

3) Tuna fish question! Create make a new dataframe which takes the previous week's prices as a variable on the same line as the current week. This would enable you to see if there is *intertemporal* substitution.
    a. There are going to be a couple of steps. First is creating a new dataframe which is like the old one except that the week variable will change by a single week

```
df1 <-oj
df1$week<-df1$week+1
# df1 now has NEXT week and not the current one.  If we merge this by
#weeks now, this is last week's price (e.g., "lagged price").
myvars <- c("price", "week", "brand","store")
df1 <- df1[myvars]
oj_with_lagged_prices <- merge(oj, df1, by=c("brand","store","week"))
```

Investigate the Df2 and rename the lagged store values needed for a lagged price within the same store

    b. Now run a regression with this week's log(quantity) on current and last week's price.
    c. What do you notice about the previous week's elasticity? Does this make sales more or less attractive from a profit maximization perspective? Why?
    d. Use cross validation to find the best model (in terms of lowest out of sample MSE) with your musical pairs program. The winning team gets a special prize!!!!