**drivetrain probability analysis**

1. Removed unwanted columns in excel: veh id etc
2. Loaded the data in R, and handpicked 39 variables as many columns had NAs
3. Performed lasso regression: <http://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html#intro>
   1. Observed that for each outcome a separate graph was drawn
   2. Got coefficient values for each outcome.
4. Performed forward/backward stepwise regression but failed to get any out put
5. Ran multinomial regression but “iterations stopped after 100 iterations”,but problem is resolved after increasing the maxit to 500
6. Based on lasso regression, 7 variables were not considered the logistic model
7. Create a training and testing data set
8. Run multinomial logistic regression.
9. Tested the accuracy of the model
10. Explained the probability for each outcome, based on individual explanatory variable
11. Variables of interest
    1. Trip count
    2. Driving average speeds
    3. Distance total
    4. Acceleration events per mile
    5. Total stops
    6. Average climbing rate
    7. Max road grade
    8. Av. Kinetic power density demand( a factor in fuel consumption)
    9. Aerodynamic speed
    10. Characteristic acceleration
    11. Kinetic intensity
12. Take pairs of explanatory variables and see the outcome possibilities

Doubt: for probability graphs, the following error is coming .

Warning messages:

1: In max(ids, na.rm = TRUE) :

no non-missing arguments to max; returning -Inf

2: In max(ids, na.rm = TRUE) :

no non-missing arguments to max; returning -Inf

library(effects)

mnl.eff <- Effect("aerodynamic\_speed",mnl.fit)

mnl.eff

library(reshape)

library(ggplot2)

transformEffectForPlot <- function(data) {

output <- melt(data$prob)

names(output) <- c('frequency','preference','probability')

output$preference <- gsub('\\.',' ',output$preference)

output$preference <- gsub('prob ','',output$preference)

output$u.probability <- melt(data$upper.prob)[,3]

output$l.probability <- melt(data$lower.prob)[,3]

return(output)

}

# max(fleetdna\_compositedata\_mod$aerodynamic\_speed) # finding the maximum value

mnl.eff.forPlot <- transformEffectForPlot(mnl.eff)

key <- data.table(frequency=1:6,aerodynamic\_speed=c(0,5,10,15,20,25)) # axis interval defined

mnl.eff.forPlot <- merge(x=mnl.eff.forPlot,y=key,by='frequency')

ggplot(mnl.eff.forPlot,aes(x=aerodynamic\_speed,colour=preference,fill=preference))+

geom\_line(aes(y=probability))+

geom\_ribbon(aes(ymin=l.probability,ymax=u.probability,linetype=NA),alpha=.15)+

xlab('aerodynamic\_speed')+

ylab('Probability')+

theme\_bw()+

scale\_color\_brewer(name='drive\_id',palette='Set1')+

scale\_fill\_brewer(name='drive\_id',palette='Set1')+

guides(col=guide\_legend(nrow=1,bycol=TRUE,title.position='top'))+

theme(legend.position='bottom',legend.text=element\_text(size=8))

***For vocation analysis***

When fitting a multinomial log model for vocations, see if we can include the driveline type also a variable

Vocations selected

1. School Buses
2. Mass Transit
3. Delivery

a. Parcel Delivery

b. Warehouse Delivery

c. Food/Beverage Delivery

EMISSION ANALYSIS

1. Load the composite fleet dna data
2. Create new columns classid,voc\_id,drive id,fuel id
3. New columns NOx,PM 2.5,PM 10
4. Run if condition , over class id and fuel id: only class 2-7considered with gasoline and diesel fuels
5. Split data into testing and training data set
6. Do forward/backward selection with lasso: NOX,PM2.5,PM10 and CO
   1. For NOX, 5 variables got deleted after lasso
7. FIT GAM MODEL,
8. Interpret result from GAM
9. Analyze the model or how good the model is ? use residual plots

Doubts

1. Is Lasso the best way to do selection of variables?
2. Can I use just the logistic regression to analyze the outcome, based the probability of each variabels and also based on the combination of variables
3. Can I use a generalized additive model here?