Regression Model to Predict Car Miles Per Gallon

Rai Chinki

California State University, East Bay

Theory Application Regression

STAT 6509

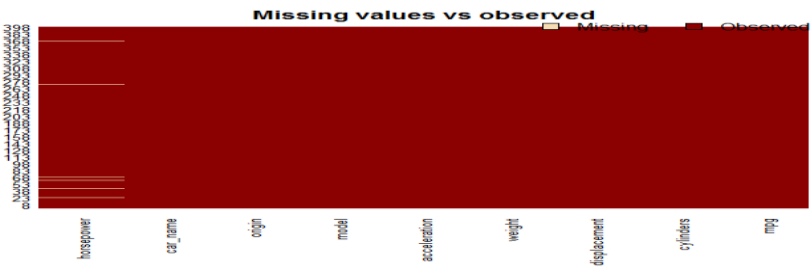
Abstract

This paper contains the development of the regression model to predict car's miles pes gallon based on the other provided variables. The dataset was taken from UCI Repository. The name of the dataset is AUTO_MPG. The dataset contains mpg, acceleration, displacement, horsepower, weight, cylinders, origin, model and car name. Initially simple model was taken from AIC and developed by using transformations of the predictors and the response variable. Validation is performed using 90% as training and 10% as the test. Model is trained with the help of training data and performed on test data to check accuracy. MAE is used for accuracy checking of prediction.

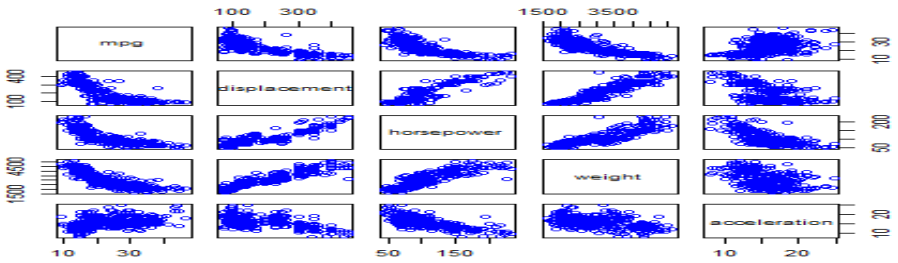
Regression Model to Predict Car Miles Per Gallon

The aim of the model is to predict miles per gallon for the cars. For the prediction model development is the very important factor. I started to develop the regression model using auto mpg data. The dataset contains 398 observations in which 305 different type of cars is present in the dataset. MPG is miles per gallon, Cylinder is the chamber where the gasoline is burned and turned into power. It could be 4,6 or 8. Displacement is the total volume of all the cylinders in an engine. Horsepower is to lift 33000 pounds one foot over period of one minute. Acceleration is the rate of change of velocity with time. For example, a car is travelling 50 km/hr starts to accelerate, 10 second after its speed 100 km/hr. Acceleration will be 1.38 m/s^2.Other variables are known by name. Dataset has 6 missing values in the horsepower column. Initially, I was planning to manipulate the missing values with the mean or median, but my data set has only 398 observations which is very small. I deleted missing information from the dataset. From my common understanding, I thought cylinder, acceleration, weight, horsepower and displacement will be significant variables to predict MPG.

To start the work on the dataset plotting missing value graph with the help of the library Amelia and get that horsepower has missing information. As dataset is small, deleting missing values is more accurate. Initially, horsepower was the factor to perform analysis converted them into the numeric variable.



Boxplot of the predictors and Response variables represent that mpg and acceleration are symmetric and other variables are skewed. Horsepower and acceleration have outliers. Scatterplot matrix is appropriate to see the relationship between predictor and response variable.



All the predictors are nonlinear. Displacement, horsepower, and weight are highly negative correlated with the mpg and acceleration are positively correlated with the mpg. Correlation matrix to confirm the relationship between predictors and Response.

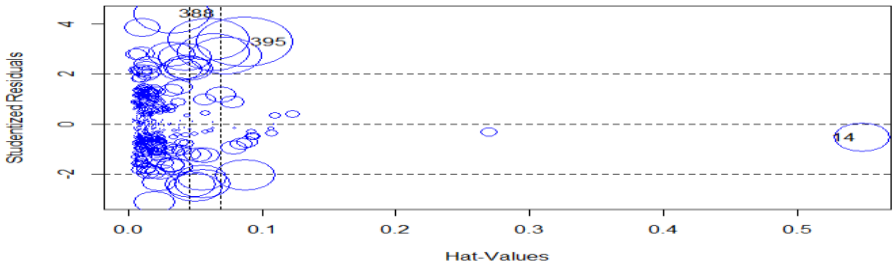
	Displacement	Horsepower	Weight	Acceleration
MPG	-0.8051269	-0.7784268	-0.832242	0.4233285

To select model, ran AIC and decided to proceed with the model based on low AIC.Low AIC model is mpg ~ weight +horsepower +acceleration + displacement + weight* horsepower + horsepower * acceleration + weight * acceleration + horsepower * displacement.

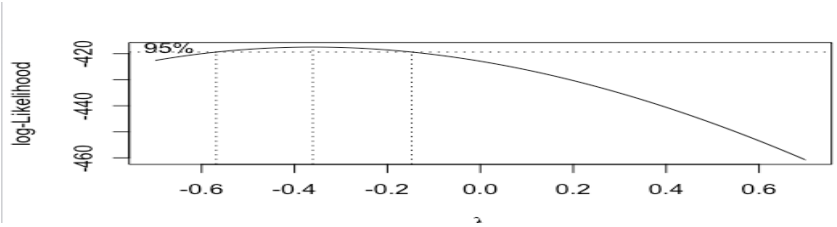
Initially, I ran the regression model on the above model got R^2 is 75% and got all the predictors, the intercept is significant and interaction between horsepower and displacement is significant.

I performed diagnostics for model validation. Constant variance assumption is not satisfying for this model. As constant variance assumption is not satisfied and scatterplot matrix of predictors and response variables is showing non-linear trade. I used transformed predictors. I used inverse transformation for weight and displacement. I used square root transformation for

horsepower and square transformation for acceleration. I plotted influence plot to check the outliers and formal test for the significant outliers. I deleted the significant outliers from the dataset and proceed with the transformed model. Significant outliers for car buick estate wagon(sw), oldmobile cutlass ciera(diesel) and VW pickup.

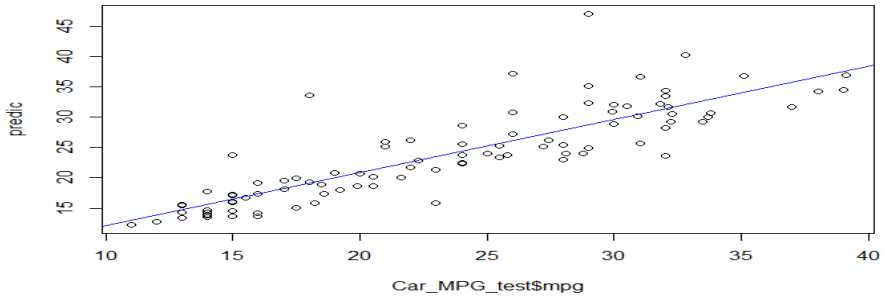


I transformed the predictor and performed regression again but constant variance assumption is not satisfied. I performed Boxcox to check response transformation.



From the Boxcox, I am getting $1/\sqrt{Y}$ transformation is appropriate. I did regression test for transformed predictor and transformed response variable. I am getting R^2 is 80.3% and all the predictors and interaction of horsepower with weight , acceleration and displacement is significant and interaction of weight with acceleration is significant.

To check the accuracy of our model, I performed the model evaluation. I divided randomly 90% dataset as training data and 10% as a test data. I performed model analysis on training data and predict mpg for test data set. Plotted scatterplot of predicted and actual value of mpg of the test data.

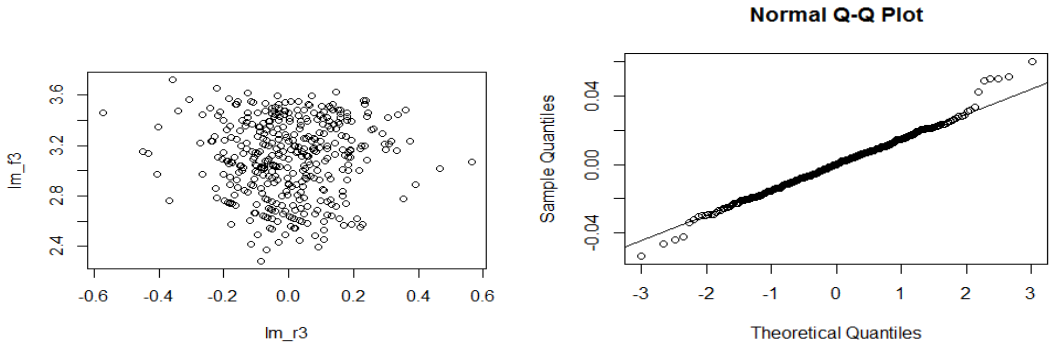


Points above the line are overfitted and points below the line are underfitted and MAE is 2.74 which indicates good predictor. To check prediction of the model for three cars. All the three cars are underfitted.

Make	Model	Estimated MPG of Factory	Predicted MPG	95% Confidence Interval
Honda	Civic	40.9	38	(36.6 , 39.4)
Honda	Accord	36.8	36	(34.7 , 37.2)
Toyota	Corolla	35.9	34.1	(33.1 , 35.0)

Result

Interaction of the weight with the horsepower and with the acceleration is significant and interaction of the horsepower with the acceleration and displacement is significant. As interaction is significant, I will keep main factors in the model. Where weight and displacement are transformed inverse, horsepower is square root transformed and acceleration is squared and mpg is transformed as inverse of square root of mpg. Residuals plot are scattered and satisfying constant variance assumption with NCV test p-value 0.877. Normality is also satisfied with the shapiro test .



Discussions

Limitations of the studies

As dataset is very small, I am getting underfitted model so this model will be accurate with the large datasets. As car model is from 1993, I cannot use this model for new cars. As in original dataset has only 305 different cars model are limited for certain cars.

Future study

Multiple regression model with the interactions is developed to predict the MPG of the cars with the help of 4 predictors, which are weight, horsepower, acceleration and displacement. Some car's mpg is overfitted and some car's mpg is underfitted with MAE 2.74, I will try to develop a model for which MAE will be near 0. For future study Regression tree is suggested. To perform accurate model result data should be big so planning to perform this model analysis on the cars data of the SASHEPL library.

References

(n.d.). Retrieved June 06, 2017, from http://archive.ics.uci.edu/ml/datasets/auto mpg

Machine Learning with R -Brett lantz

Machine Learning with R-Cookbook

Applied Linear Regression model -John Neter

Http://www.automobile-catalog.com/make/honda/accord_2gen/accord_2gen_4-door/1982.html. (n.d.).

Http://www.statmethods.net/input/missingdata.html. (n.d.)

Http://www.smartconversion.com/Articles/14.aspx. (n.d.).

(n.d.). Retrieved June 08, 2017, from http://www.web-cars.com/math/horsepower.html

What the Numbers Mean. (n.d.). Retrieved June 08, 2017, from http://www.vroomgirls.com/what-the-numbers-mean/

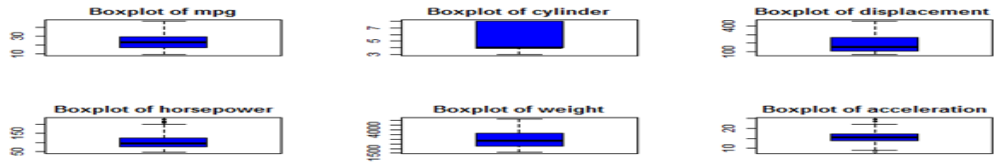
Http://www.citationmachine.net/items/427099948/copy. (n.d.).

Appendix

```
Car_MPG=read.table("C:/Computational Statistics/3rd Quater/Regression/Project/auto-
mpg.data.txt",na.strings =
T)colnames(Car_MPG)=c("mpg","cylinders","displacement","horsepower","weight","accelerati
on","model","origin","car_name")
str(Car_MPG)
Car_MPG$horsepower[Car_MPG$horsepower=="?"]=NA
table(is.na(Car_MPG$horsepower))
library(Amelia)
missmap(Car_MPG, main = "Missing values vs observed")
Car_MPG_final=Car_MPG[!(is.na(Car_MPG$horsepower)) , ]
Car_MPG_final$horsepower=as.numeric(as.character(Car_MPG_final$horsepower))
par(mfrow=c(3,3))
boxplot(Car_MPG_final[c(1)],col="Blue",main="Boxplot of mpg")
boxplot(Car_MPG_final[c(2)],col="blue",main="Boxplot of cylinder")
boxplot(Car_MPG_final[c(3)],col="blue",main="Boxplot of displacement ")
boxplot(Car_MPG_final[c(4)],col="blue",main="Boxplot of horsepower")
boxplot(Car_MPG_final[c(5)],col="blue",main="Boxplot of weight")
boxplot(Car_MPG_final[c(6)],col="blue",main="Boxplot of acceleration")
pairs(~mpg+displacement+horsepower+weight+acceleration,data = Car_MPG_final,col="blue")
scatterplotMatrix(~mpg+displacement+horsepower+weight+acceleration,data
=Car_MPG_final,ellipse=("FALSE"),smooth=F,col="blue")
cor(Car_MPG_final[,c(1,3,4,5,6)])
```

```
null_model=lm(mpg~1,data=Car_MPG_final)
Full_model=lm(mpg~weight+displacement+acceleration+horsepower+weight*displacement+we
ight*acceleration+weight*horsepower+displacement*acceleration+displacement*horsepower+ac
celeration*horsepower,data=Car_MPG_final)
step(null_model, scope=list(lower=null_model, upper=Full_model),direction="forward")
lm_model1=lm(mpg ~ weight + horsepower + acceleration + displacement +
weight:horsepower + horsepower:acceleration + weight:acceleration + horsepower:displacement
,data=Car_MPG_final)
summary(lm_model1)
influencePlot(lm_model1)
Car_MPG_final=Car_MPG_final[-c(14,388,395),]
Car_MPG_final=cbind(Car_MPG_final,weight1=1/Car_MPG_final$weight)
Car_MPG_final=cbind(Car_MPG_final,acceleration1=Car_MPG_final$acceleration^2)
Car_MPG_final=cbind(Car_MPG_final,displacement1=1/Car_MPG_final$displacement)
Car_MPG_final=cbind(Car_MPG_final,horsepower1=sqrt(Car_MPG_final$horsepower)
library(MASS)
boxcox(Car_MPG_final$mpg~Car_MPG_final$weight1+Car_MPG_final$displacement+Car_M
PG_final$acceleration+Car_MPG_final$weight1*Car_MPG_final$displacement+Car_MPG_fin
al$weight1*Car_MPG_final$acceleration+Car_MPG_final$displacement*Car_MPG_final$accel
eration, lambda = seq(-1, 1, length = 20))
Car_MPG_final=cbind(Car_MPG_final,mpg1=log(Car_MPG_final$mpg))
Car_MPG_final=cbind(Car_MPG_final,mpg2=1/sqrt(Car_MPG_final$mpg))
```

```
lm_model4=lm(mpg2  ~ weight1 + horsepower1 + acceleration1 + displacement1 +
weight1:horsepower1 + horsepower1:acceleration1 + weight1:acceleration1 +
horsepower1:displacement1,data=Car_MPG_final)
summary(lm_model4)
qqnorm(lm_r4)
qqline(lm_r4)
set.seed(123)
train_sample <- sample(398, 300)
Car_MPG_train=Car_MPG_final[train_sample,]
Car_MPG_test=Car_MPG_final[-train_sample,]
lm_model5=lm(mpg2  ~ weight1 + horsepower1 + acceleration1 + displacement1 +
weight1:horsepower1 + horsepower1:acceleration1 + weight1:acceleration1 +
horsepower1:displacement1,data=Car_MPG_train)
summary(lm_model5)
prediction=predict(lm_model5,Car_MPG_test)
predic=1/(prediction*prediction)
 plot(Car_MPG_test$mpg~predic)
abline(lm(Car_MPG_test$mpg~predic))
MAE <- function(actual, predicted) { mean(abs(actual - predicted))  }
MAE(Car_MPG_test$mpg,predic)
```



```
#Regression model with y transfromation as i/sqrt(y)
lm_model4=lm(mpg2  ~ weight1 + horsepower1 + acceleration1 + displacement1 +
   weight1:horsepower1 + horsepower1:acceleration1 + weight1:acceleration1 +
   horsepower1:displacement1,data=Car_MPG_final)
summary(lm_model4)

##
## Call:
## lm(formula = mpg2 ~ weight1 + horsepower1 + acceleration1 + displacement1 +
##     weight1:horsepower1 + horsepower1:acceleration1 + weight1:acceleration1 +
##     horsepower1:displacement1, data = Car_MPG_final)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.053316 -0.010159  0.000082  0.009552  0.059980
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.170e-01  4.990e-02   2.344 0.019598 *
## weight1                  4.387e+02  1.843e+02   2.380 0.017782 *
## horsepower1              1.843e-02  3.755e-03   4.908 1.37e-06 ***
## acceleration1           -3.927e-04  1.311e-04  -2.995 0.002928 **
## displacement1           -1.564e+01  6.370e+00  -2.455 0.014544 *
## weight1:horsepower1     -7.031e+01  1.805e+01  -3.894 0.000116 ***
## horsepower1:acceleration1  2.979e-05  8.981e-06   3.317 0.000999 ***
## weight1:acceleration1      4.215e-01  1.703e-01   2.475 0.013754 *
## horsepower1:displacement1  1.466e+00  6.576e-01   2.230 0.026337 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01582 on 381 degrees of freedom
## Multiple R-squared:  0.8214, Adjusted R-squared:  0.8177
## F-statistic: 219.1 on 8 and 381 DF,  p-value: < 2.2e-16
```