STAT 101A Kaggle Final Paper

Abstract:

1) Yu-Yuan Chang ID:205544670 (Yu-Yuan Chang lecture2 in Kaggle ranking)
2) Kaggle rank: 39
3) R square: 0.92007
4) Number for predictors used: 5
5) Number of Betas including: 37
6) BIC score: -3612.109

In this project, we have 23 predictors include numerical variable and categorical variables, it is unnecessary to use all of them to create a valid model. I divide the predictors into two groups, which are numerical and categorical. I use the function vif() to check if there is any multicollinearity issue in the model, I take out the variable with a high vif() value. Next, I want to check the interaction between predictors. I set all the predictors in the form lm(y~(x…x)^2). In this way, R language will automatically include all the possible interaction in the model, then I put it into function step(), but I encounter some problems. I found out strange that the step() function did not take out the insignificant variable in the model. I realized step() function will ignore the interaction predictors if we include all the interaction at once, so I manually type out all the possible interaction outcome and put it into step function and it works. In this way, the number of predictors and number of Betas are still large.

In the end of the quarter, I saw Professor demonstrate how does he choose the predictor to build his model and I decided to follow Professor method to do it. I put each variable into regression one by one and check the r square. I pick the variable with a relatively high r square but there are some special variables that make me pay more attention, which are

Model, Manufacturer and Ob. In this case Model and Ob is not a proper variable to predictor the Price, if we use table() function we can see that Model and Ob have many factors. If we put them in the model, we will have a high r square but an invalid model. After choosing the variable with a relatively high r square, the model only has 5 predictors. Next step, it is easier to deal with interaction because we only have a few but significant predictors in the model. I use interactionPlots() function to find out the potential interaction between predictors. I find out that Fuel.tank.capacity and AirBags have interaction between them. Finally, I have a model with r square 0.92007 and 37 Betas. The r square of my training model is 0.92007. The r square of my testing model is 0.9567.

Introduction:

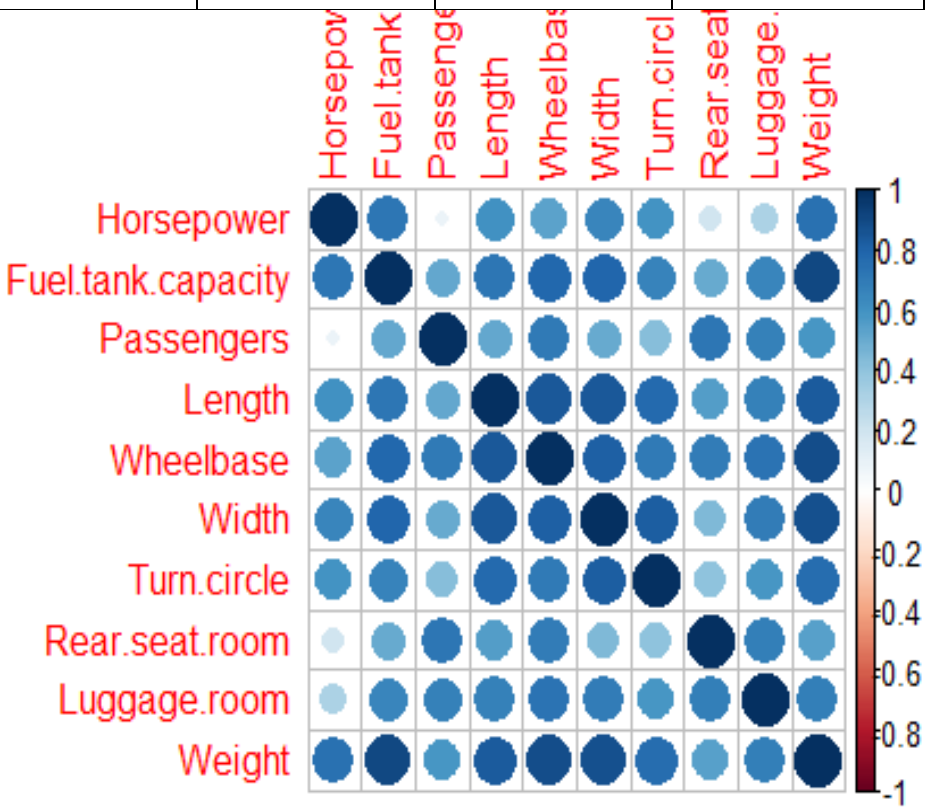
In this project, I am using PriceNew as a response variable and Horsepower, Fuel.tank.capacity, Manufacturer, AirBags, Fuel.tank.capacity*AirBags as predictor. The data size is 1500 with 25 predictors. My goal is predict the PriceNew base on the given Data set.

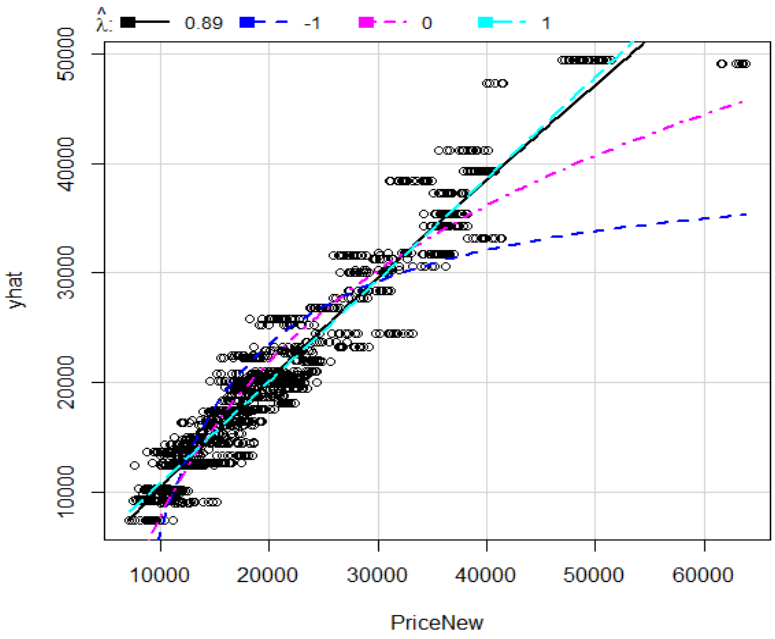
Methodology (Graph provide on below):

I use vif() function to check multicollinearity, based on the vif() chart and correlation plots attach below, some predictors are highly correlated with each other. After taking out the predictor who are highly correlated with other, we can see that the model does not have multicollinearity issue anymore. Next, I use step() function to choose variable base on AIC and BIC. I use power transfer and inverse response plot to adjust my response variable, base on the inverse response plot attach below, we can see that after adjusting the response variable, r square has increase. In order to find out the leverage point, I use leveragePlots() to check if there is any leverage point. Base on the leverage Plots attach below we can find some leverage points, to ensure they are not bad leverage point, I use a useful function

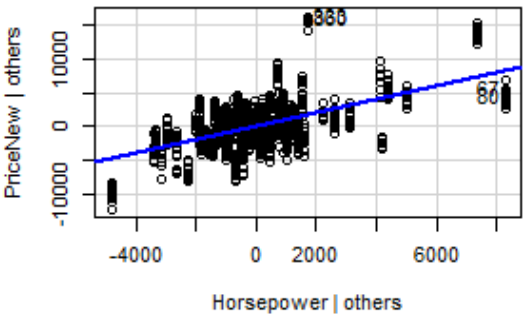
cooks.distance() to check if they are bad leverage points. They all seem to be good leverage points, so I did not remove them. I use the interaction plot to check which predictor have interaction. Based on the interaction plot below, I create a Fuel.tank.capacity*AirBags interaction term and put it into model.

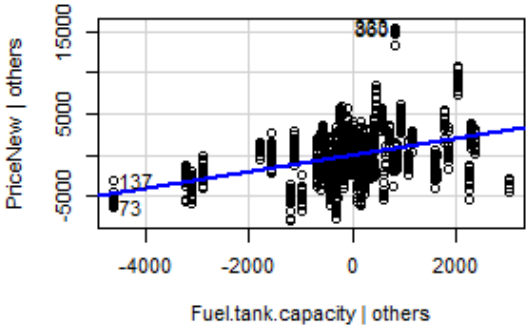
vif(final_lm)	GVIF	Df	GVIF^(1/(2*Df))
Horsepower	3.903781	1	1.975799
Fuel.tank.capacity	3.154317	1	1.776040
Manufacturer	11.527637	31	1.040219
AirBags	5.101067	2	1.502849

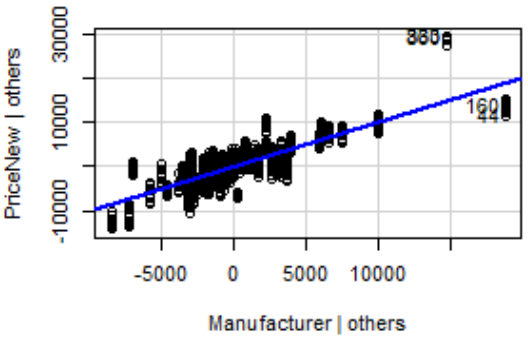


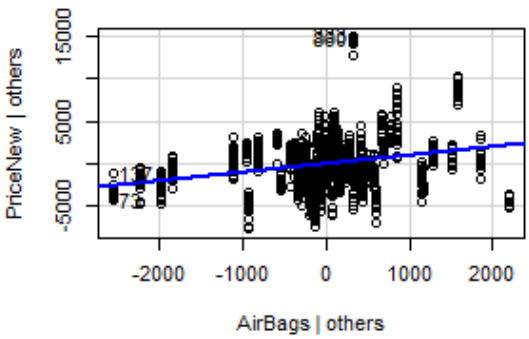


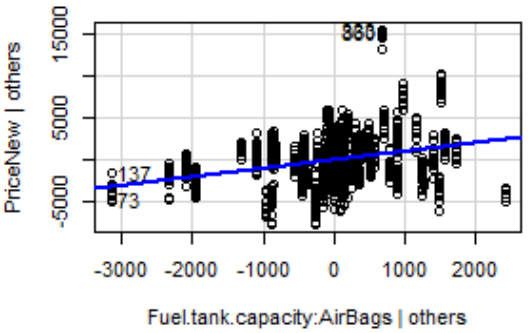
Leverage Plots

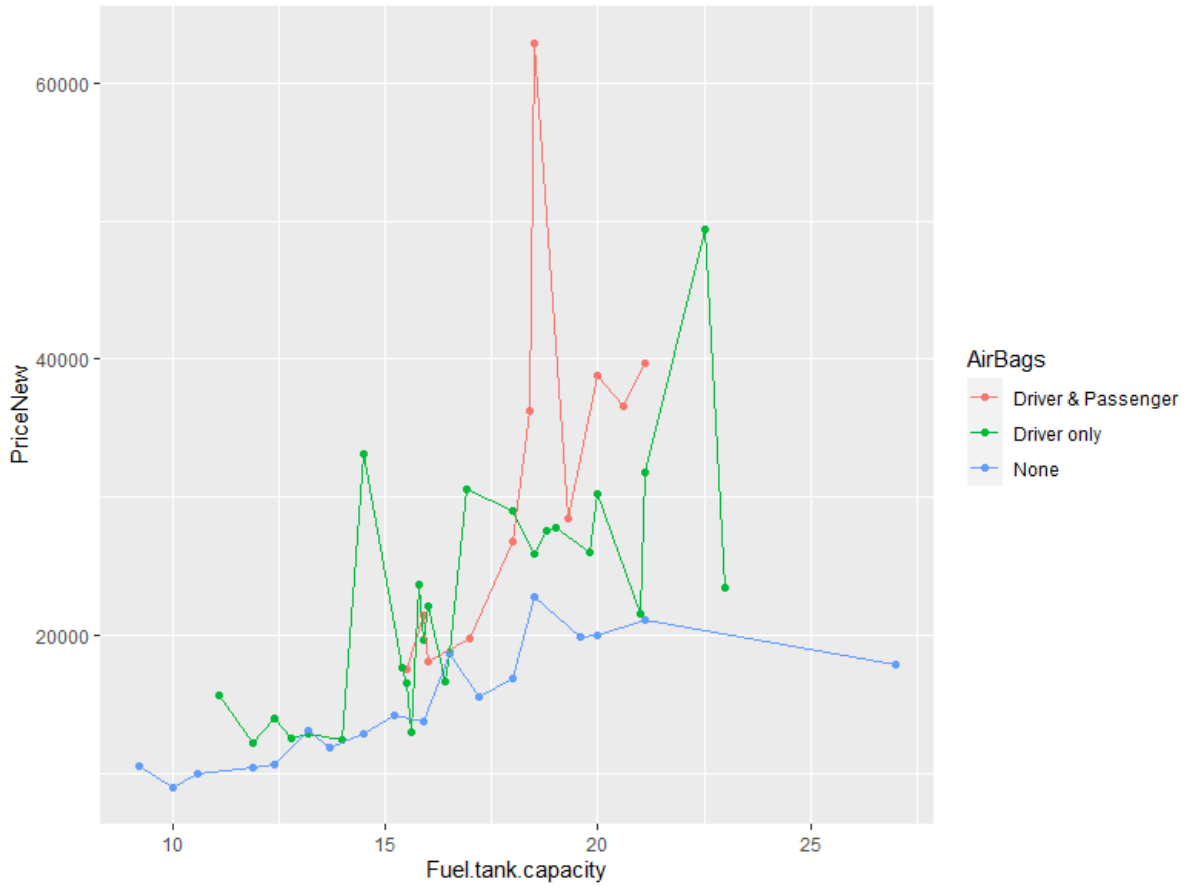












Result:

Finally, I create a model

lm(formula = PriceNew^0.2331 ~ Horsepower + Fuel.tank.capacity + Manufacturer + AirBags + Fuel.tank.capacity:AirBags, data = carsTrain)

Residual standard error: 0.277 on 1462 degrees of freedom

Multiple R-squared: 0.9282, Adjusted R-squared: 0.9263

F-statistic: 510.5 on 37 and 1462 DF, p-value: < 2.2e-16

Dicussion:

With the given Data set it is not hard to get a high r square, I spent most of the time on reduce the number of Betas. I tried several ways to reduce the predictors, but I think it is

necessary to put the predictor Manufacturer in the model. Without Manufacturer, r square drop too much.

Limitations and conclusions:

My model has a limit that the predictor Manufacturer is important, if we do not have the Manufacturer Data, then we have to consider another model. So far, I test the model on testing data and Kaggle website. It has a pretty good result.

References:

Almohalwas, Akram. (2021, March). The lecture slides.

Sheather, Simon. (1979). A Modern Approach to Regression With R.

James, Gareth. Witten, Daniela. Hastie, Trevor. Tibshirani, Robert. (2013, June) An Introduction to Statistical Learning.