**CSC 423** **Group project**

Group Member: Samir Satam, Xi Peng, Zack Larsen

**Data Source:**

<http://mlr.cs.umass.edu/ml/datasets/Automobile>

This data set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. More detail about the dataset can be found at the website above.

**Executive summary**

The dataset that we decided to use in the building of our model comes from umass.edu and contains data on automobiles. The question we asked ourselves in trying to identify the dependent variable we would use was: “which features of an automobile contribute most to its price?” We used price as our dependent variable and built a regression model to assess how well we could predict price using least-squares regression. We involved many variables in our search to come up with a good model, and some of them were qualitative while most were quantitative. A summary of the variables is listed below. We used scatterplots to help us visually assess which variables were correlated, and we used stepwise regression in R to help us identify which variables to include in the model. Throughout this document, the reader will find a thorough model-building and regression process including scatterplots, interaction plots, and residual plots where appropriate. We make a regressed model and adjust it after we check assumptions, check for outliers and assess the need for transformations, etc. Finally, we end up with our final model, we interpret the beta coefficients of the model, and end with our closing statement.

* **Correlation and Scatterplots**

We calculated the coefficients of correlation on all quantitative variables. The independent variables WHEEL, LENGTH, WIDTH and CURB are strongly correlated. The independent variables CC, HP, CTYMPG and HWYMPG are also highly correlated. We kept CC and WHEEL and removed LENGTH, WIDTH, CURB, HP, CTYMPG and HWYMPG.

* **Stepwise Selection**

The last step of stepwise selection is shown below.

*## Step: AIC=2315.16  
## PRICE ~ CC + WHEEL + MAKE\_merc + FUELSYS\_mpfi + ASPIRATION +   
## MAKE\_bmw + MAKE\_porsche + CYL\_8 + CYL\_4 + MAKE\_saab + MAKE\_jag +  
## DRIVE\_fwd + CYL\_5 + MAKE\_volvo + MAKE\_mazda + MAKE\_honda +   
## NORMLOSS + DOOR + MAKE\_nissan + RISK\_1 + HEIGHT + BODY\_hdtp +   
## MAKE\_mit*In this step, the AIC is lowest. Therefore, the variables selected in this step are the independent variables used for building model.

In order to make sure that no correlation existed between these variables, we calculated variance inflation factor (VIF).

*## CC WHEEL MAKE\_merc FUELSYS\_mpfi ASPIRATION   
## 8.366318 6.105476 4.367322 2.390864 1.411168   
## MAKE\_bmw MAKE\_porsche CYL\_8 CYL\_4 MAKE\_saab   
## 1.304744 1.142901 2.120840 3.569373 1.504818   
## MAKE\_jag DRIVE\_fwd CYL\_5 MAKE\_volvo MAKE\_mazda   
## 1.423453 2.896458 3.745741 1.794088 1.325764   
## MAKE\_honda NORMLOSS DOOR MAKE\_nissan RISK\_1   
## 1.245998 2.540819 1.842463 1.881037 1.937511   
## HEIGHT BODY\_hdtp MAKE\_mit   
## 3.304163 1.285699 1.416932*

From the result above, we found that there is no value larger than 10. Therefore, we used all the variables selected by the stepwise selection to build the model. Among these variables, CC, WHEEL, NORMLOSS and HEIGHT are quantitative variables and the rest are qualitative variables.

* **Model Building**

1. **Determine the interaction terms between qualitative variables.**

We constructed in the interaction plots by R. An example is shown below

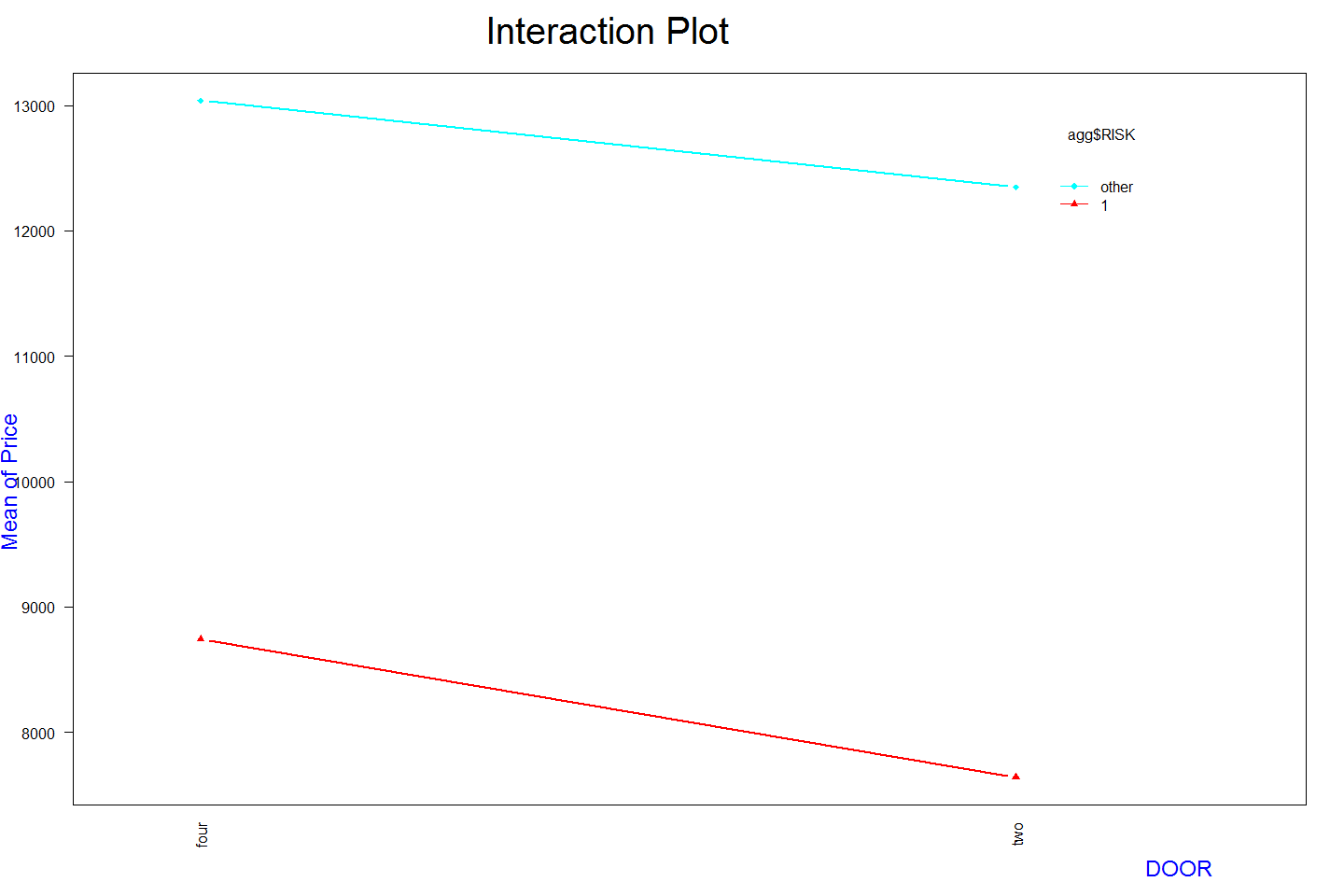


Figure 1 Interaction plot of RISK and DOOR

This plot shows the relationship between average price and the number of doors when the risk rating is ‘1’ and other. From the plot above, two lines are parallel, showing that there is no interaction term between qualitative variables RISK and DOOR.

1. **Determine the interaction terms between qualitative variables and quantitative** **variables.**

We constructed scatterplots between price and each quantitative variable. The points of the scatterplots are drawn in different colors, representing different levels of the qualitative variables. See the R code and output part for all scatterplots. Note that we didn’t construct any scatterplots based on the values of MAKE, since each level of values of MAKE didn’t have enough data to show whether the interaction term existed between MAKE and one of the quantitative variables or not.

One example is shown below.

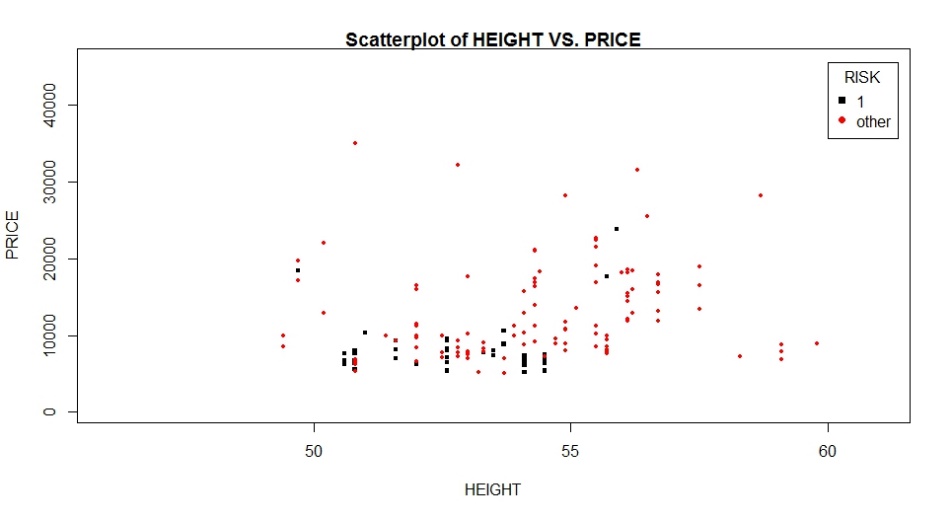


Figure 2 Scatterplot of height and price (black points representing risk rating is 1, red points representing other risk ratings)

From the plot above, we found that the value of RISK didn’t change the effect of HEIGHT on PRICE. Therefore, there is no interaction term between quantitative variable HEIGHT and qualitative variable RISK.

In general, according to the all scatterplots, we concluded that there is no interaction term between any qualitative variables and quantitative variables.

1. **Determine the interaction terms between quantitative variables.**

Since there are only four interest quantitative variables, and it is very hard to construct plots to show the interaction term between any quantitative variables. We will include all interaction terms between any quantitative variables at the beginning model and construct t-tests on corresponding beta coefficient to determine whether an interaction term exists or not.

1. **Determine the higher order terms**

We used the scatterplots between dependent variable and each independent quantitative variables to detect higher order terms before building model. Therefore, the plots used in detecting the interaction terms between qualitative and qualitative variables can also be used here.

One example is shown in Figure 2. The scatterplot of HEIGHT and PRICE shows that the relationship is not a straight line. Therefore, we should include the quadratic term of HEIGHT in the beginning model.

In general, we conclude that we should include the quadratic terms of all quantitative variables at the beginning model.

**In conclusion, the beginning model is:**

* **Model Regression and Selecting the Best Model**

We began model regression with the beginning model shown above. It turned out that global F statistic was significant and adjusted r-square was 0.9479, showing the model is statistically useful. However, the p-values of higher order terms and some interaction terms were larger than 0.05. We adjusted the model by removing these ‘bad’ terms and did least-squares regression again. After we constructed and regressed the model five times, we got the model as below

The p-value of global F statistic was less than 2.2e-16 and adjusted r-square is 0.948, showing the model was statistically useful. However, the p-value of t-test on the beta of BODY\_hdtp is still larger than 0.05. Noting there was an interaction term between BODY\_hdtp and FUELSYS\_mpfi, therefore, the beta of BODY\_hdtp was not important. However, in order to see whether we could get better model or not, we tried to remove the interaction term and did F-test to compare nested models.

*anova(reg\_model5, reg\_model6)*

*## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 135 242402402   
## 2 136 252755458 -1 -10353056 5.7659 0.0177 \**

Since the F-test is significant, we concluded that the interaction term did affect the dependent variable. Therefore, we used the model above as the best regression model.

* **Checking of Assumptions**

1. We used partial residual plots and concluded that the model satisfied the assumption that the mean of the probability distribution of random errors is 0.
2. We used scatterplot of residuals and dependent variable and concluded that the model satisfied the assumption that the variance of the probability distribution of random errors is constant for all settings of the independent variables.
3. We used histogram, density and qq-plot and concluded that the model satisfied the assumption that the probability distribution of random errors is normal.
4. We used Durbin-Vastson test and concluded that the model satisfied the assumption that the errors associated with any two different observations are independent.

* **Detecting Outliers**

outlierTest(reg\_model5)

## rstudent unadjusted p-value Bonferonni p  
## 173 4.564402 1.1209e-05 0.0017487

The outlier test showed that point NO. 173 is an outlier. The qq-plot below also showed it. Since there is only one outlier, it’s rarely caused by the lack fit of the model. Therefore, we directly removed this outlier.

* **Transformations**

In the residuals plot, the residuals are randomly around 0 without any distinctive patterns. In addition, there is only one outlier. Therefore, we believed that the model is fitted well enough so that there is no need to do any transformation.

* **The Best Regression Model and its Explanation.**

After removing the outlier, we did least-squared regression again. The partial R output is shown below

*## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12590.177 4282.500 -2.940 0.00387 \*\*   
## RISK\_1 -663.744 306.232 -2.167 0.03197 \*   
## FUELSYS\_mpfi 1331.340 307.203 4.334 2.86e-05 \*\*\*  
## ASPIRATION 2749.671 314.516 8.743 8.31e-15 \*\*\*  
## DRIVE\_fwd -1526.245 352.720 -4.327 2.93e-05 \*\*\*  
## DOOR -775.722 276.547 -2.805 0.00578 \*\*   
## BODY\_hdtp 695.826 1002.605 0.694 0.48887   
## CYL\_4 -2392.240 541.742 -4.416 2.05e-05 \*\*\*  
## CYL\_5 4094.687 938.772 4.362 2.55e-05 \*\*\*  
## CYL\_8 14509.793 1829.579 7.931 7.60e-13 \*\*\*  
## MAKE\_bmw 4494.401 724.176 6.206 6.32e-09 \*\*\*  
## MAKE\_honda 1914.473 397.844 4.812 3.97e-06 \*\*\*  
## MAKE\_jag 10208.165 1513.551 6.745 4.19e-10 \*\*\*  
## MAKE\_mazda 1832.358 447.860 4.091 7.36e-05 \*\*\*  
## MAKE\_merc 4606.580 1216.543 3.787 0.00023 \*\*\*  
## MAKE\_nissan 1239.024 432.651 2.864 0.00486 \*\*   
## MAKE\_porsche 10310.664 1347.637 7.651 3.49e-12 \*\*\*  
## MAKE\_saab 4760.126 638.903 7.450 1.03e-11 \*\*\*  
## MAKE\_volvo 2730.922 526.606 5.186 7.75e-07 \*\*\*  
## CC 22.167 9.718 2.281 0.02412 \*   
## WHEEL 311.949 47.887 6.514 1.36e-09 \*\*\*  
## NORMLOSS 12.438 4.403 2.825 0.00545 \*\*   
## HEIGHT -179.395 78.329 -2.290 0.02357 \*   
## FUELSYS\_mpfi\_BODY\_hdtp -2776.920 1288.417 -2.155 0.03293 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1251 on 134 degrees of freedom  
## Multiple R-squared: 0.9613, Adjusted R-squared: 0.9546   
## F-statistic: 144.7 on 23 and 134 DF, p-value: < 2.2e-16*

*CV=100\*(1251/mean(AutoModel\_WithoutOutliers$PRICE))  
CV*

*## [1] 10.96758*

**Therefore, the best regression model is:**

**Assess the adequacy of the best model:**

The Global F-statistic is 144.7 and the p-value is less than 2.2e-16, showing the model is statistically useful. The adjusted R-square is 0.9546, showing that about 95.46% of dependent variable error can be explained by the model. The standard error is 1251, showing about 95% of the observed prices lie within 2502 of their respective predicted price. The C.V. is 10.97%, showing that the value of standard error s for the regression model is about 11% of the value of the sample mean price.

**Model explanation / interpretation of the betas:**

Our model effectively predicts the price of an automobile using certain features of the automobile itself. Most of these features are descriptive of the engine or drivetrain of the vehicle, and some are more surface-level features of the car such as make and number of doors. The variable in our final model which contributes most to the price is CYL\_8, meaning that the number of cylinders in the engine has a dramatic effect on the price of the vehicle. Intuitively, people might prefer a more powerful car which can travel at higher speeds, and number of cylinders is an indication of power. CYL\_4 has a negative effect on the price, suggesting perhaps that people are willing to pay less for a less powerful car that has only 4 cylinders.

The variable HEIGHT has a slight negative effect on price, indicating that the taller a car is, the less money people are willing to pay for it. This might be explained by the higher risk of a taller can which can rollover more easily. Also, people living in urban areas need to be able to clear parking garage ceilings, and a shorter car is ideal for tight spaces such as that. Some people also don’t like to have to climb up into an SUV or truck, and a car that is lower to the ground will be easier for them to get into, increasing the amount they might be willing to spend on the car.

Interpreting the intercept for our model is nonsensical, because the independent variables HEIGHT, WHEEL (wheel base), CC (engine size) and NORMLOSS (average loss payment per car per insured vehicle year) cannot be 0. We can only use this model to predict price when all independent variables fall in the area of observations.

Looking at the categorical variables for car make, we see that Jaguar and Porsche have a very strong effect on the model. We would expect this, because those car manufacturers offer a line of very high-priced luxury vehicles. Even if these cars don’t have a high number of cylinders or other features that consumers desire, they can still charge a high price for brand loyalty. Therefore, we would expect that a car made by Porsche, for example, would have a higher price than, say, a Nissan, all else being equal.

**Closing statement**

We began our analysis by cleaning up the dataset to remove rows containing NA or question mark. We removed variable that had only one level of values. We used R to create dummy variables for the qualitative variables. Once this was done, we removed correlated variables, applied stepwise regression to choose the variables for model. We also calculated the VIF for the model to ensure that the VIF for each independent variable was less than 10. Next, we constructed interaction plots and scatterplots to determine the interaction terms and higher order terms in the model and formulated the beginning model. Then we ran regression and constructed partial F-test to select the best model with significant global F-statistic and adjusted R-square as well as significant t statistics on some betas of important terms. We also analyzed the residual plots. We verified that the model assumption held up, which they did. We found an outlier and removed it. Finally, we ran regression on our final model and obtained an adjusted r-squared value of 0.9546, which was very good. We can therefore predict the price of an automobile very accurately with our model. Although our model contains many variables and is not as parsimonious as we would perhaps like for it to be, we are happy with the result compared to the original 50+ variables and believe that we have performed an exhaustive search for the best model to use.