USA Car Sales Project Written Report

DATA 603 Final Project

Yong

2022-04-19

Contents

Chapter 1: Introduction

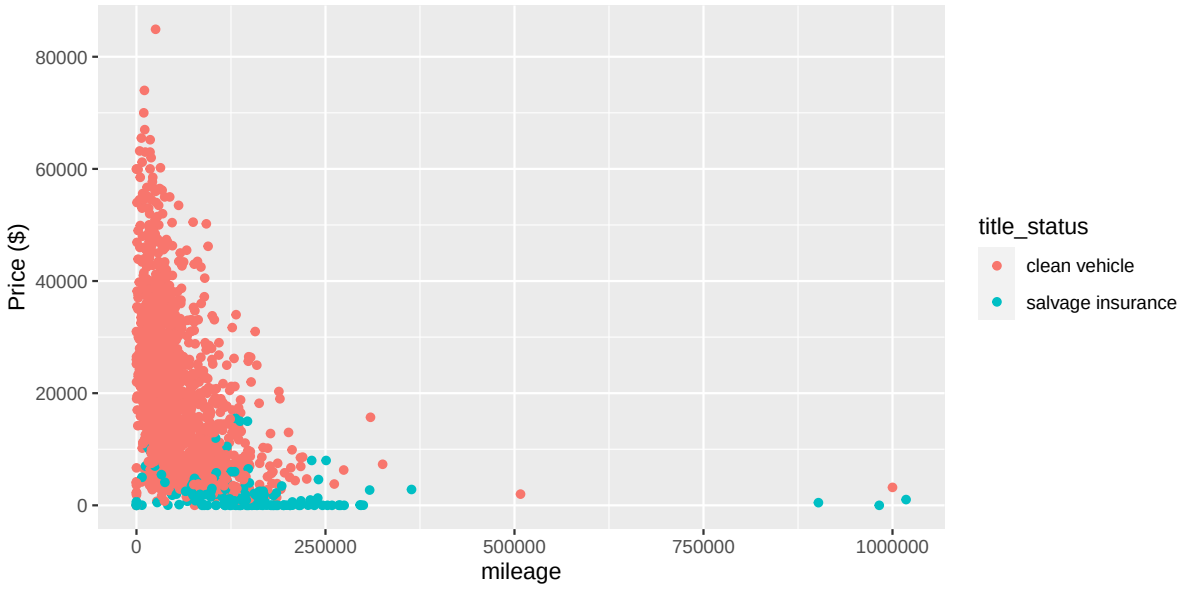
Owning a car is typically quite expensive, yet it is deemed a necessity by many households, especially in North America. In a 2021 study, 276 million vehicles were found to be registered with the US Department of Transportation. Additionally, 90% of US households have access to a vehicle and the average person was found to spend 100 minutes a day in a car (Borrelli, 2021). This highlights the significance of vehicles in our lives and the importance of selecting the correct vehicle.

The used car market is of particular interest as, according to insights from McKinsey & Company, "the US used-car market is more than twice the size of the new-car segment and is outpacing it in growth" (Ellencweig et.al, 2019). The used car market in the United States may grow by close to 4 million units in the next 4 years (Yahoo!, August 2021). Additionally, purchasing a used car involves a more complicated decision making process than a new vehicle. As such, "used-car buyers spend about 40 percent more time researching online during the buying process than new-car buyers do" (Ellencweig et.al, 2019).

There are many factors that will affect the cost of a car. In order to be conscientious consumers, we were interested in investigating the independent variables that can potentially influence the prices of cars. In this report, we investigate the various features from the "US Cars Data Set" (Alsenani, 2019) such as the vehicle registration year, state, brand, model, mileage, status (clean or used), etc.

There are a number of commonly known factors that impact an individual's decision when buying a used vehicle and affect the price. For example, a car with a clean title (no reported major accidents) will likely be worth more than one with salvage or rebuild status (has been in accident). Vehicles with higher mileage will have a lower resale price compared with vehicles with lower mileage (see the following figure). These assumptions were verified and quantified by conducting appropriate statistical tests.

```
setwd("/Users/yongpengfu/OneDrive - University of Calgary/Master in Data Science and
↪  Analytics/Courses/DATA 603 L03 - (Fall 2021) - Statistical Modelling with
↪  Data/Assignment/project")
# setwd("/Users/jessk/Dropbox/School/Data 603/Project")
cars <- as_tibble(read.csv("USA_cars_datasets.csv", header = T, row.names = 1))
ggplot(cars, aes(x = mileage, y = price, color = title_status)) + geom_point() +
↪  labs(y="Price ($)")
```



Furthermore, we were interested to see if the price will be impacted by including interaction variables. For example, can both mileage and age of car together play a key role in the price prediction of a vehicle?

All of these factors were considered using methodologies learned in DATA 603, including hypothesis testing, model diagnostics, ANOVA, etc. The results of these findings were used to find the best multiple linear regression model to predict the price of a vehicle in the United States.

Chapter 2: Methodology

Data Source

The data was collected from Kaggle and was originally scraped from AUCTION EXPORT during April 2020 (Alsenani, 2019). AUCTION EXPORT is a leading internet company specializing in the purchase and export of cars from the United States and Canada to car buyers around the world. It has served tens of thousands of active buyers around the world, allowing them to filter vehicles on a variety of attributes, including salvage/clean-title, vehicle make/model and price, before bidding. This specific data set covers a wide range of vehicles with different brands, models, colors, locations, and registration years. As such, the dataset is a good representation of how the sale price of the vehicle is affected by various factors.

Variable Explanations

The data set includes several variables that may impact vehicle prices in the US. The dependent variable is the vehicle price, while all other variables in the data set are independent variables. The 12 columns featured in the data set are described below:

1. Price = The sale price of the vehicle in the ad ($). - **Quantitative Dependent Variable**

2. Brand = The brand of car. - **Qualitative Independent Variable**

3. Model = The model of the vehicle. - **Qualitative Independent Variable**

4. Year = The vehicle registration year. - **Quantitative Independent Variable**

5. Title_status = This feature included binary classification, which are clean title vehicles and salvage insurance. - **Qualitative Independent Variable**

6. Mileage = The number of miles traveled by vehicle (in mile). - **Quantitative Independent Variable**

7. Color = Color of the vehicle. - **Qualitative Independent Variable**

8. Country = Only one variable, this data set focuses on car auction statistic in United States (USA). - **Qualitative Independent Variable**

9. State/City = The location in which the car is being available for purchase. - **Qualitative Independent Variable**

10. Vin = The vehicle identification number is a unique collection of 17 characters (digits and capital letters) for a specific car. - **Qualitative Independent Variable**

11. Lot = A lot number is an identification number assigned to a particular quantity or lot of material from a single manufacturer. For cars, a lot number is combined with a serial number to form the Vehicle Identification Number. - **Qualitative Independent Variable**

12. Condition = The time that is still available to place a bet on the auction. - **Quantitative Independent Variable**

Modeling Plan

We will use a series of methods to select the best model both conceptually and statistically. First, we will clean up the data set after doing some initial exploration. It is best practice to understand the data set and what each variable means. We will remove unrelated features based on our understanding of vehicle properties and reduce row records due to extremely low value counts. We will also use visualization methods to gain a better understanding of the data set. For example, using a scatter plot to reveal underlying trends between variables or using other visualization methods such as a boxplot or histogram to reveal the distribution of the data and detect outliers.

To build our main effects model, we will start off with all the features in the linear regression model and find features that explain the most variance of the response variable. Given the many levels for all the qualitative features, we will avoid using auto selection programs like step-wise selection and elimination procedure. Instead, we will compare R^2_{adj} and RMSE in different feature combinations. Variables that can explain a large portion of the response variance will be retained and those that only contribute a little will be discarded. The final main effects model will be confirmed by using individual t-tests for quantitative features and partial F-tests for qualitative features.

Once the main effects model is confirmed, we will use selection programs like step-wise selection and elimination procedure to test for significant interaction terms. These results will be confirmed with individual t-tests. We will then identify higher-order terms by using a pairwise scatterplot and by conducting individual t-tests.

After our final model is complete, we will use the following diagnostic tests to evaluate the model:

- Linearity Assumption - Review the "Residuals vs. Fitted plot"
- Independence Assumption - Not applicable because the data set is not time series data
- Normality Assumption - Plot a histogram of the data, review the "Q-Q plot", and use the Shapiro-Wilk normality test
- Equal Variance Assumption (heteroscedasticity) - Review the "Scale-Location plot" and use the Breusch-Pagan test
- Multicollinearity - Review ggpairs plot and variance inflation factors (VIF)
- Outliers - Review the "Residuals vs Leverage plot", check Cook's distance and leverage points

The model will be adjusted as needed if any of the assumptions above are violated. Once a final model is found, the coefficients will be interpreted. Lastly, with 20% of the data randomly retained for testing purpose, 80% will be used for the modeling training and the model will be used to predict car sale prices.

Data Cleaning and Assumptions

The data was loaded into RStudio using the "read.csv" function. The data we needed for our regression analysis included the dependent variable "Price ($)" as a function of the 12 independent variables listed above. We completed preliminary data cleaning before beginning our regression analysis.

Upon initial investigation of the data set, we identified and removed the following unrelated predictors: **vin, lot, condition**. These variables act as a unique ID for each vehicle and do not impact sale price fluctuation.

There were two levels included in the **country** predictor: US and Canada. However, Canada only made up 0.3% of the total amount of records and the US made up the remaining 99.7%. In order to simplify

the analysis, all rows with Canada recorded in the **country** column were dropped. Removing this level allowed us to avoid any effects as a result of the differences in markets between the two countries.

Finally, some records were found to be missing information (NA) or have a sale price of $0. These rows were removed to avoid complications in the analysis.

In summary, 7 out of the original 11 predictors were used in our regression analysis to predict car sale prices as follows:

```
kable(data.frame(Final_Variable = names(cars)[c(1:7, 10)],
                 Property = c("Quantitative Dependent Variable", "Qualitative
                 ↪  Independent Variable", "Qualitative Independent Variable",
                 ↪  "Quantitative Independent Variable", "Qualitative Independent
                 ↪  Variable", "Quantitative Independent Variable", "Qualitative
                 ↪  Independent Variable","Qualitative Independent Variable")))
```

Final_Variable	Property
price	Quantitative Dependent Variable
brand	Qualitative Independent Variable
model	Qualitative Independent Variable
year	Quantitative Independent Variable
title_status	Qualitative Independent Variable
mileage	Quantitative Independent Variable
color	Qualitative Independent Variable
state	Qualitative Independent Variable

Chapter 3: Results

Variable Selection Procedure

There is 1 dependent variable (**price**) with 11 independent features (**brand, model, year, title_status, mileage, color, vin, lot, state, country, condition**) in the data set. As mentioned previously, we decided to remove unrelated features after assessing the data set further:

1. We removed the **vin, lot, condition** variables as we deemed them to be unrelated to a vehicle's price. Both **vin** and **lot** are just used for unique identification purposes for a vehicle. **Condition** is not a specific property of a vehicle as it just indicates the remaining time a customer can place a bid for this vehicle.
2. Then, we also removed any rows that had an NA value in any column because NA may cause problems in regression modeling.
3. There were only two countries listed in the **country** qualitative variable: USA and Canada. However, Canada only made up 0.3% of the total amount of records and the US made up the remaining 99.7%. In order to simplify the analysis, all rows with Canada recorded in the **country** column were dropped. Removing this level allowed us to avoid any effects as a result of the differences in markets between the two countries. Another reason we excluded Canada is because we will need to bin the **state** variable into regions based on state location.
4. We will also randomly retain 20% of the data for testing purpose, while 80% is used for the modeling training.

```
setwd("/Users/yongpengfu/OneDrive - University of Calgary/Master in Data Science and
↪  Analytics/Courses/DATA 603 L03 - (Fall 2021) - Statistical Modelling with
↪  Data/Assignment/project")
cars <- as_tibble(read.csv("USA_cars_datasets.csv", header = T, row.names = 1))
cars <- cars %>% dplyr::select(-c(vin, lot, condition))
#remove any NA in the dataframe in case there is any
cars <- cars[rowSums(is.na(cars))==0,]


#Remove the rows that belong to Canada because Canada only accounts for 0.3% of the
↪  total vehicles
cars <- cars %>% filter(country != " canada")
#afterwards, country column is dropped because there is only "usa" in this column
cars <- cars %>% dplyr::select(-country)
kable(head(cars,2))
```

price	brand	model	year	title_status	mileage	color	state
6300	toyota	cruiser	2008	clean vehicle	274117	black	new jersey
2899	ford	se	2011	clean vehicle	190552	silver	tennessee

```
#set aside 20% data for testing the model eventually
set.seed(12345)
n = nrow(cars)
index = sample(1:n, round(0.8*n), replace = F) #random select 80% data points
↪  without replacement.
cars <- cars[index,] #80% of the data to build a model
test <- cars[-index,] #the rest 20% of the data for testing the model
```

After cleaning and conceptualizing the data, we selected the best variables based on statistical methods. The first step was to build the first-order model that included all variables. We hypothesized the first-order model to be as follows:

$$price = \beta_0 + \beta_1 * year + \beta_2 * mileage + \beta_3 * title_status + \beta_4 * state + \beta_5 * brand + \beta_6 * model + \beta_7 * color + \epsilon$$

We initially found that there were too many levels for some qualitative variables to run the Stepwise/Forward/Backward Selection procedures. For example, the **brand** variable had 28 levels, the **model** variable had 127 levels, the **color** variable had 49 levels, and the **state** variable had 44 levels. In order to address this issue, we attempted to group similar levels together. For example, we created bins for states based on their geographical location (Northeast, South, North Central, West). Our hypothesis was that neighbor states had approximately the same vehicle sale price fluctuation. Once the number of levels was reduced, we were able to conduct Stepwise/Forward/Backward Selection. From this analysis we found that the **year, mileage, title_status** variables were all found to be significant with very small individual p-values. However, the resulting R^2_{adj} was found to be very low (0.25-0.35) due to the loss of information when combining levels.

R^2_{adj} is an important attribute as it indicates how much variation in the sale price can be explained by using a model. Due to the low R^2_{adj} values when combining levels, we decided to retain all levels even though it mean we would be unable to use the Stepwise/Forward/Backward Selection methods. Instead, we selected the main-effects predictors based on individual t-test results with $\alpha = 0.05$, RMSE, and R^2_{adj}. We compared several models with different combinations of predictors in order to minimize the risk of overlooking significant predictors when non-significant levels of the variables resulting in some of the variance being taken away.

In each model, we always included **year, mileage, title_status** because they were found to be significant predictors for sale **price** when we conducted the selection methods with the variables grouped. We then compared different combinations of remaining predictors and analyzed their RMSE and R^2_{adj} values.

```
#define levels for each quantitative variables.
cars$title_status <- factor(cars$title_status, levels = c("salvage insurance",
↪  "clean vehicle"))
#build the first-order model with all variables included
cars_model1_full <- lm(price~ year +  mileage + factor(title_status) + factor(state)
↪  + factor(brand) + factor(model)+ factor(color), data = cars)
cars_model2_nocolor <- lm(price~ year +  mileage + factor(title_status) +
↪  factor(state) + factor(brand) + factor(model), data = cars)
cars_model3_nomodel <- lm(price~ year +  mileage + factor(title_status) +
↪  factor(state) + factor(brand) + factor(color), data = cars)
cars_model4_nostate <- lm(price~ year +  mileage + factor(title_status) +
↪  factor(brand) + factor(model)+ factor(color), data = cars)
cars_model5_nobrand <- lm(price~ year +  mileage + factor(title_status) +
↪  factor(state) + factor(model) + factor(color), data = cars)
cars_model6_nostate_color <- lm(price~ year +  mileage + factor(title_status) +
↪  factor(brand) + factor(model), data = cars)
cars_model7_nostate_brand <- lm(price~ year +  mileage + factor(title_status) +
↪  factor(model) + factor(color), data = cars)
cars_model8_nostate_model <- lm(price~ year +  mileage + factor(title_status) +
↪  factor(brand) + factor(color), data = cars)
cars_model9_nobrand_color <- lm(price~ year +  mileage + factor(title_status) +
↪  factor(state) + factor(model), data = cars)
```

```r
cars_model10_nobrand_model <- lm(price~ year +  mileage + factor(title_status) +
↪  factor(state) + factor(color), data = cars)
cars_model11_nomodel_color <- lm(price~ year +  mileage + factor(title_status) +
↪  factor(state) + factor(brand), data = cars)
cars_model12_nostate_brand_color <- lm(price~ year +  mileage + factor(title_status)
↪  + factor(model), data = cars)
cars_model13_nostate_brand_model <- lm(price~ year +  mileage + factor(title_status)
↪  + factor(color), data = cars)
cars_model14_nostate_model_color <- lm(price~ year +  mileage + factor(title_status)
↪  + factor(brand), data = cars)
cars_model15_nocolor_brand_model <- lm(price~ year +  mileage + factor(title_status)
↪  + factor(state), data = cars)
cars_model16_nostate_brand_model_color <- lm(price~ year +  mileage +
↪  factor(title_status), data = cars)

#get the summary for the above models
model_compare <- data.frame(
  model = c("Full Model", "Full Model without Color", "Full Model without Model",
  ↪  "Full Model without State", "Full Model without Brand", "Full Model without
  ↪  State_Color", "Full Model without State_Brand", "Full Model without
  ↪  State_Model", "Full Model without Brand_Color", "Full Model without
  ↪  Brand_Model", "Full Model without Model_Color", "Full Model without
  ↪  State_Brand_Color", "Full Model without State_Brand_Model", "Full Model
  ↪  without State_Model_Color","Full Model without Brand_Model_Color","Full Model
  ↪  without State_Brand_Model_Color"),

  Adjusted_R_Square = c(summary(cars_model1_full)$adj.r.squared,
  ↪  summary(cars_model2_nocolor)$adj.r.squared,summary(cars_model3_nomodel)$adj.r.squared,summary(c
  ↪  summary(cars_model10_nobrand_model)$adj.r.squared,
  ↪  summary(cars_model11_nomodel_color)$adj.r.squared,
  ↪  summary(cars_model12_nostate_brand_color)$adj.r.squared,
  ↪  summary(cars_model13_nostate_brand_model)$adj.r.squared,
  ↪  summary(cars_model14_nostate_model_color)$adj.r.squared,
  ↪  summary(cars_model15_nocolor_brand_model)$adj.r.squared,
  ↪  summary(cars_model16_nostate_brand_model_color)$adj.r.squared),

  RMSE = c(summary(cars_model1_full)$sigma,
  ↪  summary(cars_model2_nocolor)$sigma,summary(cars_model3_nomodel)$sigma,summary(cars_model4_nosta
  ↪  summary(cars_model10_nobrand_model)$sigma,
  ↪  summary(cars_model11_nomodel_color)$sigma,
  ↪  summary(cars_model12_nostate_brand_color)$sigma,
  ↪  summary(cars_model13_nostate_brand_model)$sigma,
  ↪  summary(cars_model14_nostate_model_color)$sigma,
  ↪  summary(cars_model15_nocolor_brand_model)$sigma,
  ↪  summary(cars_model16_nostate_brand_model_color)$sigma),

  stringsAsFactors = F
)

#plot Adjusted R square and RMSE
model_compare$Model_ID <- factor(seq(1,16))
```

```
kable(model_compare)
```

model	Adjusted_R_Square	RMSE	Model_ID
Full Model	0.7264025	6340.556	1
Full Model without Color	0.7176083	6441.651	2
Full Model without Model	0.4303862	9148.745	3
Full Model without State	0.6966362	6676.566	4
Full Model without Brand	0.7075274	6555.621	5
Full Model without State_Color	0.6883968	6766.627	6
Full Model without State_Brand	0.6790599	6867.257	7
Full Model without State_Model	0.3704802	9617.805	8
Full Model without Brand_Color	0.6943015	6702.209	9
Full Model without Brand_Model	0.3217828	9982.876	10
Full Model without Model_Color	0.4314005	9140.596	11
Full Model without State_Brand_Color	0.6666710	6998.546	12
Full Model without State_Brand_Model	0.2419431	10554.124	13
Full Model without State_Model_Color	0.3731452	9597.426	14
Full Model without Brand_Model_Color	0.3126719	10049.705	15
Full Model without State_Brand_Model_Color	0.2330538	10615.824	16

```
ggplot(melt(model_compare), aes(x = Model_ID, y = value, fill = variable)) +
↪  geom_bar(stat = "identity", position = "dodge") + facet_wrap(~variable,
↪  scales="free_y")
```

From the above table and plot, we can see that, in addition to the **year, mileage, title_status** predictors, the vehicle **model** variable plays a critical role in determining the sale price of a vehicle. This is demonstrated when looking at the the R^2_{adj} value. The R^2_{adj} value decreases significantly from the "Full Model" model (0.7264025) to the "Full Model without model" model (0.4303862). This indicates that approximately 30% of the model variation in sale price is explained by the vehicle **model** variable.

It was also found that the **color**, **state**, and **brand** variables do not contribute significantly to the model. The R^2_{adj} value slightly decreases from the "Full Model" model (0.7264025) to the "Full Model without Color" model (0.7176083), the "Full Model without State" model (0.6966362), and the "Full Model without Brand" model (0.7075274). There is also some increase in RMSE after dropping these three variables, but it is a small percentage. Comparing the "Full Model" model and the "Full Model

without State_Brand_Color" model, we can see that the R^2_{adj} value decreases from 0.7264025 to 0.6666710 and the RMSE value slightly increases from 6340.556 to 6998.546. Given that there are too many non-significant levels (the individual T test also shows they are not significant at 0.05 level, data not shown) in the **color**, **state**, and **brand** variables, we decided to drop them from the final model.

Individual T test and Partial F test

Individual T test at $\alpha = 0.05$

- Null hypothesis $(H_0) : \beta_i = 0$
- Alternative hypothesis $(H_A) : \beta_i \neq 0$ where i = year, mileage, title_status

```
kable(summary(cars_model12_nostate_brand_color)$coefficient[2:4,3:4])
```

	t value	Pr(>|t|)
year	4.021274	6.02e-05
mileage	-12.429551	0.00e+00
factor(title_status)clean vehicle	8.151181	0.00e+00

The above individual t-tests are used to further confirm our variable selection based on a significance level of $\alpha = 0.05$. For quantitative variables **year** and **mileage**, the t-values are 4.021274 and -12.429551, respectively and the P values are 6.017477e-05 and 3.824310e-34, respectively. Since the p-values are significantly lower than alpha at 0.05, we reject the null hypothesis and conclude that both the **year** and **mileage** variables are significant predictors for a car's sale price. For the qualitative variable **title_status**, the t-value is 8.151181 and the p-value is 6.510606e-16. Again, because the p-value is smaller than alpha at 0.05, we reject the null hypothesis and conclude that that difference between clean vehicle and salvage insurance (has been in accident) is a significant predictor for a car's sale price.

Partial F test at $\alpha = 0.05$

Since not all levels of **model** qualitative variable are significant, we will confirm them using a partial F-test.

- Null hypothesis (H_o): the coefficients for the variable levels all equal to 0
- Alternative hypothesis (H_a): at lease one coefficient level of the variable does not equal to 0

The resulting p-values after the partial F-test is completed are as follows:

```
#full model
cars_model12_nostate_brand_color <- lm(price~ year +  mileage + factor(title_status)
↪  + factor(model), data = cars)
#reduced model
cars_model12_nostate_brand_color_model <- lm(price~ year +  mileage +
↪  factor(title_status), data = cars)
#build anova table
kable(as_tibble(anova(cars_model12_nostate_brand_color_model,cars_model12_nostate_brand_color))
↪  %>% mutate_all(funs(replace_na(., " "))))
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1990	224264493293.518				

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1878	91983781196.4378	112	132280712097.08	24.1136432352398	1.15028262093547e-285

From the above partial F-test result, the **model** variable is determined to be significant based on a significance level of $\alpha = 0.05$. We therefore reject the null hypothesis: all β coefficient is 0 and accept the alternative hypothesis: at least one β coefficient is not 0. This suggests the **model** variable is a significant predictor for a car's sale price.

In conclusion, we came to the following final additive model:

$$price = \beta_0 + \beta_1 * year + \beta_2 * mileage + \beta_3 * title\_status + \beta_4 * model + \epsilon$$

Finding Interaction Terms

We then looked for interaction terms between our main effect variables. Since the **model** variable still had 43 levels after cleanup, we decided to not include this in the interaction checking to avoid overwhelming and over-complicating our model. To make sure we determined a reliable interaction model, we compared different selection procedures to find the "best" set of predictors.

Stepwise Selection

From the Stepwise Regression Procedure, two out of three significant interaction terms emerged: **mileage*year** and **mileage*title_status**. The summary of this interaction model shows the following individual t-test at $\alpha = 0.05$:

- Null hypothesis $(H_0) : \beta_i = 0$
- Alternative hypothesis $(H_a) : \beta_i != 0$. i = mileage*year, mileage*title_status

mileage*year: t = -4.373, P = 1.29e-05

mileage*title_status: t = -5.262, P = 1.59e-07

Since both p-values are much smaller than alpha at 0.05, we can reject the null hypothesis: β coefficient is 0 and accept the alternative hypothesis: β coefficient is not 0. This suggests we should include these two interaction terms in our final model as significant predictors of a car's sale price. This also makes sense within the context of motor vehicles. For example, a person may consider an older vehicle if it has very low mileage over a newer vehicle with significantly more miles. In addition to this, whether a car is a rescued one or a good condition one, it will certainly affect how often people choose to drive it as well.

The interaction model is shown as follows:

$$price = \beta_0 + \beta_1 * year + \beta_2 * mileage + \beta_3 * title\_status + \beta_4 * model + \beta_5 * mileage * year + \beta_6 * mileage * title\_status + \epsilon$$

We also tried the **Backward Elimination**, **Forward Selection** and **All Possible Regressions Selection** procedures and ended up with the same resultant model.

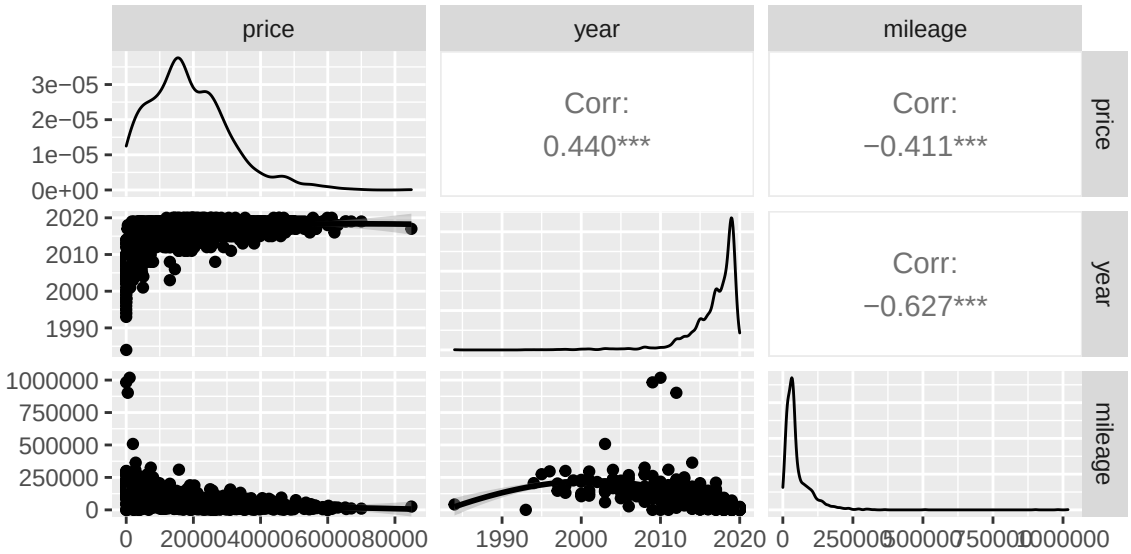
```
# Forward stepwise
# forward=ols_step_forward_p(cars_model12_nostate_brand_color_inter, pent=0.05,
↪  prem=0.3,details=F, progress = F)
# summary(forward$model)

# Backward stepwise
# backward=ols_step_backward_p(cars_model12_nostate_brand_color_inter, pent=0.05,
↪  prem=0.3,details=F, progress = F)
```

```
# summary(backward$model)

# All Possible
# bestsub=ols_step_best_subset(cars_model12_nostate_brand_color_inter, pent=0.05,
↪  prem=0.3,details=F, progress = F)
# summary(bestsub)
```

Finding Higher Order Terms

Before checking for higher-order terms, we reviewed the pairwise scatter plot to see if there was any potential quadratic relation for the quantitative predictors in our model. It appeared that both quantitative variables (**year** and **mileage**) had a higher-order relation with the response variable (**price**). We performed individual t-tests to confirm this.

```
ggpairs(cars, columns = c(1,4,6), lower = list(continuous = "smooth_loess", combo =
↪  "facethist", discrete = "facetbar", na = "na"))
```



After multiple iterations of checking and dropping the higher order terms based on individual t-tests, the adjusted R squared value, and the RMSE value, we have the following summary:

```
#interaction model
cars_model12_nostate_brand_color_inter <- lm(price~ year +  mileage +
↪  factor(title_status) + factor(model) + mileage*year +
↪  mileage*factor(title_status), data = cars)
#higher order model
cars_model12_nostate_brand_color_inter_higher0 <- lm(price~ year +  mileage +
↪  factor(title_status) + factor(model) + mileage*year +
↪  mileage*factor(title_status) + I(year^2), data = cars)
cars_model12_nostate_brand_color_inter_higher1 <- lm(price~ year +  mileage +
↪  factor(title_status) + factor(model) + mileage*year +
↪  mileage*factor(title_status) + I(year^2) + I(mileage^2), data = cars)
cars_model12_nostate_brand_color_inter_higher2 <- lm(price~ year +  mileage +
↪  factor(title_status) + factor(model) + mileage*year +
↪  mileage*factor(title_status) + I(mileage^2), data = cars)
```

```
#get the Adjusted R square and RMSE for both model
kable(summary(cars_model12_nostate_brand_color_inter)$coefficient[c(2:4,
↪  117:118),3:4])
```

	t value	Pr(>|t|)
year	6.454736	0.00e+00
mileage	4.371253	1.30e-05
factor(title_status)clean vehicle	9.284566	0.00e+00
year:mileage	-4.372884	1.29e-05
mileage:factor(title_status)clean vehicle	-5.261756	2.00e-07

```
kable(summary(cars_model12_nostate_brand_color_inter_higher0)$coefficient[c(2:4,
↪  117:119),3:4])
```

	t value	Pr(>|t|)
year	-1.419623	0.1558837
mileage	3.587432	0.0003425
factor(title_status)clean vehicle	9.380357	0.0000000
I(year^2)	1.444692	0.1487115
year:mileage	-3.587166	0.0003429
mileage:factor(title_status)clean vehicle	-5.348932	0.0000001

```
kable(summary(cars_model12_nostate_brand_color_inter_higher1)$coefficient[c(2:4,
↪  117:120),3:4])
```

	t value	Pr(>|t|)
year	-1.3592398	0.1742342
mileage	0.5426630	0.5874264
factor(title_status)clean vehicle	9.4281938	0.0000000
I(year^2)	1.3750374	0.1692842
I(mileage^2)	4.3116894	0.0000170
year:mileage	-0.5700223	0.5687309
mileage:factor(title_status)clean vehicle	-5.5354093	0.0000000

```
kable(summary(cars_model12_nostate_brand_color_inter_higher2)$coefficient[c(2:4,
↪  117:119),3:4])
```

	t value	Pr(>|t|)
year	3.5512890	0.0003928
mileage	0.9656021	0.3343679
factor(title_status)clean vehicle	9.3394763	0.0000000
I(mileage^2)	4.3356342	0.0000153
year:mileage	-0.9967411	0.3190188
mileage:factor(title_status)clean vehicle	-5.4543570	0.0000001

From the above summary of models, we made the following conclusions:

- After including $year^2$, we found that the P-value for $(year^2)$ is $= 0.1487115 > 0.05$ with small t value $= 1.444692$. Based on this, we decided to not keep the higher order term for **year** in our model.
- After including $mileage^2$, we found that although the P-value for $(mileage^2)$ is $= 0.0000153 < 0.05$, the R^2_{adj} value does not increase at all. To avoid increasing the complexity of the model unnecessarily, we decided to not keep the higher order term for **mileage** in our final model either.

The final model is therefore without any higher order terms and is shown below:

$$price = \beta_0 + \beta_1 * year + \beta_2 * mileage + \beta_3 * title\_status + \beta_4 * model + \beta_5 * mileage * year + \beta_6 * mileage * title\_status + \epsilon$$

Model Diagnostics

To ensure our model has been fit appropriately, we must perform regression diagnostics to ensure that model assumptions are valid and that there are no outliers causing unwanted influence on the analysis.

Before this is done, further analysis was completed on the data and it was determined that the cars data should be further filtered to exclude any 0 values within the price column.

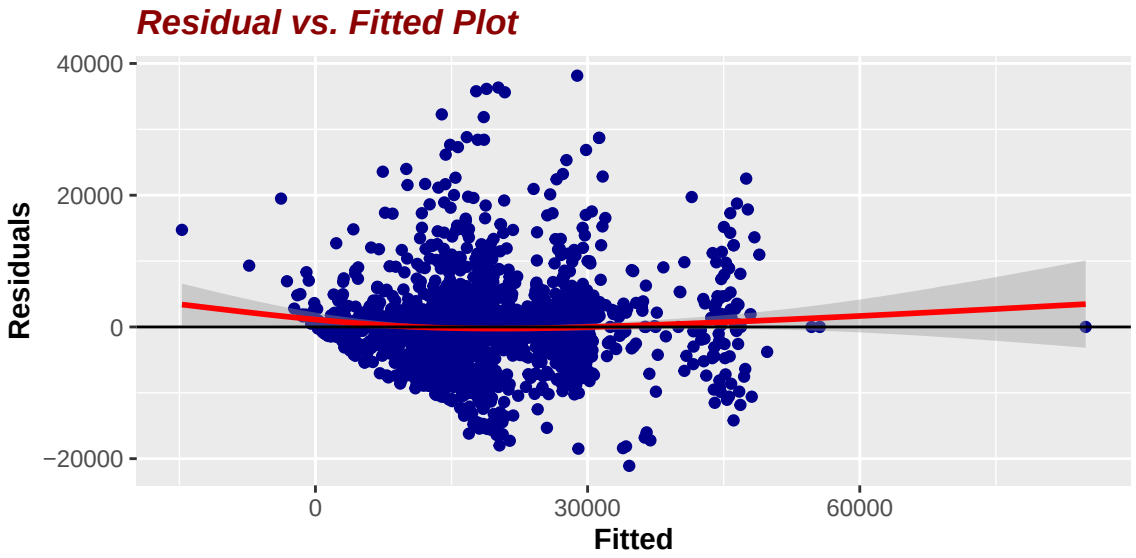
```
#remove rows where price < 0
cars1 <- filter(cars, price > 0)

cars_model12_nostate_brand_color_inter <- lm(price~ year +  mileage +
↪  factor(title_status) + factor(model) + mileage*year +
↪  mileage*factor(title_status), data = cars1)

final_model = cars_model12_nostate_brand_color_inter
#summary(final_model)
```

Linearity Assumption

The linearity assumption assumes that the relation between the predictors and the response is linear. To check if this assumptions holds, we will examine the "Residuals vs Fitted" diagnostic plot.

```
#plot(final_model, which = 1)

ggplot(final_model, aes(x=.fitted, y=.resid)) + labs(y= "Residuals", x = "Fitted") +
↪  geom_point(color = "dark blue") + geom_smooth(color = "red")+
↪  geom_hline(yintercept = 0) + ggtitle("Residual vs. Fitted Plot") +  theme (
plot.title = element_text(color="dark red", face="bold.italic")) + theme (axis.title
↪  = element_text(face="bold"))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

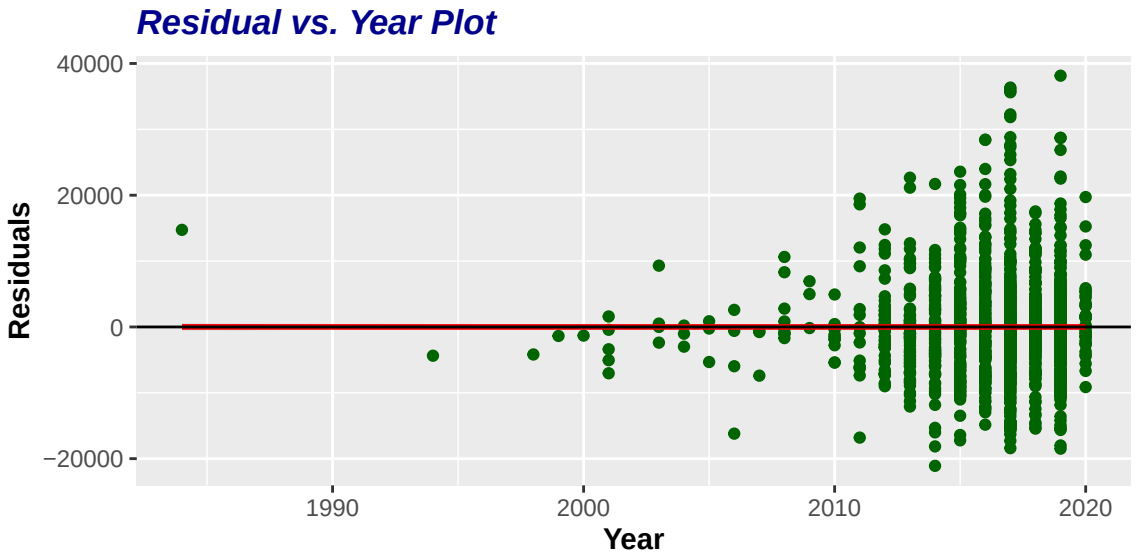


Ideally, the residual plot will show no fitted pattern & the red line should be approximately horizontal at zero. We can see that the red line is mostly horizontal and that the majority of the data points are randomly distributed around it with no particular pattern. This suggests that the relationship between the predictors and response is linear. The linearity assumption is therefore met.

Independence Assumption

The independence assumption assumes that the model's error terms are uncorrelated. This typically occurs when the data is time-series data. Since our data is not time-series dependent, we are confident that this assumption is met. To confirm this, we created a "Residuals vs. Year" plot and did not see a pattern, indicating that the assumption of independence errors is met. However, we do see an increased magnitude of the residuals along the x-axis which may indicate that there is an issue with the non-constant variances in the residuals. We will address this in the "Equal Variance Assumption" section.

```
#ggplot(data.frame(resid = residuals(final_model), Year = as.numeric(cars1$year)),
↪  aes(x=Year, y=resid)) + geom_point() + geom_smooth()+ geom_hline(yintercept = 0)

ggplot(data.frame(Residuals = residuals(final_model), Year =
↪  as.numeric(cars1$year)), aes(x=Year, y=Residuals)) + geom_point(color = "dark
↪  green") + geom_smooth(color = "red")+ geom_hline(yintercept = 0) +
↪  ggtitle("Residual vs. Year Plot") +  theme (
plot.title = element_text(color="dark blue", face="bold.italic")) + theme
↪  (axis.title = element_text(face="bold"))
```



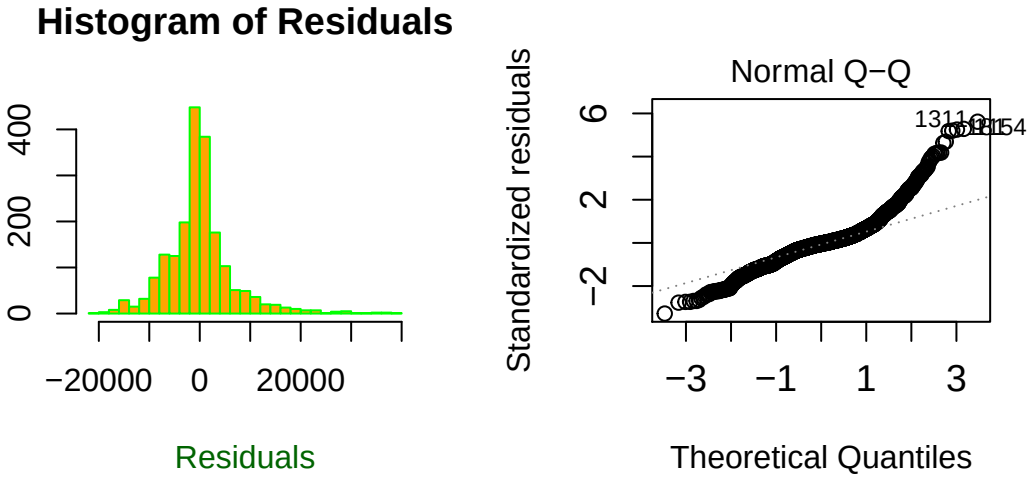
Normality Assumption

The normality assumptions assumes that the residual errors are normally distributed. We can check this assumption by looking at a histogram of the final model's residuals, looking at the "Normal Q-Q" plot, and using the Shapiro-Wilk test.

For the Shapiro-Wilk test:

- Null hypothesis (H_0): the sample data is normally distributed
- Alternative hypothesis (H_a): the sample data is not significantly normally distributed

```
par(mfrow = c(1,2))
#hist(residuals(final_model), breaks = 40, xlab ="residuals", ylab ="",main
↪  ="Histogram of residuals")
#plot(final_model, which = 2)
hist(residuals(final_model), breaks = 40, xlab ="Residuals", ylab ="",main
↪  ="Histogram of Residuals", border="Green", col="Orange", col.lab = "dark green")
↪
plot(final_model, which = 2, cex.axis = 1.15)
```



```r
shapiro.test(residuals(final_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(final_model)
## W = 0.91167, p-value < 2.2e-16
```

Based on the histogram, we can see that the residuals do not follow the typical normal distribution curve. The Q-Q plot further supports this conclusion as the majority of the data points does not follow the diagonal line, but the tails are very skewed, indicating high kurtosis. Finally, the Shapiro-Wilk normality test confirms that the residuals are not normally distributed as the p-value = < 2.2e-16 < 0.05, so we reject the null hypothesis and can conclude that there is strong evidence to suggest that the sample data is not normally distributed.

Equal Variance Assumption

The equal variance assumption assumes that the variance of residual is the same for any value of X. To test if this assumption holds, we will use the Breusch–Pagan test and review the Scale-Location diagnostic plot.

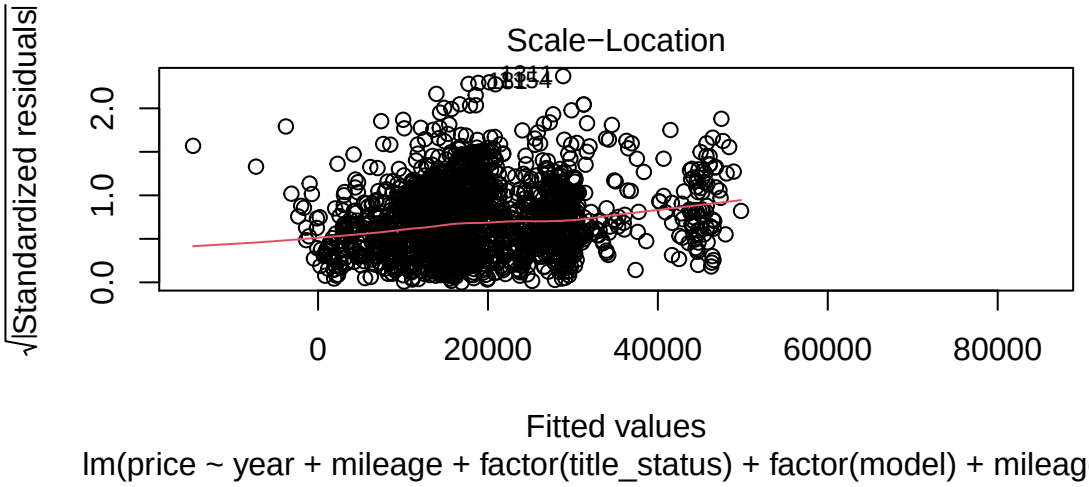
For the Breusch–Pagan test:

- Null hypothesis (H_0): heteroscedasticity is not present (homoscedasticity) ($\sigma_1^2 = \sigma_2^2) = \ldots = \sigma_n^2$
- Alternative hypothesis (H_a): heteroscedasticity is present (at least one σ_i^2 is different from the others where i = 1, 2,... n)

```r
#Breusch-Pagan test
bptest(final_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  final_model
## BP = 273.32, df = 116, p-value = 1.009e-14
```

```r
# residual plot
#plot(final_model, which = 3)
plot(final_model, which = 3)
```



Since the Breusch–Pagan value has a p-value of 1.009e-14 $< \alpha = 0.05$, we reject the null hypothesis and can conclude that we are 95% confident that there is sufficient evidence to say that heteroscedasticity is present in the regression model. From the "Scale-Location" plot, we can see that the red line is slightly increasing along the fitted value and that the spread on the graph is not uniform. This further proves that homoscedasticity is not present. The equal variance assumption is therefore not met.

Multicolinearity Tests

Multicollinearity occurs when independent variables within the model are correlated with each other. This can cause sensitivity in the coefficient estimates and weakens the statistical power of the model. To check for multicollinearity, we will use variance inflation factors (VIF). The following is a summary of the VIFs found. There is no multicolinearity detected for **year**, **mileage**, and **title_status**, but the **model** door variable has a strong correlation with the other variables. NOTE: factor(model)[remaining] refers to the remaining levels in the **model** feature except door level.

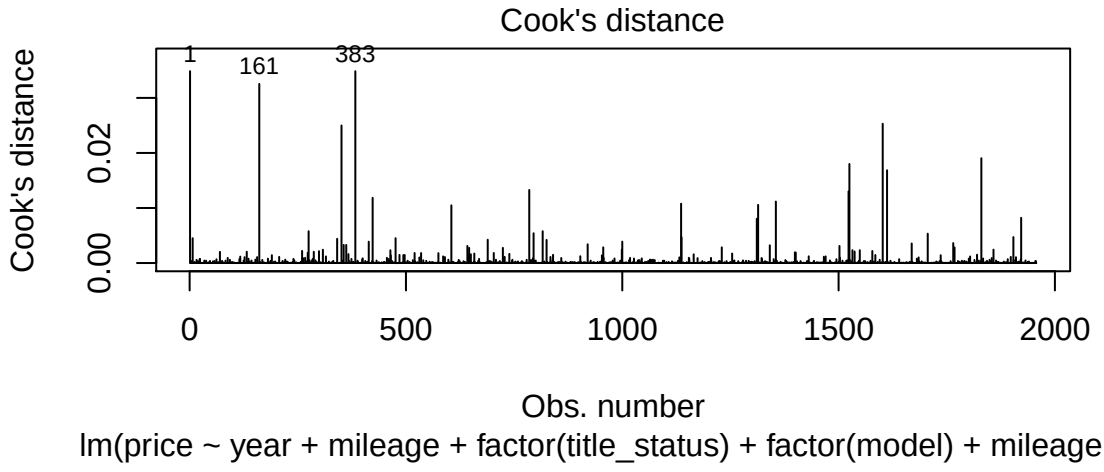
Independent Variable	VIF value	detection
year	3.2564	0
mileage	4.2126	0
factor(title_status)clean vehicle	1.6476	0
factor(model)door	14.0665	1
factor(model)[remaining]	< 2.0	0

Influential Points and Outliers

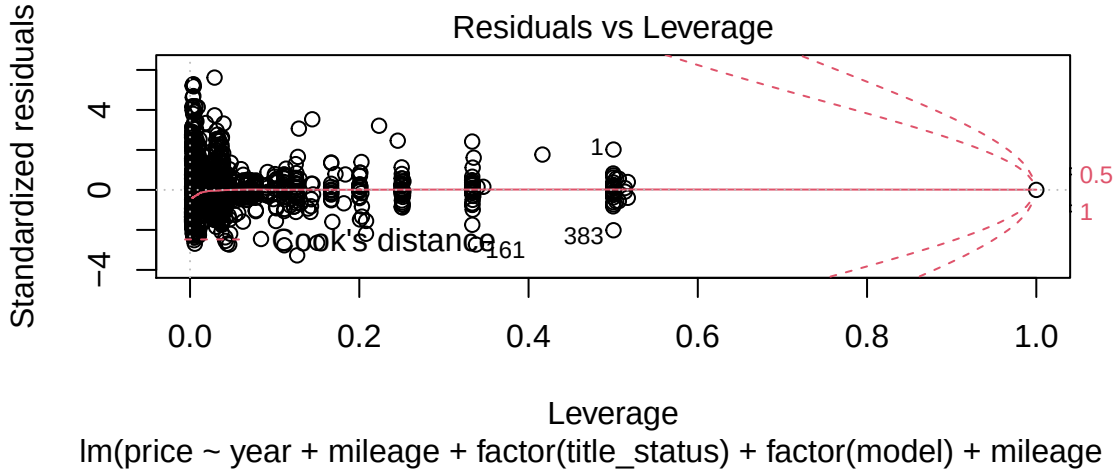
An outlier that has high influence will affect the interpretation and the results of the model. An observation is considered an outlier if it has a value for the response variable that is very different from the rest of the observations in the data set. We'll use Cook's Distance, high-leverage points, and the "Residuals vs Leverage" diagnostic plot to search for influential outliers.

Based on the Cook's distance plot, there are no points that have a Cook's distance greater than 0.5 which indicate that no points have a strong influence on the response. The "Residuals vs Leverage" plot confirms this as well.

```
# cooks distance measure plot
plot(final_model, which = 4)
```

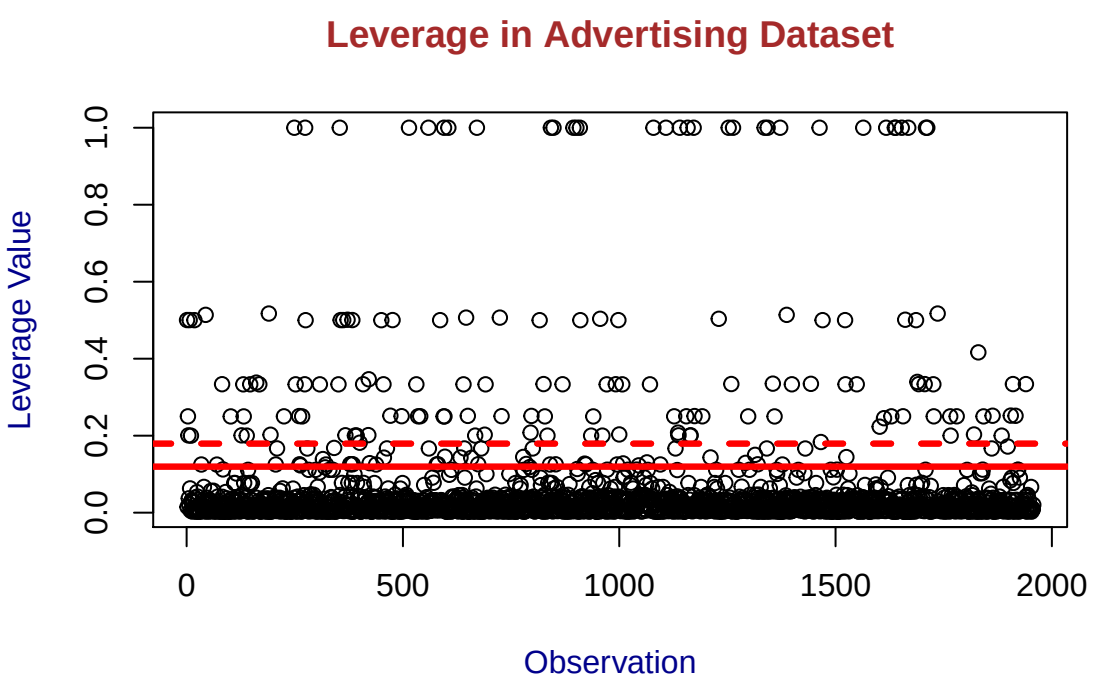



```
# residuals vs leverage plot
plot(final_model, which = 5)
```



However, if we look at the high-leverage points based on the 2p/n (solid red line) and 3p/n (dashed red line) thresholds, we can see 152 outliers above 3p/n. We will try to remove those outliers and re-do the diagnostic test as seen in the next section "Data Transformation and Diagnostic Re-test".

```
lev <- hatvalues(final_model)
p <- length(coef(final_model))
n <- nrow(cars1)
outlier <- lev[lev>(3*p/n)]
#make the plot for the data
#plot(rownames(cars1),lev, main = "Leverage in Advertising Dataset",
    # xlab = "Observation",
    # ylab = "Leverage Value", )
plot(rownames(cars1),lev, main = "Leverage in Advertising Dataset",
     xlab = "Observation",
     ylab = "Leverage Value", col.lab = "dark blue", col.main = "brown")
abline(h = 2*p/n, lty = 1, col = "red", lwd=3)
abline(h = 3*p/n, lty = 2, col = "red", lwd=3)
abline(h = 2*p/n, lty = 1, col = "red", lwd=3)
abline(h = 3*p/n, lty = 2, col = "red", lwd=3)
```

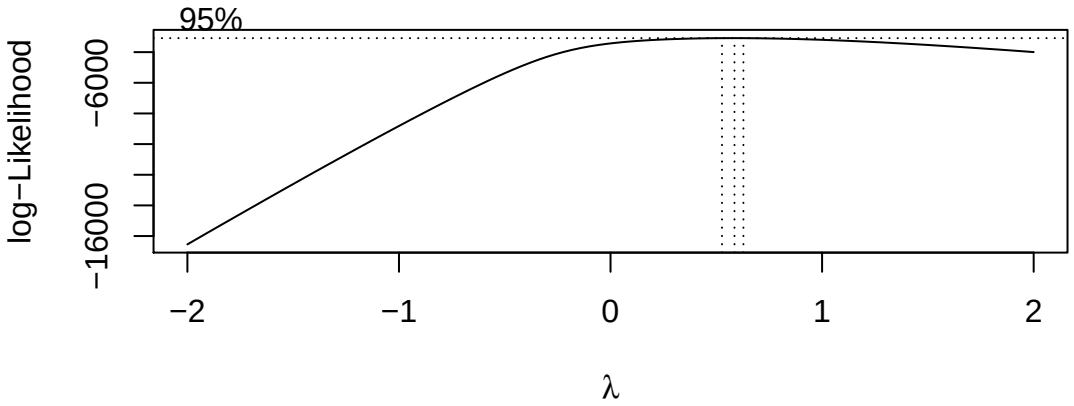


Data Transformation and Diagnostic Re-test

Since the equal variance and normality assumptions are not being met, we will try removing the outliers and transform the dependent variable using the Box-Cox Transformation.

In addition, we will also remove the door level from the **model** variable because it was detected to have strong correlation with the rest of the variables based on the tests conducted for multicolinearity.

```
# remove 3p/n outliers
# get index of outliers
outlier_index = strtoi(names(outlier))
cars2 = cars1[-c(outlier_index),]

#drop off the door level from model feature
cars2 <- cars2[!cars2$model == "door",]

# rebuild the model
cars_model13_nostate_brand_color_inter <- lm(price~ year +  mileage +
↪  factor(title_status) + factor(model) + mileage*year +
↪  mileage*factor(title_status), data = cars2)
final_model2 = cars_model13_nostate_brand_color_inter

# we will use box-cox to do the transformation
bc = boxcox(final_model2)
```




```
bestlambda=bc$x[which(bc$y==max(bc$y))]
cat("The best lambda for Box-Cox transformation is:", bestlambda)
```

```
## The best lambda for Box-Cox transformation is: 0.5858586
```

We get $\lambda = 0.5858586 \neq 0$, therefore the transformation is:

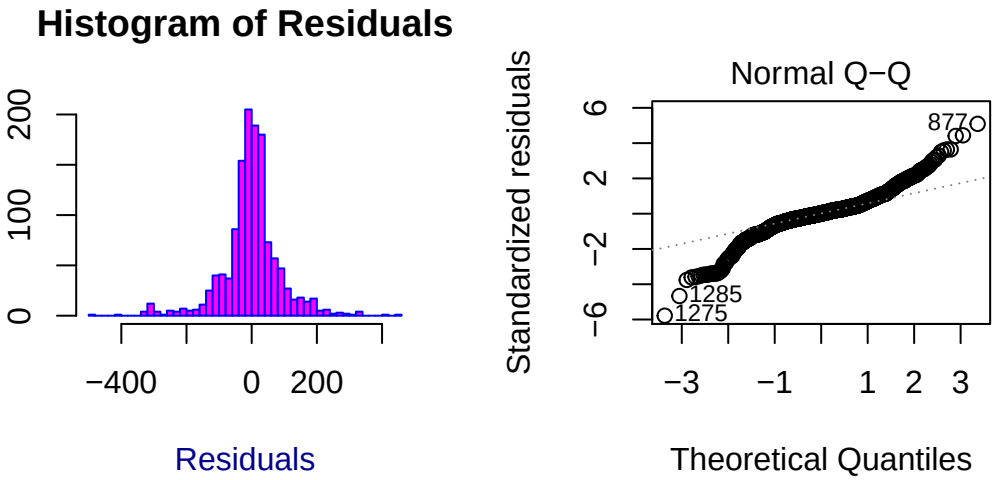
$$Y_i^{(0.5858586)} = \frac{Y^{0.5858586} - 1}{0.5858586}$$

$$Y_i^{(0.5858586)} = \beta_0 + \beta_1 * year + \beta_2 * mileage + \beta_3 * title\_status + \beta_4 * model + \beta_5 * mileage * year + \beta_6 * mileage * title\_status + \epsilon$$

```
# since lambda is not = 0, we will use the following transformation formula
final_model3 = lm((((price^bestlambda) - 1)/bestlambda) ~ year +  mileage +
↪  factor(title_status) + factor(model) + mileage*year +
↪  mileage*factor(title_status), data = cars2)
#summary(final_model3)
```

Now we can try re-doing the normality and equal variance tests.

```
par(mfrow = c(1,2))
#hist(residuals(final_model3), breaks = 40, xlab ="residuals", ylab ="",main
↪  ="Histogram of residuals")
#plot(final_model3, which = 2)

hist(residuals(final_model3), breaks = 40, xlab ="Residuals", ylab ="",main
↪  ="Histogram of Residuals", border="blue", col="magenta", col.lab = "dark blue")

plot(final_model3, which = 2)
```




```
shapiro.test(residuals(final_model3))
```

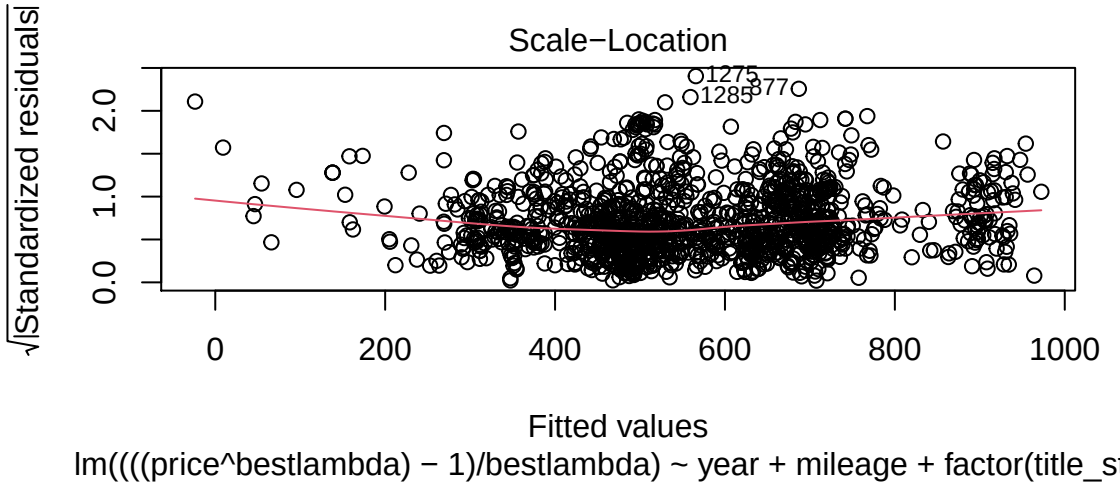
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(final_model3)
## W = 0.92169, p-value < 2.2e-16
```

We see that the histogram of residuals still does not follow a typical normal distribution and the Q-Q plot does not follow a straight line and is highly skewed at the tails. Furthermore, the Shapiro-Wilk test gives a p-value $<$ 2.2e-16 $<$ 0.05. This indicates the data does not follow a normal distribution even after the transformation and thus the normality assumption is still not met.

```
#Breusch-Pagan test
bptest(final_model3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  final_model3
## BP = 264.37, df = 46, p-value < 2.2e-16
```

```
# residual plot
plot(final_model3, which = 3)
```



We see from the "Scale-Location" plot that the red line is more horizontal but is still not even, and that the spread on the graph is not uniform. Additionally, the Breusch-Pagan test gives a p-value $<$ 2.2e-16 $<$ 0.05, indicating homoscedasticity is not present. Therefore, the equal variance assumption is still not met after the transformation.

Our Best Fitted Model

After completing transformations and re-running the model diagnostics, we continue to see that the normality assumption and equal variance assumptions are not met. To address this, the team could continue to try other transformations or could re-visit the variables selected to create the model. Due to time constraints, we will not pursue this further and will conclude that the best model we could find is the model found prior to the transformation but with outliers removed, because the $R^2_{adj} = 0.6558236$ is decent. It should be noted that since the model does not meet all assumptions, the model doesn't fully explain the data set.

The final model is therefore:

$$price = \beta_0 + \beta_1 * year + \beta_2 * mileage + \beta_3 * title\_status + \beta_4 * model + \beta_5 * mileage * year + \beta_6 * mileage * title\_status + \epsilon$$

NOTE: year and mileage are quantitative variables, while title_status (2 levels) and model (43 levels) are qualitative variables.

We can also express the final best model in the following 2 ways:

$$price = \beta_0 + \beta_1 * year + (\beta_2 + \beta_5 * year + \beta_6 * title\_status) * mileage + \beta_3 * title\_status + \beta_4 * model + \epsilon$$

$$price = \beta_0 + (\beta_1 + \beta_5 * mileage) * year + \beta_2 * mileage + (\beta_3 + \beta_6 * mileage) * title\_status + \beta_4 * model + \epsilon$$

The RMSE and R^2_{adj} of the best fitted model is as follows:

- $R^2_{adj} = 0.6558236$, this value indicates that 65.58% of the variation of the response variable sale price is explained by the final model containing the predictors **year, mileage, title_status, model**, as well as interaction terms: **mileage*year, mileage*title_status.**
- RMSE = 5727.92279, this value indicates that the standard deviation of the unexplained variation in estimation of the response variable sale price is 5727.92279$.

Interpreting Coefficients

There are a total of 6 variables (**year, mileage, title_status, model, mileage*year, mileage*title_status**) in our final model. However, because the **title_status** ("salvage insurance" as base level) and **model** ("1500" as base level) variables each have many levels, we have a total 48 coefficients as shown below:

```
# remove 3p/n outliers
# get index of outliers
outlier_index = strtoi(names(outlier))
cars2 = cars1[-c(outlier_index),]
#rebuild the model
cars_model13_nostate_brand_color_inter <- lm(price~ year +  mileage +
↪  factor(title_status) + factor(model) + mileage*year +
↪  mileage*factor(title_status), data = cars2)
kable(summary(cars_model13_nostate_brand_color_inter)$coefficient)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.559352e+06	3.362384e+05	-4.637638	0.0000038
year	7.807590e+02	1.668485e+02	4.679449	0.0000031
mileage	5.533734e+00	2.507961e+00	2.206468	0.0274802
factor(title_status)clean vehicle	1.671557e+04	1.883537e+03	8.874563	0.0000000
factor(model)2500	7.889711e+03	3.111863e+03	2.535366	0.0113191
factor(model)altima	-1.643936e+04	2.160212e+03	-7.610068	0.0000000
factor(model)caravan	-1.387532e+04	1.486622e+03	-9.333457	0.0000000
factor(model)challenger	-3.702390e+03	1.733780e+03	-2.135445	0.0328627
factor(model)charger	-8.887844e+03	1.801507e+03	-4.933560	0.0000009
factor(model)colorado	-8.234414e+03	2.533578e+03	-3.250112	0.0011754
factor(model)cutaway	-6.172142e+03	2.787653e+03	-2.214100	0.0269499
factor(model)door	-8.652708e+03	1.336436e+03	-6.474464	0.0000000
factor(model)doors	-1.103293e+04	1.428859e+03	-7.721493	0.0000000
factor(model)drw	1.490009e+04	2.912531e+03	5.115857	0.0000003
factor(model)durango	-1.603898e+03	1.589251e+03	-1.009216	0.3130098
factor(model)ecosport	-1.702901e+04	3.110350e+03	-5.474950	0.0000001
factor(model)edge	-6.836828e+03	1.883299e+03	-3.630240	0.0002913
factor(model)equinox	-1.413266e+04	2.250490e+03	-6.279812	0.0000000
factor(model)escape	-1.480442e+04	1.756800e+03	-8.426920	0.0000000
factor(model)expedition	1.296762e+04	1.868672e+03	6.939485	0.0000000

	Estimate	Std. Error	t value	Pr(>|t|)
factor(model)explorer	-2.741438e+03	1.743088e+03	-1.572748	0.1159572
factor(model)f-150	-2.396942e+03	1.382252e+03	-1.734085	0.0830785
factor(model)fiesta	-2.224679e+04	2.312648e+03	-9.619616	0.0000000
factor(model)flex	-7.313162e+03	1.911428e+03	-3.826019	0.0001348
factor(model)focus	-1.853054e+04	2.761641e+03	-6.709973	0.0000000
factor(model)frontier	-1.340759e+04	2.648983e+03	-5.061411	0.0000005
factor(model)fusion	-1.527547e+04	1.623271e+03	-9.410305	0.0000000
factor(model)impala	-1.331826e+04	2.454124e+03	-5.426891	0.0000001
factor(model)journey	-1.425306e+04	1.615770e+03	-8.821219	0.0000000
factor(model)malibu	-1.572848e+04	2.638555e+03	-5.961022	0.0000000
factor(model)max	1.467746e+04	1.791686e+03	8.191987	0.0000000
factor(model)mpv	-2.136000e+04	1.522121e+03	-14.033054	0.0000000
factor(model)mustang	-5.291699e+03	1.869425e+03	-2.830657	0.0046981
factor(model)pathfinder	-1.171604e+04	2.306774e+03	-5.078974	0.0000004
factor(model)pickup	-4.487654e+03	2.878469e+03	-1.559042	0.1191666
factor(model)rogue	-1.453920e+04	1.649276e+03	-8.815501	0.0000000
factor(model)sentra	-1.976777e+04	1.952066e+03	-10.126588	0.0000000
factor(model)series	7.823382e+03	2.766596e+03	2.827800	0.0047400
factor(model)sport	-1.609625e+04	1.781076e+03	-9.037372	0.0000000
factor(model)srw	1.468248e+04	1.846458e+03	7.951697	0.0000000
factor(model)suburban	4.305556e+03	2.198542e+03	1.958369	0.0503447
factor(model)taurus	-1.258680e+04	2.640752e+03	-4.766369	0.0000020
factor(model)transit	-7.957022e+03	1.788529e+03	-4.448919	0.0000092
factor(model)van	-1.076713e+04	1.702826e+03	-6.323094	0.0000000
factor(model)versa	-2.061827e+04	1.827211e+03	-11.284014	0.0000000
factor(model)wagon	-4.755418e+03	1.949056e+03	-2.439856	0.0147914
year:mileage	-2.752800e-03	1.247700e-03	-2.206356	0.0274880

	Estimate	Std. Error	t value	Pr(>|t|)
mileage:factor(title_status)clean vehicle	-8.142840e-02	1.522990e-02	-5.346616	0.0000001

The coefficients can be interpreted as follows: (NOTE: Since the normality assumption is not met, the coefficients interpretation may not be very accurate. However, we still provide their inferences in a way to reflect our understanding and give a general idea of how each variable may impact a car's sale price.)

- The effect of **year** is 780.8 - 0.0028x**mileage**, which means that if all other variables (**title_status** and **model**) are held constant, an increase in 1 for year (meaning the car is more recent) leads to an increase in sale price of the vehicle by 780.8 - 0.0028x**mileage** dollars.
- The effect of **mileage** is 5.53 - 0.002753x**year** - 0.08143x**title_status**, which means that if all other variables (**model**) are held constant, an increase of 1 in the number of miles traveled by the vehicle will increase the sale price of the vehicle by 5.53 - 0.002753x**year** - 0.08143x**title_status**. Note: Since **title_status** is a qualitative variable (1 = clean title vehicles and 0 = salvage insurance), we can further express the effect of **mileage** as follows:
  1. When the vehicle is salvaged, **mileage** effect = 5.53 - 0.002753x**year**
  2. When the vehicle has no damage , **mileage** effect = 5.53 - 0.002753x**year** - 0.08143 = 5.44857 - 0.002753x**year**
- The effect of **title_status** is 16716 -0.08143x**mileage**, which means that if all other variables (**year**, **model**) are held constant, the sale price difference between a clean non-damaged vehicle and a salvaged vehicle is 16716 -0.08143x**mileage** dollars.
- The effect of **model** is dependent on what model variable is used since this is a qualitative variable with a lot of levels. The base level is "1500". The coefficient for each level means that if all the other variables (**year**, **mileage**, **title_status**) are held constant, the sale price difference between a specific model and the "1500" model is the β dollars for that specific model. For example, if the car model is a caravan and if all other variables are held constant, the sale price of the vehicle will decrease by 13875 dollars compared to the "1500" model.

Predicting Car Prices

As a first check, we can look at the directional correlation of the actuals in the 20% of our dataset that we set aside and the predicted values our model generates to see if the prediction is as we expect. A higher correlation accuracy implies that the actuals and predicted values have similar directional movement, meaning that when the actuals values increase, the predicted values also increase and vice-versa. From the following result, we can see the correlation accuracy is 86.8% which indicates a good match.

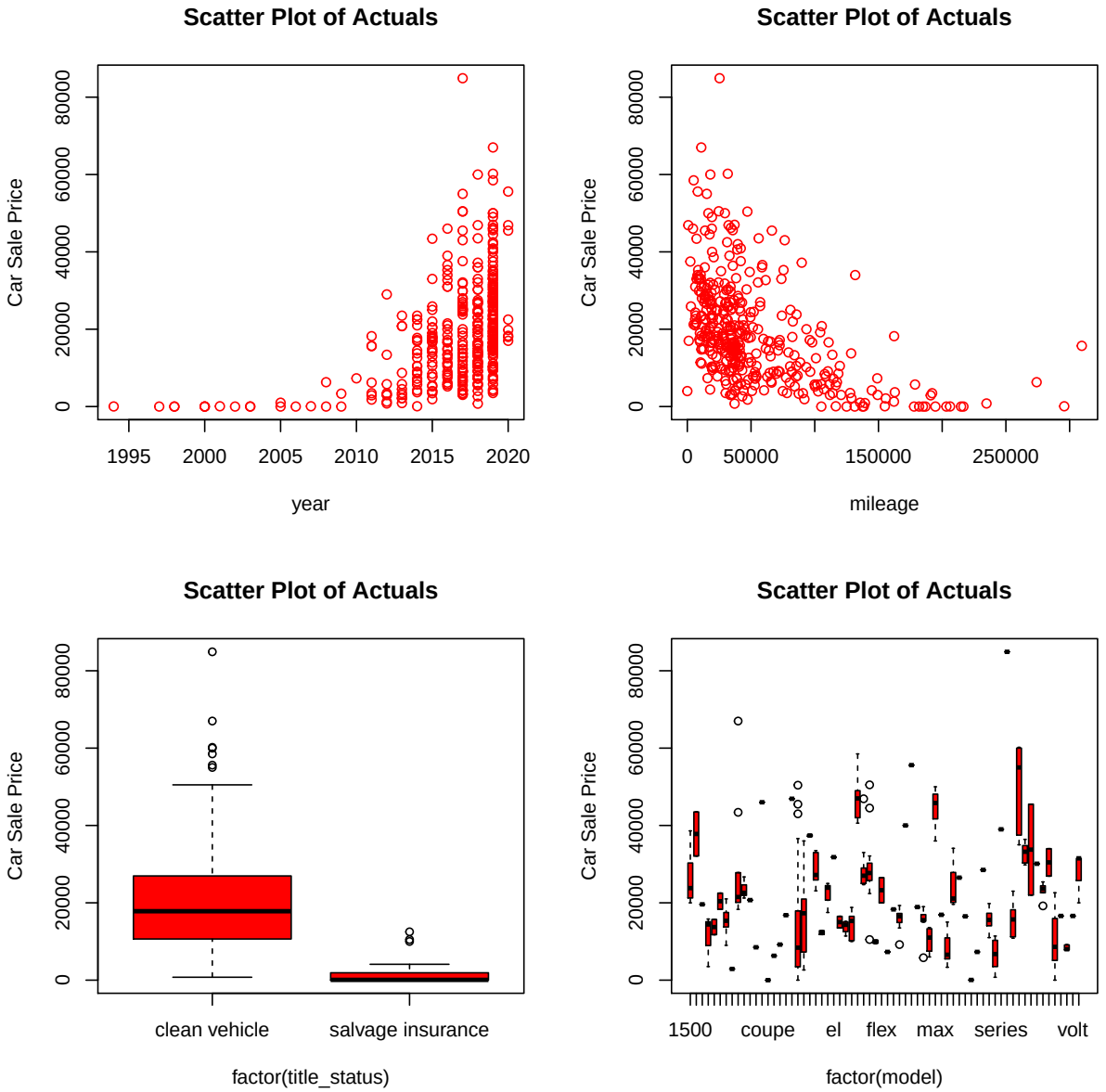
```
# Build the model on training data
test_model = lm(price~ year +  mileage + factor(title_status) + factor(model) +
↪  mileage*year + mileage*factor(title_status), data=test)  # build the model
sale_predict = predict(test_model, test)  # predict price
# summary(test_model)

# see if there is correlation between the actuals and the predicted
actuals_preds = data.frame(cbind(actuals=test$price, predicteds=sale_predict))
correlation_accuracy = cor(actuals_preds)
kable(correlation_accuracy)
```

	actuals	predicteds
actuals	1.0000000	0.8677325
predicteds	0.8677325	1.0000000

```
# head(actuals_preds)
```

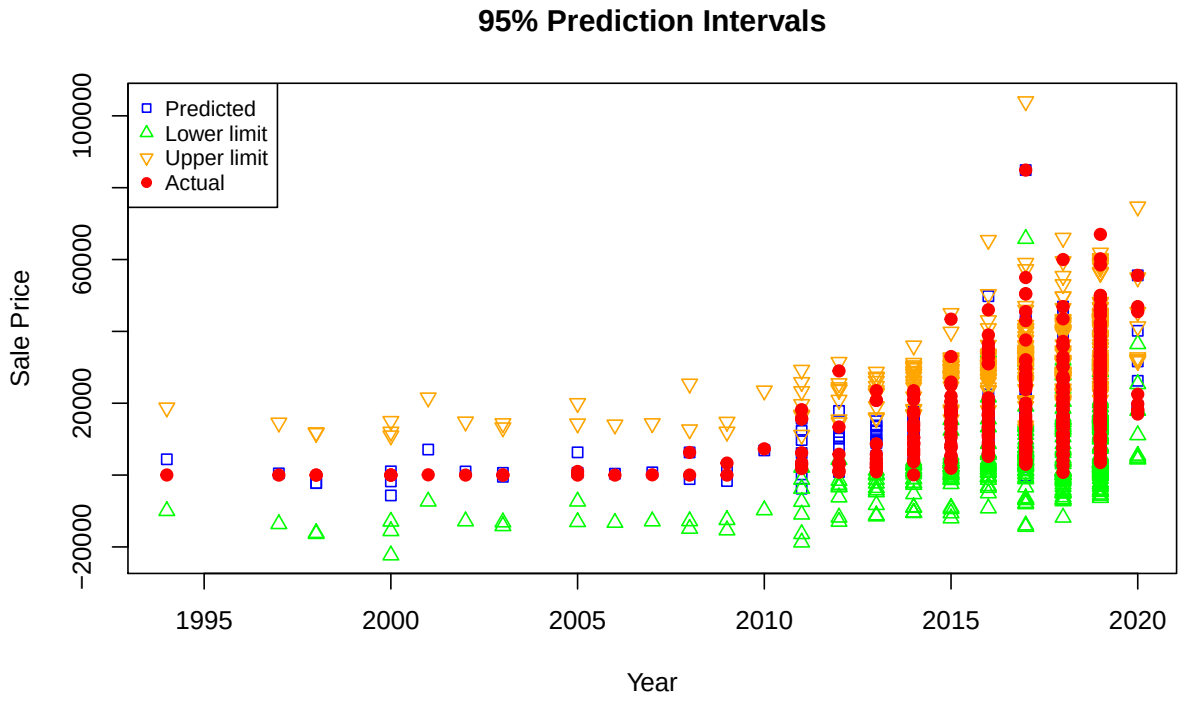
Next, we want to see how our model performs by plotting its prediction values against the actuals in the 20% of the dataset we set aside for testing purposes. As an initial step, we will show a scatterplot for each feature in the best model compared to the dependent variable "car sale price". Later on, we will compare those actual values against the predicted value.

```
# fit a linear model using 80% of the data
test20 = test
train80_fit = final_model
pred20 = predict.lm(train80_fit, data = test20)

# plot the actuals using the test data
par(mfrow = c(2,2))
plot(price~ year +  mileage + factor(title_status) + factor(model) ,
     data = test20,
     col = "red", main = "Scatter Plot of Actuals",
     ylab = "Car Sale Price")
```

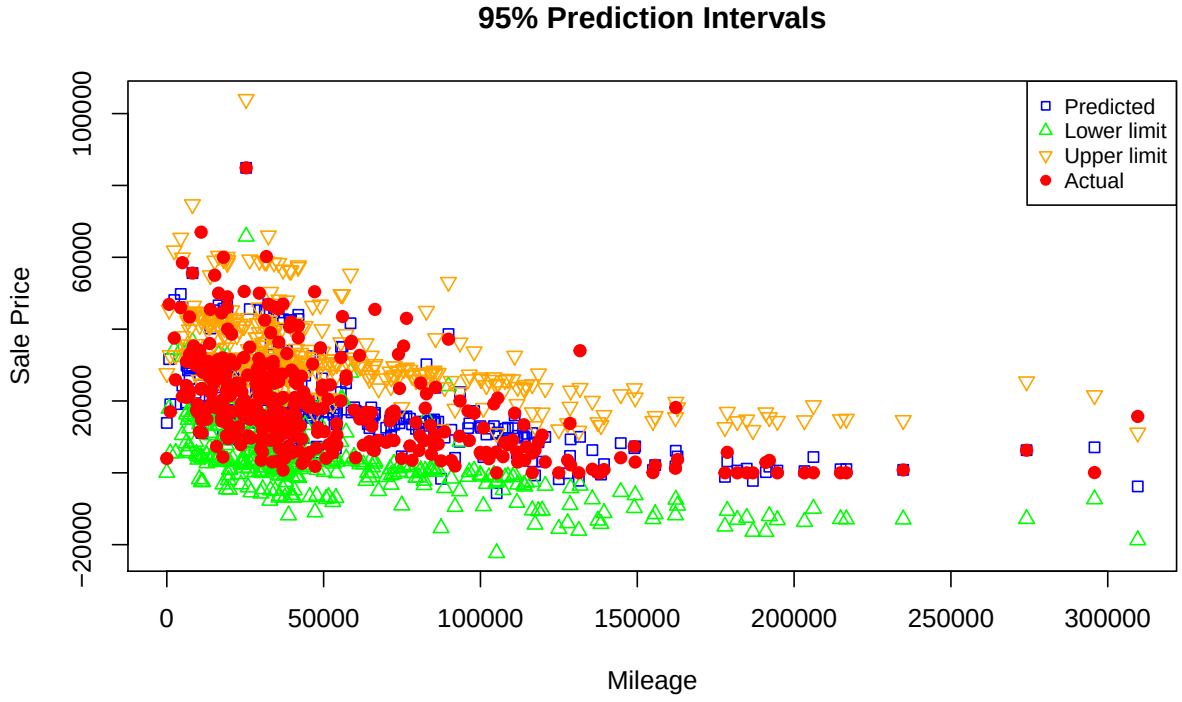


Specifically, for the quantitative variables (**year** and **mileage**), we can see that the majority of actuals fall within the 95% prediction interval. This is expected because we see a high R^2_{adj} in our training model and relatively low RMSE. Since the 20% test data points were randomly set aside, we can conclude that our model has a strong prediction power.

```
# find the prediction intervals of the final model using the test data
pred_interval = predict(train80_fit, newdata=test20, interval="prediction",
                        level = 0.95)

# plot the actuals and prediction intervals for year
matplot(test20$year, pred_interval, pch=c(22,2,6), col=c("blue","green","orange"),
 main="95% Prediction Intervals", xlab="Year", ylab="Sale Price", cex.main =1.15)
points(test20$year, test20$price, col="red", pch=19)
legend("topleft", legend=c("Predicted","Lower limit","Upper limit","Actual"),
       col=c("blue","green","orange", "red"), pch=c(22,2,6,19), cex=0.8)
```




```
# plot the actuals and prediction intervals for mileage
matplot(test20$mileage, pred_interval, pch=c(22,2,6),
↪  col=c("blue","green","orange"),
 main="95% Prediction Intervals", xlab="Mileage", ylab="Sale Price", cex.main =1.15)
points(test20$mileage, test20$price, col="red", pch=19)
legend("topright", legend=c("Predicted","Lower limit","Upper limit","Actual"),
       col=c("blue","green","orange", "red"), pch=c(22,2,6,19), cex=0.8)
```



Chapter 4: Conclusion, Discussion, and Future Plan

Conclusion and Discussion

In summary, we were aiming to build a model to predict the used car sale price in the US. Our initial data exploration showed that a variety of features can affect the price, like how old the car is, what mileage is on it, whether it is new or rebuilt, the brand, etc. However, not all of these variables contributed equally or significantly to the response variable (**price**). After removing unrelated features conceptually, we were able to identify the most statistically important main effects (**year, mileage, title_status** and **model**) based on R^2_{adj}, RMSE, and individual t-tests. For the qualitative **model** variable, we used a partial F-test to conclude whether we should keep it as a whole or discard it as a whole since it was comprised of so many levels. Interaction terms were identified using the Stepwise procedures and consistent results were obtained using the elimination procedure and all-possible-regressions selection procedure. Even though the higher-order term **mileage^2** was found to be significant (individual t-test had a p-value of 0.0000153), it did not improve the R^2_{adj} and RMSE values significantly. To avoid over complicating our model, we did not include any higher-order terms and decided to only keep the main effects and interaction terms we found. Our final model is:

$$price = \beta_0 + \beta_1 * year + \beta_2 * mileage + \beta_3 * title\_status + \beta_4 * model + \beta_5 * mileage * year + \beta_6 * mileage * title\_status + \epsilon$$

NOTE: **year** and **mileage** are quantitative variables, while **title_status** (2 levels) and **model** (43 levels) are qualitative variables.

To understand this model more practically, we looked at how each term in the model affects the response variable of car sale price.

- The **year** term tells us when the car was registered. Intuitively, the older the car is, the lower the sale price will be. Another reason we kept the **year** variable is that it also accounts for important hidden factors like the GDP, market-demand of vehicle, and consumption power of the year.

- The **mileage** variable identifies how long a car has been on the road which is a crucial factor because more wear and tear will cause high maintenance fees at the auto shop. In fact, average cars will experience transmission failure at around the 100,000 miles mark. The effect of the **mileage** variable is also intuitive - the more mileage put on the car results in a decrease of value and thus sale price of the car. However, in our model, we noticed the β coefficient for mileage is a positive value, albeit small. That might be because interaction terms with mileage take away some variance.

- Similarly but more drastically, the **title_status** variable has a great impact on the car value. If we ignore the interaction effect, the average sale price difference between a clean no-damaged vehicle and a salvaged vehicle is 16,716 dollars. The title status affects the price because salvage title cars generally have a higher insurance rate as "car insurance companies will only reimburse up to 80 percent of your car's salvage value, even if you have a full coverage salvage car insurance policy." (https://www.carinsurancecomparison.com/is-insurance-more-expensive-for-a-salvage-car/).

- Finally, the effect of the **model** variable explains a large proportion of the car price, while the **brand** variable only contributes a little. When people choose to buy a vehicle, they are more likely to based their decision on the car's model over its brand. In the US market, people prefer big trucks over smaller ones, regardless of the brand (https://www.insidehook.com/article/vehicles/why-pickup-trucks-keep-getting-bigger0.

- Our two interaction terms are: **mileage*year** and **mileage*factor(title_status)**. The **mileage*year** interaction term tells us that mileage is affected by what year the car was registered in. The more recent the car is, the less mileage would effect the car price. Similarly, the **mileage*factor(title_status)** interaction term tells us that mileage is also affected by the

status of the car. A vehicle with a clean status has less mileage and this effects a car's sale price.

Future Plan

Although our final ordinary least squares (OLS) model generates a decent $R^2_{adj} = 0.6558236$, it is not without limitations and issues.

When we were screening for the best features to explain our response variable (used car sale price), we were limited to the factors we had. Out of the 11 variables, we kept 4 of them (**year, mileage, title_status, model**). Although all 4 variables were found to be significant, the model features do not account for all of the total explainable variance in the response. If we were given more options, we could potentially find another critical factor to share the accountability, which could have resulted in a better and probably more useful OLS model.

One difficulty we encountered during our model selection and building was the overwhelming number of factor levels. The **title_status** variable only had 2 levels, while the **model** variable had more than 100 levels initially. We did not remove or group any levels to maintain the model's practicality as it demonstrates the variety of options a buyer has when selecting a car. As a consequence, we were not able to apply any stepwise procedure or elimination selection procedures when determining the main effects of the model. Instead, we compared each model manually, which would have been very tedious and time-consuming if we had to consider more than 10 features. In addition, it became very hard to interpret all the coefficients.

A more serious issue we had was with the model diagnostics. It turned out that the equal variance and normality assumptions were not met. To address this, we first compared the model with and without interaction terms but the problem persisted. Additionally, from the pairwise scatterplot, we did see some high leverage points in the "year vs. price" and "mileage vs. price" plots. In multiple linear regression, high-leverage points can alter the plane of fitting. We decided to re-build the model after removing the high-leverage points that were identified, but this did not solve the problem. Lastly, we also transformed the response using the Box-Cox method to find the best lambda transformation, with different attempts after combining interaction terms and/or removing high-leverage points. Nevertheless, none of these efforts resulted in better model diagnostics. In the end, regardless of the assumption violations, we continued to interpret the coefficients although we know that especially when Normality Assumption is violated, the inferences were not completely valid. We did this to reflect our understanding of the concept and to also give a general idea of how each feature may impact a car's sale price.

Fortunately, mathematicians make sure there are always tools in the toolbox.

- In relation to the Normality Assumption being not met, one thing that can be done in the future is to find more features to eat up more response variance. Collecting more data points can also offset effects of some trivial departures from normality. This is especially the case when the number of data points is not small; the linear regression statistic will not be significantly affected even if the population distributions are skewed. Another way we can work around the Normality Assumption is using bootstrap which is a non-parametric test that does not assume normality. Even though it is known that a non-parametric test like bootstrap does not have the same statistical power as a parametric test, it does give a better idea of the distribution and potentially would produce a better prediction model (Hayashi, Fumio, 2000).
- In relation to the Heteroscedasticity assumption being not met, the OLS estimator is actually still consistent and unbiased but probably is no longer the most efficient in the class of linear estimators. In this case we also have some useful tools as listed below:
  - If you know how the variance changes with each yi
    - Weighted Least Square Regression
  - If the variance has unknown dependence on yi

    - Other regression methods (GMM, generalized method of moments estimation) (Kutner, Michael H, 2005)
- If you know the general trend of variance change with the predictor variable, then you can transform the data
    - Log-transformation
    - Box-Cox transformations

Since some of these methods were outside of the scope of DATA 603 they were not completed for this project. In particular, the GMM estimator seems to be a strong solution for this problem, We believe our model can definitely be improved and be more useful with all of these considerations taken into account. However, because of time constraints and knowledge barriers, we conclude with the model we developed and have described our understanding of how it can be used to predict used car sale prices in the US.

References

Alsenani, D. (2019). US Cars Dataset. Kaggle. Retrieved November 18, 2021, from https://www.kaggle.com/doaaalsenani/usa-cers-dataset

Borrelli, L. (2021). Car ownership statistics. Bankrate. Retrieved November 22, 2021, from https://www.bankrate.com/insurance/car/car-ownership-statistics/

Ellencweig, B., Ezratty, S., Fleming, D., & Miller, I. (2019, June 6). Used Cars, new platforms: Accelerating sales in a digitally disrupted market. McKinsey & Company. Retrieved November 25, 2021, from https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/used-cars-new-platforms-accelerating-sales-in-a-digitally-disrupted-market.

Yahoo! (2021, August 25). United States used car market report 2021-2025 featuring TrueCar, Pendragon, CarMax, Autonation, Asbury Automotive, & Penske Automotive among others - researchandmarkets.com. Yahoo! Finance. Retrieved November 25, 2021, from https://finance.yahoo.com/news/united-states-used-car-market-120700527.html?fr=yhssrp_catchall.

Hayashi, Fumio. : "Econometrics.", Princeton University Press, 2000

Kutner, Michael H., et al. "Applied linear statistical models.", McGraw-Hill Irwin, 2005.