Predicting Ford Car Prices from Various Factors

Ashley Ibarra, Kathy Chen, Matthew Chen

California Polytechnic University, San Luis Obispo

STAT 334: Applied Linear Models

Prof. Bret Holladay

May 7, 2022

I. INTRODUCTION

Our research question asks “Is there a relationship between the price of used Ford cars between the years 1996 to 2020 with miles per gallon (mpg), mileage, engine size, model, transmission, year, and fuel type?” If there is a relationship, how would varying these variables change the price of used Ford cars? Through our analysis, we hope to obtain enough insight into the different factors that may affect the price of used Ford cars so that we can make predictions of used Ford car prices. Our findings could be beneficial to people who have an interest in purchasing used Ford cars but cannot decide which model, mpg, engine size, mileage, transmission, year, or fuel type to choose from that best matches their price range. In addition, car manufacturing industries might find our study useful when pricing cars at a certain value. As a couple of car-loving enthusiasts, my team and I found this dataset particularly interesting as we are very interested in how the various predictors, (mpg, mileage, engine size, model, transmission, year, and fuel type) have an impact on the price of the Ford car. Our two goals in this study are 1) to find which predictors have the most significant impact on the price of a Ford car and 2) to evaluate how changing our different predictors would change the price of a Ford car.

II. MATERIALS AND METHODS

Our group found the dataset on a website called Kaggle.com under the category Datasets. We typed the keyword “Car” in the search engine and the dataset “Ford Car price Prediction” was listed on page 4 of the results. The dataset was uploaded just a month ago by the user Adhurim Quku so we know that any potential outdated information would be at the absolute minimum.

We have decided to use mileage (in miles), mpg(miles per gallon), engine size (in liters), model, fuel type, transmission, and year as our predictors since we infer that they could have an association with price. Mileage, mpg, engine size, and year are the quantitative variables while model, fuel type, and transmission are categorical variables. The observational units are used Ford cars. The response variable of interest is the price (in dollars). We have found that 86% of the used cars in the dataset have manual transmission and the other 14% are either automatic or semi-automatic. We believe it does not play a huge role in predicting price range since the sample size of transmission is unbalanced. For our categorical predictors, we have 23 brands of cars and 5 types of fuel. The 23 brands of Ford cars are B-max, Edge, Focus, Grand C-max, Ka+, Mustang, S-Max, Tourneo, C-Max, Escort, Fusion, Grand Tourneo Connect, Kuga, Puma, Streetka, Transit Tourneo, EcoSport, Fiesta, Galaxy, KA, Mondeo, Ranger, and Tourneo Connect. The five types of fuel are Diesel, Electric, Hybrid, Other, and Petrol. We plan to see which variable affects price the most after adjusting for the other variable and which variable would affect price the most when other variables are not in the model. We do not know how the creator acquired the data since he did not provide the sources on the webpage and we’re also not sure if random sampling was used but we can assume that based on our understanding of observational studies, random sampling was involved. We can assume that the large dataset was generated from used car records.

III. SPLIT THE DATA

The dataset we are using in our analysis of predicting Ford car prices has 17966 observations overall. We are splitting the data into two subsets where one subset will be used for applying the formal analysis and the other will be used as a comparison to see how well our model has actually predicted the price of cars. The subset of the data that will primarily be used in our analysis will be called the training dataset and it

will have 80% of the original data in it. The training dataset has 14373 observations. The data that will be used for comparison will be called the test data and has 20% of the original data stored in it. The test dataset has 3593 observations.

To obtain the subsets of data to be used in the training dataset and the testing dataset, we used Rstudio to determine the number of observations that was 80% of the original dataset. This was 14373. We used the command set.seed(334) to ensure that the same subset of the data was generated each time we ran our code. Then we proceeded to create both of the subsets of data in Rstudio.

IV. DATA VISUALIZATION

Part A:

The associations shown in our matrix scatterplot (Figure 1) do not appear to be linear. For instance price and mpg, and price and engine size are not linear; therefore we would need to perform a transformation. None of the explanatory variables seem to be correlated with each other since there is a weak association between the explanatory variables.

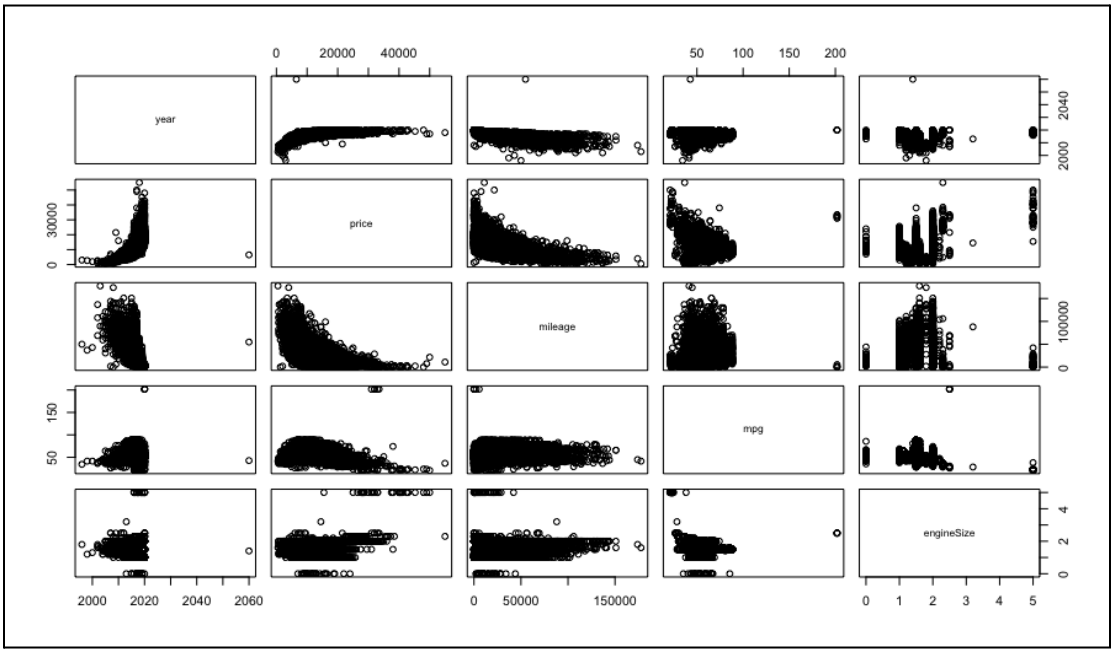

Figure 1: Matrix Scatterplot of Year, Price, Mileage, MPG, and Engine Size

There are some huge outliers in the association between price and mpg, price and engine size, mileage and mpg and year with all the predictors. Some of the associations between the explanatory variable and price were quite unexpected. We expected large engine size, more mpg, and high mileage to be associated with larger price but it seems like only engine size and price are positively correlated. The other two explanatory variables had either weak associations (price and mpg) or moderately negative associations (price and mileage).

Out of all the explanatory variables, mileage is most strongly associated with year. They have a negative, moderately strong relationship with a correlation coefficient of -0.703 according to the correlation matrix (Table 1). Price and year also have a pretty strong relationship with each other (r=0.633) but not as strong

as mileage and price. It has a correlation coefficient of 0.414 and unlike the association between mileage and price, engine size and price have a positive relationship.

	year	price	mileage	mpg	engineSize
year	1.00000000	0.6328761	-0.7028497	-0.01650403	-0.1361666
price	0.63287605	1.0000000	-0.5306912	-0.33959818	0.4141037
mileage	-0.70284970	-0.5306912	1.0000000	0.11464141	0.2156960
mpg	-0.01650403	-0.3395982	0.1146414	1.00000000	-0.2575098
engineSize	-0.13616658	0.4141037	0.2156960	-0.25750979	1.0000000

Table 1: Correlation Matrix of Year, Price, Mileage, MPG, and Engine Size

We decided to also observe the matrix scatterplot coded by the four car models, Edge, Fiesta, KA, Ka+ , as shown in Figure 1-1. We can see that transformations are needed for each predictor, but Edge cars tend to have higher prices when comparing price to year of manufacture and price to mileage.

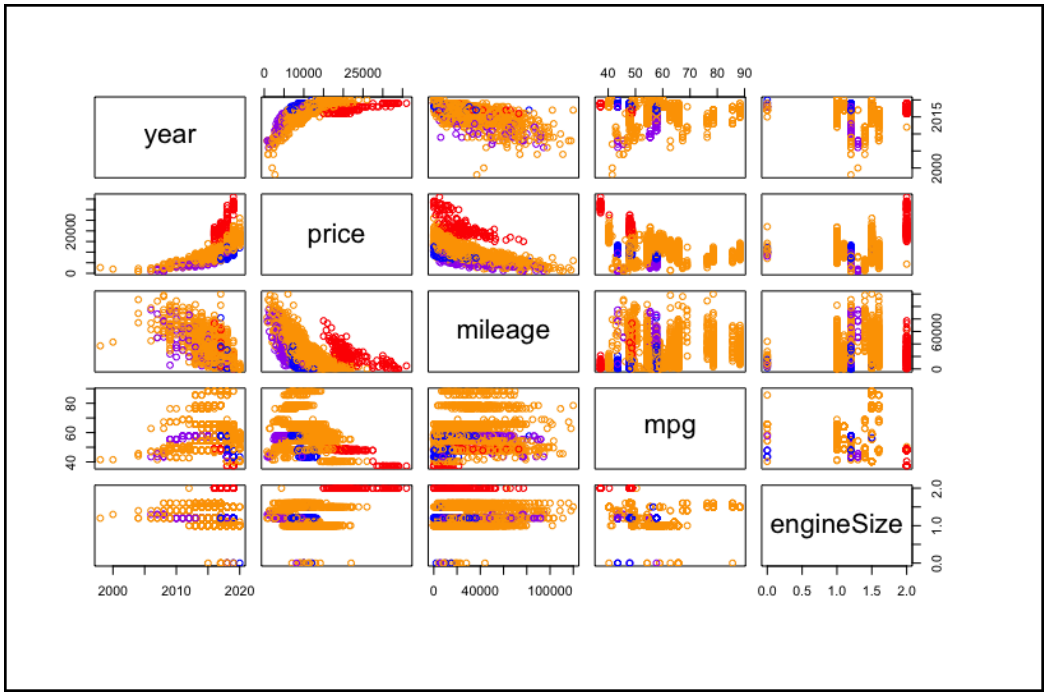

Figure 1-1: Color Coded Matrix Scatterplot of Year, Price, Mileage, MPG, and Engine Size by Car Model.

Legend of Model: Edge: Red, Fiesta: Orange, KA: Purple, Ka+: Blue

Part B:

Interaction effects show the impact one predictor has on the response based on another predictor. We decided to investigate whether there was an interaction effect on miles per gallon and car model of Edge, Fiesta, KA, and Ka+. To observe if there was an interaction between miles per gallon and car model, we first looked at a coded scatter plot, in Figure 2-1, of miles per gallon and the square root of price coded by the four car models that was included in our final model. We can see that the Edge model seems to have a steeper slope than the other models. We can also observe that the KA model may have a flatter slope than the other models. The interaction between miles per gallon and car model shows the impact of miles per gallon has on price based on different Ford car models. We can observe that the slopes between each of

the car models are not parallel. Therefore we can visually tell that there will be a statistically significant interaction between these two factors.

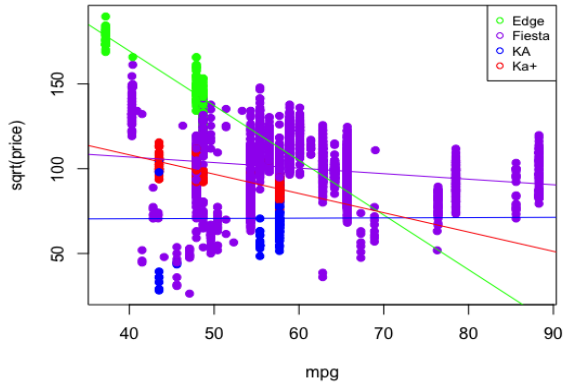

Figure 2-1: Interaction between miles per gallon and car models based on sqrt(price)

We have included an additional interaction that would investigate an interaction between two categorical variables. To further investigate the various predictors, we decided to see if there was an interaction between transmission and fuel type, based on the square root of price. We decided to remove electric from the fuel type variable since there were very few observations of electric cars in the dataset. From Figure 2-2, we can visually inspect that there may not be a significant interaction between transmission and fuel type since the lines appear to be somewhat parallel. Hybrid cars might make the impact of transmission on the price be greater than the other fuel types since the slope for hybrid cars is higher and steeper on the plot.

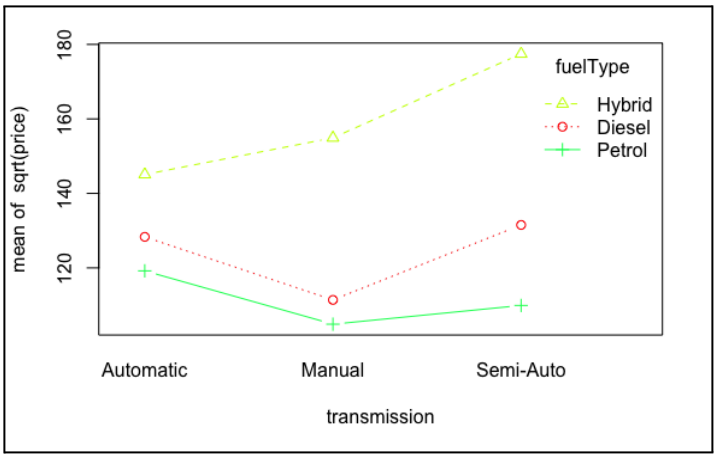

Figure 2-2: Interaction between transmission and price based on fuel type

VI. VARIABLE PRE-PROCESSING

To decide the type of transformation that would best fit our regression function of price on mileage, mpg, engine size, model, fuel type, transmission, and year we need to first examine the normality, equal variance, and linearity assumptions for that original model.

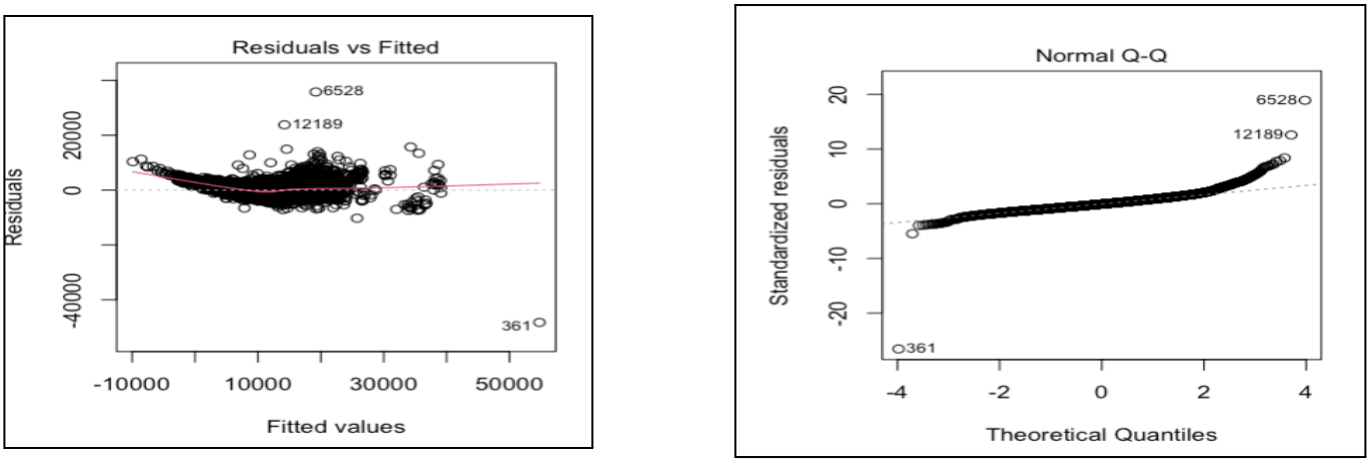

Figure 3: Residual vs Fitted and QQ plot for the full model: price~mileage+mpg+engineSize+model+fuelType+transmission+year

The residual vs fitted (figure 3) shows fanning with points spreading out on the far right which means a violation of the equal variance assumption. In addition, the QQ plot shows points deviating from the diagonal line which translates to a violation of the normality assumption. Since both normality and equal variance are most likely violated, we would do a transformation on our response variable, price. We decided that decreasing the power of price would be best because most of the scatterplots on the matrix scatterplot (figure 1) have a curve where points on the left of the scatterplot are higher. From our analysis in R, engine size has the smallest standard deviation and mileage has the most unusual observations.

The first transformation we decided to test out was a log(y) transformation. If that does not improve the assumptions and linearity in the matrix scatterplot, then we can try increasing or decreasing the power even further.

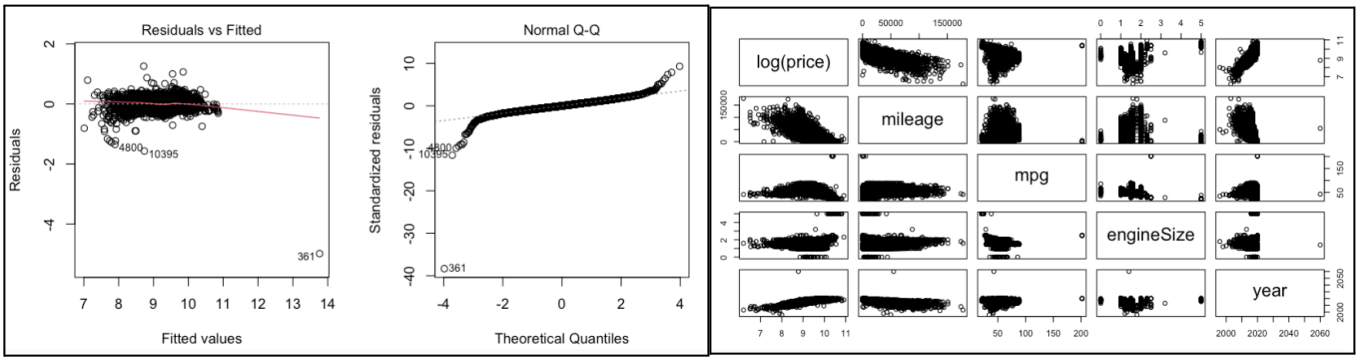

Figure 4: Residual vs fitted plot, QQ plot, correlation matrix for log(price) transformation

The log transformation improved linearity and equal variance drastically since there is no fanning or obvious trend in the residual vs predicted plot. There is one obvious outlier which is observation number 361 and less obvious ones like observations 4800 and 10395. The normality assumption still appears to be violated since points on the far right and far left of the QQ plot are straying away from the diagonal line. The QQ plot did not change much after the log(y) transformation. Because the normality assumption is violated and the scatterplot matrix doesn't show a linear line, we decided to try another transformation to better fit the regression.

Next, we decided to decrease the power even further to see if that would make the assumptions and linear trend better. It did not. Any transformation that is lower than a power of 0 (which is the log transformation) makes all the assumptions violated. For instance, a transformation of -1/sqrt(price) which has a power of -½ has fanning (indication of violation in equal variance), a trend(indication of linearity violated), and obvious deviation of points from the diagonal line in the QQ plot (indication of normality violated). (figure 5)

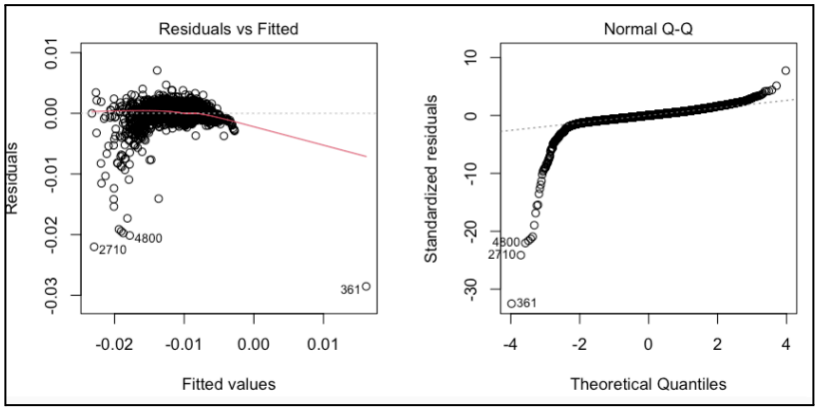

Figure 5: *Residual vs fitted plot, QQ plot for -1/sqrt(price) transformation*

Our next option is to increase the power from zero to ½ since decreasing it below 0, as we saw just earlier, causes conflicts with our normality, linearity, and equal variance assumptions. Making the power ½ which gives us the sqrt(price) transformation definitely helps with the linearity and equal variance assumption but normality is still not met because points on the far right on the QQ plot are moving away from the diagonal line. Even though normality is violated, we decided to proceed with the analysis but with caution. The sqrt made all the assumptions look the best out of all the transformations so we decided to stick with that. Interactions, polynomials, and centering did not help normality by much so we decided to stick with this simpler model by not adding them.

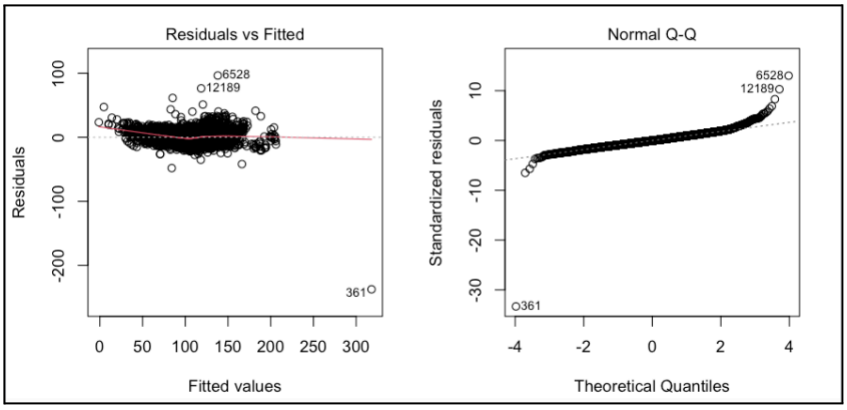

Figure 6: *Residual vs fitted plot, QQ plot for sqrt(price) transformation*

From the plots above, it seems like observation 361 is an extreme outlier with 6528 and 12189 as moderate outliers. In addition, R gives a warning, saying that observations 1448, 1908, and 3886 all have a leverage of 1 which is quite high. Because 361, 1448, 1908, 6528, 12189, and 3886 are all pretty extreme, we decided to remove these observations. Now we have 14367 observations instead of 14373.

To reduce our full model so it is more simple and less complex we decided on the five predictor model with predictors: mileage, mpg, engine size, model, and year which has a mallow's cp of 4833.77. We chose this model over the others because it had the smallest mallow's cp and was significantly smaller than the others. This means this model has the smallest combined variance and bias which is desirable.

VII. RESIDUAL ANALYSIS

Now that we have our new reduced model, we need to determine whether or not linearity, equal variance, and normality are met. According to figure 7, all the assumptions are violated which means we should transform the response variable. There are a few outliers like observations 13010, 4842, and 6401 so we decided to remove them.

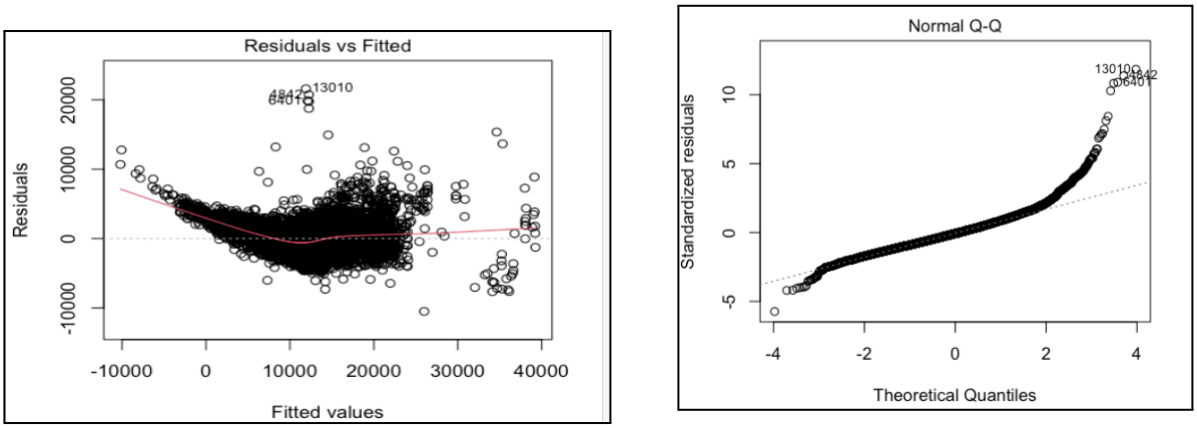

Figure 7: Residual vs Fitted and QQ plot for the reduced model: price~mileage+mpg+engineSize+model+year

This time I tested the sqrt transformation (figure 8) on our reduced model first since it looked the best on our full model. The sqrt transformation helped the equal variance and linearity assumption look better since there is no fanning or trend so I think they are met. Normality did not improve or get worse with the sqrt transformation. We think that assumption is still violated.

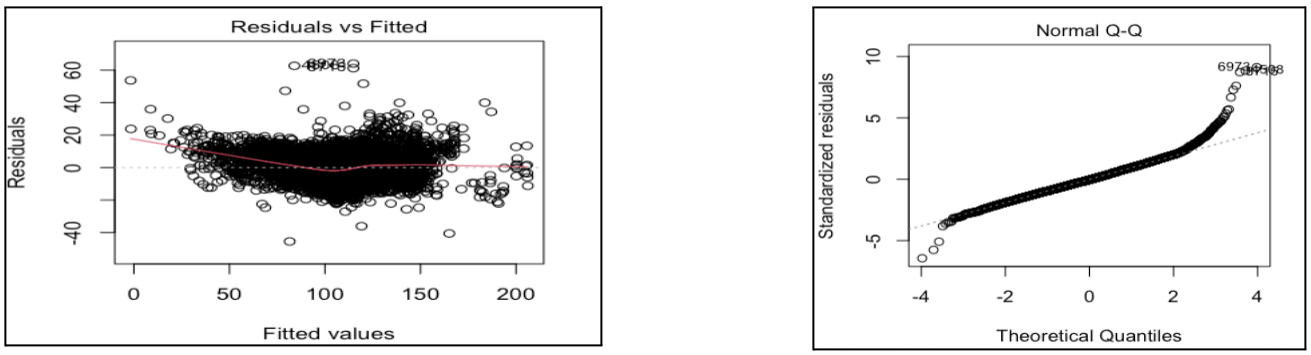

Figure 8: Residual vs Fitted and QQ plot for sqrt transformation on the reduced model

We wanted to see if the other transformations would make the normality assumption better looking so the next transformation we tried was log(y).

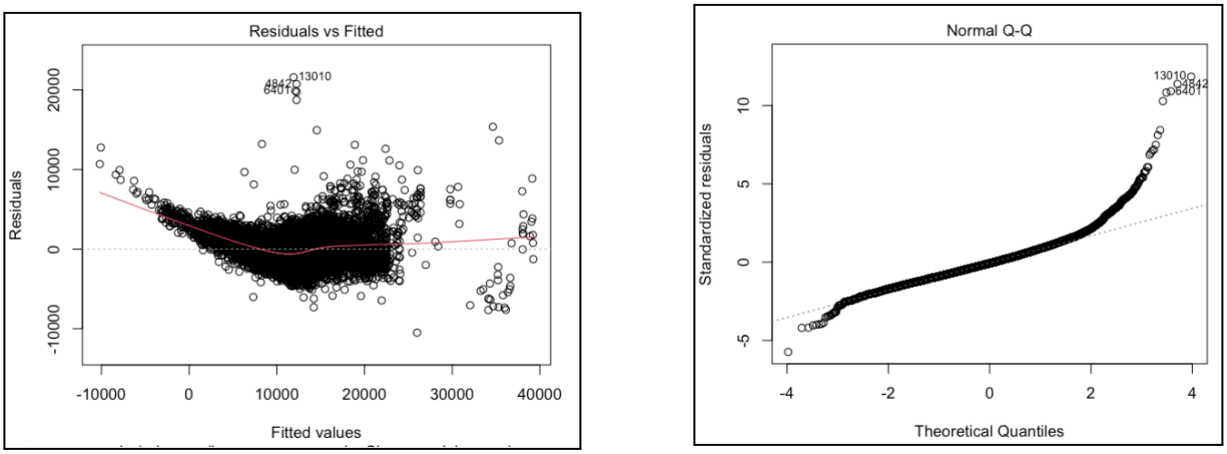

Figure 9: Residual vs Fitted and QQ plot for log(price) transformation on the reduced model

The log(y) transformation made equal variance and linearity look way worse but normality looked approximately the same as the sqrt transformation. Because of the violation in all the assumptions, we decided to stick with the sqrt transformation. We tried other transformations as well like the -1/sqrt(price), -1/(price), and price^2, and those just made the linearity and equal variance worse.

Since the linearity assumption and equal variance assumption look met on the sqrt transformation, we want to conduct some formal tests to determine if they are really met.

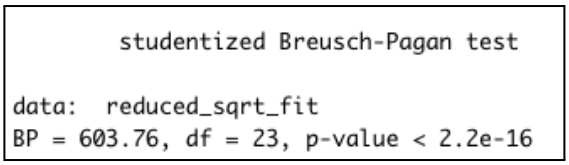

Figure 10: equal variance test for sqrt(price) on reduced model

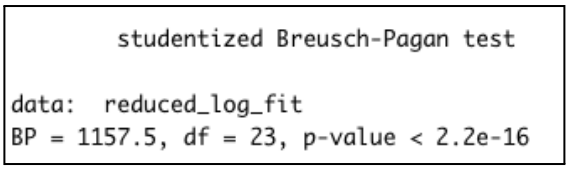

Figure 11: equal variance test for log(price) on reduced model

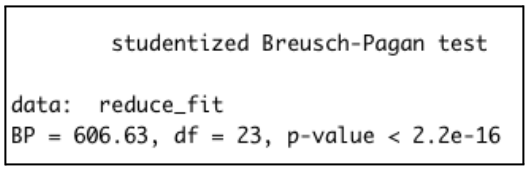

Figure 12: equal variance test for -1/sqrt(price)

The equal variance assumption is violated (figure 10) with a small p-value of essentially zero for the sqrt transformation which was surprising to us because the residual vs fitted plot looks fine. We then performed the Breusch Pagan test for the log(y) transformation and -1/sqrt(y) and they were also violated (figure 11 and 12). The BP is lowest for the sqrt transformation (603.76 vs 1157 and 606.63) which makes the equal variance assumption better than the other transformation but regardless it is still not met because of a p-value of essentially zero.

We then added polynomials (on variables year alone, then mpg, and mileage) to the sqrt model to attempt to improve equal variance but the BP test did not produce a large pvalue which we want and instead produced a large BP which is bad. For example, in figure 13, we did a quadratic polynomial on year and the p-value was essentially zero. So, we decided that because the BP got larger with the polynomials, then we shouldn't add them. We ended up sticking with our original sqrt transformation model without polynomials since that model was simpler and had a smaller BP.

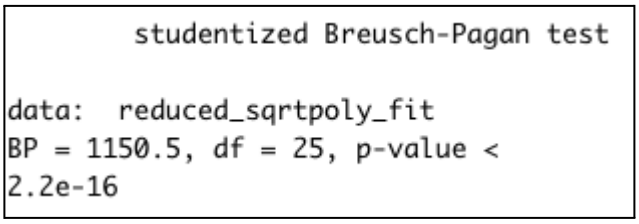

Figure 13: equal variance test for sqrt transformation with quadratic polynomial on year

A Shapiro Wilk test for normality could not be conducted because our sample size for our dataset was above 5000 observations so R could not run it. Although, we expect the Shirpo Wilk test to produce a very small p value anyway because normality looks violated according to the plots above.

We also performed a lack of fit test to test linearity since there are replicates in the variables mpg and engine size as shown in figure 14. Unfortunately with a very small p-value of almost zero, we have to conclude that linearity is violated.

```
Analysis of Variance Table

Model 1: sqrt(price) ~ mileage + mpg + engineSize + model + year
Model 2: sqrt(price) ~ mileage + as.factor(mpg) + as.factor(engineSize) +
    model + year
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1  14337 721268
2  14241 495961 96    225307 67.39 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 14: Lack of fit test for linearity with replicates in engineSize and mpg

Even though equal variance, linearity, and normality are all violated, we decided to proceed with the rest of our analysis with caution. We tried our best to find the best transformation and to fix equal variance with polynomials but those did not work.

VIII. FIT A LINEAR MODEL

Regression Model:

predicted square root(price) = -9,453 - 0.0002511(mileage) - 0.02797(mpg) + 2.760 (engine size) + -40.38 (model Fiesta) - 55.05(model KA) - 57.43(model Ka +) + 4.766(year)

The final linear model we have selected has 5 predictors. The four quantitative predictors we have are mileage, miles per gallon, engine size, and year. The one categorical predictor we have is the car model which has four levels that are Fiesta, Edge, KA, and Ka+. The model predicts the price of Ford cars based on these 5 predictors.

The mileage predictor lowers the price of the car since cars with high mileage have more wear and tear. The miles per gallon predictor increases the predicted price since it indicates whether a car is more fuel-efficient. The engine size predictor increases the predicted price since it is often more expensive to build a bigger sized engine. The year predictor increases the predicted price since cars built more recently are newer therefore cost more than older cars. The model predictor indicates that each model of the car decreases the price. The overall behavior of the model does well to predict the price. Eighty-four percent of the variability of price is explained by mileage, mpg, engine size, year and the brand of the car. The typical deviation in the actual price of Ford cars compared to the predicted price given values for mileage, miles per gallon, engine size, year of manufacture, and the car model is $6.71.

When mileage, miles per gallon, engine size, and year are all zero, the predicted price of Edge Ford cars is -$9,452.54. When mileage, miles per gallon, engine size, and year are all zero, the predicted price of Fiesta Ford cars is -$9,492.92. When mileage, miles per gallon, engine size, and year are all zero, the predicted price of KA Ford cars is -$9,507.60. When mileage, miles per gallon, engine size, and year are all zero, the predicted price of Ka+ Ford cars is -$9,509.98.

We estimate the mean price for Ford cars that are the Fiesta model is $40.38 lower than the Edge model for any given mileage, mpg, engine size, and year. For every 1 year increase for the year of manufacture, the estimated mean increase in the price of a car is $4.77 after adjusting for the other predictors in the model.

To observe if there is a problem with multicollinearity, we calculated the $[GVIF^{1/(2*df)}]^2$ for each predictor and obsessed the values to determine if the values were less than 5 which would indicate moderate multicollinearity or if the values were less than 10 which would indicate severe multicollinearity. From Table 2, we can observe that the VIF for each predictor is less than 5 which would indicate that there are no issues in the predictors that are caused by multicollinearity

	Mileage	MPG	Engine Size	Model	Year
$[GVIF^{1/(2*df)}]^2$	2.086	1.217	1.972	1.286	2.161

Table 2: GVIF for Mileage, MPG, Engine Size, Model, and Year

IX. STATISTICAL INFERENCE

We decided to conduct the model utility test to decide whether this model is statistically significant or not. Our null hypothesis will be that none of the variables is useful and our alternative hypothesis will be that at least one of the variables is useful. As we could see from the very small overall p-value of the summary output of the training data set minus row 4 (figure 15), we have very strong evidence to reject the null hypothesis in favor of the alternative hypothesis and conclude that at least one of the predictors is significantly useful to the model since this p-value which is derived from the partial F-test is much smaller than 0.05 and therefore this model is useful and statistically significant. This small p-value corresponds to a very big F-statistic in the summary output. In conclusion, we could say that this model with all these predictors present is better than the model with no predictors as this model with all these predictors has achieved a significant reduction in Sum of Squared Error.

```
Call:
lm(formula = sqrt(price) ~ mileage + mpg + engineSize + model +
    year, data = traindata_minusRows4)

Residuals:
    Min      1Q  Median      3Q     Max
-21.153  -4.363  -0.237   4.037  39.694

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -9.453e+03  1.294e+02 -73.047  < 2e-16 ***
mileage      -2.511e-04  7.725e-06 -32.509  < 2e-16 ***
mpg          -2.797e-01  1.176e-02 -23.796  < 2e-16 ***
engineSize    2.760e+00  5.242e-01   5.264 1.46e-07 ***
model Fiesta -4.038e+01  7.182e-01 -56.226  < 2e-16 ***
model KA     -5.505e+01  8.942e-01 -61.570  < 2e-16 ***
model Ka+    -5.743e+01  7.417e-01 -77.427  < 2e-16 ***
year          4.766e+00  6.401e-02  74.461  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.706 on 5921 degrees of freedom
Multiple R-squared:  0.8469,    Adjusted R-squared:  0.8467
F-statistic:  4678 on 7 and 5921 DF,  p-value: < 2.2e-16
```

Figure 15: Overall model test aka model utility test

We will compare our final linear model with a model with only year as the predictor using the single coefficient test. To decide whether the year predictor significantly improves the model after adjusting for all other variables, we decided to look at the F-statistic and p-value of the single variable year from the Anova Type II table (figure 16) to make our decision. We state our null hypothesis as the variable year is not different from zero after adjusting for the other variables while our alternative hypothesis will be that the variable year is different from zero after adjusting for the other variables. Due to a very small p-value of less than 0.01 and a very large F-statistic of 5544.46, we do have very strong evidence to conclude that year does significantly improve the model containing all the other predictors. In other words, we reject the null hypothesis in favor of the alternative hypothesis.

```
Analysis of Variance Table

Response: sqrt(price)
             Df Sum Sq Mean Sq F value    Pr(>F)
mileage       1 645609  645609 14357.8 < 2.2e-16 ***
mpg           1  45830   45830  1019.2 < 2.2e-16 ***
engineSize    1  88420   88420  1966.4 < 2.2e-16 ***
model         3 443382  147794  3286.8 < 2.2e-16 ***
year          1 249311  249311  5544.5 < 2.2e-16 ***
Residuals  5921 266242      45
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Anova Table (Type II tests)

Response: sqrt(price)
           Sum Sq   Df  F value    Pr(>F)
mileage     47523    1 1056.867 < 2.2e-16 ***
mpg         25462    1  566.258 < 2.2e-16 ***
engineSize   1246    1   27.714 1.456e-07 ***
model      327196    3 2425.519 < 2.2e-16 ***
year       249311    1 5544.455 < 2.2e-16 ***
Residuals  266242 5921
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 16: Type I & Type II Anova table

Next, we want to test if the interaction between model and mpg is statistically significant. Our null hypothesis is testing if the interaction coefficient is equal to zero while our alternative hypothesis will be that at least one of the interaction coefficients is nonzero. Due to a very small p-value of less than 0.01 calculated using 5921 degrees of freedom with a large F-statistic of 58.106, we have very strong evidence to conclude that the interaction between mpg and model is statistically significant. In conclusion, the effect of mpg is not the same across all the 4 car models. The significance of this test matches with what I

have learned in question 5 part B because these lines in the coded scatterplots are not parallel as some are so much steeper than the other so they do not have the same slope and therefore there is an interaction.

```
Analysis of Variance Table

Model 1: sqrt(price) ~ mpg + model
Model 2: sqrt(price) ~ mpg + model + mpg:model
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1   5924 1138043
2   5921 1105497  3     32546 58.106 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 17: Anova interaction between mpg and model

We are also interested in whether or not the interaction between mpg and engine size is statistically significant. Our null hypothesis for this test would be that the interaction coefficient is zero while our alternative hypothesis would be that the interaction coefficient will not be zero.

```
lm(formula = sqrt(price) ~ mpg + engineSize + mpg:engineSize,
    data = traindata_minusRows4)

Residuals:
    Min      1Q  Median      3Q     Max
-77.311 -10.068  -0.056   9.476  87.243

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -18.69125    7.06338  -2.646  0.00816 **
mpg             1.97937    0.11927  16.596  < 2e-16 ***
engineSize    108.88006    5.03456  21.627  < 2e-16 ***
mpg:engineSize  -1.82934    0.08679 -21.078  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16 on 5925 degrees of freedom
Multiple R-squared:  0.1279,    Adjusted R-squared:  0.1274
F-statistic: 289.6 on 3 and 5925 DF,  p-value: < 2.2e-16
```

Figure 18: Summary output with an interaction between mpg and engine size

Due to a very small p-value of less than 0.01 and a very small t-statistic of - 21.078 of the interaction term between mpg and engine size from figure 18, we can conclude that this interaction term is statistically significant so we reject the null hypothesis in favor of the alternative hypothesis. In conclusion, the negative interaction term means that the effect of miles per gallon decreases as engine size increases and vice versa. This makes sense in context because cars with bigger engine size usually eat up more fuel than cars with smaller engine size.

After finding out that year has the largest positive coefficient, we decided to carry out the confidence interval for a mean response value of year and the prediction interval for a future predicted value of year to see how helpful it is in predicting a ford car price.

```
> predict(sqrtpriceyearfit, new.year, interval = 'confidence', level =.95)
      fit      lwr      upr
1 107.4776 107.1327 107.8225
```

Figure 19: Confidence interval for predicting mean price with the 4 car models during year 2018

```
> predict(sqrtpriceyearfit, new.year, interval = 'prediction', level =.95)
       fit      lwr      upr
1 107.4776 84.60869 130.3465
```

Figure 20: Prediction interval for individual Ford car price with the 4 models during year 2018

Since the response variable in figure 19 and figure 20 is square-rooted, we need to square it to interpret the confidence and prediction intervals. After squaring, the confidence interval should be (11477.42,11625.69) and our prediction interval should be (7158.68,16990.21). Therefore, we are 95% confident that the mean price for used Ford cars in 2018 with models KA, ka+, Fiesta, and Edge is between $11477.42 and $11625.69. We are 95% confident that individual Ford cars with model KA, ka+, Fiesta, and Edge in year 2018 will have a price between $7158.68 and $16990.21.

We choose the year 2018 because this is the 3rd quartile of the variable, year, and also occurs the most out of all the years as shown in figure 21 so we think that this is one of the reasons why this variable year is useful for predicted car prices because we could see that price and year are strongly positively correlated as newer cars are more popular with the consumers so there are more new cars in the market and the price went up as the car model got newer.

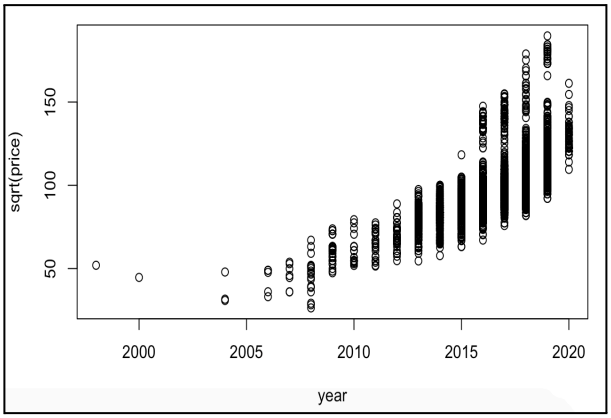

Figure 21: Scatter Plot of Sqrt(Price) vs. Year of Manufacture

X. MODEL VALIDATION

To determine if our chosen model performs well, we are comparing the mean squared prediction error (MSPE) from our model to the mean square error (MSE) of the test data that was set aside from the earlier

section. From our model, the observed MSPE value of 47.596 is fairly close to the MSE of the test data which was 45. This is an indication that the model we have chosen for our dataset does not have an issue with overfitting and the predictive ability of this model is acceptable.

XI. CONCLUSION

We have determined that our final chosen model which has 5 predictors that are mileage, miles per gallon, engine size, year, and model is somewhat valid. When fitting the model, we used a square root transformation on the predicted price since that was the transformation that gave the most visually appropriate residual plots of price fitted against the 5 predictors. The final model, however, does not formally meet any of the standard assumptions of linearity, equal variance, or normality. So we proceeded to use our chosen model with caution when continuing with the analysis.

We determined that our model was statistically significant in predicting the price of Ford cars and that the predicted price of used Ford cars is not very high. One of the weaknesses in our model was that we removed most of the car models and decided to only use Edge, Fiesta, KA, and Ka+ since we determined that the four car models improved our chosen model the most. However, this would mean that we cannot apply our results to a wide variety of Ford models and only to these four car models. In an ideal model, we would want to be able to predict the price of Ford cars along with a wide range of Ford models. Another weakness in our model is that we were not able to include more categorical factors of cars such as fuel type or transmission. For future investigations, we would suggest including more characteristics of cars such as car size, for example, whether the car was a sedan, a hatchback, or a SUV.

One of the strengths of our chosen model was that all five of the predictors were statistically significant. Therefore we concluded that the chosen factors that we used to predict price were helpful in the model. Another strength of our model was that 84% of the total variability in the square root of price was explained by our five predictors of mileage, miles per gallon, engine size, year, and model and the MSE was low as well. The low MSE indicates that the observed points did not largely deviate from the regression line from our model. If the observational study was done again, there should be more clarity on where the data was gathered from such as if the observations were gathered from online sources or directly from used car dealerships. There should also be random sampling introduced when collecting data of used cars so that the number of observations for each categorical factor is as similar as possible to help deal with any potential violations of the standard assumptions for linearity, equal variance, and normality. A take-away from our model could be that the predictors of mileage, miles per gallon, engine size, and year of manufacture could be useful in future studies of predicting car prices.

Overall, despite our chosen model performing well and having significant predictors, we would not recommend our model be used if looking to predict the price of Ford cars. We would want people to observe our model with caution and understand that we did run into several issues.

Appendix A

Datafile and Contents

The original dataset, Ford.csv, came from Kaggle.com and was uploaded by Adhurim Quku. 17966 observations and 9 variables

The categorical variables are:

➢ Model: Type of car

➢ Transmission: Type of transmission a car has

➢ Fuel Type: Type of fuel the car uses

The quantitative variables are:

➢ Year: year the car was manufactured

➢ Price: price of a car measured in U.S currency

➢ Mileage: number of miles a car has traveled

➢ MPG: miles per gallon that car has

➢ EngineSize: size of a car's engine measured in liters

Appendix B

R Source code file