|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Market Price and Mileage of Craigslist Cars  DSC 323 |  |
|  |  |  |  |
|  | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | |
|  |  | TABLE OF CONTENTS [DSC 323 1](#_Toc198936288)  [TABLE OF CONTENTS 2](#_Toc1448262333)  [INTRODUCTION 3](#_Toc2064407315)  [Methodology 3](#_Toc1624071338)  [Data Exploration 5](#_Toc805828697)  [Data Analysis 8](#_Toc1619278120)  [5-Fold cross-validation 14](#_Toc249314311)  [Survey selected training and testing 17](#_Toc443557863)  [APPENDIX 20](#_Toc839343567) |  |
|  |  |  |  |

|  |  |  |
| --- | --- | --- |
| INTRODUCTION |  | |
| With the introduction of COVID-19 in 2020, the world is trying to come to a new sense of normality. As the world is rebuilding the economy and bringing back jobs mean people are starting to return to in person lifestyles again. The return and creation of jobs means a personal decision whether to make their commute in public or personal ways of transportation for work or going to a store for supplies. It is a decision of riding to work taking a risk with cheaper priced public transportation but while COVID-19 is still around a chance of being infected. The other option is buying a personal transportation where they can control their vehicle who enters and exits and what comes in contact. “Covid-19 has led to an increase in used car sales as people avoid mass transportation and are more sensitive to auto cost in the [recession](https://www.cnbc.com/2020/10/15/used-car-boom-is-one-of-hottest-coronavirus-markets-for-consumers.html#:~:text=Covid%2D19%20has%20led%20to,auto%20cost%20in%20the%20recession.&text=Consumers%20can%20now%20find%20greater,demand%20is%20driving%20up%20prices.)” (Rossenbaum). Instead of buying straight from realtors and brand companies, individuals are finding cheaper and affordable way of transportation. “Consumers can now find greater inventory of used vehicles online, and dealers are accelerating digital efforts too, but it can come at a premium and demand is driving up [prices](https://www.cnbc.com/2020/10/15/used-car-boom-is-one-of-hottest-coronavirus-markets-for-consumers.html#:~:text=Covid%2D19%20has%20led%20to,auto%20cost%20in%20the%20recession.&text=Consumers%20can%20now%20find%20greater,demand%20is%20driving%20up%20prices.)” (Rossenbaum). With the economy being down and many individuals not earning a stable income car shopping on websites such as Craigslist is an affordable solution. Going on a website like Craigslist to find used cars while they list general information being able to compare per each listing which has the efficient odometer and price. The other independent variables that help construct an answer of affordable vehicles such as car type, car drive, car transmission, car condition, car fuel. | |  |

|  |  |  |
| --- | --- | --- |
| Methodology |  | |
| We as a group found the dataset that we will be using in our project from this link: <https://www.kaggle.com/austinreese/craigslist-carstrucks-data?select=vehicles.csv>. We cleaned our data through several steps. The first thing we did was change cylinder from a text field into a numerical value. Before changing the data, the column for cylinders was “x cylinders” where x was the number of cylinders the car had. After changing the cells, the column would instead hold the number of cylinders as a number. The next thing we did was look for inconsistencies. We only used observations that had all the columns filled in without any missing information. Lastly, the remaining text variables were coded into dummy variables.  This report will heavily focus on two models Odometer and Price and their data exploration, final models, predictions, validation, and SAS results. In the data exploration section, it will explain both Odometer and Price using a five-number summary, histograms, transformed histogram using the log transformation on price, scatterplots, and correlation.    The analysis section will include model validation of Price and Odometer with a full model, model selection, a final model, predictions.    The model validation section will be testing Price and Odometer models using 5-fold cross-validation and checking the performance statistics of a training and testing model with a 75% split. Each of the training models will be selected using the stepwise selection method to reflect the same process we will go through to create our final models. Additionally, each of the splits will use a unique seed to avoid using repeated observations in the training and testing models for price and odometer. | |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Data Exploration | |  | |
| For Data Exploration stage of our analysis our team utilized multiple techniques to find outliers, multicollinearity, and correlation between variables. All our techniques helped us make the data more symmetric and remove the -obvious outliers. Starting from 5-num summary and histogram for both Price, and Odometer.  Price  Histogram and 5 Num Summary  we used 5 Num Summary and Histogram to get a general idea for the distribution of data, as shown in Figure 1 and Figure 2. In Figure 1, we can see that the distribution of data is very skewed to the right, we can also see that it is due to couple datapoints at around 69000. When get a more details about those data points on Figure 2, which tells us that there are 3 values that are causing the skewness. Upon removing those values, we can visually see in Figure 1.1 the skewness has decreased, we can also see that the skewness is still caused by 1 data point with value around 58000. So, we decided to do another around of data cleaning. Upon removing that data point we can see that we have minimized the skewness by a lot. However, we decided that it can be improved by making doing a log transformation. Upon doing the log transformation, we can see in Figure 1.3, that the skewness is eliminated and now it looks a lot like a normal bell curve. Our findings are reflected in the 5 num summary, we can see that the spread between each 5% increment is tighter in Figure 2.1 than Figure 2. After doing the log transformation the spread is even tighter. | | |  |
| |  |  | | --- | --- | | Scatter Plots  To visualize and analyze the association between price and 8 variables, our team used scatter plots (Figure 3). Upon examining we determined that the Price and Odometer has a negative linear association and Price, and Year has a positive linear association. We could not determine the association between other 6 variables, due to them being dummy variables. We decided upon doing a linear regression model for our data, since the two variables we could determine association for were linear.  Pearson Correlation Coefficient  After scatterplots, our team analyzed the Pearson correlation coefficient values using Proc corr procedure (Figure3 & 4). This procedure allowed us to determine the correlation between price and other 8 variables. Starting from price and odometer the correlation is -0.42203, which is indicates a weak negative correlation. The correlation value between price and transmission is -0.01637, which also indicated a weak negative correlation. The correlation value between price and drive is 0.1515, which indicates a weak positive correlation between the two. The correlation value between price and type is 0.2384 which indicates a weak positive correlation. The correlation value between price and fuel is 0.1063, which indicates a weak positive correlation. The correlation value between price and condition is -0.03773, which indicates a weak negative correlation. The correlation between price and year is 0.2932 which indicates a weak positive correlation. The correlation between price and cylinders is 0.1504, which also indicates a weak positive correlation. After analyzing the correlation for price and other 8 variables, our team checked for the multicollinearity between other variables, and all the variables appeared to be independent, so we moved on to cleaning up odometer variable. |  | | | |  |
|  |

Odometer

Like Price, our team decided to follow similar steps for odometer, starting from data visualization, followed by cleanup, association, correlation, and finally multicollinearity.

Histogram and 5 Num Summary: -

We used Histogram and 5 Num Summary to visualize the data and get an idea on the spread of the data. Starting from Histogram (Figure 6), we can see that the skewness is caused by couple variables located at around 2000000, and 1000000. Once we saw the skewness, we looked at the 5num summary of the data (Figure 7), which confirmed our assumption that it is caused due to just two variables. The 5 Num Summary told us that the min of the dataset is at 0, Q1 is at 71000, Median is at 110000, Q3 is at 149000, and the max is at 425000. As a team we agreed to remove the variables that are acting as influential points, and upon removing those two variables, we can see in (Figure 8), that the skewness is minimized, but not eliminated, therefore our team decided to do another round of removing outliers. Upon removing that one outlier we can see that we have eliminated the skewness and now the graph looks like a normal bell shape (Figure 10). The findings are reflected on the 5 Num summary (Figure 11) as, our max went from being at 425000 to 293000, while other values like Q3, Median, Q1, and mean stayed the same. After visualizing and removing influential points, our team decided to find the association using scatterplots.

Scatter Plots

Like Price we used scatterplots, to visualize the association between variables (Figure 12), our findings were as expected with Odometer and Price having a negative linear association, followed by Odometer and year also having negative linear association. We could not determine the association of the other 6 variables due to them being dummy variables. Using the information, we got from the scatterplot our team moved on to finding out about the correlation between each variable.

Pearson Correlation Coefficient

We used the proc corr procedure to get the Pearson Correlation Coefficient between odometer and other 8 variables. This procedure also allowed to quickly see if there is any multicollinearity between any variables. Our findings are as follows, odometer and price have a collinearity of -0.4268, which indicates a weak negative correlation. Odometer and transmission have correlation of -0.05875, which indicates a weak negative correlation. Odometer and drive have correlation of -0.0587, which also indicates a weak negative correlation. Odometer and drive have a correlation of 0.1009, which indicates a weak positive correlation. Odometer and type have a correlation of -0.00304, which indicates a weak negative correlation. Odometer and fuel have a correlation of -0.0553, which indicates a weak negative correlation. Odometer and condition have a correlation of 0.05505 which indicates a weak positive correlation. Odometer and year have a correlation of -0.2182 which indicates a weak negative correlation. Lastly odometer and cylinders have a correlation of 0.0896, which indicates a weak positive correlation.

After looking at the correlation numbers for each variable against odometer, our team decided to check for correlation numbers for the other variables against each other, to track down the multicollinearity. Upon a detailed examination of the correlation values, we did not find any strong correlation of any variable between each other.

|  |  |  |
| --- | --- | --- |
| |  |  | | --- | --- | | Data Analysis |  |   Selection Method  Our group picked a stepwise selection method would fit the best for this dataset. After running our stepwise function there was four variables that should be included in our final model. Stepwise will rank the significance per the Odometer model for us which ended up being price, d\_type, d\_drive, and cylinders (Figure 13).  Significant Values  To determine in the Odometer model what variables are significant to we looked in the parameter estimates and determined an alpha value=0.05 to test against the Pr > |t| in Odometer. After running the code section, it was found that d\_fuel with 0.4643 needed to be removed first since it was the highest value above the alpha test of 0.05 (Figure 14.1). After that removal re-running the code again it was found that year was the next highest insignificant value at 0.4505 which is higher than 0.05 on the alpha test (Figure 14.2). After re-running the code again, it was found that d\_condition was insignificant at 0.2104 which is higher than the alpha 0.05 (Figure 14.3). The next part that needed to be removed was d\_transmission at 0.1730 which is higher than the alpha test at 0.05 (Figure 14.4). After all those removals the remaining variables left was price <.0001, d\_drive <.0001, d\_type 0.0003, and cylinders <.0001 which were all under the alpha test 0.05 (Figure 14.5).  Outliers and Influential Points  We ran the graph to find outlier and influential points. We determined as a group what to remove from the dataset was on the graph was if there is a red arrowhead for outlier and blue arrowhead for influential point. In the code we used /influence r at to show the graph. After running that code section there was a handle of influential and outlier points in our dataset. Those data points consisted of 25, 35, 333, 475, 570, and 578. Our group then decided to re-run our code after removing we did find more outliers and influential point arrowheads. Our group made the decision to check our original Ajd-R^2 value with the original dataset which is 0.3249 or 32.49% (Figure 15.1), we then compared the next Adj-R^2 with the first set of removal is 0.3131 or 31.31% (Figure 15.2). Our group’s determining factor was if there was no increase after each removal of data set in the Adj-R^2 then that is a tell us as a group that we can stop removing data point from the dataset. With our group found for the Odometer model that since after the first removal of data and there was no increase in Adj-R^2 per that removal that we needed to stop. We also wanted to reference that we tried to remove points but will not be including as the final dataset. From the original data set to the first removal dataset for the Odometer model there is no increase of change of Adj-R^2 so we will keep the original dataset for the Odometer model.  Goodness of Fit Test  To test goodness of fit per the Odometer model a null hypothesis is needed. Our group decided that our Null Hypothesis was none of the X-variables included in the Odometer model have any association with Y. While creating a null hypothesis we also created an alternative hypothesis which was at least one x-variable has a significant effect on changes in Y. To test the Null and Alternative hypotheses we needed to have a test statistic which is F=MSR/MSE. When doing our code section, we needed to find the Mean Squares Model and Error results. The Mean Square Model result was 1.539508E11 and Error result was 2,060,449,452. Now we hand did the calculations MSR is going to equal Model result and MSE is going to equal Error result. That would make the equation 1.636251E1/2,056,581,376 that would equal F=74.72. To now double check our hand calculations with SAS we then looked at the F value to find SAS calculated that it is 74.72 (Figure 16). Then we looked at the Pr > F value for the F value which was <.0001. That means based on the alpha test=0.05 means this F value is valid. The conclusion is that the null hypothesis says there is no association between any x-variable and y and should be rejected. The F-test gives strong support to the fitted model meaning since the F-test found that 74.72 is <.0001 when testing alpha=0.05, <.0001 is <0.05 meaning there is at least one X-variable that has a significant effect on Y accepting the alternative hypothesis.  Multicollinearity  Our group decided putting our determination if a variable is over the multicollinearity threshold of >=10 VIF. After running our code section, the results found that all our independent variables were all less than 1.31907 meaning no further removal of variables were needed in the Odometer model (Figure 17).  Residual Plots  For the predicted residual plot, it was found that there is a pattern of a funnel. This plot before 50,000 on the x-axis not much data was plotted. After 50,000 to 150,000 most data were plotted on the plot then. From 50,000 on the x-axis from 100,000 to -100,000 on the y-axis data is starting to spread. From 50,000 and beyond on the x-axis 100,000 on the y-axis the data is increasing creating a positive slope. From 50,000 and beyond on the x-axis -100,000 on the y-axis the data is decreasing creating a negative slope. Putting those two shaped slopes together creates that funnel pattern (Figure 18.1).  Looking at the individual x-variables residual plots our group decided that there is a pattern in the residual plot for price in the Odometer residual plot. When looking at the pattern much of the data is plotted right between 0-10,000 on the price axis and between 100,000 and -100,000 on the y-axis. After passing that section of 10,000 on the x-axis that data starts to be plotted less and starts to diagonally come closer to the 0 line on the y-axis. Looking form >10,000 on the price axis the data from 100,000 data plotted creates this slope down trend. For data -100,000 plotted that is created making a positive slope up so creating that funnel shaped pattern. The other residual plots because they are dummy variables or cylinders that they are all plotted vertically, and conclusions cannot be made about those plots (Figure 18.2).  The normal plot of residuals does have a few outliers in the 150,000 and 300,000 range. Most of the plotted data is linear. It is acknowledged that there are some distant points, but it is not so far off where it is felt that even though it may look far it is not straying away from most of the spread of data to be alarming (Figure 18.3). |

Selection Method for Price

To find the final model for price, our group decided to use the Stepwise selection method to remove any insignificant variables from the model. We decided to use Stepwise since it was the most accurate selection method when it comes to finding the price of used cars. Through the Stepwise selection method findings, we conclude that there are a total of six significant variables in the final model for predicting price. The significant variables for the final model are as follows, odometer, d\_transmission, d\_drive, d\_type, year, and cylinders. The Stepwise selection method tells us that these six variables are all significant for the final model. Figure 19 is the output we got from the selection method and it provides us with the final model variables.

Final model for Price

After concluding that condition is an insignificant variable when it comes to predicting price, we recreated the final model with all significant variables listed above from the Stepwise selection method. Using the influence r function, we found that there are a few outliers and influential points that we decided to remove from the data. As shown in figure 21, we decided to remove a total of six data points from the set that we considered to be outliers or influential points. After running the PROC REG procedure with the removed data points, our residual plots (figure 22 and figure 23) tell us there are no signs of outlier and influential data points. The PROC REG procedure output from figure 20 also gives us a adjusted R^2 value of 36.67%. The goodness of fit test gives us a F-Value of 64.01 and a P-Value of <.0001 making this a good model and proving that all variables in the model are significant. Since the P-Value is less than the alpha of .05, we know that it is a good model. We also used the VIF option in the PROC REG procedure and the output (figure 20) showed us that none of the variables have a VIF value of >10.0 this tells us that there are no issues with multicollinearity in the final model and we can use it to accurately compute predicted price of a used car.

Price Final Model Result

After removing any insignificant variables, outliers, and influential points, we can compute our final model using the output of the PROC REG procedure from figure 20. We concluded our final model for predicting price is as follows,

ln\_price = -66.81541 – 0.00000653\*(odometer) + 0.31549\*(d\_transmission) + 0.16229\*(d\_drive) + 0.17182\*(d\_type) + 0.03716\*(year) + 0.10893\*(cylinders)

This model tells us that as the reading for the odometer increases even by 1, the ln\_price of the car will decrease by 0.00000653 therefore, the higher the odometer reading, the less the car is worth. Same goes for transmission, drivetrain, and the type of car it is, as transmission type changes from automatic to manual, the ln\_price will go up by 0.31549 and if the drivetrain goes from rwd to fwd to 4wd, the ln\_price of the car will also increase by 0.16229. As the type of the car changes from sedan, to suv, to van, to coupe, and to truck, the ln\_price will also increase by 0.17182 for each type change. We can also predict that as the model year for the car increases by 1, so does the ln\_price by 0.03716. We can also see that as the number of cylinders in the car increases, so does the ln\_price by 0.10893. Therefore, the more cylinders a car has and the newer the car is, the higher the ln\_price prediction will be.

Predictions

For our first prediction of price, we used a price of $6,500, 80000 miles on the odometer, automatic transmission, real wheel drive, sedan type, diesel fuel, like-new condition, model year of 2014, 8 cylinder engine, and lm\_price prediction of 3.8129. As shown in figure 24, this gave us a prediction value of 8.8551 with a 95% confidence interval of (8.6614,9.0488) and a prediction interval of (7.2783,10.4319).

For our second prediction of price, we used a price of $85,000, 20 miles on the odometer, manual transmission, front wheel drive, van type, diesel fuel, excellent condition, 2021 model year, 10 cylinders, and ln\_price prediction of 4.92. Figure 24 shows us the predicted value for this prediction was 10.2449 with a 95% confidence interval of (9.9779,10.5119) and a prediction interval of (7.2783,10.4319).

|  |  |  |
| --- | --- | --- |
| 5-Fold cross-validation |  | |
| Procedure  For both of the 5-fold cross-validation testing for ln\_price and odometer, 75% of the data set was used for the training models, and 25% was used for the test models. The total number of observations in the data set was 648, so that meant about 480 was for training and about 165 for testing - plus or minus a few observations depending on how the seed splits them. During the model selection process, the final models were generated using the stepwise selection method. An average square error (ASE) plot and fit criteria plots were created for ln\_price and odometer.  Price  Final Model  The 7 predictors given to the stepwise model selection were: odometer, d\_type, year, d\_drive, d\_fuel, and d\_condition. Out of these predictors, 5 were selected to be in the final model. Those were odometer, d\_type, year, d\_drive, and cylinders (Figure 26). Additionally, the seed used to split the data between training and testing was: 86225000 (Figure 25). The final expression for the final training model was (Figure 27):  Ln\_price = -61.13 - 0.0000062(odometer) + 0.186(d\_drive) + 0.179(d\_type) + 0.035(year) + 0.078(cylinders)  The RMSE was 0.73486 which is generally low, however, the range of ln\_price is between [range of ln\_price]. The r-square and adjusted r-square were 0.3386 and 0.3316 respectively (Figure 28). This m5eans that the final model can represent about 33% of the variance of the observations that were used to train the model.  Comparison  The validation between the training model and testing can be compared using the ASE terms and the generated plots. The ASE term for training was 0.533 and the ASE term for testing was 0.587 (Figure 28). The ASE term for testing was 10% higher than for training. Ideally, the testing ASE term would be lower than training’s ASE term.  The fit criteria graphs for ln\_price show ideal progression. AIC, AICC, SBC, and CV Press plots show a negative trend for all the chosen predictors. The adjusted r-square increases throughout each of the predictors. The most change happens for the first 4 predictors before showing slight improvement for the last two predictors (Figure 29).  The ASE plot shows the progression of the training and testing ASE terms as each predictor gets added. Throughout each of the predictors, the testing ASE term was higher than the training ASE term. Additionally, the testing ASE term increased on the 4th predictor rather than declining like the training model. The ASE terms had a difference of about 0.05 when the cross-validation was completed. Although the training ASE term was lower than the testing ASE term, there was not a significant difference between training and testing (Figure 30).  Odometer  Final Model  The same full model predictors as ln\_price was given to the stepwise model selection to generate a final model. The 4 predictors that were chosen were: price, d\_type, d\_drive, and cylinders (Figure 32). The seed that was used to generate this 5-fold cross-validation was: 86701001 (Figure 31). The final expression that was built using the model selection was (Figure 33):  Odometer = 77,917 - 3.897(price) + 12,650(drive) + 5,930.35(d\_type) + 5,463.45(cylinders)  This final model resulted in an RMSE value of 44186 which is high since the odometer’s range of values is between [range of odometer’s values]. The r-square was 0.3565 and the adjusted r-square was 0.3511 which means about 35% of the variance in the training dataset could be explained using this model (Figure 34).  Comparison  The final model resulted in high ASE terms for both training and testing. The terms were 1,932,211,142 and 2,390,212,022 for training and testing respectively. Testing’s ASE term was about 24% higher than training’s ASE term. There is a clear discrepancy between the training and testing observations (Figure 34).  The fit criteria plots display expected behavior throughout each of the predictors. The largest change for the criteria terms happens at the first step after the intercept. Afterward, the trend tapers and there is a slight improvement made to each of the terms (Figure 35).  The ASE plot for the odometer is not ideal. The testing ASE term finishes much higher than the training ASE term. Ideally, the testing ASE term would be lower than training’s ASE term or at least close to it. However, each of the ASE terms progresses through the cross-validation sequence in a similar pattern and the testing terms do not peak at any of the predictors like in ln\_price’s ASE plot (Figure 36). The cross-validation finished with a difference of 458,000,880 between each of the ASE terms (Figure 34).   |  |  |  | | --- | --- | --- | | Survey selected training and testing |  | | | Procedure  The survey select method was used in order to split the data set for training and testing. For both of the y variables, ln\_price and odometer, 75% of the data set were used for training and 25% for testing. Each split was done using their own unique seed. During the model selection process, the stepwise selection method was used to select the predictors for the final model. In the testing dataset, the y variable was predicted (denoted as yhat) in order to compare the performance between the training and testing models. Additionally, correlation values were generated for both ln\_price and odometer in order to check the cross-validation r-square(CV-r-^2).  Price  Final Model  The seed used to split the dataset into training and testing sets was: 137287 (Figure 37). After the stepwise selection method was finished, the chosen predictors to be used in the training model were d\_type, year, cylinders, d\_drive, and d\_transmission. However, d\_transmission was not found to be significant at the 0.05 level, so it was removed before creating the final training model (Figure 38).  The r-square for this model was 0.3848 and the adjusted r-square was 0.3771 meaning that about 38% of the variance in the training data set can be explained using this model. The RMSE was 0.726. The F-value was 49.93 and the training model passed the goodness-of-fit test (Figure 39). The residual plots do not pass all assumptions (Figure 41). They do show some signs of constant variance and independence. However, the normality plot does not show linearity. There is a group of values in the lower range of ln\_price that skew the plot to make the first third of values into a curve rather than a 45-degree line. Furthermore, the expression of this model can be given by (Figure 39):  Ln\_price = -82.57 - 0.00000614(odometer) + 0.232(d\_transmission) + 0.131(d\_drive) + 0.171(d\_type) + 0.045(year) + 0.117(cylinders)  Comparison  The RMSE for the testing model was 0.783 which is about 0.057 (8%) higher than the training model’s RMSE. The MAE was 0.502 and the r-square was 0.2717. The testing model’s r-square was 0.1131 (42%) less than training’s r-square. That means that the testing model can explain about 11% less of the overall variance in the testing data set (Figure 40). Ideally, the cross-validated r-square (CV-r^2) would be less than or equal to 0.3. Since the CV-r^2 is 0.11, it does pass this claim.  Odometer  Final Model  For the odometer training and testing datasets, the seed: 146486 (Figure 43) was used to split the observations between them (75% and 25%). After giving the stepwise selection method the full model, the predictors it chose were price, d\_type, d\_drive, cylinders, and d\_fuel (Figure 44).  After running the final model, it had an r-square of 0.3281 and an adjusted r-square of 0.3225. So, about 32% of the variation in the training observations could be explained by the model. The RMSE was 44,717 which is high for the range of values in this data set. The F-value is 58.72 and the model does pass the goodness-of-fit test (Figure 45). The residuals only pass half of the assumptions. The normality plot does indicate a strong 45-degree line displaying linearity. However, the other residual plots do not show independence or constant variance. The points create a funnel shape (Figure 42). Lastly, the expression given by this model is (Figure 45):  Odometer = 85,065 - 3.802(price) + 5,534.33(d\_type) + 11,499(d\_drive) + 4,296.35(cylinders)  Comparison  The RMSE for the training model was 47,583.76. This was about 2,867 (6%) higher compared to the training model’s RMSE. Furthermore, the MAE for the testing model was 36,403.3. Then, the r-square was 0.276 which is 0.0521 (19%) lower than the training model r-square (Figure 46). This means that the CV-r^2 (0.05) easily passes the 0.3 threshold. | |  | | |  |

|  |  |
| --- | --- |
| APPENDIX |  |

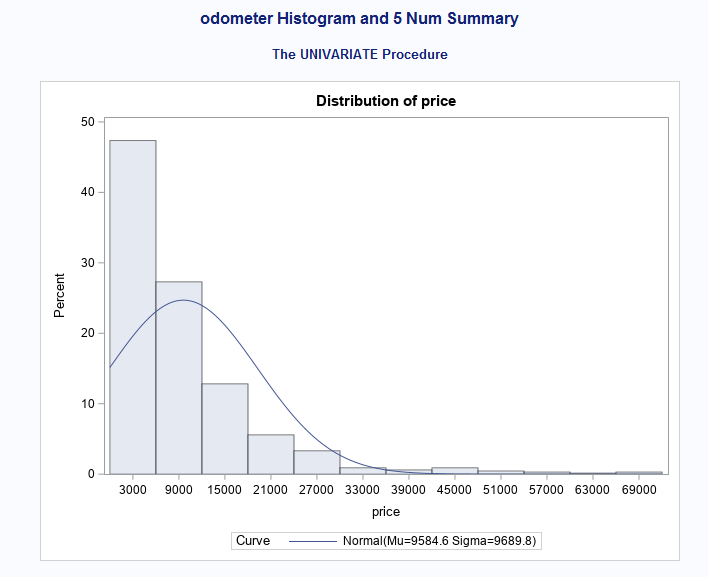


Figure 1

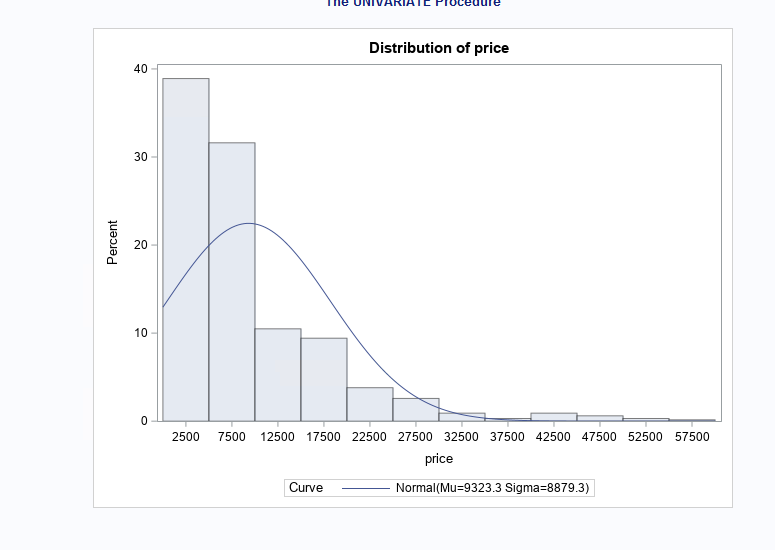


Figure 1.1

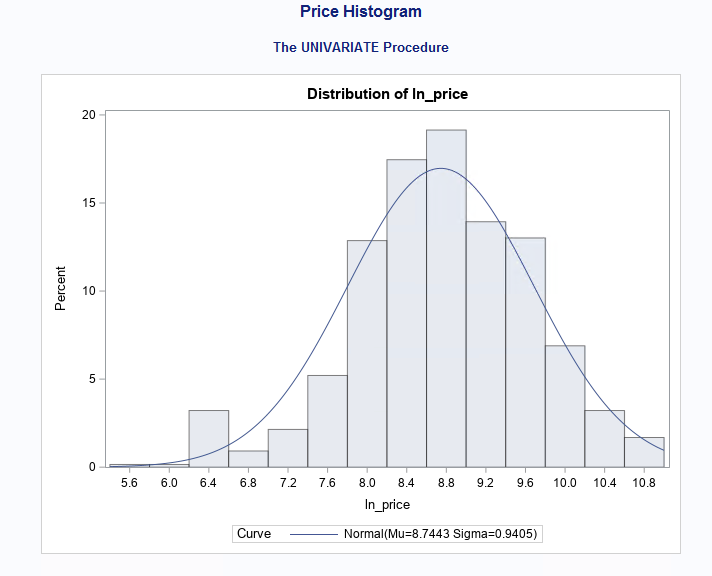


Figure 1.3

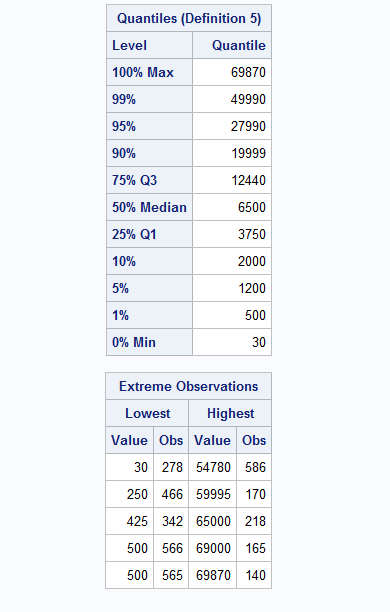


Figure 2

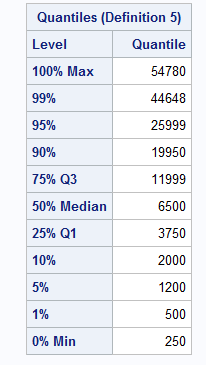


Figure 2.1

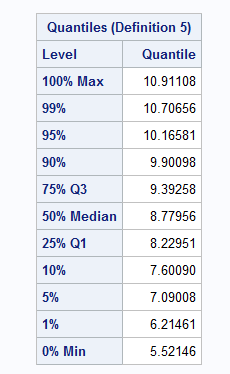


Figure 2.2

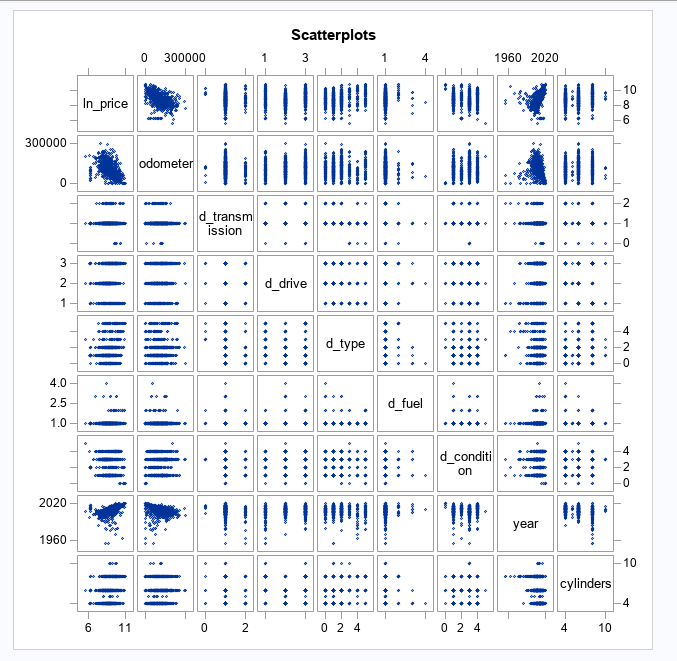


Figure 3

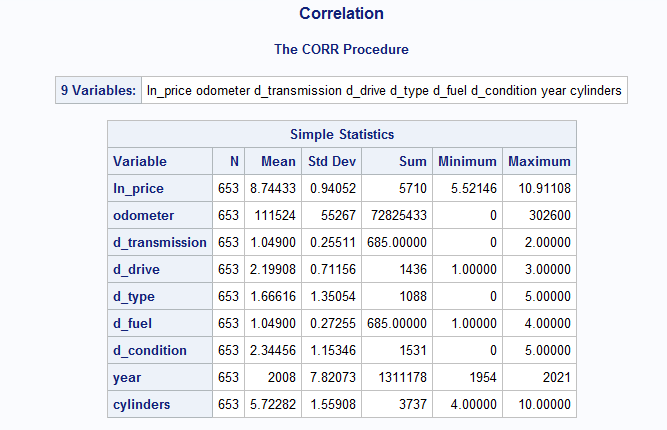


Figure 4

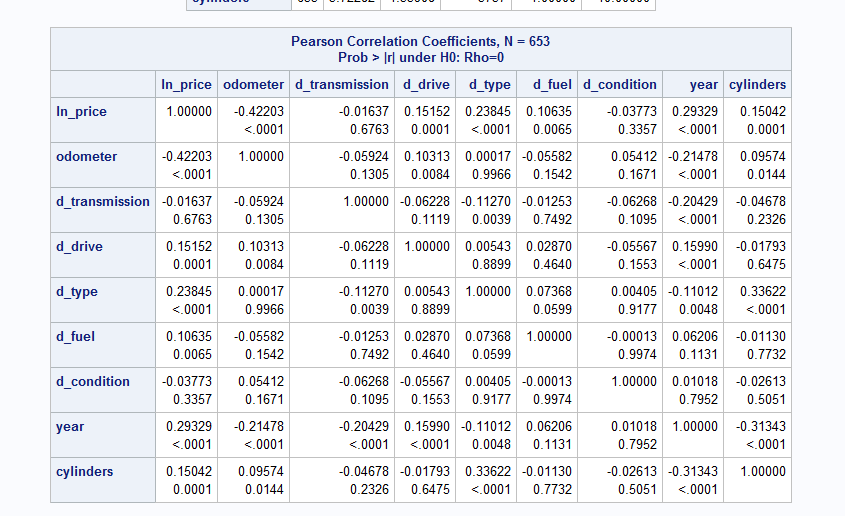


Figure 5

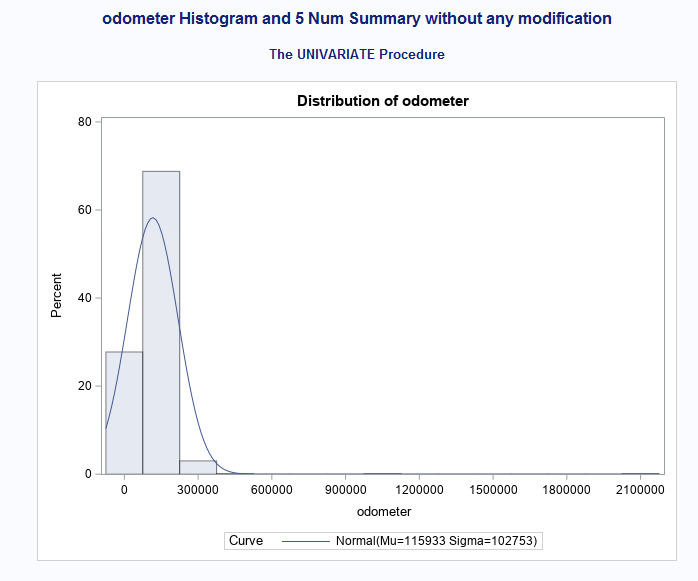


Figure 6



Figure 7

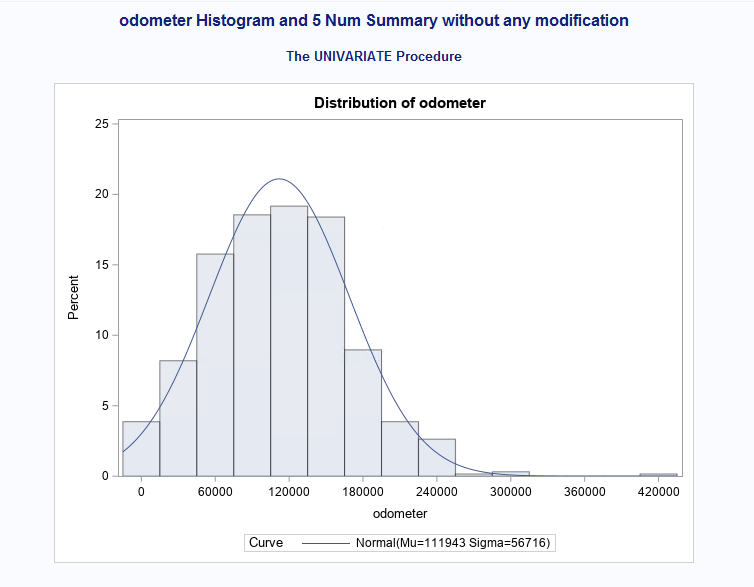


Figure 8



Figure 9

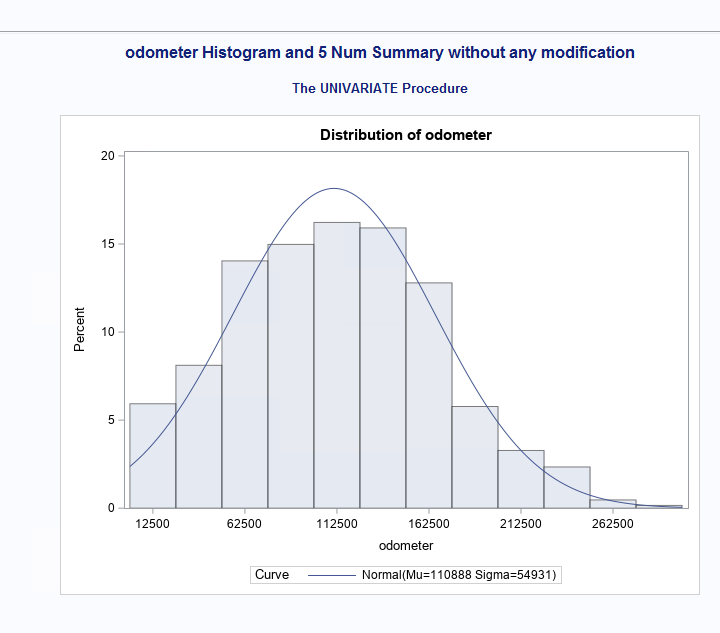


Figure 10

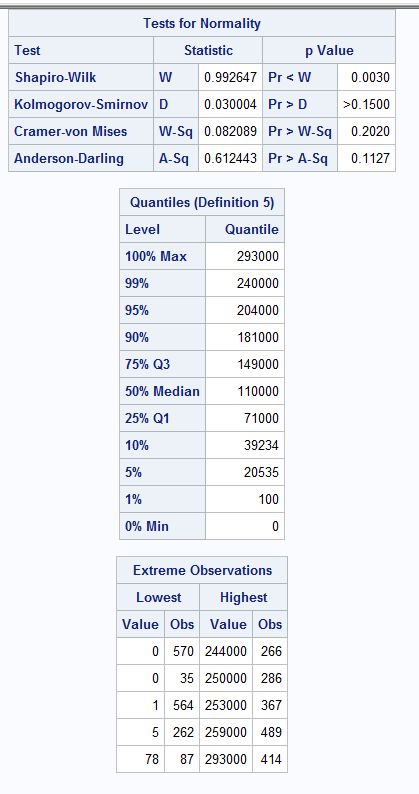


Figure 11

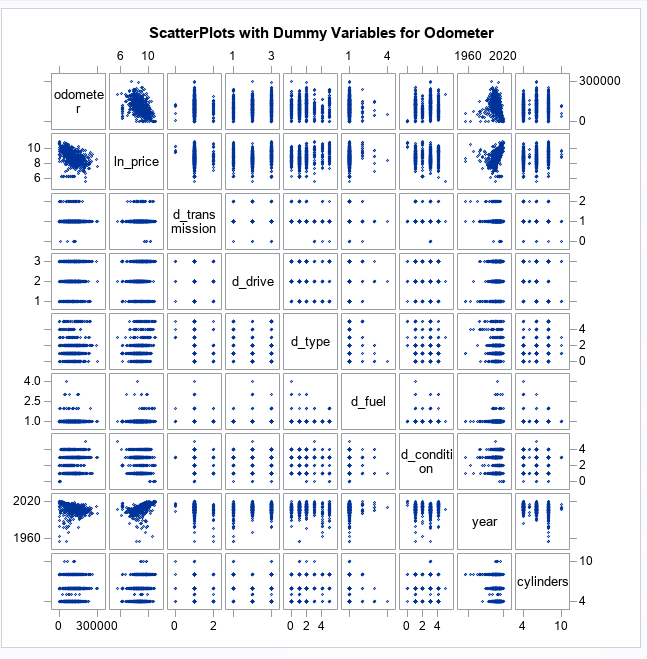


Figure 12

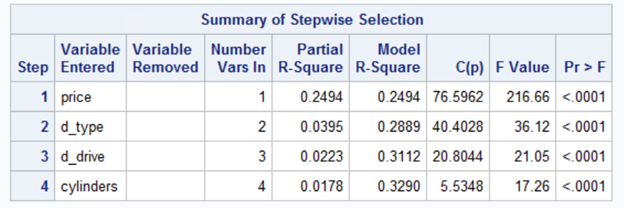


Figure 13

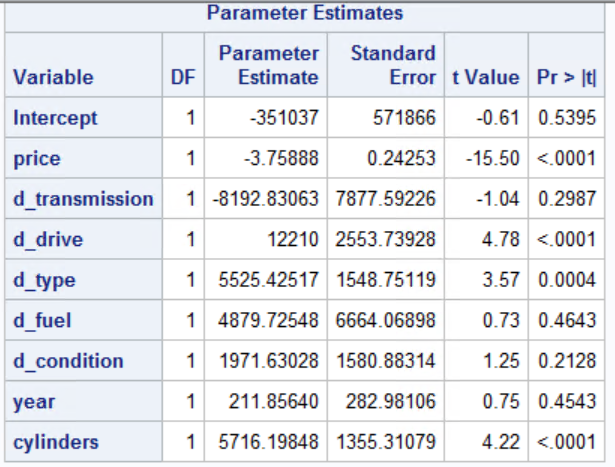


Figure 14.1

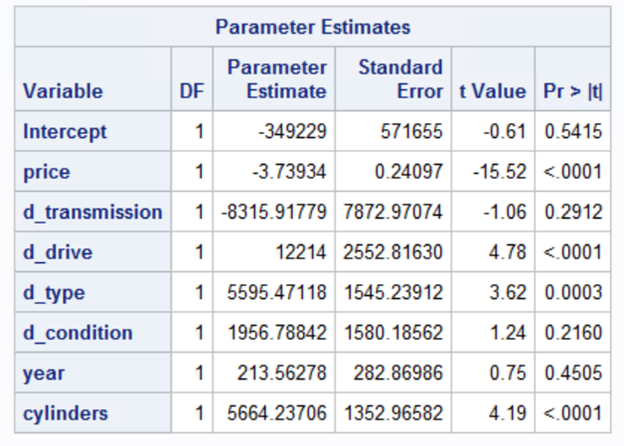


Figure 14.2

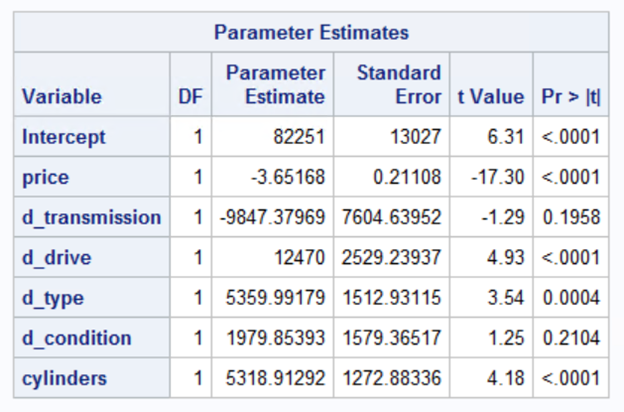


Figure 14.3

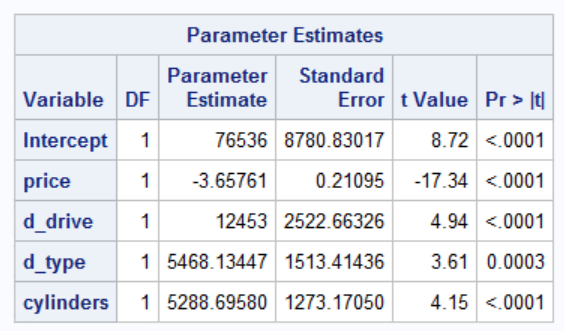


Figure 14.4

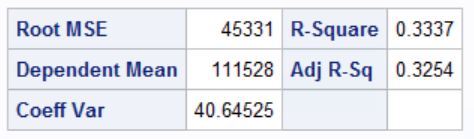


Figure 15.1

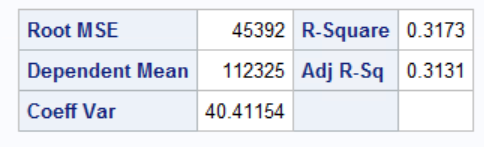


Figure 15.2

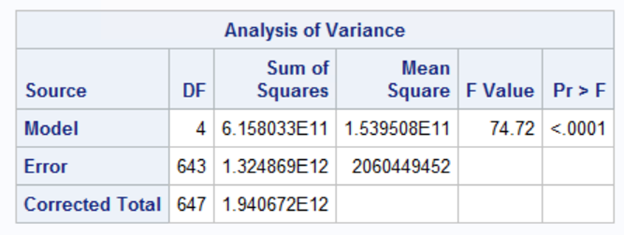


Figure 16

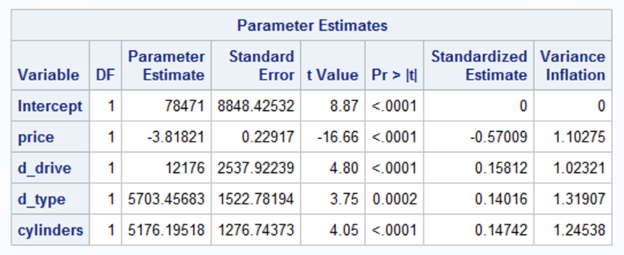


Figure 17

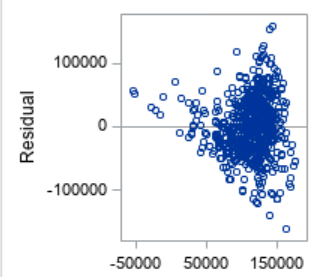
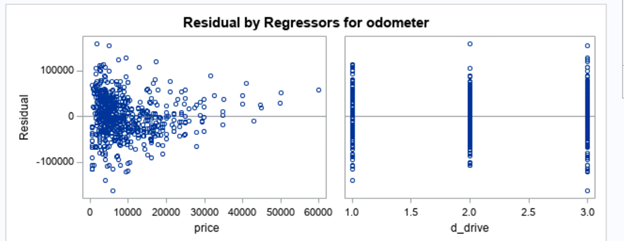


Figure 18.1



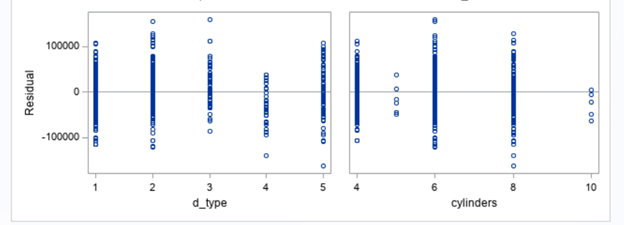


Figure 18.2

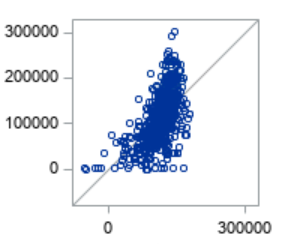


Figure 18.3

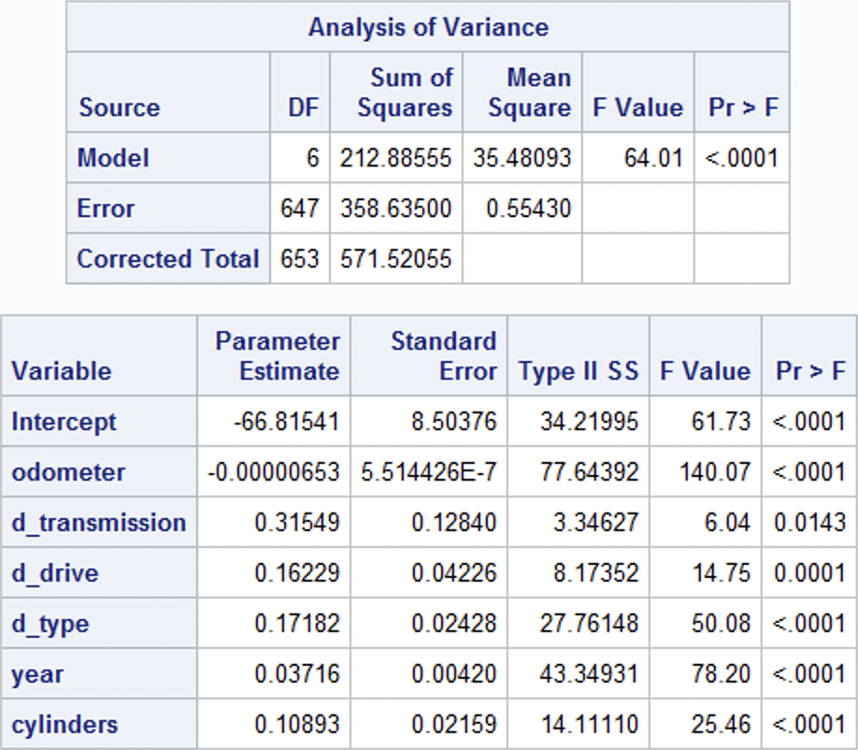


Figure 19

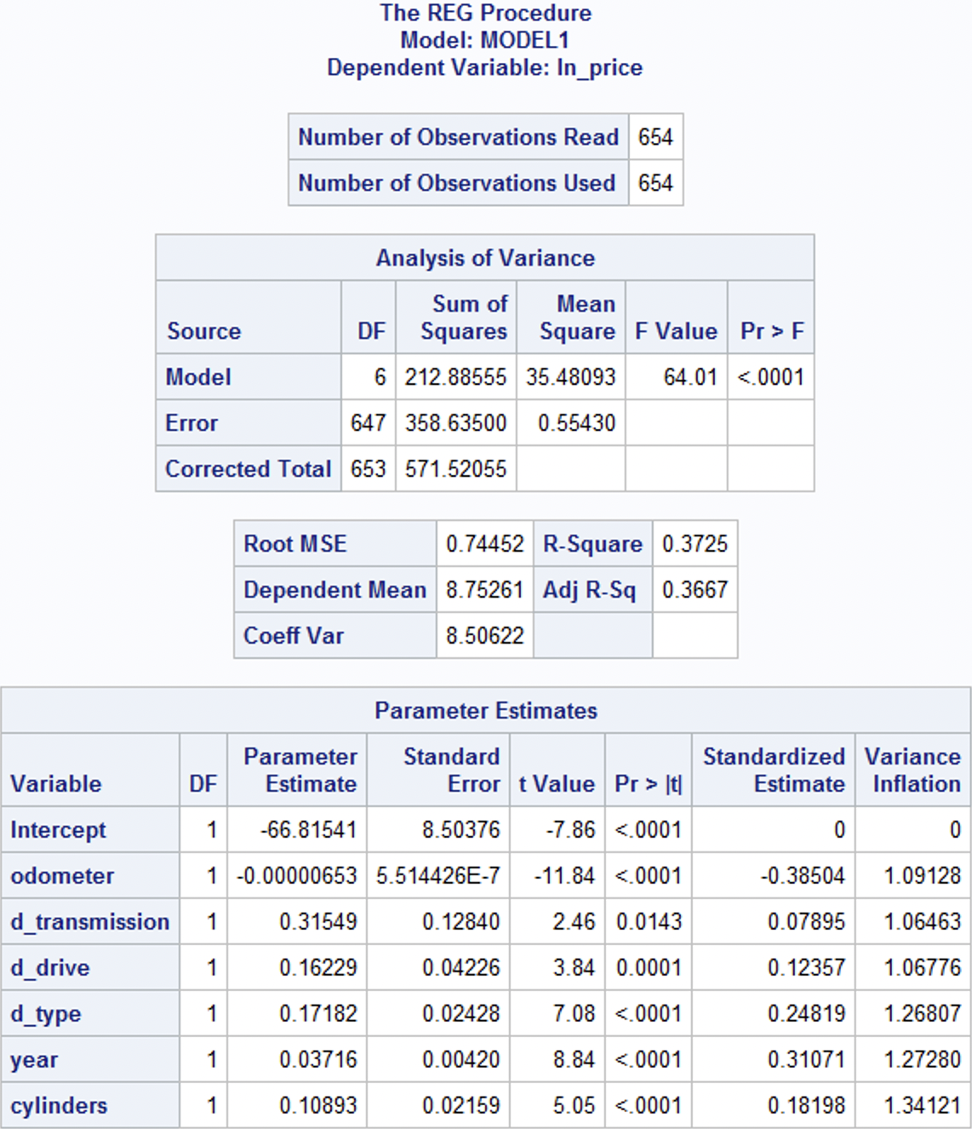


Figure 20

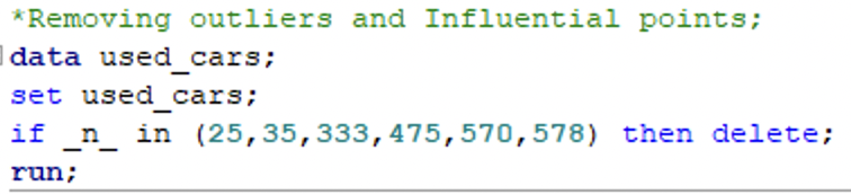


Figure 21

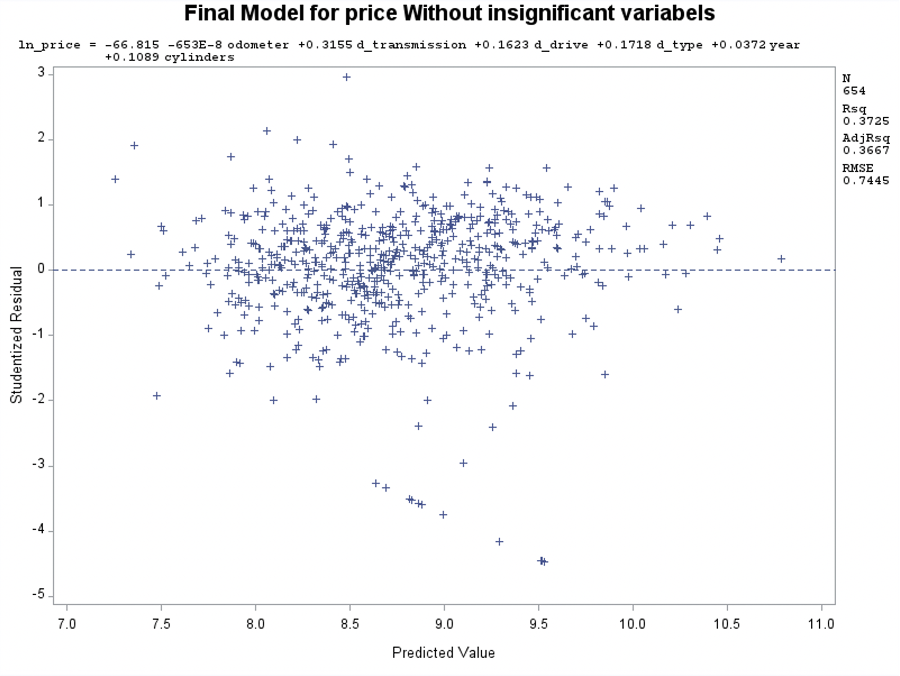


Figure 22



Figure 23

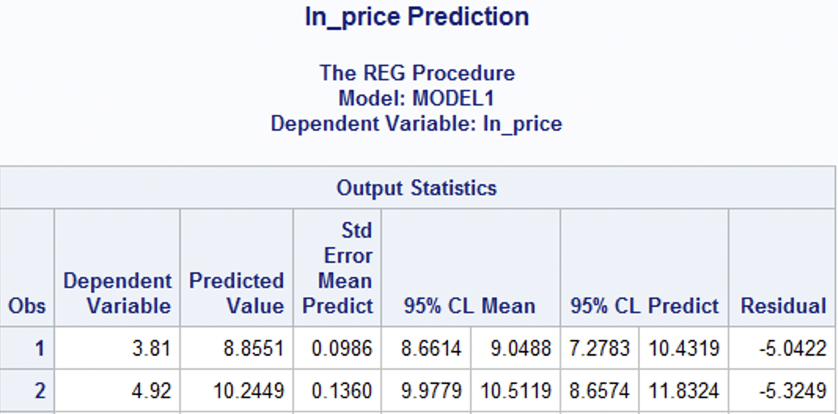


Figure 24

|  |
| --- |
| Figure 25 |

|  |
| --- |
| Figure 26 |

|  |
| --- |
| Figure 27 |

|  |
| --- |
| Figure 28 |

|  |
| --- |
| Figure 29 |

|  |
| --- |
| Figure 30 |

|  |
| --- |
| Figure 31 |

|  |
| --- |
| Figure 32 |

|  |
| --- |
| Figure 33 |

|  |
| --- |
| Figure 34 |

|  |
| --- |
| Figure 35 |

|  |
| --- |
| Figure 36 |

|  |
| --- |
| Figure 37 |

|  |
| --- |
| Figure 38 |

|  |
| --- |
| Figure 39 |

|  |
| --- |
| Figure 40 |

|  |
| --- |
| Figure 41 |

|  |
| --- |
| Figure 42 |

|  |
| --- |
| Figure 43 |

|  |
| --- |
| Figure 44 |

|  |
| --- |
| Figure 45 |

|  |
| --- |
| Figure 46 |