Multiple Linear Regression Model of Vehicle Price

Abiha Rahman

2024-07-02

Multiple Linear Regression Model of Vehicle Data

Date: April 24, 2023

Abstract:

The purpose of this analysis is to gain insights about vehicle features and prices so that consumers can make informed decisions when purchasing a vehicle. Our analysis seeks to create a regression model that can predict Vehicle price based on its features to give consumers an idea of what the price of a vehicle with their specifications would be. Our regression equation is log(Car price) = 2.47 + 0.0054 * x1 (horsepower) + 0.301x2 (Europe) -0.0471x3 (US) + 0.39x4 (Curb weight) - 0.0155x5 (Width) -0.0067x6 (Wheelbase) + 0.0096x7(Fuel efficiency) - 0.0028571x8 (Fuel Capacity). Based on their individual T tests, the significant predictors of vehicle price were Horsepower, Continent/region of Origin, specifically Europe, Curb weight, width, and wheelbase, with the most significant being horsepower.

Background Information:

The Dataset we used contains information about a variety of cars from major car manufacturers. Some of these features have an impact on vehicle price, and some of these are important to consumers' wants and needs in a vehicle. Continent of Origin specifies which Continent/Region the vehicle brand is from. This is a big difference between the vehicle brands, which we seek to explore further. Vehicle type is defined here as passenger and car. Passenger refers to smaller cars, often classified as sedans or compact cars. Car refers to bigger cars, often classified as SUVs in North America. Our engine size is measured in Liters, measuring the cylinders in the car engine. Horsepower measures the power a car's engine produces. Wheelbase is the difference between the center of the front wheels and center of the rear wheels of a car in inches. These variables give a lot of insight about a vehicle's performance and feel, which is something considered in both consumer's wants as well as a vehicle's price. Curb weight is the car's weight without anything inside in thousands of pounds. Fuel capacity measures how many gallons of fuel a car's tank can carry. Fuel efficiency shows how many miles a car can travel with one gallon of fuel.

```
# import Packages
# library(car)
# library(PMCMRplus)
# library(pastecs)
# library(ggplot2)
# library(pgirmess)
# library(olsrr)

# Read File
# CarData <- read.csv("CarSalesNM + new.csv")
# Make categorical into factors
# CarData$Manufacturer <- as.factor(CarData$Manufacturer)
# CarData$Country.origin <- as.factor(CarData$Country.origin)
# CarData$Continnent.origin <- as.factor(CarData$Continnent.origin)
```

- Predict Vehicle Price based on Car features using a multiple linear regression model.
- Reason: It is very helpful for consumers to know how their feature preferences in vehicles can impact the price of a vehicle, and be able to predict the price.
- Put all variables believed to be relevant to vehicle price into linear model and perform model selection to determine best variables to include. Check assumptions to determine fit and Perform F test and t tests to determine significance of model and individual predictors. Interpret model as well.

1. Predict Car Price Based on Car Features

- We do not want to use the variables Manufacturer or Model or Country of Origin because these are just identifiers, and not a lot of samples. We do not want to use Sales or resale values because this does not help predict price, rather, it is a result of it. We also don't want to use Latest Launch because the years are very similar and not too important to us.
- We will put in all variables relevant to price in our full model to see how model selection chooses.

Create Linear Model

```
# create original full model with variables that are relevant to price for our predictors.

FullLm <- lm(Price_in_thousands ~ Continnent.origin + Vehicle_type +
Engine_size + Horsepower + Wheelbase + Width + Length + Curb_weight + Fuel_capacity
+ Power_perf_factor + Fuel_efficiency , data = CarData )

# Problem: Perfect model, solution: check variables

# matrix

matrix <- cor(CarData[, c("Price_in_thousands","Engine_size","Horsepower","Wheelbase", "Width", "Len

cor(CarData$Horsepower, CarData$Power_perf_factor)
```

```
## [1] 0.9940706
```

```
# Error Found! Horsepower and power perf factor are perfectly correlated, need to remove one!
# revised model:

NewFullLm <- lm(Price_in_thousands ~ Continnent.origin +
Vehicle_type + Engine_size + Horsepower + Wheelbase + Width + Length + Curb_weight + Fuel_capacity

summary(NewFullLm)
```

```
##
## Call:
## lm(formula = Price_in_thousands ~ Continnent.origin + Vehicle_type +
##     Engine_size + Horsepower + Wheelbase + Width + Length + Curb_weight +
##     Fuel_capacity + Fuel_efficiency, data = CarData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6742 -3.5386  0.0689  2.5554 17.2267
```

```
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             14.21084   16.86740   0.843 0.401421
## Continnent.originEurope 10.50701    1.92175   5.467 3.11e-07 ***
## Continnent.originUS     -0.25424    1.38819  -0.183 0.855037
## Vehicle_typePassenger    0.44476    2.27336   0.196 0.845269
## Engine_size             -1.31714    1.45449  -0.906 0.367236
## Horsepower               0.20461    0.02199   9.305 2.22e-15 ***
## Wheelbase               -0.21608    0.16556  -1.305 0.194690
## Width                   -0.50201    0.26669  -1.882 0.062554 .
## Length                  -0.05466    0.10853  -0.504 0.615595
## Curb_weight              8.13006    2.31367   3.514 0.000653 ***
## Fuel_capacity            0.39081    0.31203   1.253 0.213168
## Fuel_efficiency          0.50442    0.25610   1.970 0.051517 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.401 on 105 degrees of freedom
## Multiple R-squared:  0.8681, Adjusted R-squared:  0.8543
## F-statistic: 62.83 on 11 and 105 DF,  p-value: < 2.2e-16
```

- While not printed out to conserve space, in the summary, our original model had an adjusted R squared of 1, which indicates an error.
- After looking at matrix correlation, it was found that power perf factor and horsepower are almost perfectly correlated at .9904, so we removed power perf factor so it wouldn't ruin our model. This has to do with multicollinearity, which we officially check later in the report, but had to address now due to possible innacuracies in our model selection methods.
- Our new full model has an adjusted R^2 of .8543, which is much more reasonable.

```
# Model Selection

# Forward Selection

ols_step_forward_aic(NewFullLm , details=FALSE)
```

```
##
##                               Selection Summary
## ------------------------------------------------------------------------------
## Variable                 AIC        Sum Sq         RSS         R-Sq      Adj. R-Sq
## ------------------------------------------------------------------------------
## Horsepower             804.561    16916.629    6308.194    0.72839      0.72602
## Continnent.origin      746.800    19503.884    3720.939    0.83979      0.83553
## Curb_weight            746.513    19575.921    3648.903    0.84289      0.83728
## Width                  740.136    19828.015    3396.808    0.85374      0.84715
## Wheelbase              737.775    19952.299    3272.525    0.85909      0.85141
## Fuel_efficiency        735.481    20070.232    3154.591    0.86417      0.85545
## Fuel_capacity          735.110    20133.530    3091.293    0.86690      0.85704
## ------------------------------------------------------------------------------
```

```r
forward.model <- lm(Price_in_thousands ~ Horsepower + Continnent.origin + Curb_weight + Width + Whee

# Backwards Selection

ols_step_backward_aic(NewFullLm, details= FALSE)
```

```
##
##
##                          Backward Elimination Summary
## ------------------------------------------------------------------------
## Variable          AIC        RSS        Sum Sq       R-Sq      Adj. R-Sq
## ------------------------------------------------------------------------
## Full Model      740.037    3063.094    20161.729    0.86811      0.85429
## Vehicle_type    738.080    3064.210    20160.613    0.86806      0.85562
## Length          736.331    3070.786    20154.037    0.86778      0.85666
## Engine_size     735.110    3091.293    20133.530    0.86690      0.85704
## ------------------------------------------------------------------------
```

```r
backward.model <- lm(Price_in_thousands ~ Continnent.origin + Horsepower + Wheelbase + Width + Curb_

# Step wise Selection

ols_step_both_aic(NewFullLm, details = FALSE)
```

```
##
##
##                                      Stepwise Summary
## ---------------------------------------------------------------------------------------
## Variable              Method       AIC         RSS         Sum Sq        R-Sq      Adj. R-Sq
## ---------------------------------------------------------------------------------------
## Horsepower           addition    804.561    6308.194    16916.629    0.72839       0.72602
## Continnent.origin    addition    746.800    3720.939    19503.884    0.83979       0.83553
## Curb_weight          addition    746.513    3648.903    19575.921    0.84289       0.83728
## Width                addition    740.136    3396.808    19828.015    0.85374       0.84715
## Wheelbase            addition    737.775    3272.525    19952.299    0.85909       0.85141
## Fuel_efficiency      addition    735.481    3154.591    20070.232    0.86417       0.85545
## Fuel_capacity        addition    735.110    3091.293    20133.530    0.86690       0.85704
## ---------------------------------------------------------------------------------------
```

```r
both.model <- lm(Price_in_thousands ~ Horsepower + Continnent.origin + Curb_weight + Width + Wheelba

# Compare AIC Values

AIC(forward.model)
```

```
## [1] 735.1096
```

```r
AIC(backward.model)
```

```
## [1] 735.1096
```

```
AIC(both.model)
```

```
## [1] 735.1096
```

- All 3 model selection methods gave me the same AIC of 735.1, with the model using variables: Horsepower + Continnent.origin + Curb_weight + Width + Wheelbase + Fuel_efficiency + Fuel_capacity to predict price.

```
# Chosen model
ChosenLM <- lm(Price_in_thousands ~ Horsepower +
Continnent.origin + Curb_weight + Width + Wheelbase +
Fuel_efficiency + Fuel_capacity, data = CarData)

# Generalized F test
anova(ChosenLM, NewFullLm)
```

```
## Analysis of Variance Table
##
## Model 1: Price_in_thousands ~ Horsepower + Continnent.origin + Curb_weight +
##     Width + Wheelbase + Fuel_efficiency + Fuel_capacity
## Model 2: Price_in_thousands ~ Continnent.origin + Vehicle_type + Engine_size +
##     Horsepower + Wheelbase + Width + Length + Curb_weight + Fuel_capacity +
##     Fuel_efficiency
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    108 3091.3
## 2    105 3063.1  3    28.199 0.3222 0.8093
```

- p value of generalized F test is 0.8093, so we fail to reject null hypothesis and reduced model is sufficient.

CHECK ASSUMPTIONS

Assumption 5

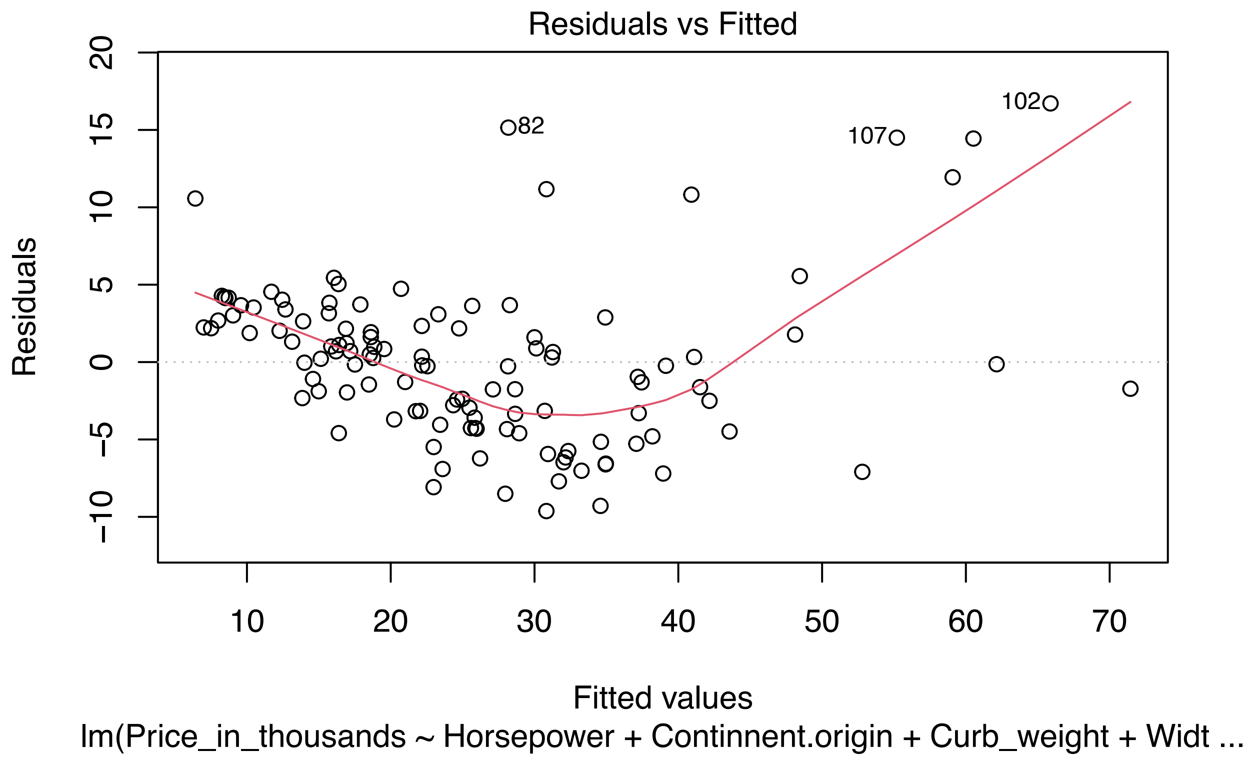
```
# test for multicollinearity
vif(ChosenLM)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## Horsepower        1.992985  1        1.411731
## Continnent.origin 1.389145  2        1.085643
## Curb_weight       6.224502  1        2.494895
## Width             3.114685  1        1.764847
## Wheelbase         2.764275  1        1.662611
## Fuel_efficiency   4.098182  1        2.024397
## Fuel_capacity     4.833923  1        2.198618
```

- All values under 10, No multicollinearity, assumption is valid

Assumption 2

```r
# Check for Linearity
plot(ChosenLM,1)
```



- This asssumption fails, as there is a clear trend curve in our plot.

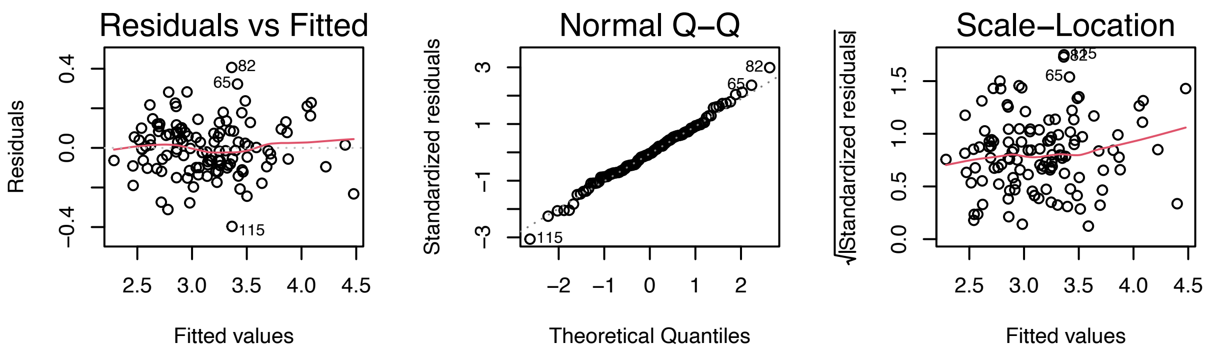
REMEDY: Transformation

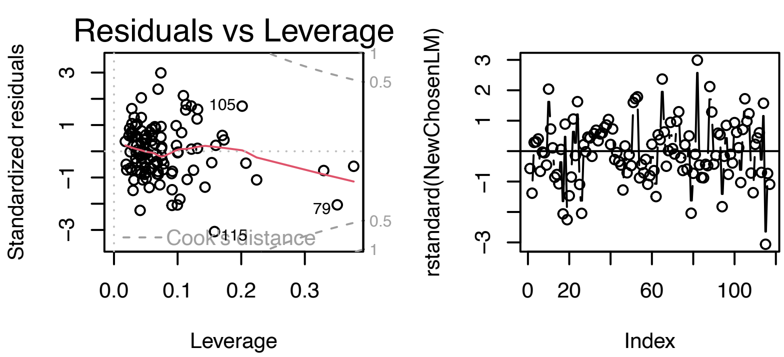
```r
# Transform using log transformation

NewChosenLM <- lm(log(Price_in_thousands) ~ Horsepower +
Continnent.origin + Curb_weight + Width +
Wheelbase + Fuel_efficiency + Fuel_capacity, data = CarData)
```

```r
# New Plots
par(mfrow=c(2,3))
plot(NewChosenLM)
plot(rstandard(NewChosenLM), type="b")
abline(h=0)

# Recheck Assumption 2 - linearity

# Recheck Assumption 5 - multicollinearity

vif(NewChosenLM)
```

```
##                          GVIF Df GVIF^(1/(2*Df))
```

```
## Horsepower        1.992985  1        1.411731
## Continnent.origin 1.389145  2        1.085643
## Curb_weight       6.224502  1        2.494895
## Width             3.114685  1        1.764847
## Wheelbase         2.764275  1        1.662611
## Fuel_efficiency   4.098182  1        2.024397
## Fuel_capacity     4.833923  1        2.198618
```





-Log Transformation makes random scatter of residuals, Linearity assumption now holds, multicollinerity assumption check still holds.

Assumption 1

```
# Test for independent errors - autocorrelation
durbinWatsonTest(NewChosenLM)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1223742      1.744814    0.11
##  Alternative hypothesis: rho != 0
```

- Visually, doesn't seem to be an apparent pattern, random scatter above and below line. Durbin watson test p value is 0.108, so above 0.05, so autocorrelation assumption holds.

Assumption 3

```r
# check for Homoscedascity
```

- With new transformed model, Variance looks even, so Assumption holds!

Assumption 4

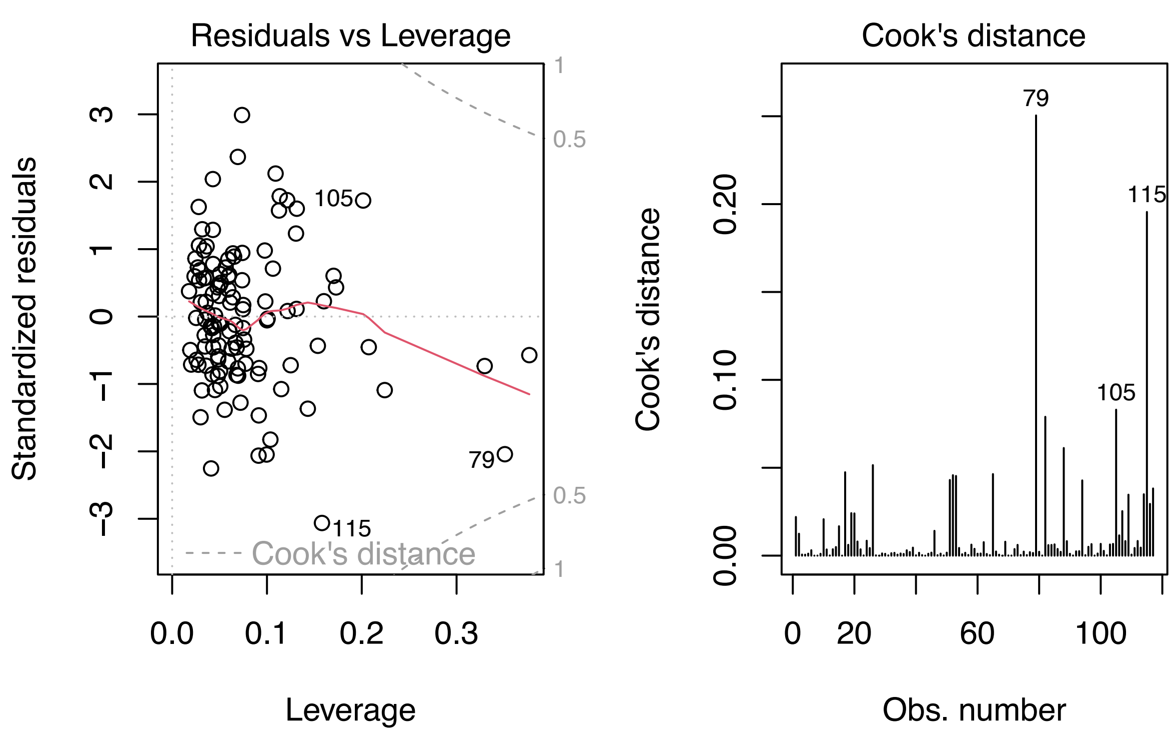
```r
# Check for normality
shapiro.test(resid(NewChosenLM))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(NewChosenLM)
## W = 0.99469, p-value = 0.9392
```

- QQ plot looks not exactly normal, but points generally close to line
- Shapiro Wilk's test statistic is 0.94 , greater than 0.05, so normality Holds!

Check for Outliers and influential points

```r
# check for outliers  and influential points in plot
par(mfrow=c(1,2))
myplot <- plot(NewChosenLM,5)
plot(NewChosenLM, 4)
```



```r
# check for outliers
outliers <- which(rstandard(NewChosenLM)>2 | rstandard(NewChosenLM)< -2)
```

- The outliers we found were the points 10,17,19,26,65,79,82,88,and 115 as these points had standardized residual values above and below 2. It is very concerning to me that we have this many, So I will do a sensitivity analysis to find changes.
- No influential points, all cook's distance values below 1

```r
SensitivityAnalysis<- lm(log(Price_in_thousands) ~ Horsepower + Continnent.origin + Curb_weight + Wi

summary(SensitivityAnalysis)
```

```
##
## Call:
## lm(formula = log(Price_in_thousands) ~ Horsepower + Continnent.origin +
##     Curb_weight + Width + Wheelbase + Fuel_efficiency + Fuel_capacity,
##     data = CarData[-c(10, 17, 19, 26, 65, 79, 82, 88, 115), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21929 -0.08650 -0.00311  0.08174  0.21749
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.7376019  0.3412845   8.021 2.16e-12 ***
## Horsepower              0.0057917  0.0002825  20.502  < 2e-16 ***
## Continnent.originEurope  0.3178634  0.0352243   9.024 1.47e-14 ***
## Continnent.originUS    -0.0342395  0.0269287  -1.271 0.206535
## Curb_weight             0.2949680  0.0466740   6.320 7.55e-09 ***
## Width                  -0.0189382  0.0052672  -3.596 0.000507 ***
## Wheelbase              -0.0039342  0.0023078  -1.705 0.091377 .
## Fuel_efficiency         0.0042222  0.0049500   0.853 0.395731
## Fuel_capacity           0.0011526  0.0061099   0.189 0.850763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1094 on 99 degrees of freedom
## Multiple R-squared:  0.9436, Adjusted R-squared:  0.9391
## F-statistic: 207.2 on 8 and 99 DF,  p-value: < 2.2e-16
```

- We ran a sensitivity analysis (created new regression model omitting outlier points). Just slightly changed slopes and a slightly higher coefficient of determination value of 0.9391, which is interesting. Additionally, in the model without outliers, wheelbase is not a significant predictor, when in the original model it is. We will keep the outliers because they do not alter our model too much and we would like to have as many data points included as possible.

MAKING INFERENCES

Checking How Well the Model Fits the Data

F Test and Adjusted R^2

```r
# F - test
summary(NewChosenLM)
```

```
##
## Call:
## lm(formula = log(Price_in_thousands) ~ Horsepower + Continnent.origin +
##     Curb_weight + Width + Wheelbase + Fuel_efficiency + Fuel_capacity,
##     data = CarData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39670 -0.09103 -0.00627  0.08322  0.40618
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.4787181  0.4223313   5.869 4.88e-08 ***
## Horsepower              0.0054146  0.0003158  17.144  < 2e-16 ***
## Continnent.originEurope  0.3010551  0.0427048   7.050 1.77e-10 ***
## Continnent.originUS    -0.0471701  0.0329651  -1.431   0.1553
## Curb_weight             0.3912624  0.0547635   7.145 1.11e-10 ***
## Width                  -0.0154747  0.0065532  -2.361   0.0200 *
## Wheelbase              -0.0066564  0.0027071  -2.459   0.0155 *
## Fuel_efficiency         0.0096409  0.0060248   1.600   0.1125
## Fuel_capacity          -0.0028571  0.0075949  -0.376   0.7075
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1412 on 108 degrees of freedom
## Multiple R-squared:  0.911,  Adjusted R-squared:  0.9044
## F-statistic: 138.1 on 8 and 108 DF,  p-value: < 2.2e-16
```

```r
NewChosenLM
```

```
##
## Call:
## lm(formula = log(Price_in_thousands) ~ Horsepower + Continnent.origin +
##     Curb_weight + Width + Wheelbase + Fuel_efficiency + Fuel_capacity,
##     data = CarData)
##
## Coefficients:
##             (Intercept)               Horsepower  Continnent.originEurope
##                2.478718                 0.005415                 0.301055
##     Continnent.originUS              Curb_weight                    Width
##               -0.047170                 0.391262                -0.015475
##               Wheelbase          Fuel_efficiency            Fuel_capacity
##               -0.006656                 0.009641                -0.002857
```

H0: The Beta coefficients of the predictor variables all equal 0 HA: At least one beta coefficient does not equal 0

- Result: F statistic is 138.1 and p value is less than 2.2e^-16, which is less than 0.05, so we reject the null hypothesis. There is at least one predictor variable whose beta coefficient is not equal to 0. This means that the model is significant, and there is a significant relationship between at least one of the predictor variables and The price of a vehicle.

- The Adjusted R squared Value is .9044, which means that 90.44% of the Variability in Car Price is determined by its linear relationship with Horsepower, Continent/Region of Origin, Curb weight, Width, Wheelbase, Fuel efficiency, and fuel capacity. This Means that enough of the variability is explained by our model.

INTERPRETING AND LOOKING AT INDIVIDUAL PREDICTORS

Interpreting our linear model

$$log(y)_{price} = 2.47 + 0.0054x_{horsepower} + .301x_{Europe} - 0.0471x_{US} + 0.39x_{Curbweight} - 0.0155x_{width} - 0.0067x_{Wheelbase} + 0.0096$$

- Our equation is log(y) = 2.47 + 0.0054x1 + 0.301x2 -0.0471x3 + 0.39x4 - 0.0155x5 -0.0067x6 + 0.0096x7 - 0.0028571x8
- The intercept of our model is 2.48, so for a car From Asia, when the horsepower, curb weight, width, wheelbase, fuel efficiency, and fuel capacity is 0, we expect the average log of car price to be 2.48 thousand dollars.
- For each additional hp of horsepower, we expect the log of the car price to increase by about .005415, keeping all other variables constant. According to its t test p value of less than 2e^-16, after adjusting for all other variables, horspoewer has a significant relationship with car price.
- For a car from Europe, we expect the log of the base selling price to increase by .301 thousand dollars, holding all other variables constant.
- For a car from the US, we expect the log of the base selling price to decrease by .047 thousand dollars, holding all other variables constant.
- For each additional thousand pounds of curb weight, we expect the log of the car price to increase by about 0.391 thousand dollars, keeping all other variables constant. According to its t test p value of less than 1.11e-10, after adjusting for all other variables, Curb weight has a significant relationship with car price.
- For each additional inch in width, we expect the log of the car price to decrease by about 0.015 thousand dollars, keeping all other variables constant. According to its t test p value of 0.0200, after adjusting for all other variables, Width has a significant relationship with car price.
- For each additional inch in wheelbase, we expect the log of the car price to decrease by about 0.0067 thousand dollars, keeping all other variables constant. According to its t test p value of 0.0155, after adjusting for all other variables, Wheelbase has a significant relationship with car price.
- For each additional mile per gallon in Fuel efficiency, we expect the log of the car price to increase by about 0.0096 thousand dollars, keeping all other variables constant. According to its t test p value of .1125, Fuel_efficiency does not have a significant relationship with price.
- For each additional gallon of fuel capacity, we expect the log of the car price to decrease by about 0.0029 thousand dollars, keeping all other variables constant. According to its t test p value of 0.7075, Fuel Capacity does not have a significant relationship with price.

T tests: - Based on individual t tests, our significant predictors of price are Horsepower, Continent/region of Origin, specifically Europe, Curb weight, width, and wheelbase.

SUMMARY OF FINDINGS:

- In analysis 1, through our non parametric kruskal wallis test, it was found that there was a significant difference in price between vehicles from Asia, Europe, and the US. Specifically, Vehicles from Europe had a higher average price than vehicles from Asia and the US. This tells us that consumers who prefer high end vehicles should look at European brands, and consumers on a budget should look at Asia and US brands. In analysis 2, in our regression model to predict price based on vehicle features, after using model selection methods and performing a log transformation to verify normality, we found through the F test that our model was significant and the significant predictors of price were horsepower, Continent/region of Origin, specifically Europe, Curb weight, width, and wheelbase. This is an important consideration for consumers when budgeting out their wants, as it can help them determine how their features will affect the car price. When determining a budget, consumers should focus on the significant predictors of vehicle price determined by our model.