PROJECT REPORT

MA642B – Regression Analysis

SUBMITTED TO
PROF. VICTOR XIE

SUBMITTED BY
DILAWAR SINGH

SUBMISSION DATE - AUGUST 4, 2020

Introduction

This report discusses the development of a regression model to predict the retail price of 2005 GM used cars. The structure of this data set allowed me to work through the entire process of model building and assessment.

For this data set, a representative sample of over eight hundred 2005 GM used cars were selected, then retail price was calculated from the tables provided in the 2005 Central Edition of the Kelly Blue Book. It containing the following variables:

- **Price:** suggested retail price of the used 2005 GM car in excellent condition. The condition of a car can greatly affect price. All cars in this data set were less than one year old when priced and considered to be in excellent condition.
- **Mileage:** number of miles the car has been driven
- **Make:** manufacturer of the car such as Saturn, Pontiac, and Chevrolet
- **Model:** specific models for each car manufacturer such as Ion, Vibe, Cavalier
- **Trim (of car):** specific type of car model such as SE Sedan 4D, Quad Coupe 2D
- **Type:** body type such as sedan, coupe, etc.
- **Cylinder:** number of cylinders in the engine
- **Liter:** a more specific measure of engine size
- **Doors:** number of doors
- **Cruise:** indicator variable representing whether the car has cruise control (1 = cruise)
- **Sound:** indicator variable representing whether the car has upgraded speakers (1 = upgraded)
- **Leather:** indicator variable representing whether the car has leather seats (1 = leather)

Goals of Regression

It is important to note that multiple regression analysis can be used to serve different goals. The goals will influence the type of analysis that is conducted. The most common goals of multiple regression are to describe, predict, or confirm.

Describe: A model may be developed to describe the relationship between multiple explanatory variables and the response variable.

Predict: A regression model may be used to generalize to observations outside the sample. Just as in simple linear regression, explanatory variables should be within the range of the sample data to predict future responses.

Confirm: Theories are often developed about which variables or combination of variables should be included in a model. For example, is mileage useful in predicting retail price? Inferential techniques can be used to test if the association between the explanatory variables and the response could just be due to chance.

Activity 1 - A Simple Linear Regression Model

1. **Produce a scatterplot from the Cars data set to display the relationship between mileage (Mileage) and suggested retail price (Price). Does the scatterplot show a strong relationship between Mileage and Price?**

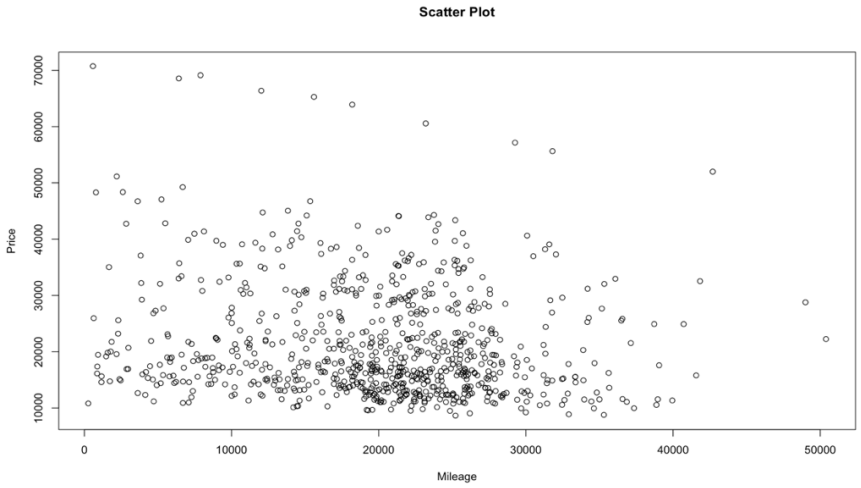

Figure: 1.1

No, It is difficult to say from scatterplot(Figure 1.1) that there is a strong relationship between Price and Mileage. We can see there is so much variation of price with respect to mileage

2. **Calculate the least square line, b0 + b1*Mileage. Report the regression model, the R squared value, the correlation coefficient, the t-statistics, and the p-value for the estimated model coefficients(the intercept and slope). Based on these statistics, can you conclude Mileage is a strong indicator of Price? Explain your reasoning in a few sentences.**

Regression Model: $Price = 24764.56 - 0.172 * Mileage$

R-Squared Value: 0.02046

Correlation Coefficient: -0.1430

	T-statistics	p-value
Intercept	27.383	$< 2e-16$
Mileage	-4.093	$4.68e-05$

Table: 2.1

If we look at the R-Square value it seems mileage is not able to account for the variation in the model but its coefficient is significant as we can see from p-value. This is a case where other independent variables explain the rest of variation but mileage is a significant variable and a strong indicator of price as the scale of mileage is going to be in thousands.

3. **The first car in this data set is a Buick Century with 8221 miles. Calculate the residual value for this car.**

   Residual: -6032.165

4. **Comment on the limitations of using simple linear regression for car price.**

   Simple linear regression will not be good choice for predicting car price for this data set as from the statistics seen in question 2, it is clear that the model is unable to explain the variance in the price. R-squared is 0.02 which is low number. We need further more independent variables to explain the variation in residuals. So, if we use the simple linear regression for predicting the price the residuals will be larger.

Activity 2 - Comparing Variable Selection Techniques

5. **Use the Cars data to conduct stepwise regression analysis.**

   a. **Calculate seven regression models, each with one of the following explanatory variables: Cyl, Liter, Doors, Cruise, Sound, Leather, and Mileage. Identify the explanatory variable that corresponds to the model with the largest R square value. Call this variable X1.**

      After calculating seven regression models with Cyl, Liter, Doors, Cruise, Sound, Leather, and Mileage. We got the R-Squared values as follow:

Model	R-Squared
Price ~ Cyl	0.323859
Price ~ Liter	0.311526
Price ~ Doors	0.019251
Price ~ Cruise	0.185633
Price ~ Sound	0.015462
Price ~ Leather	0.024710
Price ~ Mileage	0.020463

Table: 5.1

      Therefore, Cyl is having the largest R-squared value as seen in Table 5.1. So, we will consider Cyl as the X1 variable.

   b. **Calculate six regression models. Each model should have two explanatory variables, X1 and one of the other six explanatory variables. Find the two-variable model that has the largest R square value. How much did R square improve when this second variable was included?**

      After calculating six regression model with X1 variable as Cyl and X2 as one of the Liter, Doors, Cruise, Sound, Leather, and Mileage, We got the following R-Squared values:

Model	R-Squared
Price ~ Cyl + Liter	0.325915
Price ~ Cyl + Doors	0.343460
Price ~ Cyl + Cruise	0.383949
Price ~ Cyl + Sound	0.329275
Price ~ Cyl + Leather	0.336980
Price ~ Cyl + Mileage	0.339820

Table: 5.2

We found that model with Cyl and Cruise as explanatory variable gave the highest R-Squared value as seen in the Table 5.2. Therefore, X2 variable is Cruise.

The R-squared value changed from 0.323859 to 0.383949 which is approximately 18.55% increase from the initial value.

c. Instead of continuing this process to identify more variables, use the software to conduct a stepwise regression analysis. List each of the explanatory variables in the model suggested by the stepwise regression procedure.

For identifying more variables using stepwise regression procedure I used the function stepAIC() from the MASS library. I did this by specifying the intercept model as $Price \sim 1$ and total model as $Price \sim Cyl + Liter + Doors + Cruise + Sound + Leather + Mileage$. After this, applying the stepAIC function to get the final model. StepAIC() function uses the AIC to decide the variable which will be added or removed. I used the forward method to calculate the final model which starts from intercept model and proceeding further by adding or removing one variable at a time.

The final model suggested by the software is

$$Price = 7323.16 + 3200.12 * Cyl + 6205.51 * Cruise + 3327.14 * Leather - 0.17 * Mileage - 1463.40 * Doors - 2024.40 * Sound$$

This does not include Liter variable as it increases the AIC for the model. This specific model gives the lowest AIC that's why this model is selected.

6. Use the software to develop a model using best subsets techniques for the whole data set. Notice that stepwise regression simply states which model to use, while best subsets provides much more information and requires the user to choose how many variables to include in the model. In general, statisticians select models that have a large R square, and a relatively small number of explanatory variables. Based on the output from nest subsets, which several explanatory variables should be included in a regression model?

To perform best subsets technique, I used the regsubsets() function from the leaps library. This function categorizes various subsets on the basis of number of variables

in the model. Hence, it gives the best subset with 1 variable, 2 variables and so on. I have specified the nvmax = 11 which will give subsets of variables up to size 11.

Also, the regsubsets() function does not work well when the number of variables are large as it becomes difficult to check all the combination of variables. The number of subsets(2^n where n is the number of variables) gets large as number of variable increases. If data consists of Factor objects with large number of levels then regsubsets consider all the levels as each different variable for the regression. For this dataset, the number of variable become 96 is we use regsubsets function. Therefore, I had to do some data wrangling to make use of regsubsets() on the data. I have recoded four columns that are Make, Model, Trim and Type from Factor to Numeric type.

The function summary() reports the best set of variables for each model size. In the table 6.1, an asterisk specifies that a given variable is included in the corresponding model.

Size	Mileage	Make	Model	Trim	Type	Cyl	Liter	Doors	Cruise	Sound	Leather
1						*					
2				*		*					
3				*		*			*		
4	*			*		*			*		
5	*		*	*		*			*		
6	*		*	*		*			*		*
7	*	*	*	*		*			*		*
8	*	*	*	*	*	*			*		*
9	*	*	*	*	*	*			*	*	*
10	*	*	*	*	*	*	*		*	*	*
11	*	*	*	*	*	*	*	*	*	*	*

Table: 6.1

Now, we need to choose size of subset. To do so, I have used the Adjusted R-Square and Mallow's Cp.

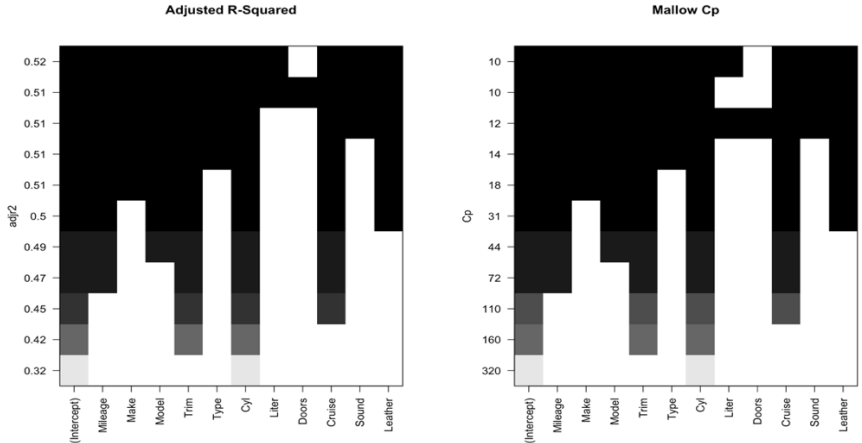

Figure: 6.1

Both of the Adjusted R-Squared and Mallow Cp's plot(Figure 6.1) suggest that the model with 10 variables will give the highest adjusted r-squared and mallow's cp to the approximately equal to the number of predictors.

Hence the model is,

$$Price \sim Mileage + Make + Model + Trim + Type + Cyl + Liter + Cruise + Sound + Leather$$

This does not include the variable Doors.

7. Compare the regression models in question 5 and 6. Summarize your findings by commenting on the following questions.

a. Are different explanatory variables considered important?

Yes, both techniques suggested different variables to include which can be seen in the table 7.1 below.

Technique	Mileage	Make	Model	Trim	Type	Cyl	Liter	Doors	Cruise	Sound	Leather
Stepwise	*					*		*	*	*	*
Best Subset	*	*	*	*	*	*	*		*	*	*

Table: 7.1

Make, Model, Trim and Type were not suggested by the Stepwise regression as we did not include them in the analysis. But there was difference in suggestion of Liter and Doors.

Stepwise suggested that Doors variable is important for the model while Best Subset suggested that Liter is important for the model.

b. Did the stepwise regression in Question 5 provide any indication that Liter could be useful in predicting Price? Did the best subsets output in Question 6 provide any indication that Liter might be useful in predicting Price? Explain why best subsets techniques can be more informative than sequential techniques.

No, the Stepwise regression did not indicate that Liter could be useful in predicting Price. While, the best subset selection's final model suggested that Liter is useful in predicting the Price.

Both procedures build models from a set of predictors that you specify. Stepwise does not assess all models but constructs a model by adding or removing one predictor at a time. Best Subsets does assess all possible models

and it presents you with the best candidates. Stepwise yields a single model, which can be simpler. Best subsets provides more information by including more models, but it can be more complex to choose one. Because Best Subsets assesses all possible models, large models may take a long time to process.

Activity 3 - Checking the model assumptions

8. Using the regression equation obtained in Question 5, create plots of the residuals versus each explanatory variable in the model. Also create a plot of the residuals versus the predicted retail price(often called a residual versus fit plot).

Below are the residual versus variable plots:

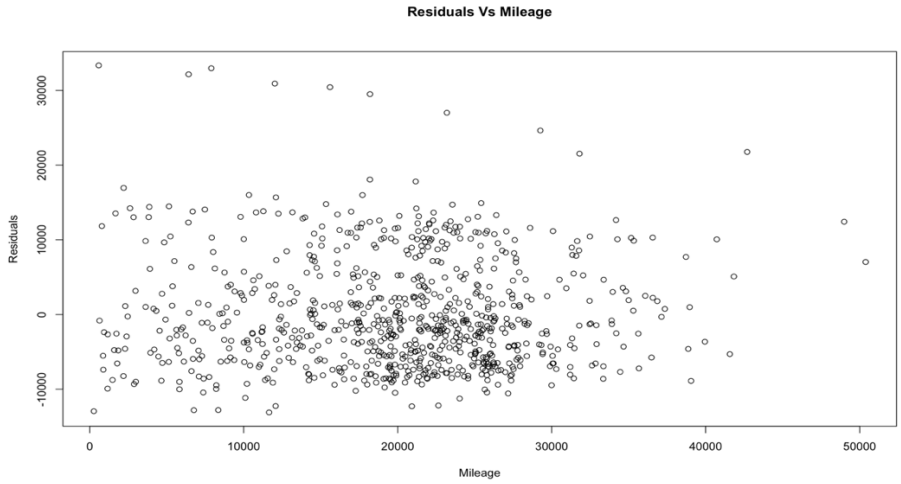

Figure: 8.1

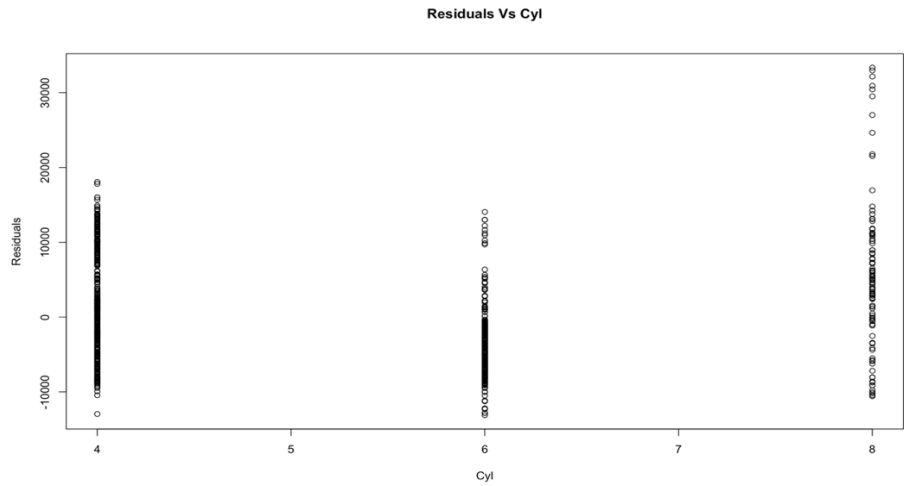

Figure: 8.2

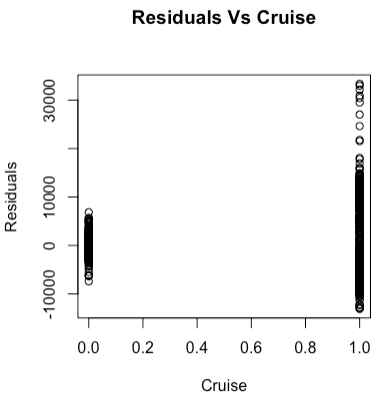

Figure: 8.3

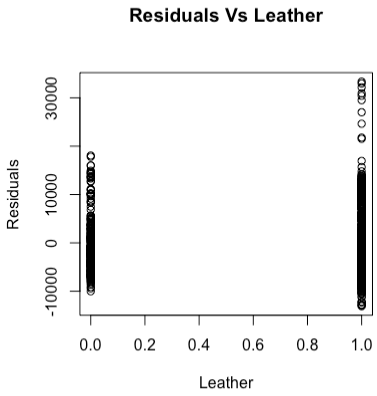

Figure: 8.4

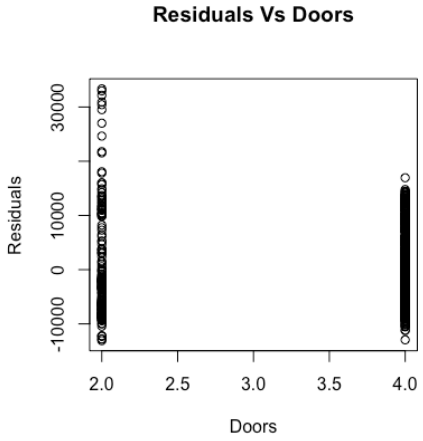

Figure: 8.5

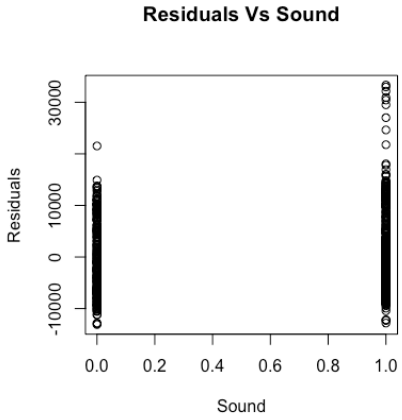

Figure: 8.6

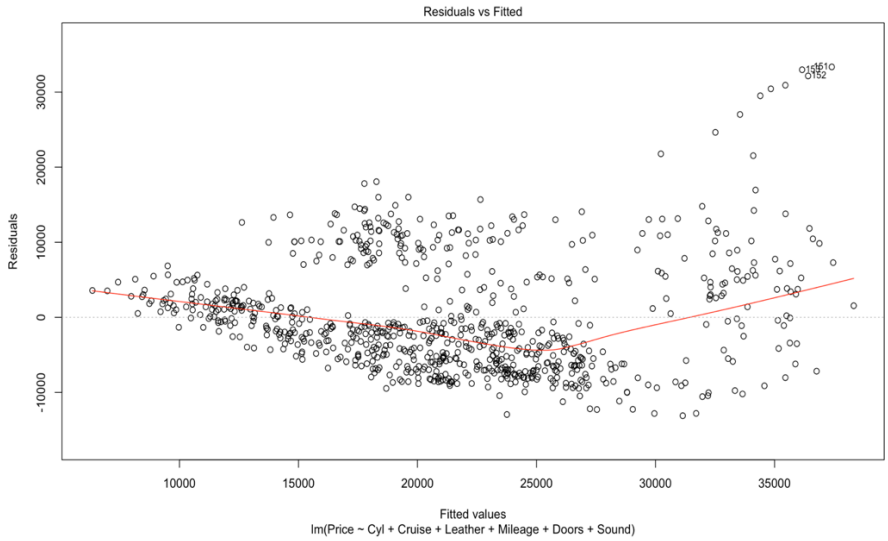

Figure: 8.7

a. Does the size of the residuals tend to change as mileage changes?

No, the size of residuals do not change with the change in mileage. We can observe that in Figure 8.1 that the residuals remain almost constant throughout the different mileage values.

b. Does the size of residuals tend to change as the predicted retail price changes? You should see patterns indicating heteroskedasticity(non-constant variance)

Yes, the size of residuals tend to change as the predicted retail price changes. As seen in Figure 8.7, the red line is going down and then increases which show non-constant variance which is also known as heteroskedasticity. Also, we can observe the funnel shape of the residuals in the plot which suggests we can use transformations like log.

c. Another pattern that may not be immediately obvious from these residual plots is the right skewed. To see the pattern, look at just one vertical slice of this plot. With a pencil, draw a vertical line corresponding to mileage equal to 8000. Are the points in the residual plots balanced around the line Y=0?

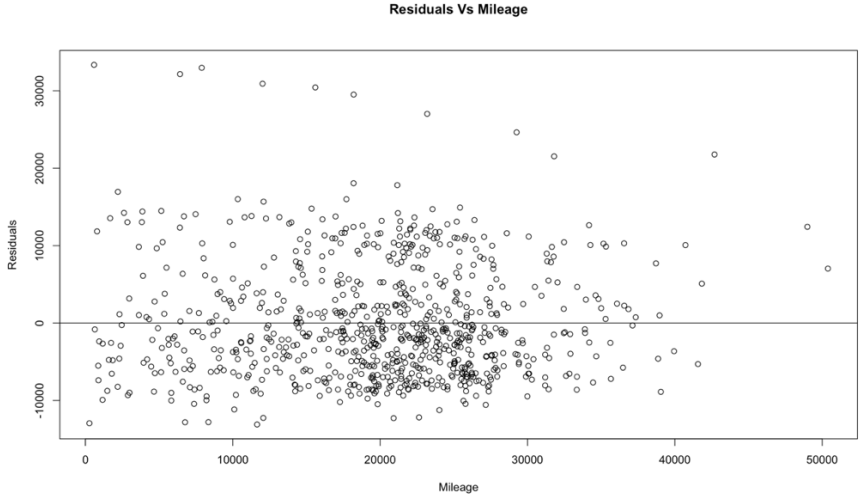

Figure: 8.8

After drawing line y=0 in Figure 8.8, we can see that residuals are not balanced around the line. It can be better seen by making density plot of residuals(Figure 8.9) that residuals are right skewed. It can also be seen in the Q-Q plot(figure 8.10) that residuals are right skewed.

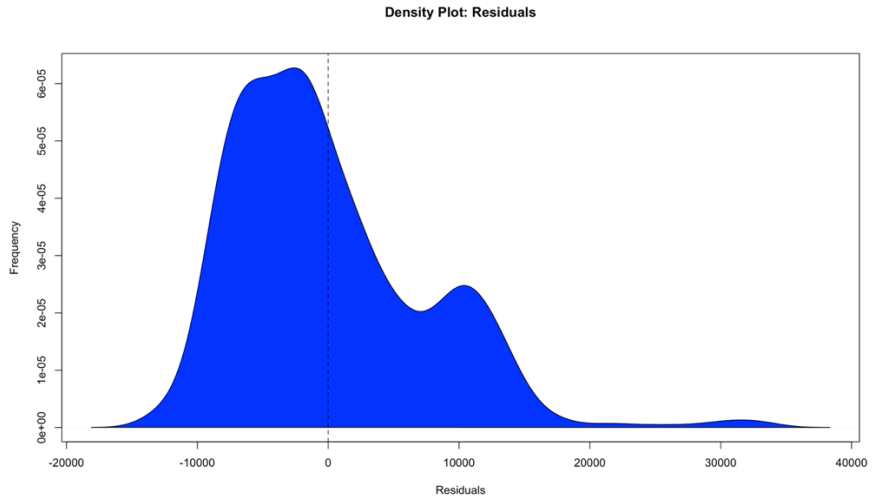

Figure: 8.9

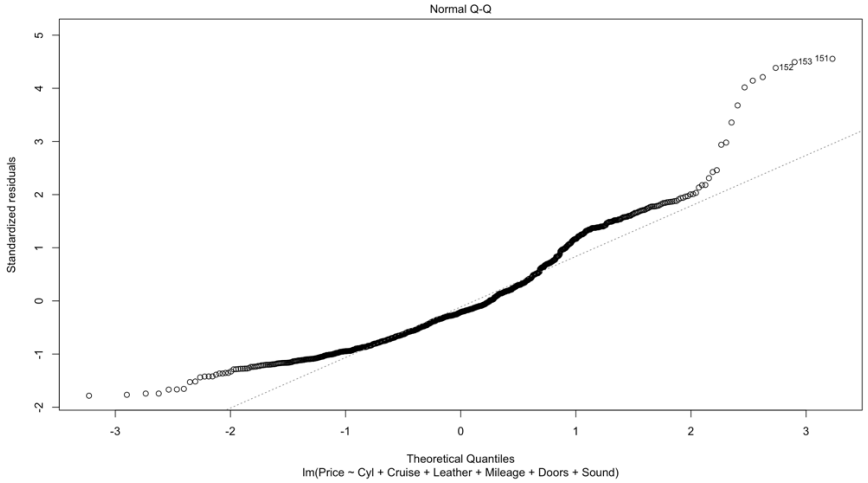

Figure: 8.10

d. Describe any patterns seen in the other residual plots.

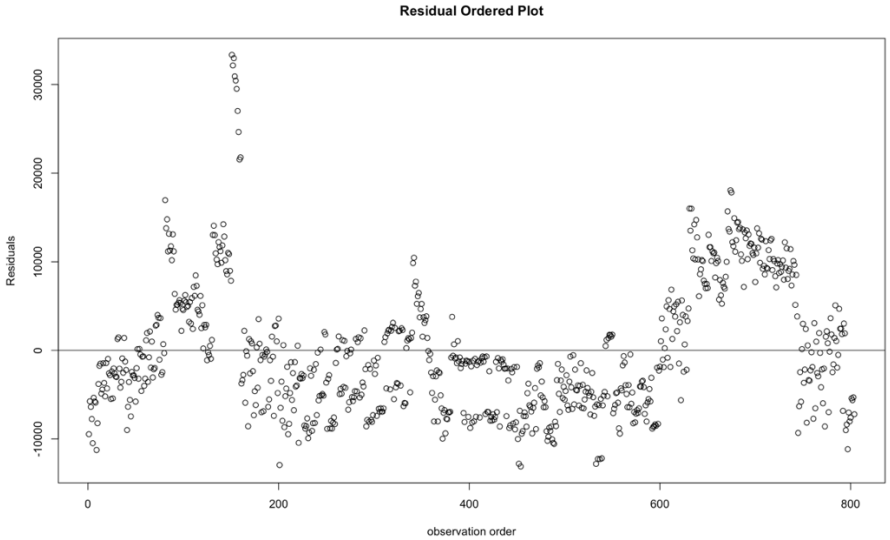

Figure: 8.11

From the ordered residual plot in figure 8.11, we can see that there is clustering of residuals according to the order. The order in the dataset is by Make, Model, Trim and Type in the respective order.

9. **Transform the suggested retail price to $log(Price)$ and $\sqrt{Price}$. Create regression models and residual plots for these transformed response variables using the explanatory variables selected in the Question 5.**

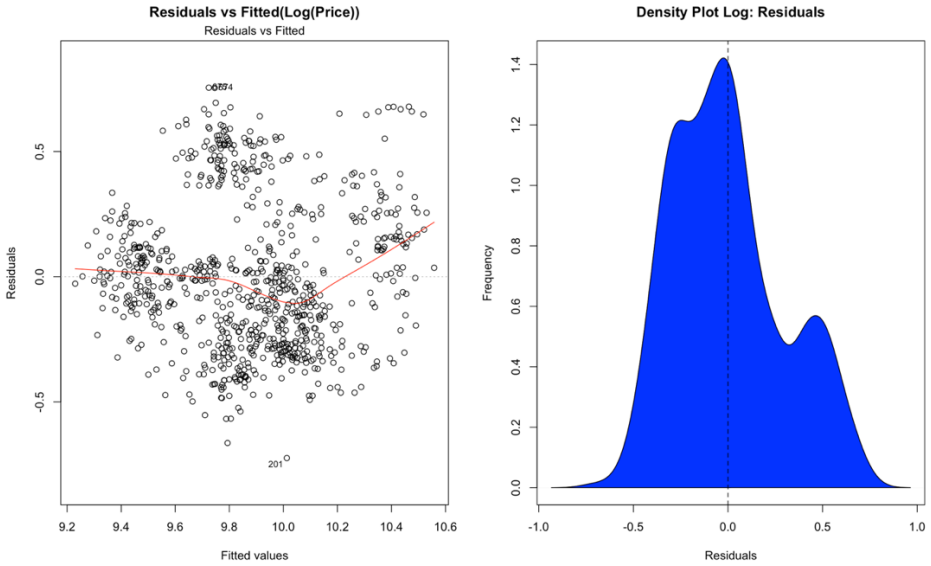

Figure: 9.1

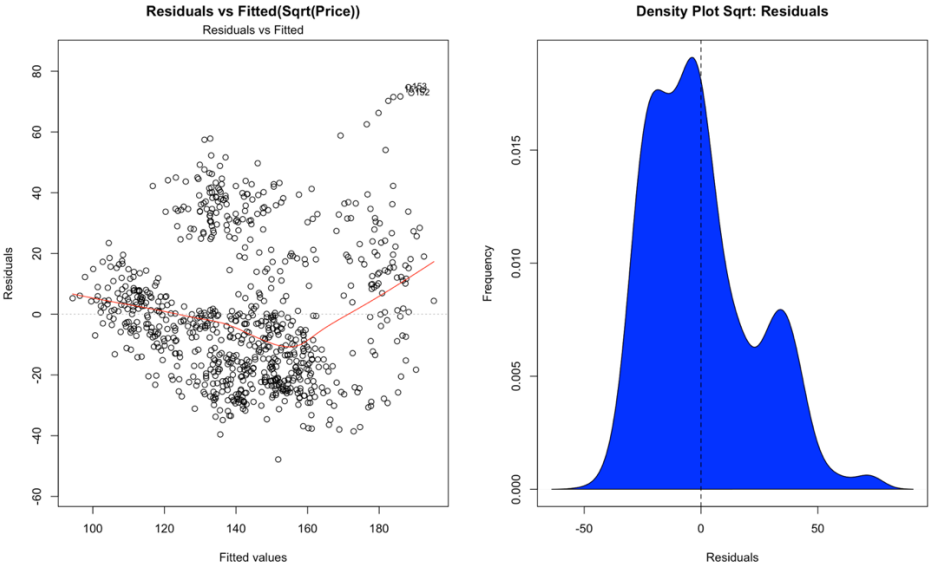

Figure: 9.2

a) Which transformation did best job of reducing the heteroskedasticity and skewness in the residual plots? Give R square for both new models.

$$R - Square\ for\ Log\ Model\ =\ 0.4836$$

$$R - Square\ for\ Sqrt\ Model\ =\ 0.4689$$

Log transformation did a good job in reducing the heteroskedasticity and skewness in the residual plots, we can observe that from the comparison of figure 9.1 and 9.2, by looking at y=0 we can see that for log model the residuals are spread evenly on both sides of y=0 line than that of sqrt model.

On the other hand, density plot also show that sqrt model have more skewness than that of log model as sqrt model is having long right tail.

b) Do the best residual plots correspond to the best R-Square values? Explain.

Yes, the best residual plots correspond to the best R-Square value. This is due to the fact heteroskedasticity mean 'unlike variance'. Therefore, variance changes with the change in input. So, the model will predict better where residual variance is low while the predictions will be worse where residual variance is high. Hence, R-squared will be better for the model which reduces heteroskedasticity.

c) Compare the results obtained from the transformations and without transformations and report your findings.

$$R - Square\ obtained\ from\ the\ linear\ model\ =\ 0.4457$$

And

$$R - Square\ for\ Log\ Model\ =\ 0.4836$$

It can be seen that R-Square for log model is improved. We can compare the r-square here because only dependent variable is transformed. The r-square was able to improve because log-transformation made the residuals to be normally distributed which is our assumption for the model. From the residual vs fitted plot of the without transformed model, we can see the funnel like shape which suggests that log transformation will make the residuals normally distributed.

Also, the coefficients for the model changed due to the fact our dependent variable is log-transformed.

10. Calculate a regression equation using the explanatory variables suggested in Question 5 and Price as the response. Identify any residuals (or cluster of residuals) that don't seem to fit the overall pattern in the residual versus fit and residual versus mileage plots. Any data values that don't seem to fit the general pattern of the data set are called outliers.

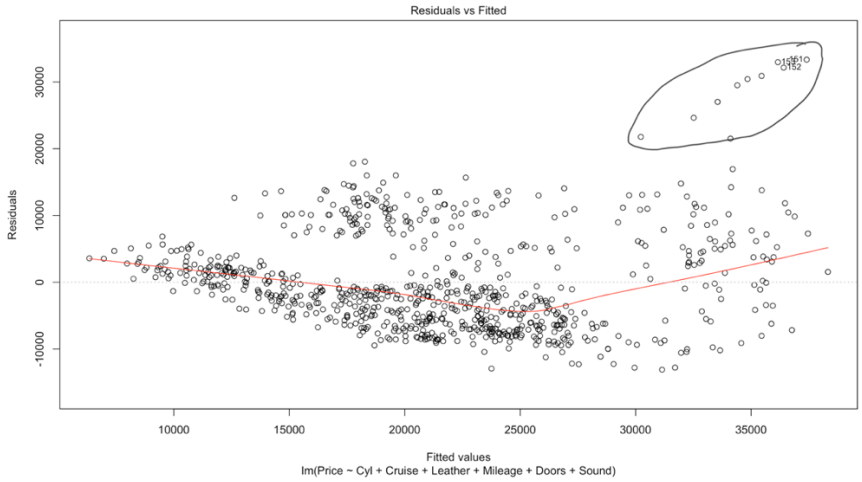

Figure: 10.1

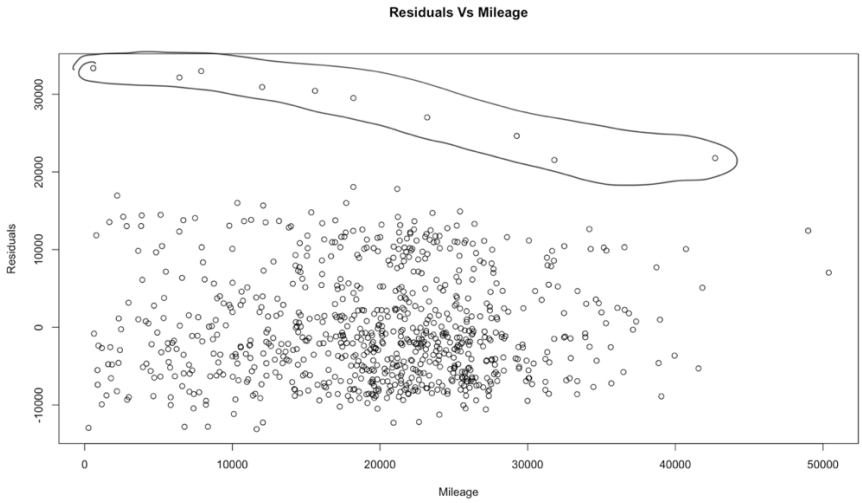

Figure: 10.2

The circled observations in Figure 10.1 and 10.2 are the outliers in the dataset.

a) Identify specific rows of data that represent these points. Are there any consistencies that you can find?

After looking into the data, it was found that these outliers range from observation number 151 to 160. These are the Cadillac's XLR-V8.

b) Is this cluster of outliers helpful in identifying the patterns that were found in the ordered residual plots? Why or Why not?

Yes, these are the same observations which show in the ordered residual plots in figure 8.11, the outliers for which the value of residuals is higher than other cluster corresponds to these outliers.

11. Run the analysis with and without the largest cluster of potential outliers. Use price as the response. Does the cluster of the outliers influence the coefficients in the line?

Analysis with outliers in data:

Variable	Coefficient
Intercept	7323.16
Cyl	3200.12
Cruise	6205.51
Leather	3327.14
Mileage	-0000.17
Doors	-1463.40
Sound	-2024.40

Analysis without outliers in data:

Variable	Coefficient
Intercept	7194.42
Cyl	2614.49
Cruise	6355.60
Leather	3102.43
Mileage	-0000.16
Doors	-0640.83
Sound	-2476.09

From the analysis, it seems that outliers influence the coefficients. To verify this, let's do the analysis further. We will plot Residuals vs Leverage plot to check if the observations are influential or not.

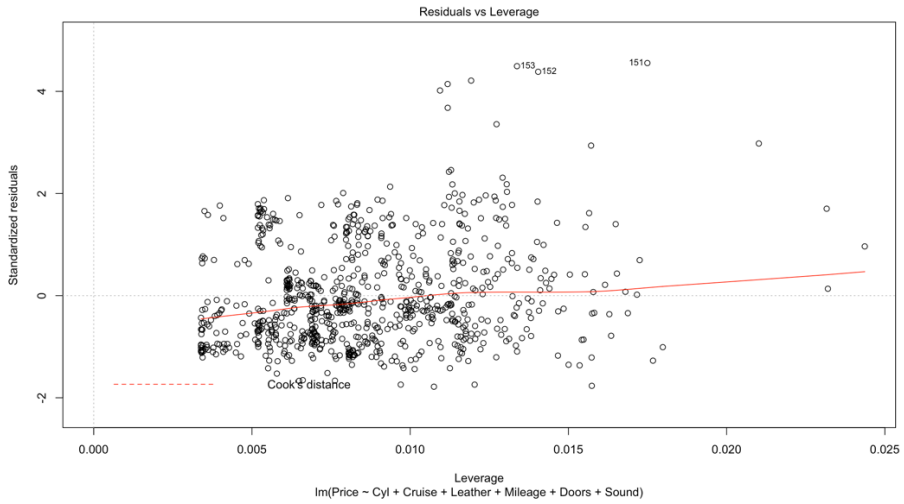

Figure: 11.1

From figure 11.1, it is evident the outliers are not influential in this case. As we can see, there is not even cook's distance lines visible in this plot. Therefore, the observations are not influential.

Final Model

To suggest a final model, I used stepwise regression to select variables from whole dataset using forward approach. In this, stepAIC considered each level of factor variables as separate variable. Also, I used log transformation on response variable to handle the heteroskedasticity and skewness.

So, the suggested model by stepAIC was as follows-

$$\log(Price) \sim Model + Mileage + Trim + Leather + Sound + Cruise$$

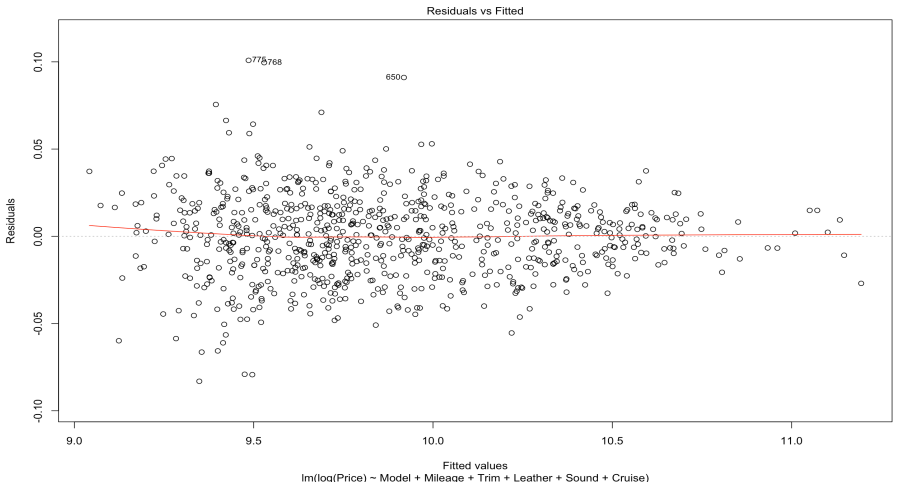

Figure: 12.1

In the above figure(12.1), we can observe the model does not have any heteroskedasticity and it explains almost all variation as $R^2 = 0.9966$. Also, the model is significant.

Appendix

R-Code

```
#Packages to be loaded
library(MASS)
library(leaps)

#Import data from csv file
Cars_Data <- read.csv(file='Project_data_Cars.csv')


#Activity 1
#--1
plot(Cars_Data$Mileage,Cars_Data$Price, xlab = "Mileage", ylab = "Price", main="Scatter Plot")

#--2
Corr_Mileage_Price = cor(Cars_Data$Mileage,Cars_Data$Price)
linear_model_Mileage <- lm(Price ~ Mileage, data = Cars_Data)
summary(linear_model_Mileage)

#--3
linear_model_Mileage$residuals[1]

#Activity 2
#--5
#----a
linear_model_Cyl <- lm(Price ~ Cyl, data = Cars_Data)
linear_model_Liter <- lm(Price ~ Liter, data = Cars_Data)
linear_model_Doors <- lm(Price ~ Doors, data = Cars_Data)
linear_model_Cruise <- lm(Price ~ Cruise, data = Cars_Data)
linear_model_Sound <- lm(Price ~ Sound, data = Cars_Data)
linear_model_Leather <- lm(Price ~ Leather, data = Cars_Data)

R_Squares_X1 <- list()
R_Squares_X1["Mileage"] <- summary(linear_model_Mileage)$r.squared
R_Squares_X1["Cyl"] <- summary(linear_model_Cyl)$r.squared
R_Squares_X1["Liter"] <- summary(linear_model_Liter)$r.squared
R_Squares_X1["Doors"] <- summary(linear_model_Doors)$r.squared
R_Squares_X1["Cruise"] <- summary(linear_model_Cruise)$r.squared
```

```
R_Squares_X1["Sound"] <- summary(linear_model_Sound)$r.squared
R_Squares_X1["Leather"] <- summary(linear_model_Leather)$r.squared

which.max(R_Squares_X1)

#----b
linear_model_Cyl_Mileage <- lm(Price ~ Cyl + Mileage, data = Cars_Data)
linear_model_Cyl_Liter <- lm(Price ~ Cyl + Liter, data = Cars_Data)
linear_model_Cyl_Doors <- lm(Price ~ Cyl + Doors, data = Cars_Data)
linear_model_Cyl_Cruise <- lm(Price ~ Cyl + Cruise, data = Cars_Data)
linear_model_Cyl_Sound <- lm(Price ~ Cyl + Sound, data = Cars_Data)
linear_model_Cyl_Leather <- lm(Price ~ Cyl + Leather, data = Cars_Data)

R_Squares_X1_X2 <- list()
R_Squares_X1_X2["Mileage"] <-
summary(linear_model_Cyl_Mileage)$r.squared
R_Squares_X1_X2["Liter"] <-
summary(linear_model_Cyl_Liter)$r.squared
R_Squares_X1_X2["Doors"] <-
summary(linear_model_Cyl_Doors)$r.squared
R_Squares_X1_X2["Cruise"] <-
summary(linear_model_Cyl_Cruise)$r.squared
R_Squares_X1_X2["Sound"] <-
summary(linear_model_Cyl_Sound)$r.squared
R_Squares_X1_X2["Leather"] <-
summary(linear_model_Cyl_Leather)$r.squared

which.max(R_Squares_X1_X2)

#----c
linear_model_StepWise_intercept = lm(Price~1, data=Cars_Data)
linear_model_StepWise_total =
lm(Price~Cyl+Mileage+Liter+Doors+Cruise+Sound+Leather,data=Cars_D
ata)
linear_model_StepWise_final =
stepAIC(linear_model_StepWise_intercept, direction = "forward",
scope=formula(linear_model_StepWise_total))
linear_model_StepWise_final$anova
```

```
#--6

#Modified Car data so that it can be used by regsubsets

Mod_Cars_Data = Cars_Data
Mod_Cars_Data$Make = factor(Mod_Cars_Data$Make, labels=1:6)
Mod_Cars_Data$Make = as.numeric(Mod_Cars_Data$Make)
Mod_Cars_Data$Model = factor(Mod_Cars_Data$Model, labels=1:32)
Mod_Cars_Data$Model = as.numeric(Mod_Cars_Data$Model)
Mod_Cars_Data$Trim = factor(Mod_Cars_Data$Trim, labels=1:47)
Mod_Cars_Data$Trim = as.numeric(Mod_Cars_Data$Trim)
Mod_Cars_Data$Type = factor(Mod_Cars_Data$Type, labels=1:5)
Mod_Cars_Data$Type = as.numeric(Mod_Cars_Data$Type)

linear_model_BestSub = regsubsets(Price~., data = Mod_Cars_Data,
nvmax = 11)
linear_model_BestSub_summ = summary(linear_model_BestSub)
which.max(linear_model_BestSub_summ$adjr2)
which.min(linear_model_BestSub_summ$cp)
par(mfrow=c(1,2))
plot(linear_model_BestSub, scale = "adjr2", main="Adjusted R-Squared")
plot(linear_model_BestSub, scale = "Cp", main="Mallow Cp")


#Activity 3
#--8
par(mfrow=c(1,1))
plot(Cars_Data$Cyl,linear_model_StepWise_final$residuals, xlab = "Cyl",
ylab="Residuals", main="Residuals Vs Cyl")
plot(Cars_Data$Cruise,linear_model_StepWise_final$residuals, xlab =
"Cruise", ylab="Residuals", main="Residuals Vs Cruise")
plot(Cars_Data$Leather,linear_model_StepWise_final$residuals, xlab =
"Leather", ylab="Residuals", main="Residuals Vs Leather")
plot(Cars_Data$Mileage,linear_model_StepWise_final$residuals, xlab =
"Mileage", ylab="Residuals", main="Residuals Vs Mileage")
plot(Cars_Data$Doors,linear_model_StepWise_final$residuals, xlab =
"Doors", ylab="Residuals", main="Residuals Vs Doors")
plot(Cars_Data$Sound,linear_model_StepWise_final$residuals, xlab =
"Sound", ylab="Residuals", main="Residuals Vs Sound")
```

```
plot(linear_model_StepWise_final$fitted.values,
linear_model_StepWise_final$residuals, xlab = "Predicted Price",
ylab="Residuals", main="Residuals Vs Predicted Price")
plot(linear_model_StepWise_final,1)

#----c
plot(density(linear_model_StepWise_final$residuals), main="Density
Plot: Residuals", xlab="Residuals", ylab="Frequency")
polygon(density(linear_model_StepWise_final$residuals), col="blue")
abline(v = 0, lty = 2)

plot(Cars_Data$Mileage,linear_model_StepWise_final$residuals, xlab =
"Mileage", ylab="Residuals", main="Residuals Vs Mileage")
abline(a = 0, b=0)

#----d
plot(c(1:804),linear_model_StepWise_final$residuals, xlab="observation
order", ylab="Residuals", main="Residual Ordered Plot")
abline(a=0,b=0)

#--9
par(mfrow=c(1,2))

linear_Model_Log = lm(log(Price) ~ Cyl + Cruise + Leather + Mileage +
Doors + Sound, data=Cars_Data)
summary(linear_Model_Log)
plot(linear_Model_Log,1, main = "Residuals vs Fitted(Log(Price))")

plot(density(linear_Model_Log$residuals), main="Density Plot Log:
Residuals", xlab="Residuals", ylab="Frequency")
polygon(density(linear_Model_Log$residuals), col="blue")
abline(v = 0, lty = 2)

linear_Model_Sqrt = lm(sqrt(Price) ~ Cyl + Cruise + Leather + Mileage +
Doors + Sound, data=Cars_Data)
summary(linear_Model_Sqrt)
plot(linear_Model_Sqrt,1, main = "Residuals vs Fitted(Sqrt(Price))")

plot(density(linear_Model_Sqrt$residuals), main="Density Plot Sqrt:
Residuals", xlab="Residuals", ylab="Frequency")
```

```
polygon(density(linear_Model_Sqrt$residuals), col="blue")
abline(v = 0, lty = 2)


#Activity 4
#--10
par(mfrow=c(1,1))
plot(linear_model_StepWise_final,2)

#--11

cooksd = cooks.distance(linear_model_StepWise_final)
plot(linear_model_StepWise_final, which = 4)
Mod_Cars_Data2 = Cars_Data[-c(151:160),]

lm_without_outlier = lm(Price~Cyl + Cruise + Leather + Mileage + Doors +
Sound, data=Mod_Cars_Data2)
summary(lm_without_outlier)

#FINAL MODEL

fm_intercept = lm(log(Price)~1, data=Cars_Data)
fm_total = lm(log(Price)~., data=Cars_Data)
fm_final = stepAIC(fm_intercept, direction = "forward", scope =
formula(fm_total))
summary(fm_final)
```