MID-TERM PROJECT

Charan B

2024-03-18

# INTRODUCTION

* In this project, we will analyze the “mtcars” dataset to investigate the link between different automobile features and quarter-mile timings (qsec). We use regression analysis to study how parameters such as mpg, cyl, disp, hp, drat, and wt impact qsec. Our objective is to build a regression model that best predicts qsec, taking into account first-order effects, second-order interactions, data transformation, influential points, leverage, and multicollinearity. By doing so, we want to gain insight into the major factors of automobile acceleration and create a prediction model for quarter-mile timings.

# DATA

* The’mtcars’ collection provides statistics on 32 vehicles, including fuel efficiency, cylinders, horsepower, and more. It is often used to evaluate vehicle performance and attributes.

## Considering the Inbuilt data Mtcars

#Loading the data  
data <- mtcars

## We have taken the mtcars data into a variable called data and checking the top 6 data using head function

#Getting the top 6 data  
head(data)

## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1  
## Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1  
## Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2  
## Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

## Checking the structure of the data

#Checking the structure of the data  
str(data)

## 'data.frame': 32 obs. of 11 variables:  
## $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...  
## $ cyl : num 6 6 4 6 8 6 8 4 4 6 ...  
## $ disp: num 160 160 108 258 360 ...  
## $ hp : num 110 110 93 110 175 105 245 62 95 123 ...  
## $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...  
## $ wt : num 2.62 2.88 2.32 3.21 3.44 ...  
## $ qsec: num 16.5 17 18.6 19.4 17 ...  
## $ vs : num 0 0 1 1 0 1 0 1 1 1 ...  
## $ am : num 1 1 1 0 0 0 0 0 0 0 ...  
## $ gear: num 4 4 4 3 3 3 3 4 4 4 ...  
## $ carb: num 4 4 1 1 2 1 4 2 2 4 ...

* We can see all the variables are in num so we don’t have to do any transformations to the data

## Checking if there is any NA values in the data

#Checking for NA values in the data  
sum(is.na(data))

## [1] 0

* We can see there is no NA values in the data that we have considered

## Checking the summary statistics of the whole data (mean,median,max etc)

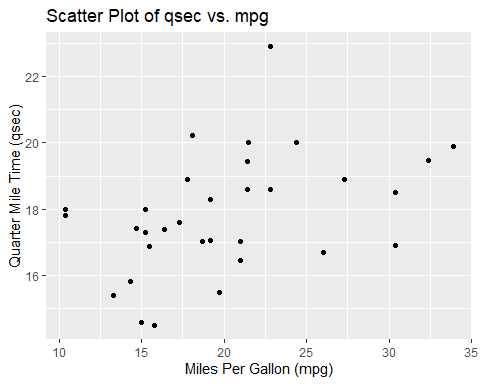
#checking the summary  
summary(data)

## mpg cyl disp hp   
## Min. :10.40 Min. :4.000 Min. : 71.1 Min. : 52.0   
## 1st Qu.:15.43 1st Qu.:4.000 1st Qu.:120.8 1st Qu.: 96.5   
## Median :19.20 Median :6.000 Median :196.3 Median :123.0   
## Mean :20.09 Mean :6.188 Mean :230.7 Mean :146.7   
## 3rd Qu.:22.80 3rd Qu.:8.000 3rd Qu.:326.0 3rd Qu.:180.0   
## Max. :33.90 Max. :8.000 Max. :472.0 Max. :335.0   
## drat wt qsec vs   
## Min. :2.760 Min. :1.513 Min. :14.50 Min. :0.0000   
## 1st Qu.:3.080 1st Qu.:2.581 1st Qu.:16.89 1st Qu.:0.0000   
## Median :3.695 Median :3.325 Median :17.71 Median :0.0000   
## Mean :3.597 Mean :3.217 Mean :17.85 Mean :0.4375   
## 3rd Qu.:3.920 3rd Qu.:3.610 3rd Qu.:18.90 3rd Qu.:1.0000   
## Max. :4.930 Max. :5.424 Max. :22.90 Max. :1.0000   
## am gear carb   
## Min. :0.0000 Min. :3.000 Min. :1.000   
## 1st Qu.:0.0000 1st Qu.:3.000 1st Qu.:2.000   
## Median :0.0000 Median :4.000 Median :2.000   
## Mean :0.4062 Mean :3.688 Mean :2.812   
## 3rd Qu.:1.0000 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :1.0000 Max. :5.000 Max. :8.000

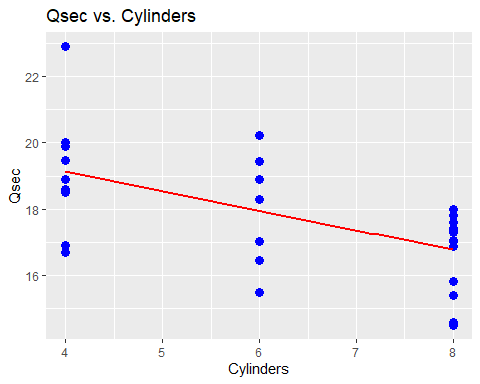
* we can see the whole summary statistics of the data
* The’mtcars’ dataset does not require power modifications because its distribution looks to be about normal.

## Plotting graphs using the above data

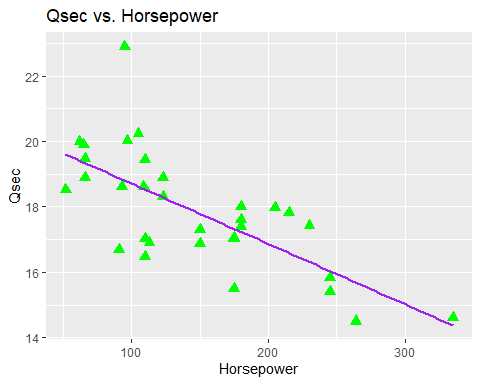
# Load required libraries  
library(ggplot2)  
  
# Scatter plots of all the variables comparing to qsec  
  
ggplot(data, aes(x = mpg, y = qsec)) +  
 geom\_point() +  
 ggtitle("Scatter Plot of qsec vs. mpg") +  
 xlab("Miles Per Gallon (mpg)") +  
 ylab("Quarter Mile Time (qsec)")



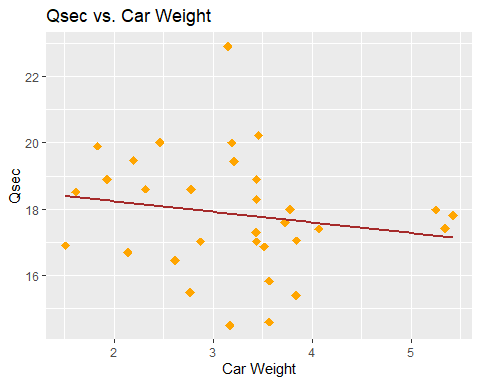
ggplot(data, aes(x = cyl, y = qsec)) +  
 geom\_point(color = "blue", size = 3) +  
 geom\_smooth(method = "lm", se = FALSE, color = "red") +  
 labs(title = "Qsec vs. Cylinders", x = "Cylinders", y = "Qsec")



ggplot(data, aes(x = hp, y = qsec)) +  
 geom\_point(color = "green", size = 3, shape = 17) +  
 geom\_smooth(method = "lm", se = FALSE, color = "purple") +  
 labs(title = "Qsec vs. Horsepower", x = "Horsepower", y = "Qsec")



ggplot(data, aes(x = wt, y = qsec)) +  
 geom\_point(color = "orange", size = 3, shape = 18) +  
 geom\_smooth(method = "lm", se = FALSE, color = "brown") +  
 labs(title = "Qsec vs. Car Weight", x = "Car Weight", y = "Qsec")



**Regression** **Analysis**

## Fitting the model along with the interactions

#Fitting the model  
fit <- lm(qsec ~ mpg + cyl + disp + hp + drat + wt + mpg:cyl + mpg:disp + mpg:hp + mpg:drat + mpg:wt + cyl:disp + cyl:hp + cyl:drat + cyl:wt + disp:hp + disp:drat + disp:wt + hp:drat + hp:wt + drat:wt, data = data)

## Checking the summary of the fit and rsquare value

#summary of fit  
summary(fit)

##   
## Call:  
## lm(formula = qsec ~ mpg + cyl + disp + hp + drat + wt + mpg:cyl +   
## mpg:disp + mpg:hp + mpg:drat + mpg:wt + cyl:disp + cyl:hp +   
## cyl:drat + cyl:wt + disp:hp + disp:drat + disp:wt + hp:drat +   
## hp:wt + drat:wt, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.01422 -0.31671 -0.07481 0.39584 1.31717   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.139e+01 1.309e+02 0.316 0.758   
## mpg -5.740e-02 3.473e+00 -0.017 0.987   
## cyl -1.115e+01 1.061e+01 -1.052 0.318   
## disp 1.758e-01 2.556e-01 0.688 0.507   
## hp -4.658e-01 3.358e-01 -1.387 0.196   
## drat -8.239e+00 2.559e+01 -0.322 0.754   
## wt 2.588e+01 2.370e+01 1.092 0.300   
## mpg:cyl -4.101e-02 3.507e-01 -0.117 0.909   
## mpg:disp -3.442e-03 4.538e-03 -0.759 0.466   
## mpg:hp 1.117e-02 8.586e-03 1.301 0.222   
## mpg:drat 9.536e-02 6.681e-01 0.143 0.889   
## mpg:wt -2.267e-01 2.925e-01 -0.775 0.456   
## cyl:disp 1.822e-02 1.025e-02 1.777 0.106   
## cyl:hp 8.976e-03 1.908e-02 0.471 0.648   
## cyl:drat 3.035e+00 2.862e+00 1.060 0.314   
## cyl:wt -1.720e+00 1.606e+00 -1.071 0.309   
## disp:hp -2.726e-04 2.700e-04 -1.010 0.336   
## disp:drat -1.178e-02 3.959e-02 -0.298 0.772   
## disp:wt -4.564e-02 2.361e-02 -1.933 0.082 .  
## hp:drat -1.155e-02 2.353e-02 -0.491 0.634   
## hp:wt 9.341e-02 5.766e-02 1.620 0.136   
## drat:wt -3.086e+00 5.117e+00 -0.603 0.560   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9518 on 10 degrees of freedom  
## Multiple R-squared: 0.9085, Adjusted R-squared: 0.7163   
## F-statistic: 4.727 on 21 and 10 DF, p-value: 0.0075

summary(fit)$r.squared

## [1] 0.9084855

* From the above summary of fit we can observe that p-value is 0.0075.
* We can see that there is only one significant value that is interaction of disp:wt
* Now we are gonna check with the second order interaction to dertermine the best significant model.

## Checking the t and f test of fit Using anova

#anova of fit  
anova(fit)

## Analysis of Variance Table  
##   
## Response: qsec  
## Df Sum Sq Mean Sq F value Pr(>F)   
## mpg 1 17.3523 17.3523 19.1550 0.0013846 \*\*   
## cyl 1 19.8719 19.8719 21.9364 0.0008627 \*\*\*  
## disp 1 3.4117 3.4117 3.7662 0.0809925 .   
## hp 1 20.8555 20.8555 23.0222 0.0007256 \*\*\*  
## drat 1 4.6800 4.6800 5.1662 0.0463389 \*   
## wt 1 8.8910 8.8910 9.8147 0.0106389 \*   
## mpg:cyl 1 0.0039 0.0039 0.0043 0.9487767   
## mpg:disp 1 1.9328 1.9328 2.1336 0.1748010   
## mpg:hp 1 1.9611 1.9611 2.1648 0.1719541   
## mpg:drat 1 0.0011 0.0011 0.0013 0.9723903   
## mpg:wt 1 0.1864 0.1864 0.2058 0.6597902   
## cyl:disp 1 0.0238 0.0238 0.0262 0.8745764   
## cyl:hp 1 0.0516 0.0516 0.0569 0.8162097   
## cyl:drat 1 1.3994 1.3994 1.5448 0.2422555   
## cyl:wt 1 3.2778 3.2778 3.6184 0.0863086 .   
## disp:hp 1 0.0973 0.0973 0.1074 0.7498639   
## disp:drat 1 1.1768 1.1768 1.2991 0.2809467   
## disp:wt 1 1.3428 1.3428 1.4823 0.2513693   
## hp:drat 1 0.1704 0.1704 0.1881 0.6737423   
## hp:wt 1 2.9123 2.9123 3.2149 0.1032208   
## drat:wt 1 0.3295 0.3295 0.3637 0.5598662   
## Residuals 10 9.0588 0.9059   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* From the anova table we can observe that the interaction terms are not significant.
* The regression study was performed using R's linear model function, with the dependent variable "qsec" regressed on numerous predictor variables such as "mpg," "cyl," "disp," "hp," "drat," and "wt," as well as their interactions. The analysis produced estimates, standard errors, t-values, and p-values for each predictor variable. Notably, the variables "hp" and "drat" have statistically significant associations with "qsec," as seen by their low p-values. The whole model has a rather high adjusted R-squared value of 0.7163, suggesting that the model's predictors explained roughly 71.63% of the variation in "qsec". The F-statistic (p < 0.01) indicated that the entire model accurately predicted "qsec."

## Fitting the model along with the Quadratic interactions

#fitting the model1  
fit1 <- lm(qsec ~ mpg + cyl + disp + hp + drat + wt + I(mpg^2)+ I(cyl^2)+ I(disp^2)+ I(hp^2) + I(drat^2) + I(wt^2), data = data)

## Checking the summary of the fit1 and rsquare value

#summary of fit1  
summary(fit1)

##   
## Call:  
## lm(formula = qsec ~ mpg + cyl + disp + hp + drat + wt + I(mpg^2) +   
## I(cyl^2) + I(disp^2) + I(hp^2) + I(drat^2) + I(wt^2), data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.68896 -0.37954 -0.00563 0.44964 2.20705   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.995e+01 1.016e+01 2.947 0.00827 \*\*  
## mpg -2.650e-01 3.372e-01 -0.786 0.44165   
## cyl -2.420e+00 1.469e+00 -1.647 0.11591   
## disp 7.505e-03 2.404e-02 0.312 0.75826   
## hp -2.010e-02 2.352e-02 -0.855 0.40344   
## drat -3.980e+00 5.520e+00 -0.721 0.47971   
## wt 4.624e+00 2.283e+00 2.025 0.05711 .   
## I(mpg^2) 7.673e-03 7.070e-03 1.085 0.29137   
## I(cyl^2) 1.227e-01 1.251e-01 0.981 0.33887   
## I(disp^2) -1.603e-05 3.931e-05 -0.408 0.68800   
## I(hp^2) 1.551e-05 5.000e-05 0.310 0.75982   
## I(drat^2) 4.320e-01 7.298e-01 0.592 0.56083   
## I(wt^2) -4.089e-01 3.124e-01 -1.309 0.20623   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9462 on 19 degrees of freedom  
## Multiple R-squared: 0.8281, Adjusted R-squared: 0.7196   
## F-statistic: 7.63 on 12 and 19 DF, p-value: 5.752e-05

summary(fit1)$r.squared

## [1] 0.8281429

* From the above summary we can observe that
* The Adjusted R-squared: 0.7196 and the p-value: 5.752e-05 is small and the model is adequate.

##checking the t and f test of fit1

#anova of fit1  
anova(fit1)

## Analysis of Variance Table  
##   
## Response: qsec  
## Df Sum Sq Mean Sq F value Pr(>F)   
## mpg 1 17.3523 17.3523 19.3802 0.0003061 \*\*\*  
## cyl 1 19.8719 19.8719 22.1943 0.0001521 \*\*\*  
## disp 1 3.4117 3.4117 3.8104 0.0658307 .   
## hp 1 20.8555 20.8555 23.2929 0.0001173 \*\*\*  
## drat 1 4.6800 4.6800 5.2270 0.0338929 \*   
## wt 1 8.8910 8.8910 9.9301 0.0052583 \*\*   
## I(mpg^2) 1 0.0082 0.0082 0.0092 0.9247939   
## I(cyl^2) 1 1.1988 1.1988 1.3389 0.2615719   
## I(disp^2) 1 3.7534 3.7534 4.1921 0.0546942 .   
## I(hp^2) 1 0.4152 0.4152 0.4637 0.5041030   
## I(drat^2) 1 0.0048 0.0048 0.0054 0.9421545   
## I(wt^2) 1 1.5336 1.5336 1.7128 0.2062320   
## Residuals 19 17.0118 0.8954   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Comparing Anova of fit and fit1  
anova(fit,fit1)

## Analysis of Variance Table  
##   
## Model 1: qsec ~ mpg + cyl + disp + hp + drat + wt + mpg:cyl + mpg:disp +   
## mpg:hp + mpg:drat + mpg:wt + cyl:disp + cyl:hp + cyl:drat +   
## cyl:wt + disp:hp + disp:drat + disp:wt + hp:drat + hp:wt +   
## drat:wt  
## Model 2: qsec ~ mpg + cyl + disp + hp + drat + wt + I(mpg^2) + I(cyl^2) +   
## I(disp^2) + I(hp^2) + I(drat^2) + I(wt^2)  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 10 9.0588   
## 2 19 17.0118 -9 -7.953 0.9755 0.5102

* Model 1 contains interaction terms between predictors, whereas Model 2 includes quadratics. The F-test indicates that the quadratic components in Model 2 do not significantly improve model fit over Model 1. Thus, Model 1 with interaction terms is favored for predicting ‘qsec’.
* By comparing all the models eliminating the non significant terms and making the final fit with the only significant terms.

## The fit2 is mentioned below

#fitting the model fit2  
fit2 <- lm(qsec ~ mpg + cyl + disp + hp + drat + wt + cyl:wt , data = data)

## Checking the summary of the fit2 and rsquare value

#summary  
summary(fit2)

##   
## Call:  
## lm(formula = qsec ~ mpg + cyl + disp + hp + drat + wt + cyl:wt,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.63829 -0.44372 0.01414 0.44389 2.49497   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.1394440 5.8192676 2.773 0.0106 \*  
## mpg 0.1084893 0.0826365 1.313 0.2016   
## cyl -0.0556096 0.5198243 -0.107 0.9157   
## disp -0.0003244 0.0049490 -0.066 0.9483   
## hp -0.0127062 0.0053915 -2.357 0.0269 \*  
## drat -0.8974424 0.5344670 -1.679 0.1061   
## wt 3.0580317 1.2296149 2.487 0.0202 \*  
## cyl:wt -0.2260005 0.1644141 -1.375 0.1820   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9613 on 24 degrees of freedom  
## Multiple R-squared: 0.7759, Adjusted R-squared: 0.7106   
## F-statistic: 11.87 on 7 and 24 DF, p-value: 1.889e-06

summary(fit2)$r.squared

## [1] 0.7759362

**Checking the t and f test of fit2**

#anova of fit2  
anova(fit2)

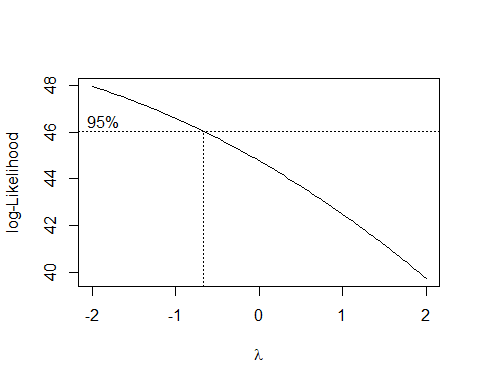
## Analysis of Variance Table  
##   
## Response: qsec  
## Df Sum Sq Mean Sq F value Pr(>F)   
## mpg 1 17.3523 17.3523 18.7764 0.0002264 \*\*\*  
## cyl 1 19.8719 19.8719 21.5028 0.0001045 \*\*\*  
## disp 1 3.4117 3.4117 3.6917 0.0666370 .   
## hp 1 20.8555 20.8555 22.5671 7.832e-05 \*\*\*  
## drat 1 4.6800 4.6800 5.0641 0.0338627 \*   
## wt 1 8.8910 8.8910 9.6207 0.0048673 \*\*   
## cyl:wt 1 1.7462 1.7462 1.8895 0.1819612   
## Residuals 24 22.1797 0.9242   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* We can see all the predictor variables are significant.
* R-squared: 0.7106 and p-value: 1.889e-06 so the fit is adequate.
* Next we are Trying power transformation on the data to check whether the fit will be more significant.

## Boxcox Transformation

* using boxcox transformation transforming the data and fitting model to check whether the model will be more significant after transforming.

# Load the required library  
library(MASS)  
  
# Box-Cox transformation  
b <- boxcox(fit2)



lambda <- b$x  
likelihood <- b$y  
best\_lambda <- lambda[which.max(likelihood)]  
best\_lambda

## [1] -2

# Transform the response variable using the best lambda  
data$qsec\_trans <- (data$qsec^best\_lambda)

# Fit the model with transformed response variable  
fit2\_trans <- lm(qsec\_trans ~ mpg + cyl + disp + hp + drat + wt + cyl:wt, data = data)  
  
# Summary of the transformed model  
summary(fit2\_trans)

##   
## Call:  
## lm(formula = qsec\_trans ~ mpg + cyl + disp + hp + drat + wt +   
## cyl:wt, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.710e-04 -1.396e-04 -9.290e-06 1.131e-04 5.828e-04   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.364e-03 1.749e-03 1.352 0.189003   
## mpg -2.773e-05 2.483e-05 -1.117 0.275164   
## cyl 1.279e-04 1.562e-04 0.819 0.421008   
## disp 7.506e-07 1.487e-06 0.505 0.618384   
## hp 6.859e-06 1.620e-06 4.233 0.000292 \*\*\*  
## drat 3.470e-04 1.606e-04 2.160 0.040966 \*   
## wt -6.590e-04 3.695e-04 -1.783 0.087165 .   
## cyl:wt 1.531e-05 4.941e-05 0.310 0.759288   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0002889 on 24 degrees of freedom  
## Multiple R-squared: 0.8454, Adjusted R-squared: 0.8003   
## F-statistic: 18.75 on 7 and 24 DF, p-value: 2.671e-08

#performing the anova of transformed fit2  
anova(fit2\_trans)

## Analysis of Variance Table  
##   
## Response: qsec\_trans  
## Df Sum Sq Mean Sq F value Pr(>F)   
## mpg 1 2.2666e-06 2.2666e-06 27.1573 2.436e-05 \*\*\*  
## cyl 1 2.1662e-06 2.1662e-06 25.9547 3.272e-05 \*\*\*  
## disp 1 4.5550e-07 4.5550e-07 5.4578 0.0281621 \*   
## hp 1 4.2362e-06 4.2362e-06 50.7563 2.304e-07 \*\*\*  
## drat 1 6.1640e-07 6.1640e-07 7.3857 0.0120096 \*   
## wt 1 1.2039e-06 1.2039e-06 14.4243 0.0008766 \*\*\*  
## cyl:wt 1 8.0000e-09 8.0000e-09 0.0961 0.7592877   
## Residuals 24 2.0031e-06 8.3500e-08   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## comparing the r square values of the fit2 and fit2\_trans

#summary of fit2  
summary(fit2)$r.squared

## [1] 0.7759362

#summary of transformedc fit2  
summary(fit2\_trans)$r.squared

## [1] 0.8453929

* We can observe that by doing the boxcox transformation the r square value has been increased.
* The transformed fit is more significant so we are gonna do our further analysis using the transformed fit
* From the above anova table of fit2\_trans we can observe that the interaction is not significant.
* Eliminating the interaction and creating the new fit using the remaining variables.

## Creating new fit by removing the interaction

# Fit the model with transformed response variable  
fit3 <- lm(qsec\_trans ~ mpg + cyl + disp + hp + drat + wt , data = data)  
  
# Summary of the fit3  
summary(fit3)

##   
## Call:  
## lm(formula = qsec\_trans ~ mpg + cyl + disp + hp + drat + wt,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.992e-04 -1.544e-04 -6.210e-06 1.348e-04 5.807e-04   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.964e-03 1.157e-03 1.697 0.102102   
## mpg -2.436e-05 2.192e-05 -1.111 0.276946   
## cyl 1.685e-04 8.361e-05 2.015 0.054778 .   
## disp 9.209e-07 1.357e-06 0.679 0.503578   
## hp 6.930e-06 1.575e-06 4.401 0.000176 \*\*\*  
## drat 3.521e-04 1.568e-04 2.245 0.033877 \*   
## wt -5.538e-04 1.432e-04 -3.869 0.000694 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0002836 on 25 degrees of freedom  
## Multiple R-squared: 0.8448, Adjusted R-squared: 0.8075   
## F-statistic: 22.68 on 6 and 25 DF, p-value: 5.534e-09

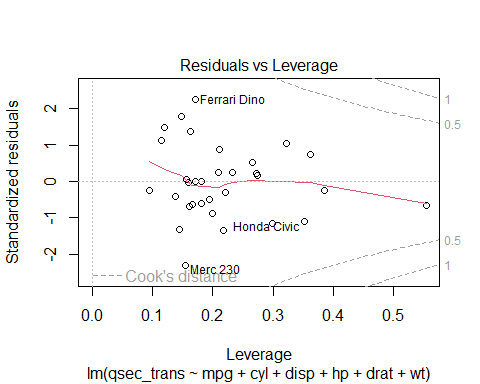
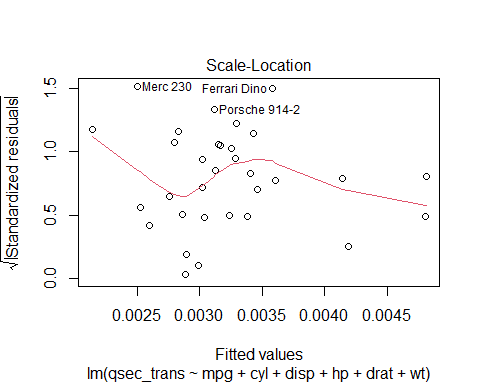
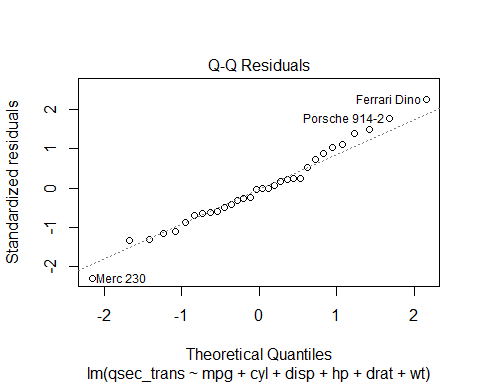
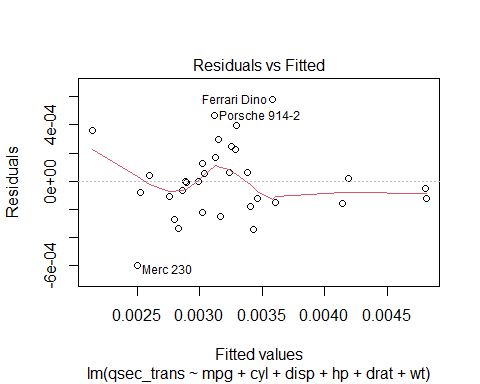
#performing the anova of fit3  
anova(fit3)

## Analysis of Variance Table  
##   
## Response: qsec\_trans  
## Df Sum Sq Mean Sq F value Pr(>F)   
## mpg 1 2.2666e-06 2.2666e-06 28.1761 1.683e-05 \*\*\*  
## cyl 1 2.1662e-06 2.1662e-06 26.9283 2.286e-05 \*\*\*  
## disp 1 4.5550e-07 4.5550e-07 5.6626 0.0252763 \*   
## hp 1 4.2362e-06 4.2362e-06 52.6603 1.326e-07 \*\*\*  
## drat 1 6.1640e-07 6.1640e-07 7.6628 0.0104609 \*   
## wt 1 1.2039e-06 1.2039e-06 14.9654 0.0006939 \*\*\*  
## Residuals 25 2.0111e-06 8.0400e-08   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* From the above summary we can observe that R-squared: 0.8075 and p-value: 5.534e-09 so the model is adequate.
* From the anova table we can observe that all the variables are significant.
* We are gonna go further and check for outliners, leverage and influential points.

## Checking fot outliners, levarge and influential points

#plotting the fit3  
plot(fit3)



* From the above diagnostic plots we can observe that,
* Variance is constant and error is normally distributed.
* In the Standardized residual plot suggests observations 3 points are the outliers.
* I the Residuals vs leverage plot, we can say that observation 3 points have leverage.
* And finally from Cook’s distance shows that no observation is influential.
* we are gonna remove merc230,honda civic,ferarri dino leverage point to check whether the fit will be more significant.

## Removing Point Merc 230 Data

#removing the points  
new\_data<-data[-c(9,19,30),]  
head(new\_data)

## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1  
## Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1  
## Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2  
## Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1  
## qsec\_trans  
## Mazda RX4 0.003690968  
## Mazda RX4 Wag 0.003452080  
## Datsun 710 0.002887402  
## Hornet 4 Drive 0.002646107  
## Hornet Sportabout 0.003452080  
## Valiant 0.002445894

str(new\_data)

## 'data.frame': 29 obs. of 12 variables:  
## $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 19.2 17.8 ...  
## $ cyl : num 6 6 4 6 8 6 8 4 6 6 ...  
## $ disp : num 160 160 108 258 360 ...  
## $ hp : num 110 110 93 110 175 105 245 62 123 123 ...  
## $ drat : num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...  
## $ wt : num 2.62 2.88 2.32 3.21 3.44 ...  
## $ qsec : num 16.5 17 18.6 19.4 17 ...  
## $ vs : num 0 0 1 1 0 1 0 1 1 1 ...  
## $ am : num 1 1 1 0 0 0 0 0 0 0 ...  
## $ gear : num 4 4 4 3 3 3 3 4 4 4 ...  
## $ carb : num 4 4 1 1 2 1 4 2 4 4 ...  
## $ qsec\_trans: num 0.00369 0.00345 0.00289 0.00265 0.00345 ...

## Creating the model using the new data

# Fit the model with transformed response variable  
fit4 <- lm(qsec\_trans ~ mpg + cyl + hp + drat + wt , data = new\_data)  
  
# Summary of the fit4  
summary(fit4)

##   
## Call:  
## lm(formula = qsec\_trans ~ mpg + cyl + hp + drat + wt, data = new\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.658e-04 -7.543e-05 -3.244e-05 9.656e-05 4.123e-04   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.055e-03 9.920e-04 1.064 0.298358   
## mpg -1.747e-05 1.771e-05 -0.987 0.334137   
## cyl 1.949e-04 6.592e-05 2.957 0.007074 \*\*   
## hp 6.488e-06 1.308e-06 4.960 5.16e-05 \*\*\*  
## drat 4.645e-04 1.446e-04 3.212 0.003862 \*\*   
## wt -3.999e-04 9.423e-05 -4.244 0.000307 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0002318 on 23 degrees of freedom  
## Multiple R-squared: 0.8791, Adjusted R-squared: 0.8528   
## F-statistic: 33.45 on 5 and 23 DF, p-value: 8.046e-10

#performing the anova of fit4  
anova(fit4)

## Analysis of Variance Table  
##   
## Response: qsec\_trans  
## Df Sum Sq Mean Sq F value Pr(>F)   
## mpg 1 1.8378e-06 1.8378e-06 34.212 5.830e-06 \*\*\*  
## cyl 1 1.4914e-06 1.4914e-06 27.764 2.398e-05 \*\*\*  
## hp 1 3.5986e-06 3.5986e-06 66.990 2.892e-08 \*\*\*  
## drat 1 1.0884e-06 1.0884e-06 20.261 0.0001614 \*\*\*  
## wt 1 9.6750e-07 9.6750e-07 18.010 0.0003066 \*\*\*  
## Residuals 23 1.2355e-06 5.3700e-08   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Comparing the fit3 and fit4

#summary of fit3  
summary(fit3)

##   
## Call:  
## lm(formula = qsec\_trans ~ mpg + cyl + disp + hp + drat + wt,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.992e-04 -1.544e-04 -6.210e-06 1.348e-04 5.807e-04   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.964e-03 1.157e-03 1.697 0.102102   
## mpg -2.436e-05 2.192e-05 -1.111 0.276946   
## cyl 1.685e-04 8.361e-05 2.015 0.054778 .   
## disp 9.209e-07 1.357e-06 0.679 0.503578   
## hp 6.930e-06 1.575e-06 4.401 0.000176 \*\*\*  
## drat 3.521e-04 1.568e-04 2.245 0.033877 \*   
## wt -5.538e-04 1.432e-04 -3.869 0.000694 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0002836 on 25 degrees of freedom  
## Multiple R-squared: 0.8448, Adjusted R-squared: 0.8075   
## F-statistic: 22.68 on 6 and 25 DF, p-value: 5.534e-09

summary(fit3)$r.squared

## [1] 0.8447741

#summary of fit4  
summary(fit4)

##   
## Call:  
## lm(formula = qsec\_trans ~ mpg + cyl + hp + drat + wt, data = new\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.658e-04 -7.543e-05 -3.244e-05 9.656e-05 4.123e-04   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.055e-03 9.920e-04 1.064 0.298358   
## mpg -1.747e-05 1.771e-05 -0.987 0.334137   
## cyl 1.949e-04 6.592e-05 2.957 0.007074 \*\*   
## hp 6.488e-06 1.308e-06 4.960 5.16e-05 \*\*\*  
## drat 4.645e-04 1.446e-04 3.212 0.003862 \*\*   
## wt -3.999e-04 9.423e-05 -4.244 0.000307 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0002318 on 23 degrees of freedom  
## Multiple R-squared: 0.8791, Adjusted R-squared: 0.8528   
## F-statistic: 33.45 on 5 and 23 DF, p-value: 8.046e-10

summary(fit4)$r.squared

## [1] 0.8790989

* By comparing the both fit3 and fit4, fit4 is more significant than the fit 3 with the R-squared: 0.8790989
* We can observe that by removing the leverage points the model significane is increased.

## Performing the f test and t test using anova

anova(fit4)

## Analysis of Variance Table  
##   
## Response: qsec\_trans  
## Df Sum Sq Mean Sq F value Pr(>F)   
## mpg 1 1.8378e-06 1.8378e-06 34.212 5.830e-06 \*\*\*  
## cyl 1 1.4914e-06 1.4914e-06 27.764 2.398e-05 \*\*\*  
## hp 1 3.5986e-06 3.5986e-06 66.990 2.892e-08 \*\*\*  
## drat 1 1.0884e-06 1.0884e-06 20.261 0.0001614 \*\*\*  
## wt 1 9.6750e-07 9.6750e-07 18.010 0.0003066 \*\*\*  
## Residuals 23 1.2355e-06 5.3700e-08   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* From the above anova table we can see that all the variables are significant.
* so fit4 is our final fit we are gonna do check for multicollinearity using vif function.

## Checking multicollinearity using VIF function

* We will check the VIF values to check the multicollinearity.

#install required packages  
#install.packages("car")  
  
#Setting the Library  
library(car)

## Loading required package: carData

#Using the Function VIF  
vif(fit4)

## mpg cyl hp drat wt   
## 5.879606 7.163337 4.231466 2.679013 4.422203

* All VIF values are less than 10, showing that predictors are not highly correlated. However, to increase the model’s significance, we will delete the predictor with the greatest VIF value.

## Removing the mpg

#creating new\_fit1 by removing mpg  
new\_fit <- lm(qsec\_trans ~ cyl + hp + drat + wt , data = new\_data)  
  
# Checking the summary of new\_fit  
summary(new\_fit)

##   
## Call:  
## lm(formula = qsec\_trans ~ cyl + hp + drat + wt, data = new\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.690e-04 -1.028e-04 -1.175e-05 7.956e-05 4.248e-04   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.174e-04 7.517e-04 0.555 0.583838   
## cyl 2.111e-04 6.380e-05 3.309 0.002948 \*\*   
## hp 6.795e-06 1.270e-06 5.350 1.72e-05 \*\*\*  
## drat 4.570e-04 1.443e-04 3.166 0.004167 \*\*   
## wt -3.475e-04 7.780e-05 -4.467 0.000161 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0002316 on 24 degrees of freedom  
## Multiple R-squared: 0.874, Adjusted R-squared: 0.853   
## F-statistic: 41.61 on 4 and 24 DF, p-value: 1.842e-10

# Performing anova on new\_fit  
anova(new\_fit)

## Analysis of Variance Table  
##   
## Response: qsec\_trans  
## Df Sum Sq Mean Sq F value Pr(>F)   
## cyl 1 3.2232e-06 3.2232e-06 60.069 5.505e-08 \*\*\*  
## hp 1 3.2391e-06 3.2391e-06 60.366 5.273e-08 \*\*\*  
## drat 1 1.3985e-06 1.3985e-06 26.064 3.184e-05 \*\*\*  
## wt 1 1.0706e-06 1.0706e-06 19.951 0.0001612 \*\*\*  
## Residuals 24 1.2878e-06 5.3700e-08   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## We removed the mpg because by removing the cyl the significance of the model is decreasing.

## We decided to remove the mpg instead of cyl to improve the model efficiency.

## Checking the Vif for the new\_fit

#performing vif on new\_fit  
vif(new\_fit)

## cyl hp drat wt   
## 6.718145 3.992542 2.671449 3.017913

* The VIF values are less than 10 These values show low multicollinearity between these predictors, as they are significantly lower than the usually recognized requirements of 10. As a result,the predictors ‘hp’ and ‘wt’ ,‘drat’ ,‘cyl’ are generally independent of one another in the regression model.

## Final fit

* After all the analysis and consideration the final and best fit is the new fit.
* And I am gonna rewrite the new fit as my best fit.

#creating best\_fit  
best\_fit <- lm(qsec\_trans ~ cyl + hp + drat + wt , data = new\_data)  
  
# Checking the summary of best\_fit  
summary(best\_fit)

##   
## Call:  
## lm(formula = qsec\_trans ~ cyl + hp + drat + wt, data = new\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.690e-04 -1.028e-04 -1.175e-05 7.956e-05 4.248e-04   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.174e-04 7.517e-04 0.555 0.583838   
## cyl 2.111e-04 6.380e-05 3.309 0.002948 \*\*   
## hp 6.795e-06 1.270e-06 5.350 1.72e-05 \*\*\*  
## drat 4.570e-04 1.443e-04 3.166 0.004167 \*\*   
## wt -3.475e-04 7.780e-05 -4.467 0.000161 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0002316 on 24 degrees of freedom  
## Multiple R-squared: 0.874, Adjusted R-squared: 0.853   
## F-statistic: 41.61 on 4 and 24 DF, p-value: 1.842e-10

# Getting R squared value of best fit  
summary(best\_fit)$r.squared

## [1] 0.8739829

# Performing anova on best\_fit  
anova(best\_fit)

## Analysis of Variance Table  
##   
## Response: qsec\_trans  
## Df Sum Sq Mean Sq F value Pr(>F)   
## cyl 1 3.2232e-06 3.2232e-06 60.069 5.505e-08 \*\*\*  
## hp 1 3.2391e-06 3.2391e-06 60.366 5.273e-08 \*\*\*  
## drat 1 1.3985e-06 1.3985e-06 26.064 3.184e-05 \*\*\*  
## wt 1 1.0706e-06 1.0706e-06 19.951 0.0001612 \*\*\*  
## Residuals 24 1.2878e-06 5.3700e-08   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Getting confidence intervals and prediction intervals of the best fit.

# Getting confidence intervals of best fit  
conf\_intervals <- predict(best\_fit, newdata = new\_data, interval = "confidence")  
conf\_intervals

## fit lwr upr  
## Mazda RX4 0.003303208 0.003135670 0.003470746  
## Mazda RX4 Wag 0.003214591 0.003051752 0.003377429  
## Datsun 710 0.002846870 0.002682267 0.003011474  
## Hornet 4 Drive 0.002721721 0.002559353 0.002884089  
## Hornet Sportabout 0.003539423 0.003375080 0.003703766  
## Valiant 0.002456375 0.002225758 0.002686991  
## Duster 360 0.003997315 0.003810836 0.004183795  
## Merc 240D 0.002260769 0.002055411 0.002466127  
## Merc 280 0.003115718 0.002945244 0.003286191  
## Merc 280C 0.003115718 0.002945244 0.003286191  
## Merc 450SE 0.003317904 0.003182996 0.003452813  
## Merc 450SL 0.003436061 0.003291952 0.003580169  
## Merc 450SLC 0.003418685 0.003277207 0.003560162  
## Cadillac Fleetwood 0.003013733 0.002790033 0.003237434  
## Lincoln Continental 0.003053203 0.002806454 0.003299951  
## Chrysler Imperial 0.003287685 0.003040156 0.003535213  
## Fiat 128 0.002810209 0.002655113 0.002965306  
## Toyota Corolla 0.002994234 0.002819607 0.003168861  
## Toyota Corona 0.002755115 0.002567481 0.002942749  
## Dodge Challenger 0.003163529 0.002949286 0.003377771  
## AMC Javelin 0.003371285 0.003176455 0.003566116  
## Camaro Z28 0.004141109 0.003959713 0.004322504  
## Pontiac Firebird 0.003366691 0.003224423 0.003508958  
## Fiat X1-9 0.002902302 0.002742087 0.003062516  
## Porsche 914-2 0.003160875 0.002973685 0.003348064  
## Lotus Europa 0.003226661 0.002970046 0.003483276  
## Ford Pantera L 0.004726965 0.004443311 0.005010619  
## Maserati Bora 0.004759667 0.004413034 0.005106300  
## Volvo 142E 0.002914544 0.002729036 0.003100052

# Getting prediction intervals of best fit  
predict\_intervals <- predict(best\_fit, newdata = new\_data, interval = "prediction")  
predict\_intervals

## fit lwr upr  
## Mazda RX4 0.003303208 0.002796616 0.003809800  
## Mazda RX4 Wag 0.003214591 0.002709533 0.003719648  
## Datsun 710 0.002846870 0.002341241 0.003352499  
## Hornet 4 Drive 0.002721721 0.002216816 0.003226627  
## Hornet Sportabout 0.003539423 0.003033879 0.004044967  
## Valiant 0.002456375 0.001925573 0.002987176  
## Duster 360 0.003997315 0.003484148 0.004510483  
## Merc 240D 0.002260769 0.001740444 0.002781094  
## Merc 280 0.003115718 0.002608147 0.003623288  
## Merc 280C 0.003115718 0.002608147 0.003623288  
## Merc 450SE 0.003317904 0.002821148 0.003814660  
## Merc 450SL 0.003436061 0.002936727 0.003935394  
## Merc 450SLC 0.003418685 0.002920104 0.003917265  
## Cadillac Fleetwood 0.003013733 0.002485900 0.003541567  
## Lincoln Continental 0.003053203 0.002515196 0.003591210  
## Chrysler Imperial 0.003287685 0.002749320 0.003826050  
## Fiat 128 0.002810209 0.002307595 0.003312824  
## Toyota Corolla 0.002994234 0.002485254 0.003503214  
## Toyota Corona 0.002755115 0.002241527 0.003268704  
## Dodge Challenger 0.003163529 0.002639633 0.003687424  
## AMC Javelin 0.003371285 0.002855024 0.003887546  
## Camaro Z28 0.004141109 0.003629767 0.004652451  
## Pontiac Firebird 0.003366691 0.002867885 0.003865496  
## Fiat X1-9 0.002902302 0.002398084 0.003406519  
## Porsche 914-2 0.003160875 0.002647449 0.003674301  
## Lotus Europa 0.003226661 0.002684058 0.003769263  
## Ford Pantera L 0.004726965 0.004171064 0.005282866  
## Maserati Bora 0.004759667 0.004169141 0.005350193  
## Volvo 142E 0.002914544 0.002401729 0.003427359

* So i conclude that best\_fit is the final model with R-squared: 0.853 and p-value: 1.842e-10
* The model with variables ‘hp’ and ‘wt’ ‘cyl’ ‘drat’ has statistically significant coefficients (p < 0.05) and accounts for about 85% of the variability in ‘qsec’ (Adjusted R-squared: 0.853 ). Despite being simpler than prior models, it gives a reasonable match to the data.
* The model, which only uses ‘hp’ and ‘wt’ ‘cyl’ ‘drat’ as predictors, properly describes the variance in ‘qsec’. It has statistically significant coefficients and an acceptable level of explanatory power. This simplified model provides a realistic and successful technique to forecasting ‘qsec’

# CONCLUSION

* The ‘mtcars’ dataset undergone a first exploratory data analysis (EDA).
* Linear regression models were built and tested with different combinations of predictor variables.
* ANOVA tests were used to examine the significance of the various models.
* The model ‘best\_fit’ emerged as the best fit, with ‘cyl ’hp’ ‘drat’ and ‘wt’ as predictor variables.
* ‘best\_fit’ has a strong fit with significant predictor values, making it the most efficient model for predicting ‘qsec’.
* The model ‘best\_fit’ is selected as the best fit for forecasting ‘qsec’. It uses ‘cyl’,‘drat’,‘hp’ (horsepower) and ‘wt’ (vehicle weight) as predictions. This model is highly significant (p < 0.001) and has a high coefficient of determination (R-squared = 0.853), indicating it well explains the variation in ‘qsec’. ‘best\_fit’ efficiently captures the link between predictors and ‘qsec’, making it the best option for predictive modeling.

## Total code

#loading data  
data <- mtcars  
#getting the top 6 data  
head(data)  
#checking the structure of the data  
str(data)  
#checking for na values  
sum(is.na(data))  
#getting summary  
summary(data)  
# Load required libraries  
library(ggplot2)  
# Scatter plots of all the variables comparing to qsec  
# Scatter plot of qsec vs. mpg  
# ggplot(data, aes(x = mpg, y = qsec)) +  
# geom\_point() +  
# ggtitle("Scatter Plot of qsec vs. mpg") +  
# xlab("Miles Per Gallon (mpg)") +  
# ylab("Quarter Mile Time (qsec)")  
  
# Scatter plot of qsec vs. cyl with linear regression line  
# ggplot(data, aes(x = cyl, y = qsec)) +  
# geom\_point(color = "blue", size = 3) +  
# geom\_smooth(method = "lm", se = FALSE, color = "red") +  
# labs(title = "Qsec vs. Cylinders", x = "Cylinders", y = "Qsec")  
  
# Scatter plot of qsec vs. hp with linear regression line  
# ggplot(data, aes(x = hp, y = qsec)) +  
# geom\_point(color = "green", size = 3, shape = 17) +  
# geom\_smooth(method = "lm", se = FALSE, color = "purple") +  
# labs(title = "Qsec vs. Horsepower", x = "Horsepower", y = "Qsec")  
  
# Scatter plot of qsec vs. wt with linear regression line  
# ggplot(data, aes(x = wt, y = qsec)) +  
# geom\_point(color = "orange", size = 3, shape = 18) +  
# geom\_smooth(method = "lm", se = FALSE, color = "brown") +  
# labs(title = "Qsec vs. Car Weight", x = "Car Weight", y = "Qsec")  
  
#Fitting the model  
fit <- lm(qsec ~ mpg + cyl + disp + hp + drat + wt + mpg:cyl + mpg:disp + mpg:hp + mpg:drat + mpg:wt + cyl:disp + cyl:hp + cyl:drat + cyl:wt + disp:hp + disp:drat + disp:wt + hp:drat + hp:wt + drat:wt, data = data)  
#summary of fit  
summary(fit)  
summary(fit)$r.squared  
#anova of fit  
anova(fit)  
#fitting the model1  
fit1 <- lm(qsec ~ mpg + cyl + disp + hp + drat + wt + I(mpg^2)+ I(cyl^2)+ I(disp^2)+ I(hp^2) + I(drat^2) + I(wt^2), data = data)  
#summary of fit1  
summary(fit1)  
summary(fit1)$r.squared  
#anova of fit1  
anova(fit1)  
# Comparing Anova of fit and fit1  
anova(fit,fit1)  
#fitting the final model as final\_fit  
fit2 <- lm(qsec ~ mpg + cyl + disp + hp + drat + wt + cyl:wt , data = data)  
#summary  
summary(fit2)  
summary(fit2)$r.squared  
#anova of fit2  
anova(fit2)  
# Load the required library  
library(MASS)  
# Box-Cox transformation  
#b <- boxcox(fit2)  
lambda <- b$x  
likelihood <- b$y  
best\_lambda <- lambda[which.max(likelihood)]  
best\_lambda  
# Transform the response variable using the best lambda  
data$qsec\_trans <- (data$qsec^best\_lambda)  
# Fit the model with transformed response variable  
fit2\_trans <- lm(qsec\_trans ~ mpg + cyl + disp + hp + drat + wt + cyl:wt, data = data)  
#summary of fit2  
summary(fit2)$r.squared  
#summary of transformedc fit2  
summary(fit2\_trans)$r.squared  
# Fit the model with transformed response variable  
fit3 <- lm(qsec\_trans ~ mpg + cyl + disp + hp + drat + wt , data = data)  
# Summary of the fit3  
summary(fit3)  
#performing the anova of fit3  
anova(fit3)  
#plotting the fit3  
#plot(fit3)  
#removing the points  
new\_data<-data[-c(9,19,30),]  
head(new\_data)  
str(new\_data)  
# Fit the model with transformed response variable  
fit4 <- lm(qsec\_trans ~ mpg + cyl + hp + drat + wt , data = new\_data)  
# Summary of the fit4  
summary(fit4)  
#performing the anova of fit4  
anova(fit4)  
#summary of fit3  
summary(fit3)  
summary(fit3)$r.squared  
#summary of fit4  
summary(fit4)  
summary(fit4)$r.squared  
anova(fit4)  
#install required packages  
#install.packages("car")  
#Setting the Library  
library(car)  
#Using the Function VIF  
vif(fit4)  
#creating new\_fit1 by removing mpg  
new\_fit <- lm(qsec\_trans ~ cyl + hp + drat + wt , data = new\_data)  
# Checking the summary of new\_fit  
summary(new\_fit)  
# Performing anova on new\_fit  
anova(new\_fit)  
#performing vif on new\_fit  
vif(new\_fit)

#creating best fit after all the analysis  
best\_fit <- lm(qsec\_trans ~ cyl + hp + drat + wt , data = new\_data)  
# Checking the summary of new\_fit  
summary(best\_fit)

# Getting R squared value of best fit  
summary(best\_fit)$r.squared  
# Performing anova on new\_fit  
anova(best\_fit)

# Getting confidence intervals of best fit  
conf\_intervals <- predict(best\_fit, newdata = new\_data, interval = "confidence")  
conf\_intervals  
# Getting prediction intervals of best fit  
predict\_intervals <- predict(best\_fit, newdata = new\_data, interval = "prediction")  
predict\_intervals