Regression Analysis Peer-reviewed Project

Nirmal Ghimire

11/10/2020

Summary of the Research

I am a data scientist at MTcars and currently I am working on a piece for a magazine called Motor Trend about the automobile industry. This project looks at a data set of a collection of cars as my team is interested in exploring the relationship between a set of variables and how they impact fuel efficiency using the miles they travel per one gallon of traditional fuel (MPG). This research briefly progresses through following steps:

- Exploratory Analysis
- Regression Analysis and Model Fit
- Results and Assumption Checks
- Conclusion

This research will attemp to answer following research questions:

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions.

Data and Methods

This study will be conducted using the Motor Trend Car Road Tests (mtcars) data set. These data were extracted from the 1974 Motor Trend Magazine. The dataframe comprises fuel consumption by thirty-two automobile types based on ten different aspects.

Here are further information about the variables:included in the dataset:

- **mpg**: Miles/US Gallon
- **cyl**: Number fo Cylinder
- **disp**: Displacement/cubic inch
- **hp**: Gross Horsepower
- **drat**: Rear Axle Ratio
- **wt**: Weight (in 1000 Lbs.)
- **qsec**: 1/4 Mile Time
- **vs**: Engine (0=V-shaped, 1=Straight)
- **am**: Transmission Type (0 = Automatic, 1 = Manual)
- **gear**: Number of Forward Gears
- **carb**: Number of Carburetors

1. Exploratory Analysis

Let’s begin by invoking the library and selecting the dataframe

```r
library(datasets)
data(mtcars)
summary(mtcars)
```

```
##       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
```

Based on the summary, we have following variables with following specifications:

```
## Warning: package ’knitr’ was built under R version 4.0.3
```

Table 1: Table Showing Automobile Types and Specifications

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4

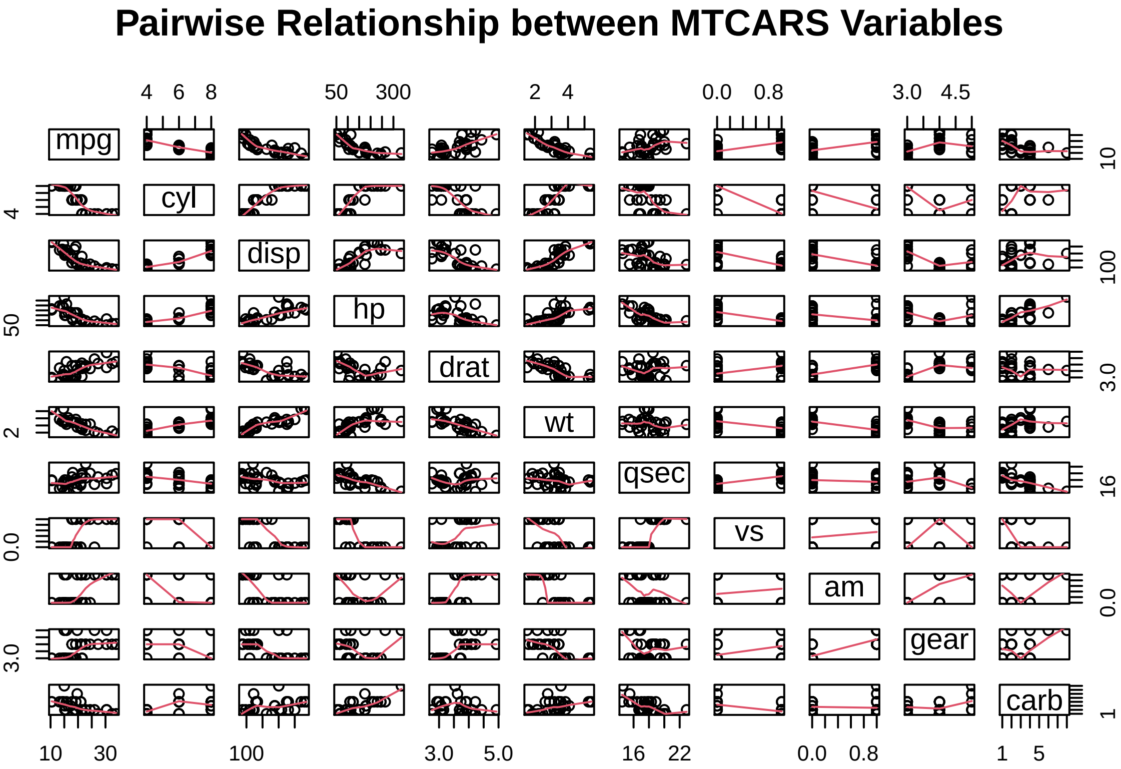
	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

We can clearly see that the vehicles have either 4, or 6, or 8 cylinders. Based on the research questions, we need to pay attention to *cyl* variable to identify how they differ in fule efficiency by cylinder types. In addition, we are also interested in the fuel efficiency based on automobiles' transmission types. The vehicles are identified to have either automatic (coded 0) or manual (coded 1) transmissions.

Along the way we will also see how the impact of these variables change when we include other variables in the model for example, the engine types, weight, gross horsepower, number of forward gears, or number of carburetors. The ***vs*** is a bionomial variable, ***gear*** and ***carb*** are categorical variables suggesting number of forward gears and number of carburetors, respectively. Similarly, ***wt*** (the weight of vehicles), ***hp*** (gross horsepower)and ***mpg*** miles per US gallon are continuous variables.

Let's check the correlation between the variables included in the dataset:

```r
pairs(mtcars, panel=panel.smooth, main="Pairwise Relationship between MTCARS Variables")
```



The graph shows that most of the variables do have some sort of linear relationship. Looks like the automobiles with higher number of cylinders give lower miles per gallon. Similarly, cars with automatic transmissions have lower mpg compared to cars with manual transmissions. Likewise, horsepower, weight, displacement, and number of carburetors seem to have lower fuel efficiency when the values increase. Conversely, the rear axle ratio, 1/4 mile time, and engine type seem to have positive linear relationships with mpg.

Let's check if they really behave the way they appear to be

2. Regression Analysis and Model Fit

There are some categorical variables in the dataset. We need to change them into factor variables before we are able to do conduct a regression analysis.

```
transmission<-as.factor(mtcars$am)
levels(transmission)<-c("Automatic", "Manual")
cylinder<-as.factor(mtcars$cyl)
levels(cylinder)<-c("4cyl", "6cyl","8cyl")
engine<-as.factor(mtcars$vs)
levels(engine)<-c("V-Shaped", "Straight")
GEAR<-as.factor(mtcars$gear)
CARB<-as.factor(mtcars$carb)
```

Research Question 1: "Is an automatic or manual transmission better for MPG"

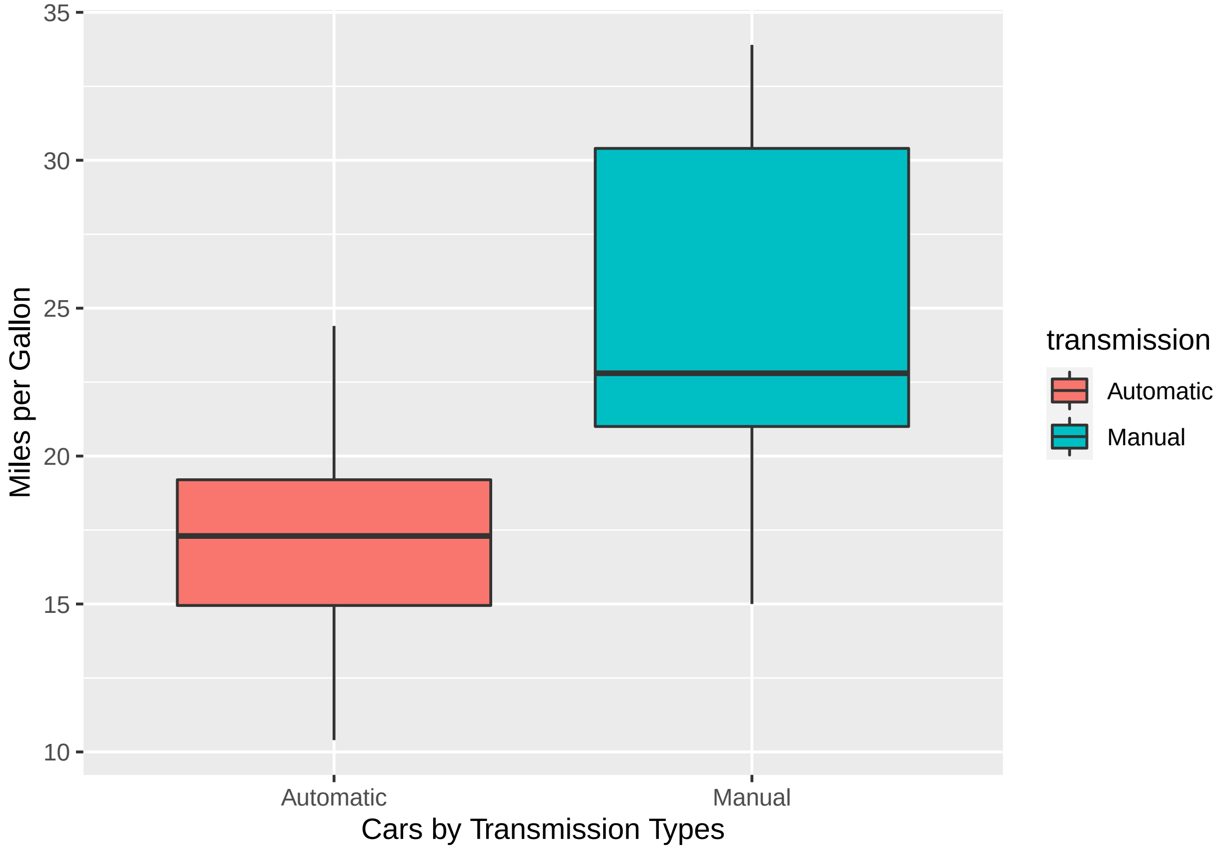
```
fit<-lm(mpg~am, data=mtcars)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

As can be seen in the results above, the average miles per gallon are 17.147 across all cars and the model is statistically significant. Similarly, compared to the cars with the automatic transmission, the cars with manual transmission had better mpg per gallon. In other words, per one gallon traditional fuel is linked to 7.245 more miles in the cars with manual transmission, and this increase was statistically significantly higher than zero.

Following box plot strengthen the above finding.

```
library(ggplot2)
g<-ggplot(mtcars, aes(x=transmission,y=mpg))
g=g+xlab("Cars by Transmission Types")
g=g+ylab("Miles per Gallon")
g=g+geom_boxplot(aes(fill=transmission))
print(g)
```

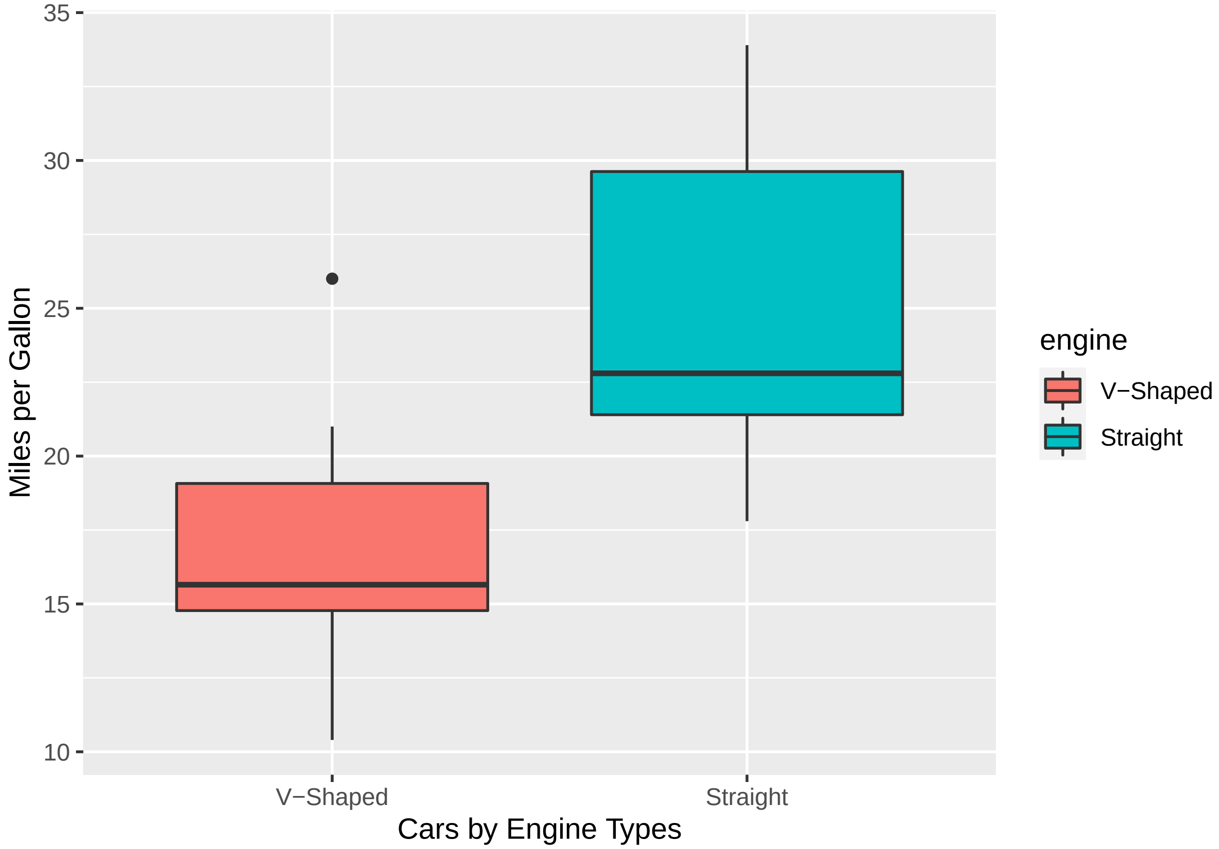


Clearly, automatic automobiles give approximately 17 miles per gallon, while the manual automobiles have slightly more than 22 miles per gallon. ***If you want to ride a car and save some dimes at the same time, you gotta walk away with manual transmission***.

```
library(ggplot2)
fit1<-lm(mpg~vs, data=mtcars)
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ vs, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.757 -3.082 -1.267  2.828  9.383
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   16.617      1.080  15.390 8.85e-16 ***
## vs             7.940      1.632   4.864 3.42e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.581 on 30 degrees of freedom
## Multiple R-squared:  0.4409, Adjusted R-squared:  0.4223
## F-statistic: 23.66 on 1 and 30 DF,  p-value: 3.416e-05
```

```r
g1<-ggplot(mtcars, aes(x=engine,y=mpg))
g1=g1+xlab("Cars by Engine Types")
g1=g1+ylab("Miles per Gallon")
g1=g1+geom_boxplot(aes(fill=engine))
print(g1)
```



The the results show that average miles per gallon across the cars regardless of their engine types was 16.617 miles. Compared to the cars with v-shaped engine, straight engine cars have statistically significant higher fuel efficiency, i.e., 7.94 miles per gallon. This result has been strengthened by the box plot. We can clearly see that straight engine automobiles have slightly more than 22 miles per gallon rate, while the v-shaped cars give little more than 15 miles per gallon.

Results and Assumption Checks

Now, lets put transmission and engine types as the predictors of the mpg,and see how the aforementioned statistics change.

```r
fit2<-lm(mpg~am+vs, data=mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ am + vs, data = mtcars)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.1905 -2.5988  0.2222  2.7315  6.3095 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  14.5944     0.9262  15.758 9.35e-16 ***
## am            6.0667     1.2748   4.759 4.96e-05 ***
## vs            6.9294     1.2621   5.490 6.50e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.491 on 29 degrees of freedom
## Multiple R-squared:  0.6861, Adjusted R-squared:  0.6644 
## F-statistic: 31.69 on 2 and 29 DF,  p-value: 5.056e-08
```

The results shows that this model is statistically significantly. Comparing R-squared among these models we can see the model with both predictors has better value, i.e., 0.6861. In addition, the average miles per gallon regardless of car types has been slightly dropped to 14.59. Both transmission types and engine types are statistically significant predictors of fuel efficiency in a car.

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.0.3
```

```
car::vif(fit2)
```

```
##       am       vs 
## 1.029167 1.029167
```

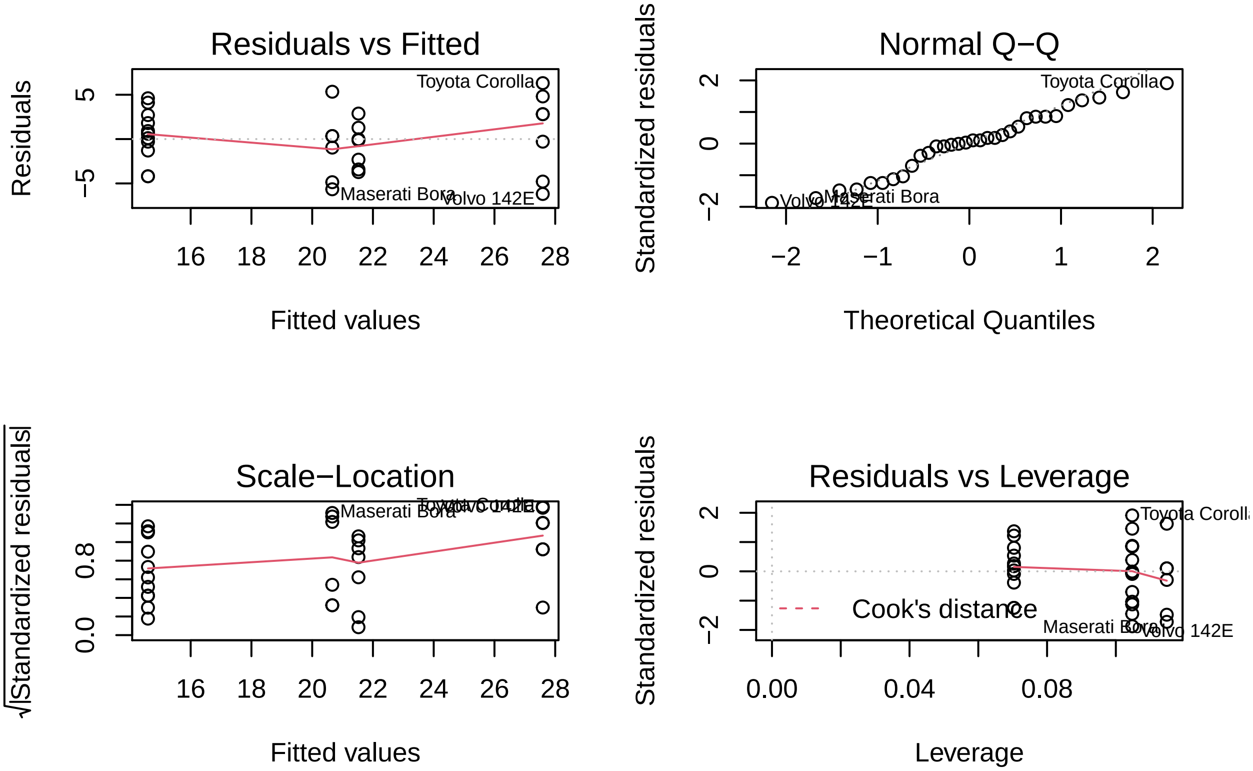
Based on the variance inflation factors which suggests the increase in the variance for the second variance in our final model. In reality we don't see any rate of inflation after we include vs in the model. So, it's okay, in fact better that we put them in the model.

```
anova(fit, fit1, fit2)
```

```
## Analysis of Variance Table
## 
## Model 1: mpg ~ am
## Model 2: mpg ~ vs
## Model 3: mpg ~ am + vs
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
## 1     30 720.90                                  
## 2     30 629.52  0    91.377                     
## 3     29 353.49  1   276.033 22.646 4.958e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above decision that the two predictor model was better than the one predictor model only, has been confirmed by further ANOVA analysis. Based on the results, the two predictors model is statistically significant, showing it is a better fitting model.

```
par(mfrow=c(2,2))
plot(fit2)
```



Looking at the upper left hand plot, i.e., Residuals vs. Fitted value, the horizontal reference line at 0 both rises and drops below 0 suggesting the sum around 0. Like in residual plots, the Normal Q-Q plot shows roughly 45 degree angle which suggests normality.

Conclusion

1. The model with two predictors was better fitting model than the one predictor models.
2. The rate of Fuel efficiency in an automobile is directly related to their engine and transmission types.
3. Cars with manual transmission are more fuel efficient compared to the cars with automatic transmission. And
4. Cars with straight engine types are more economical in terms of fuel consumption compared to the cars with v-shaped engines.

Thanks