ALY 6015: INTERMEDIATE ANALYTICS

Assignment 4: Feature Selection in R

Submitted to

Prof. Fatemeh Ahmadi Abkenari

Submitted by

Abhilash Dikshit

Mrityunjay Gupta

Siddharth Alashi

Smit Parmar

Assignment 4: Feature Selection in R

Abhilash Dikshit, Siddharth Alashi ,Mrityunjay Gupta, Smit Parmar
College of Professional Studies
Northeastern University
Vancouver, Canada

I. Abstract:

A built-in dataset in R called "mtcars" provides measurements for 32 distinct cars over 11 different attributes. In this paper we will be summarizing methods to optimize model using feature selection techniques. Forwards selection techniques and Both direction regression method helps us to select the best regression model on this dataset. The 'mtcars' dataset is split into Train and Test dataset with the ratio of 70/30. Furthermore, we have made visualizations and descriptive analysis which describes the comparison between the variables. The importance of the models functioning best amongst them is then summarized by ANOVA TEST. Finally, the references provides a support for our arguments, ideas, and opinions.

II. Introduction

The data was extracted from the 1974 Motor Trend US magazine and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models)

This dataset shall consist of 11 columns and 32 observations which are labelled below.

mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	1/4 mile time
vs	Engine (0 = V-shaped, 1 = straight)
am	Transmission (0 = automatic, 1 = manual)

gear	Number of forward gears
carb	Number of carburetors

It is frequently required and beneficial to divide the data set into training and testing sets. The model will be trained on the training set of data, and the test set will be used to evaluate the model. This makes sure that we are not overfitting the model and that it functions properly with new data. It is extremely usual to utilise a 70/30 split, where 70% of the observations are used for the training set and 30% are used for testing. Fig1. and Fig2. describes the structure of Train and Test dataset

```
> str(sample_train)
'data.frame':   22 obs. of  11 variables:
 $ mpg : num  21 21.4 18.7 18.1 22.8 17.8 16.4 17.3 15.2 14.
 $ cyl : num  6 6 8 6 4 6 8 8 8 8 ...
 $ disp: num  160 258 360 225 141 ...
 $ hp  : num  110 110 175 105 95 123 180 180 180 230 ...
 $ drat: num  3.9 3.08 3.15 2.76 3.92 3.92 3.07 3.07 3.07 3.
 $ wt  : num  2.88 3.21 3.44 3.46 3.15 ...
 $ qsec: num  17 19.4 17 20.2 22.9 ...
 $ vs  : num  0 1 0 1 1 1 0 0 0 0 ...
 $ am  : num  1 0 0 0 0 0 0 0 0 0 ...
 $ gear: num  4 3 3 3 4 4 3 3 3 3 ...
 $ carb: num  4 1 2 1 2 4 3 3 3 4 ...
```

Fig1. Structure of Train dataset

```
> str(sample_test)
'data.frame':   10 obs. of  11 variables:
 $ mpg : num  21 22.8 14.3 24.4 19.2 10.4 10.4 32.4 33.9 15
 $ cyl : num  6 4 8 4 6 8 8 4 4 8
 $ disp: num  160 108 360 147 168 ...
 $ hp  : num  110 93 245 62 123 205 215 66 65 335
 $ drat: num  3.9 3.85 3.21 3.69 3.92 2.93 3 4.08 4.22 3.54
 $ wt  : num  2.62 2.32 3.57 3.19 3.44 ...
 $ qsec: num  16.5 18.6 15.8 20 18.3 ...
 $ vs  : num  0 1 0 1 1 0 0 1 1 0
 $ am  : num  1 1 0 0 0 0 0 1 1 1
 $ gear: num  4 4 3 4 4 3 3 4 4 5
 $ carb: num  4 1 4 2 4 4 4 1 1 8
```

Fig2. Structure of Test dataset

III. Descriptive Analysis

The summary in Fig3. Concludes the introduction of the statistics to 9 variables in the dataset.

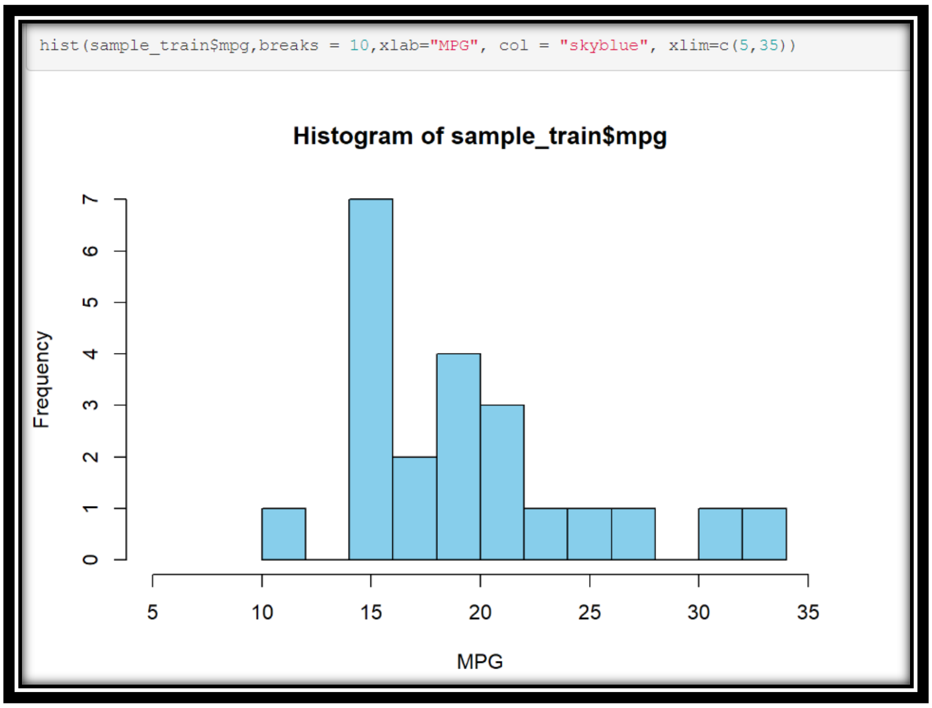
1. we find that the average miles per gallon for 22 cars is 19.96mpg. 25% of cars have 15.96 mpg as their average and 75% of the cars have 21.48mpg as their average.
2. The average number of cylinders among 22 cars is 6 cylinder but the maximum number of cylinders in the cars are 8 cylinder cars.
3. The average horse-power (hp) of the cars is 144.3 hp, whereas the minimum horsepower of the car is only 52hp and the maximum horse-power is 264hp.
4. Average weight (wt) of cars is 3.16 tons. the heaviest cars are of 5.34 tons.

```
> summary(sample_train)
      mpg              cyl             disp      
 Min.   :13.30   Min.   :4.000   Min.   : 75.7  
 1st Qu.:15.95   1st Qu.:4.000   1st Qu.:126.0  
 Median :18.95   Median :6.000   Median :241.5  
 Mean   :19.96   Mean   :6.273   Mean   :229.9  
 3rd Qu.:21.48   3rd Qu.:8.000   3rd Qu.:314.5  
 Max.   :30.40   Max.   :8.000   Max.   :440.0  
       hp             drat            wt       
 Min.   : 52.0   Min.   :2.76   Min.   :1.513  
 1st Qu.:106.0   1st Qu.:3.08   1st Qu.:2.772  
 Median :136.5   Median :3.66   Median :3.325  
 Mean   :144.3   Mean   :3.58   Mean   :3.161  
 3rd Qu.:178.8   3rd Qu.:3.92   3rd Qu.:3.678  
 Max.   :264.0   Max.   :4.93   Max.   :5.345  
      qsec             vs               am        
 Min.   :14.50   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:16.93   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :17.41   Median :0.0000   Median :0.0000  
 Mean   :17.83   Mean   :0.4091   Mean   :0.3636  
 3rd Qu.:18.82   3rd Qu.:1.0000   3rd Qu.:1.0000  
```

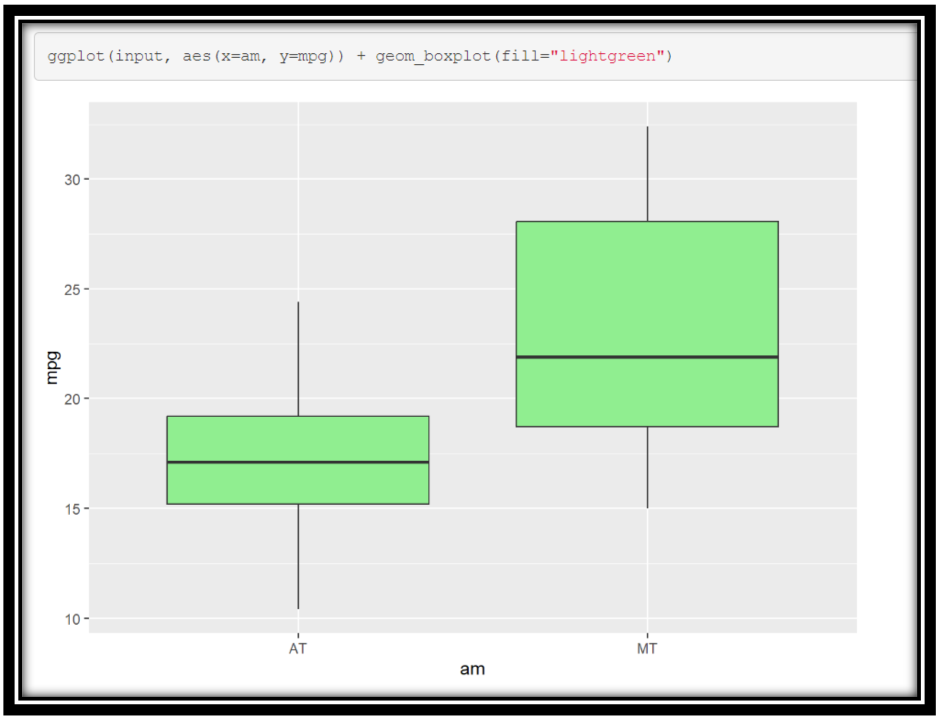
Fig3. Summary of 'mtcars' dataset

III. Exploratory Analysis

1. The distribution of the outcome variable (mpg) is plotted using a histogram which suggests a resemblance with normal distribution. Furthermore, the maximum number of cars has 15 mpg as their average, almost 45% of the cars have an average more than the Mean calculated.

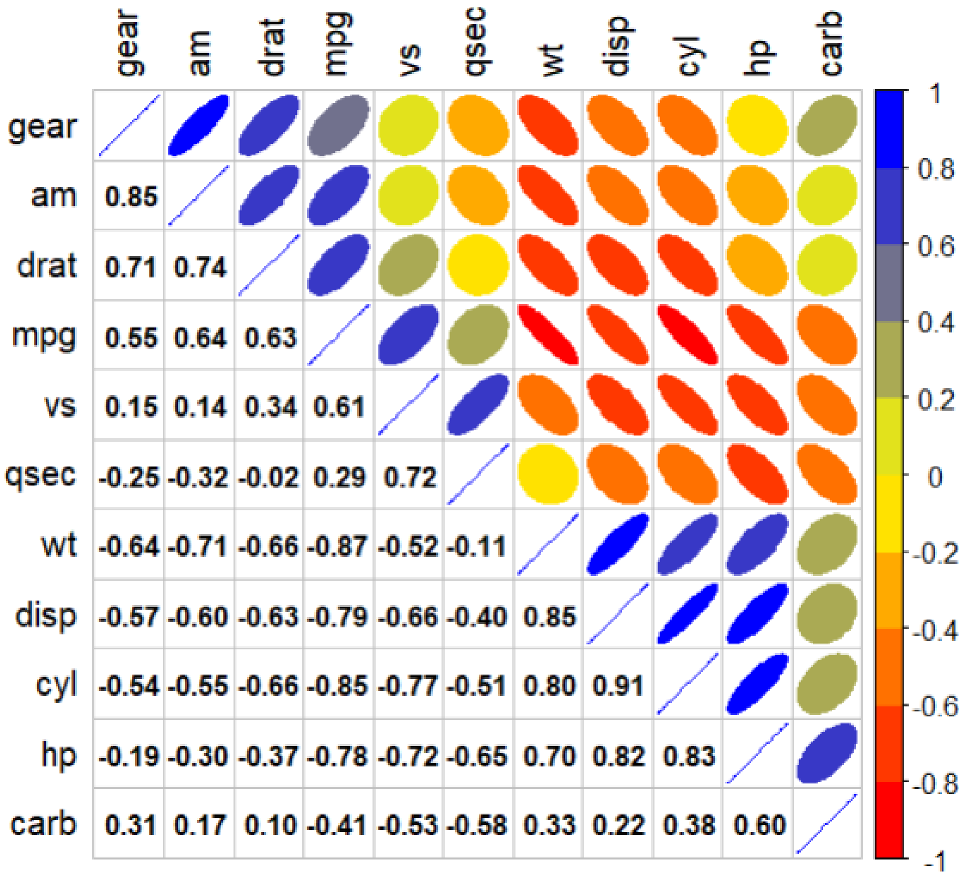


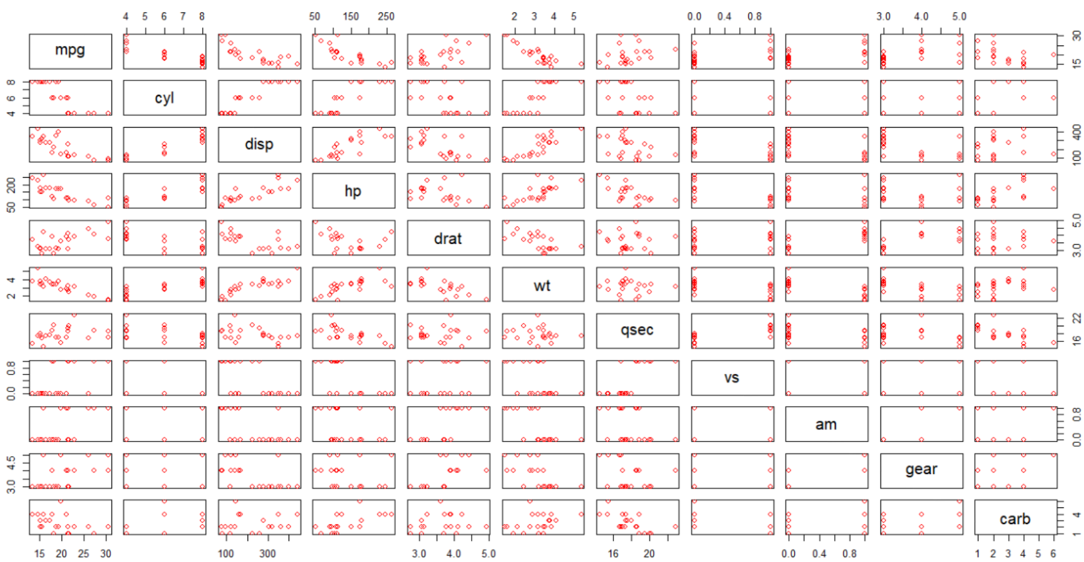
2. A boxplot of the outcome variable (mpg) is plotted with (am). It suggests manual transmission is better for mpg as compared to an automatic transmission.



3. To check the co-linearity between the variables a **Pair plot** is plotted. The Pair plot shows a strong relationship between different variables and miles per gallon. we can conclude from the Fig6. and Fig7.

- Gear has strong positive linear relationship between Transmission , real axel ratio and Negative weakly linear relationship with weight, displacement, Cylinder, and horsepower.
- Transmission has weak positive linear relationship with carburetors. Also, it has weak negative linear relationship with qsec, horsepower(hp), cylinder(cyl).
- Miles per gallon (mpg) has strong positive linear co-relationship with Engine(vs). Whereas, it has Strong Negative Linear relationship with weight(wt), displacement(disp), cylinder(cyl), horsepower(hp).





Feature Selection Method

By adding and removing predictors from the model progressively until there is no longer a statistically legitimate reason to add or remove any more, stepwise regression is a technique we may use to create a regression model from a set of predictor variables.

With this model we have used Stepwise regression as

- Forward Selection
- Both-Direction Selection.

Forward selection

The first method is the forward selection method. In this case, we start with no predictors and then add the predictor with the highest correlation with the response variable.

By including the variable, we ensure that the model has actually improved.

If it has, repeat the process. When there are no more improvements that can be made by adding variables to the model, the process will end. By setting the 'direction' parameter to "forward," we select the step() function's forward selection method.

```
step(lm(mpg ~ 1, data = mtcars), direction = 'forward', scope = ~ disp + hp + drat + wt + qsec)
```

```
## Start:  AIC=115.94
## mpg ~ 1
##
##        Df Sum of Sq     RSS     AIC
## + wt    1    847.73  278.32  73.217
## + disp  1    808.89  317.16  77.397
## + hp    1    678.37  447.67  88.427
## + drat  1    522.48  603.57  97.988
## + qsec  1    197.39  928.66 111.776
## <none>              1126.05 115.943
##
## Step:  AIC=73.22
## mpg ~ wt
##
##        Df Sum of Sq    RSS    AIC
## + hp    1    83.274 195.05 63.840
## + qsec  1    82.858 195.46 63.908
## + disp  1    31.639 246.68 71.356
## <none>              278.32 73.217
## + drat  1     9.081 269.24 74.156
##
## Step:  AIC=63.84
## mpg ~ wt + hp
##
##        Df Sum of Sq    RSS    AIC
## <none>              195.05 63.840
## + drat  1   11.3659 183.68 63.919
## + qsec  1    8.9885 186.06 64.331
## + disp  1    0.0571 194.99 65.831
```

```
##
## Call:
## lm(formula = mpg ~ wt + hp, data = mtcars)
##
## Coefficients:
## (Intercept)           wt           hp
##    37.22727     -3.87783     -0.03177
```

Here we can see that after applying the Forward Selection method we found the best model

mpg ~ wt + hp

Here the above conclusion comes up based on their AIC values which are minimum(minimum AIC gives the best model) for the above equation.

When we performed Linear regression of the above equation with a given data set our output is justified

Linear regression of Forward Selection Methods Result

```
model_forward <- lm(formula = mpg ~ wt + hp, data = mtcars)
summary(model_forward)
```

```
##
## Call:
## lm(formula = mpg ~ wt + hp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.941 -1.600 -0.182  1.050  5.854
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 37.22727    1.59879  23.285 < 0.0000000000000002 ***
## wt          -3.87783    0.63273  -6.129           0.00000112 ***
## hp          -0.03177    0.00903  -3.519              0.00145 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.593 on 29 degrees of freedom
## Multiple R-squared:  0.8268, Adjusted R-squared:  0.8148
## F-statistic: 69.21 on 2 and 29 DF,  p-value: 0.000000000009109
```

Here we have performed Linear regression validating our output of the forward selection method as we can see the important predictors are identified in the output with intercept

Both-direction Stepwise regression

The following code and the output pasted show the performance of the Both-Direction stepwise selection method.

~~~~
```{r}
options(scipen = 100)
model_step <- step(lm(mpg ~ ., data = mtcars), direction =
'both')
summary(model_step)
```
~~~~

```
Start:  AIC=70.9
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear

       Df Sum of Sq    RSS    AIC
- cyl   1    0.0799 147.57 68.915
- vs    1    0.1601 147.66 68.932
- carb  1    0.4067 147.90 68.986
- gear  1    1.3531 148.85 69.190
- drat  1    1.6270 149.12 69.249
- disp  1    3.9167 151.41 69.736
- hp    1    6.8399 154.33 70.348
- qsec  1    8.8641 156.36 70.765
<none>              147.49 70.898
- am    1   10.5467 158.04 71.108
- wt    1   27.0144 174.51 74.280

Step:  AIC=68.92
mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq    RSS    AIC
- vs    1    0.2685 147.84 66.973
- carb  1    0.5201 148.09 67.028
- gear  1    1.8211 149.40 67.308
- drat  1    1.9826 149.56 67.342
- disp  1    3.9009 151.47 67.750
- hp    1    7.3632 154.94 68.473
<none>              147.57 68.915
- qsec  1   10.0933 157.67 69.032
- am    1   11.8359 159.41 69.384
+ cyl   1    0.0799 147.49 70.898
- wt    1   27.0280 174.60 72.297

Step:  AIC=66.97
mpg ~ disp + hp + drat + wt + qsec + am + gear + carb

       Df Sum of Sq    RSS    AIC
- carb  1    0.6855 148.53 65.121
- gear  1    2.1437 149.99 65.434
- drat  1    2.2139 150.06 65.449
- disp  1    3.6467 151.49 65.753
- hp    1    7.1060 154.95 66.475
<none>              147.84 66.973
- am    1   11.5694 159.41 67.384
- qsec  1   15.6830 163.53 68.200
+ vs    1    0.2685 147.57 68.915
+ cyl   1    0.1883 147.66 68.932
- wt    1   27.3799 175.22 70.410

Step:  AIC=65.12
mpg ~ disp + hp + drat + wt + qsec + am + gear

       Df Sum of Sq    RSS    AIC
- gear  1     1.565 150.09 63.457
- drat  1     1.932 150.46 63.535
<none>              148.53 65.121
- disp  1    10.110 158.64 65.229
- am    1    12.323 160.85 65.672
- hp    1    14.826 163.35 66.166
+ carb  1     0.685 147.84 66.973
+ vs    1     0.434 148.09 67.028
+ cyl   1     0.414 148.11 67.032
- qsec  1    26.408 174.94 68.358
- wt    1    69.127 217.66 75.350

Step:  AIC=63.46
mpg ~ disp + hp + drat + wt + qsec + am

       Df Sum of Sq    RSS    AIC
- drat  1     3.345 153.44 62.162
- disp  1     8.545 158.64 63.229
<none>              150.09 63.457
- hp    1    13.285 163.38 64.171
+ gear  1     1.565 148.53 65.121
+ cyl   1     1.003 149.09 65.242
+ vs    1     0.645 149.45 65.319
+ carb  1     0.107 149.99 65.434
- am    1    20.036 170.13 65.466
- qsec  1    25.574 175.67 66.491
- wt    1    67.572 217.66 73.351

Step:  AIC=62.16
mpg ~ disp + hp + wt + qsec + am

       Df Sum of Sq    RSS    AIC
- disp  1     6.629 160.07 61.515
<none>              153.44 62.162
- hp    1    12.572 166.01 62.682
+ drat  1     3.345 150.09 63.457
+ gear  1     2.977 150.46 63.535
+ cyl   1     2.447 150.99 63.648
+ vs    1     1.121 152.32 63.927
+ carb  1     0.011 153.43 64.160
- qsec  1    26.470 179.91 65.255
- am    1    32.198 185.63 66.258
- wt    1    [illegible] [illegible] 72.051
```

```
Call:
lm(formula = mpg ~ wt + qsec + am, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4811 -1.5555 -0.7257  1.4110  4.6610 

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)    
(Intercept)   9.6178     6.9596   1.382   0.177915    
wt           -3.9165     0.7112  -5.507 0.00000695 ***
qsec          1.2259     0.2887   4.247   0.000216 ***
am            2.9358     1.4109   2.081   0.046716 *  
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared:  0.8497,	Adjusted R-squared:  0.8336 
F-statistic: 52.75 on 3 and 28 DF,  p-value: 0.0000000000121
```

The procedure information for the **Both-direction** Stepwise regression is:

As with the forward-stepwise selection, we added predictors to the model successively. After including each predictor, we did, however, also delete any predictors that were no longer improving the model's fit.

The final model turns out to be:

mpg ~ 9.62 – 3.92*wt + 1.23*qsec + 2.94*am

The formula describes that with 1% change such as increase in miles per gallon (mpg) will result in -3.92 % decrease in weight and 1.23% increase in qsec , lastly 2.94% increase in Transmission.

Residual Plots and Diagnostics

Plot analysis from left to right in :

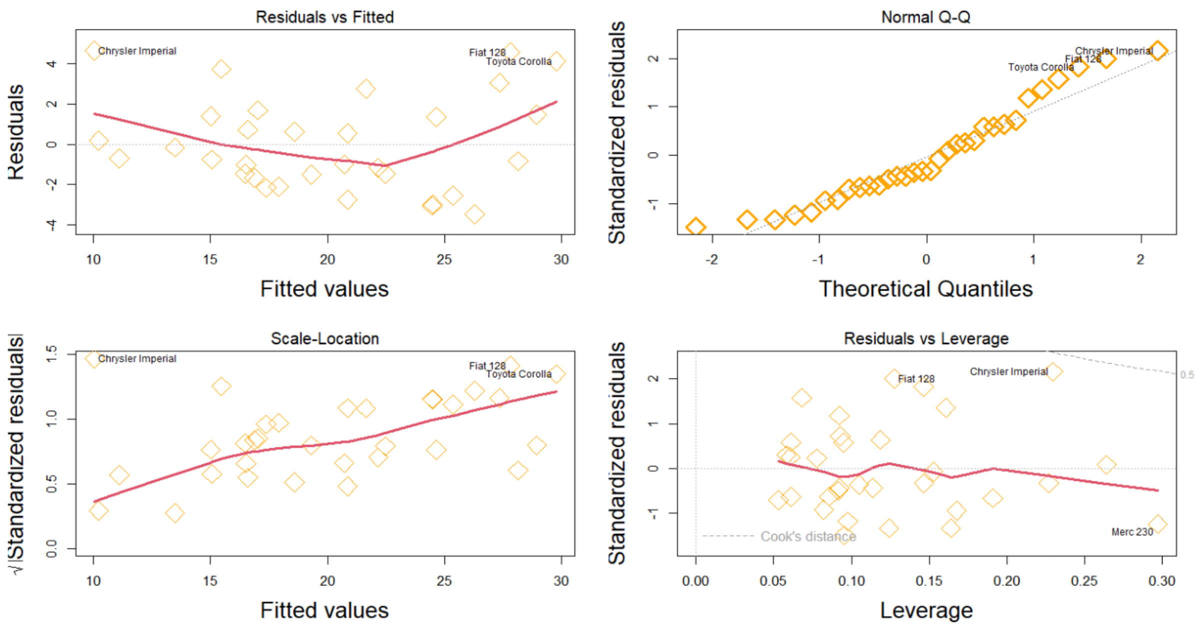
1) The residuals, distance of a point to the regression line, do not show a pattern as they have a random scatter about the dotted line.

2) The residuals in the Quantile/Quantile plot for the most part follow the line and can be assumed to be normally distributed,

3) The red line is fairly flat demonstrating homoschedasity, the residuals are not affected by explanatory variables

4) None of the residuals have a Cook's distance of greater than 0.5.

In conclusion, the type of car transmission that achieves better fuel efficiency is uncertain as other car attributes; horsepower, car weight and number of cylinders, may be a better indication of fuel efficiency. This model could be further refined through such techniques such as reducing any covariance between variables such as horsepower and number of cylinders or weight.



Model Comparison

We are performing the model comparison of the Results of the Forward Selection method and Stepwise Selection Methods to determine which method provides the better Selection.

We are performing three comparison methods namely ANOVA, AIC, and BIC

Compare Models With Anova

```
fit1 <- lm(formula = mpg ~ wt, data = mtcars)
fit2 <- lm(formula = mpg ~ wt + hp, data = mtcars)
anova(fit1, fit2)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	30	278.3219	NA	NA	NA	NA
2	29	195.0478	1	83.27418	12.38133	0.001451229

2 rows

The Annova value of Fit 1(forward Selection) has the Anova null whereas the value of Annova for Fot 2 is 0.014 so, clearly, we can see fit 2 gives the best result than model 1 stating that the forward selection method is accurate with the given dataset.

Compare models with AIC

```
AIC(fit1, fit2)
```

	df <dbl>	AIC <dbl>
fit1	3	166.0294
fit2	4	156.6523

2 rows

The AIC value of Fit 1 is 166.0294 and the AIC value of Fit is 156.6523 So, we can say that compared models using AIC methods which establish our result about fit 2 or forward selection was the best for the given dataset

Compare models with BIC

Compare models with BIC

```
BIC(fit1, fit2)
```

	df <dbl>	BIC <dbl>
fit1	3	170.4266
fit2	4	162.5153

2 rows

We have performed model testing By using BIC and the obtained value of fit 1 is170.4266 and the fit 2 value is 162.5193 So, clearly, we got the output validating our previous two testing methods stating that fit 2 or forward selection was perfect to give an accurate model for provided dataset

Dataset -2 Hitters Dataset.

This Hitters data collection was obtained via the Carnegie Mellon University-maintained StatLib library. This is a portion of the data that was used in the poster session for the 1988 ASA Graphics Section. The pay information was first published in Sports Illustrated on April 20, 1987. The 1987 Baseball Encyclopedia Update, published by Collier Books, Macmillan Publishing Company, New York, provided the 1986 and career statistics.

AtBat	Number of times at bat in 1986
Hits	Number of hits in 1986
HmRun	Number of home runs in 1986
Runs	Number of runs in 1986
RBI	Number of runs batted in in 1986
Walks	Number of walks in 1986
Years	Number of years in the major leagues
CAtBat	Number of times at bat during his career
CHits	Number of hits during his career
CHmRun	Number of home runs during his career
CRuns	Number of runs during his career
CRBI	Number of runs batted in during his career
CWalks	Number of walks during his career
League	A factor with levels A and N indicating player's league at the end of 1986
Division	A factor with levels E and W indicating player's division at the end of 1986

PutOuts	Number of put outs in 1986
Assists	Number of assists in 1986
Errors	Number of errors in 1986
Salary	1987 annual salary on opening day in thousands of dollars
NewLeague	A factor with levels A and N indicating player's league at the beginning of 1987

From the Fig below we Summarize the dataset in the following conclusion:

1. On an average 380 players had come on bat, and the maximum of them who had come on bat are 600 players. Also, as many as 7 years the players played the major leagues.
2. The maxim of 8 rounds the number of players have batted (RBI).
3. On an average the average salary of the players is roughly estimated to 535.6 thousand dollars. The maximum paid salary was of 2460 thousand dollars.
4. On average 8 times the players have made errors, but its surprising there are as many as 32 errors made.
5. 106.9 times the Assist were provided to these professional players. the maximum number of assist provided is 492 times.

```
     AtBat            Hits          HmRun      
 Min.   : 16.0   Min.   :  1   Min.   : 0.00  
 1st Qu.:255.2   1st Qu.: 64   1st Qu.: 4.00  
 Median :379.5   Median : 96   Median : 8.00  
 Mean   :380.9   Mean   :101   Mean   :10.77  
 3rd Qu.:512.0   3rd Qu.:137   3rd Qu.:16.00  
 Max.   :687.0   Max.   :238   Max.   :40.00  

      Runs             RBI             Walks       
 Min.   :  0.00   Min.   :  0.00   Min.   :  0.00  
 1st Qu.: 30.25   1st Qu.: 28.00   1st Qu.: 22.00  
 Median : 48.00   Median : 44.00   Median : 35.00  
 Mean   : 50.91   Mean   : 48.03   Mean   : 38.74  
 3rd Qu.: 69.00   3rd Qu.: 64.75   3rd Qu.: 53.00  
 Max.   :130.00   Max.   :121.00   Max.   :105.00  

     Years            CAtBat            CHits       
 Min.   : 1.000   Min.   :   19.0   Min.   :   4.0  
 1st Qu.: 4.000   1st Qu.:  816.8   1st Qu.: 209.0  
 Median : 6.000   Median : 1928.0   Median : 508.0  
 Mean   : 7.444   Mean   : 2648.7   Mean   : 717.6  
 3rd Qu.:11.000   3rd Qu.: 3924.2   3rd Qu.:1059.2  
 Max.   :24.000   Max.   :14053.0   Max.   :4256.0  

     CHmRun            CRuns             CRBI        
 Min.   :  0.00   Min.   :   1.0   Min.   :   0.00  
 1st Qu.: 14.00   1st Qu.: 100.2   1st Qu.:  88.75  
 Median : 37.50   Median : 247.0   Median : 220.50  
 Mean   : 69.49   Mean   : 358.8   Mean   : 330.12  
 3rd Qu.: 90.00   3rd Qu.: 526.2   3rd Qu.: 426.25  
 Max.   :548.00   Max.   :2165.0   Max.   :1659.00  

     CWalks          League  Division    PutOuts      
 Min.   :   0.00   A:175   E:157     Min.   :   0.0  
 1st Qu.:  67.25   N:147   W:165     1st Qu.: 109.2  
 Median : 170.50                     Median : 212.0  
 Mean   : 260.24                     Mean   : 288.9  
 3rd Qu.: 339.25                     3rd Qu.: 325.0  
 Max.   :1566.00                     Max.   :1378.0  

    Assists          Errors          Salary       NewLeague
 Min.   :  0.0   Min.   : 0.00   Min.   :  67.5   A:176    
 1st Qu.:  7.0   1st Qu.: 3.00   1st Qu.: 190.0   N:146    
 Median : 39.5   Median : 6.00   Median : 425.0            
 Mean   :106.9   Mean   : 8.04   Mean   : 535.9            
 3rd Qu.:166.0   3rd Qu.:11.00   3rd Qu.: 750.0            
 Max.   :492.0   Max.   :32.00   Max.   :2460.0            
                                 NA's   :59                
```

The regsubsets() method from the 'leaps' package identifies the best model that contains a specified number of predictors, where best is measured using RSS, and conducts best subset selection. The syntax is identical to that of lm (). For each model size, the summary() command returns the ideal set of variables.

```
Subset selection object
Call: regsubsets.formula(Salary ~ ., data = Hitters, nvmax = 19)
19 Variables  (and intercept)
           Forced in Forced out
AtBat          FALSE      FALSE
Hits           FALSE      FALSE
HmRun          FALSE      FALSE
Runs           FALSE      FALSE
RBI            FALSE      FALSE
Walks          FALSE      FALSE
Years          FALSE      FALSE
CAtBat         FALSE      FALSE
CHits          FALSE      FALSE
CHmRun         FALSE      FALSE
CRuns          FALSE      FALSE
CRBI           FALSE      FALSE
CWalks         FALSE      FALSE
LeagueN        FALSE      FALSE
DivisionW      FALSE      FALSE
PutOuts        FALSE      FALSE
Assists        FALSE      FALSE
Errors         FALSE      FALSE
NewLeagueN     FALSE      FALSE
1 subsets of each size up to 19
Selection Algorithm: exhaustive
          AtBat Hits HmRun Runs RBI Walks Years CAtBat
1  ( 1 )  " "   " "  " "   " "  " " " "   " "   " "
2  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "
3  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "
4  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "
5  ( 1 )  "*"   "*"  " "   " "  " " " "   " "   " "
6  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "
7  ( 1 )  " "   "*"  " "   " "  " " "*"   " "   "*"
8  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "
9  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   "*"
10  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   "*"
11  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   "*"
12  ( 1 ) "*"   "*"  " "   "*"  " " "*"   " "   "*"
13  ( 1 ) "*"   "*"  " "   "*"  " " "*"   " "   "*"
14  ( 1 ) "*"   "*"  "*"   "*"  " " "*"   " "   "*"
15  ( 1 ) "*"   "*"  "*"   "*"  " " "*"   " "   "*"
16  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   " "   "*"
17  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   " "   "*"
18  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   "*"   "*"
19  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   "*"   "*"
          CHits CHmRun CRuns CRBI CWalks LeagueN DivisionW
1  ( 1 )  " "   " "    " "   "*"  " "    " "     " "
2  ( 1 )  " "   " "    " "   "*"  " "    " "     " "
3  ( 1 )  " "   " "    " "   "*"  " "    " "     " "
4  ( 1 )  " "   " "    " "   "*"  " "    " "     "*"
5  ( 1 )  " "   " "    " "   "*"  " "    " "     "*"
6  ( 1 )  " "   " "    " "   "*"  " "    " "     "*"
7  ( 1 )  "*"   "*"    " "   " "  " "    " "     "*"
8  ( 1 )  " "   "*"    "*"   " "  "*"    " "     "*"
9  ( 1 )  " "   " "    "*"   "*"  "*"    " "     "*"
10  ( 1 ) " "   " "    "*"   "*"  "*"    " "     "*"
11  ( 1 ) " "   " "    "*"   "*"  "*"    "*"     "*"
12  ( 1 ) " "   " "    "*"   "*"  "*"    "*"     "*"
13  ( 1 ) " "   " "    "*"   "*"  "*"    "*"     "*"
14  ( 1 ) " "   " "    "*"   "*"  "*"    "*"     "*"
15  ( 1 ) "*"   " "    "*"   "*"  "*"    "*"     "*"
16  ( 1 ) "*"   " "    "*"   "*"  "*"    "*"     "*"
17  ( 1 ) "*"   " "    "*"   "*"  "*"    "*"     "*"
18  ( 1 ) "*"   " "    "*"   "*"  "*"    "*"     "*"
19  ( 1 ) "*"   "*"    "*"   "*"  "*"    "*"     "*"
          PutOuts Assists Errors NewLeagueN
1  ( 1 )  " "     " "     " "    " "
2  ( 1 )  " "     " "     " "    " "
3  ( 1 )  "*"     " "     " "    " "
4  ( 1 )  "*"     " "     " "    " "
5  ( 1 )  "*"     " "     " "    " "
6  ( 1 )  "*"     " "     " "    " "
7  ( 1 )  "*"     " "     " "    " "
8  ( 1 )  "*"     " "     " "    " "
9  ( 1 )  "*"     " "     " "    " "
10  ( 1 ) "*"     "*"     " "    " "
11  ( 1 ) "*"     "*"     " "    " "
12  ( 1 ) "*"     "*"     " "    " "
13  ( 1 ) "*"     "*"     "*"    " "
14  ( 1 ) "*"     "*"     "*"    " "
15  ( 1 ) "*"     "*"     "*"    " "
16  ( 1 ) "*"     "*"     "*"    " "
17  ( 1 ) "*"     "*"     "*"    "*"
18  ( 1 ) "*"     "*"     "*"    "*"
19  ( 1 ) "*"     "*"     "*"    "*"
[1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"
[7] "outmat" "obj"
```

A variable is marked with an asterisk ("*") if it is present in the associated model. For instance, this result shows that Hits and CRBI are the only two variables in the optimal two-variable model. Regsubsets() by default only presents results for the top-performing eight-variable model. However, it is possible to return as many variables as needed by using the nvmax option. Here, we fit a model with up to 19 variables.

```
> names(reg.summary)
[1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"
[7] "outmat" "obj"
> reg.summary$cp
 [1] 104.281319  50.723090  38.693127  27.856220  21.613011
 [6]  14.023870  13.128474   7.400719   6.158685   5.009317
[11]   5.874113   7.330766   8.888112  10.481576  12.346193
[16]  14.187546  16.087831  18.011425  20.000000
> reg.summary$adjr2
 [1] 0.3188503 0.4208024 0.4450753 0.4672734 0.4808971
 [6] 0.4972001 0.5007849 0.5137083 0.5180572 0.5222606
[11] 0.5225706 0.5217245 0.5206736 0.5195431 0.5178661
[16] 0.5162219 0.5144464 0.5126097 0.5106270
> reg.summary$bic
 [1]  -90.84637 -128.92622 -135.62693 -141.80892 -144.07143
 [6] -147.91690 -145.25594 -147.61525 -145.44316 -143.21651
[11] -138.86077 -133.87283 -128.77759 -123.64420 -118.21832
[16] -112.81768 -107.35339 -101.86391  -96.30412
```

By looking at the output below, we can see that the model with 6 variables performs the best overall, according to BIC. There are ten variables in Cp. The adjusted R2 hints that 11 might be the ideal. A model with 5 or less predictors is insufficient, whereas a model with more than 12 predictors is overfitting. Again, no one measure will provide us with an absolutely correct picture.

```
> which.min(reg.summary$cp)
[1] 10
> which.max(reg.summary$adjr2)
[1] 11
> which.min(reg.summary$bic)
[1] 6
```

```
> backward = regsubsets(Salary ~ ., data = Hitters, method = "backward")
> reg.summary <- summary(backward)
> reg.summary
```

```
Subset selection object
Call: regsubsets.formula(Salary ~ ., data = Hitters, method = "backward")
19 Variables  (and intercept)
           Forced in Forced out
AtBat          FALSE      FALSE
Hits           FALSE      FALSE
HmRun          FALSE      FALSE
Runs           FALSE      FALSE
RBI            FALSE      FALSE
Walks          FALSE      FALSE
Years          FALSE      FALSE
CAtBat         FALSE      FALSE
CHits          FALSE      FALSE
CHmRun         FALSE      FALSE
CRuns          FALSE      FALSE
CRBI           FALSE      FALSE
CWalks         FALSE      FALSE
LeagueN        FALSE      FALSE
DivisionW      FALSE      FALSE
PutOuts        FALSE      FALSE
Assists        FALSE      FALSE
Errors         FALSE      FALSE
NewLeagueN     FALSE      FALSE
        AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI
1  ( 1 ) " "   " "  " "   " "  " " " "   " "   " "    " "   " "    "*"   " " 
2  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    "*"   " " 
3  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    "*"   " " 
4  ( 1 ) "*"   "*"  " "   " "  " " " "   " "   " "    " "   " "    "*"   " " 
5  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   " " 
6  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   " " 
7  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   " " 
8  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   "*" 
        CWalks LeagueN DivisionW PutOuts Assists Errors NewLeagueN
1  ( 1 ) " "    " "     " "       " "     " "     " "    " "       
2  ( 1 ) " "    " "     " "       " "     " "     " "    " "       
3  ( 1 ) " "    " "     " "       "*"     " "     " "    " "       
4  ( 1 ) " "    " "     " "       "*"     " "     " "    " "       
5  ( 1 ) " "    " "     " "       "*"     " "     " "    " "       
6  ( 1 ) " "    " "     "*"       "*"     " "     " "    " "       
7  ( 1 ) "*"    " "     "*"       "*"     " "     " "    " "       
8  ( 1 ) "*"    " "     "*"       "*"     " "     " "    " "       
> names(reg.summary)
[1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"   
> which.max(reg.summary$adjr2)
[1] 8
```

The 8 variable model is preferred, as seen in the result below, according to the Adjusted R2.

References

1. Bevans (2022, November 11);Datanovia (2019, December 26);
2. *Linear Regression Example in r Using Lm() Function* (n.d.);Zach (2021, September 29);John (2023, January 25);Rithika (2022, December 29)
3. Bevans, R. 2022, November 11. *Hypothesis Testing | a Step-by-Step Guide with Easy Examples*. https://www.scribbr.com/statistics/hypothesis-testing/.
4. Datanovia. 2019, December 26. *How to Do a t-Test in r: Calculation and Reporting*. https://www.datanovia.com/en/lessons/how-to-do-a-t-test-in-r-calculation-and-reporting/.
5. *Linear Regression Example in r Using Lm() Function*. n.d. https://www.learnbymarketing.com/tutorials/linear-regression-in-r/.
6. Rithika, S. 2022, December 29. *Building a Churn Prediction Model on Retail Data Simplified: The Ultimate Guide 101. Learn | Hevo.* https://hevodata.com/learn/churn-prediction-model/.
7. Zach, Z. 2021, September 29. *How to Perform Logistic Regression in r (Step-by-Step)*. https://www.statology.org/logistic-regression-in-r/.

Appendix

```
data('mtcars')
head(mtcars)

?mtcars

set.seed(100)
trainIndex <- sort(sample(x = nrow(mtcars), size = nrow(mtcars) * 0.7))
sample_train <- mtcars[trainIndex,]
sample_test <- mtcars[-trainIndex,]
head(sample_train)
head(sample_test)

summary(sample_train)

hist(sample_train$mpg,breaks = 10,xlab="MPG", col = "skyblue", xlim=c(5,35)
)

input<- sample_train
```

```
input$am <- as.factor(input$am)
levels(input$am) <-c("AT", "MT")

table(input$am)

dim(input)

library(ggplot2)
library(caret)
ggplot(input, aes(x=am, y=mpg)) + geom_boxplot(fill="lightgreen")

pairs(mpg ~ ., data = sample_train, col= "red")

options(scipen = 100)
model_step <- step(lm(mpg ~ ., data = mtcars), direction = 'both')
summary(model_step)

step(lm(mpg ~ 1, data = mtcars), direction = 'forward', scope = ~ disp + hp
+ drat + wt + qsec)
model_forward <- lm(formula = mpg ~ wt + hp, data = mtcars)
summary(model_forward)

par(mfrow=c(2,2))
plot(model_step,pch=23,col="orange",cex=2.5,cex.lab=1.6,lwd=3)

fit1 <- lm(formula = mpg ~ wt, data = mtcars)
fit2 <- lm(formula = mpg ~ wt + hp, data = mtcars)
anova(fit1, fit2)

AIC(fit1, fit2)

BIC(fit1, fit2)

library(leaps)
library(ISLR)
library(dplyr)
```

```
summary(Hitters)
Hitters <- Hitters %>% na.omit()

best_subset = regsubsets(Salary ~ ., data = Hitters, nvmax = 19)
reg.summary <- summary(best_subset)
reg.summary
names(reg.summary)

reg.summary$cp
reg.summary$adjr2
reg.summary$bic

which.min(reg.summary$cp)
which.max(reg.summary$adjr2)
which.min(reg.summary$bic)

backward = regsubsets(Salary ~ ., data = Hitters, method = "backward")
reg.summary <- summary(backward)
reg.summary
names(reg.summary)

which.max(reg.summary$adjr2)

## NA
```