MotorTrend Data Analysis - EDA & Regression

Willianto Asalim

14/06/2020

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

Motor Trend Data Analysis

1. Executive Summary

Motor Trend, an automobile industry magazine is looking at a data set of a collection of cars. The company is interested in exploring the relationship between a set of variables and the outcome of miles per gallon (MPG). They are particularly interested in the following two questions:

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

Regression models and exploratory data analysis will be used to answer these two questions.

Information about the Motor Trend dataset can be found at the following link: mtcars info <-click here

2. Exploratory Data Analysis

Load the mtcars data and perform some basic exploratory data analyses

```
library(datasets)
data(mtcars) ##Load mtcars dataset
```

From the fig. 1 in the appendix we know that the mtcars dataset contains 32 observations and 11 variables: mpg, cyl, disp, hp, drat, wt, qsec, vs, am, gear, carb. The fig. 2 boxplot shows the relationship between MPG and AM, from the boxplot we can see that the cars with manual transmission yield a better miles per gallon than the car with automatic transmission.

However we need to test whether it is true that the cars with different transmissions yield better MPG outcome by conducting a hyphothesis testing (T Test). The NULL HYPOTHESIS is the different transmissions does not impact the outcome of MPG and the ALTERNATIVE HYPOTHESIS is the different tranmissions will impact the outcome of MPG. To reject the null hypothesis, a scientific standard of more than 95% confidence interval is used because anything less is no significance difference for scientific studies. Hence the P value (critical value) must be less than 5% (.05) for a significance difference. If the P value is less than .05, it is likely that the transmission has impact on MPG. If the P value is more than .05, it is unlikely that the transmission has impact on MPG.

```
##T test to show whether transmission has impact on MPG - Appendix fig 2
hTest <- t.test(mpg~am, data=mtcars, paired=F, var.equal=T, conf.level=0.95)
hTest$p.value ##Getting the p value of T test
```

```
## [1] 0.0002850207
```

The P value results of the hypothesis T test conducted shows the p value is 2.8502074×10^{-4} which is less than .05. Therefore we reject the null hypothesis and we can conclude that tranmission type has an impact on the outcome of MPG. The T Test result in Fig. 2 appendix of the mtcars dataset also shows that the mean for automatic tranmission is 17.15 MPG and the mean for manual transmission is 24.39 MPG.

3. Regression Models Analysis

We need to examine further whether the AM variable is the biggest factor in determining the impact of MPG (outcome) or perhaps there are other variables in the mtcars data that we should explore further. In the fig. 3 you can look at the correlation relationship between all the variables.

The first regression is the relationship between AM and MPG and in this instance we are using SIMPLE LINEAR REGRESSION (single variable).

```r
##Fit simple linear regression model - Appendix fig 4
linReg <- lm(mpg ~ am, data = mtcars)
```

Please refer to the fig. 4 in the appendix for the simple linear regression output that shows the Multiple R Square and the Adjusted R Square of 0.36 and 0.34 respectively which is pretty low. AM varible might not be the the best single variable to determine the outcome of MPG.

The second regression will be the MULTIVARIABLE REGRESSION where we will include all the 11 variables in the mtcars.

```r
##Fit multivariable regression model - Appendix fig 5
mulReg <- lm(mpg~., data = mtcars)
```

Please refer to the fig. 5 in the appendix for the multiple regression output that shows the R Square and the Adjust R Square of 0.87 and 0.81 respectively which is higher than the first simple linear regression. However the Variance Inflation Factor (VIF) is very high (more than 5) for a number of variables and the p Value is more than .05 for a number of variables.

The third regression will be the STEPWISE REGRESSION in which the choice of predictive variables is carried out by an automatic procedure. Please look at the appendix for the stepwise regression method explanation.

```r
##Fit Stepwise regression using bidirectional method - Appendix fig 6
stepReg <- step(lm(mpg~., data=mtcars), direction = "both")
```

Please refer to Fig. 6 in the appendix for the stepwise regression output that shows the three variables matters to MPG outcome namely WT, QSEC and AM. The Multiple R Square and the Adjusted R Square of 0.85 and 0.835 respectively which is better than the simple linear and multivariable regression. The Variance Inflation Factor (VIF) is very good as it is low (less than 5) and the p Value is significance which is less than .05.

4. Conclusion

The simple linear regression model with one variable, AM is not strong enough to determine the MPG as its multiple R Square lower than the other two models. The second model, multivariable regression model with all the 11 variables included in the model is not ideal due to the high VIF and poor p value in some variables. Thus the stepwise regression model with three variables (WT, QSEC and AM) produces a better R Square, higher F statistics, lower VIF and siginificance in p value than the other two models. Hence the stepwise regression model is superior than linear and multivariable regression models for determining MPG (outcome).

Best model: Stepwise Regression Model

5. Appendix

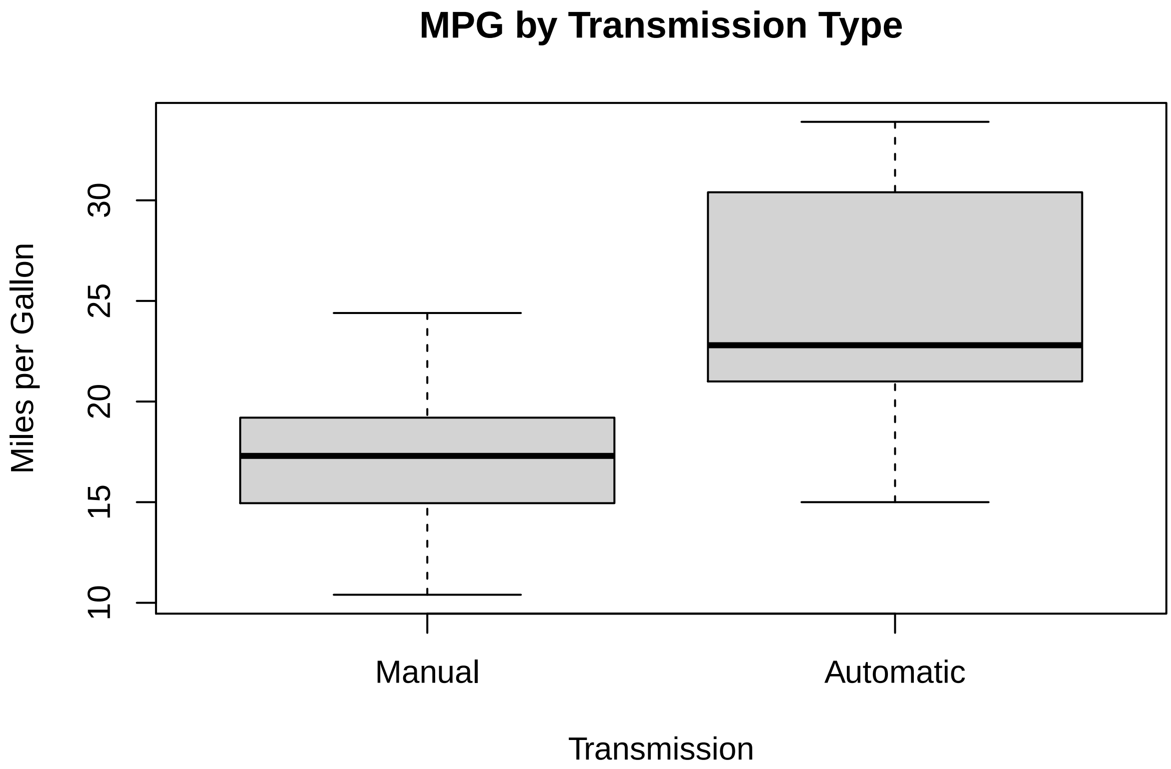
Figure 1: Data Summary

```r
summary(mtcars) ##Summary of mtcars dataset
dim(mtcars) ##Number of observations and variables in mtcars dataset
names(mtcars) ##Names of the variables in mtcars dataset
```

```
##       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  
## [1] 32 11
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

Figure 2: Boxplot of MPG and AM relationship + T Test

```r
##Exploratory Data Analysis by looking at the relationship of MPG and AM using boxplot
boxplot(mpg~am, data = mtcars,
        names = c("Manual", "Automatic"),
        xlab = "Transmission",
        ylab = "Miles per Gallon",
        main = "MPG by Transmission Type")
```

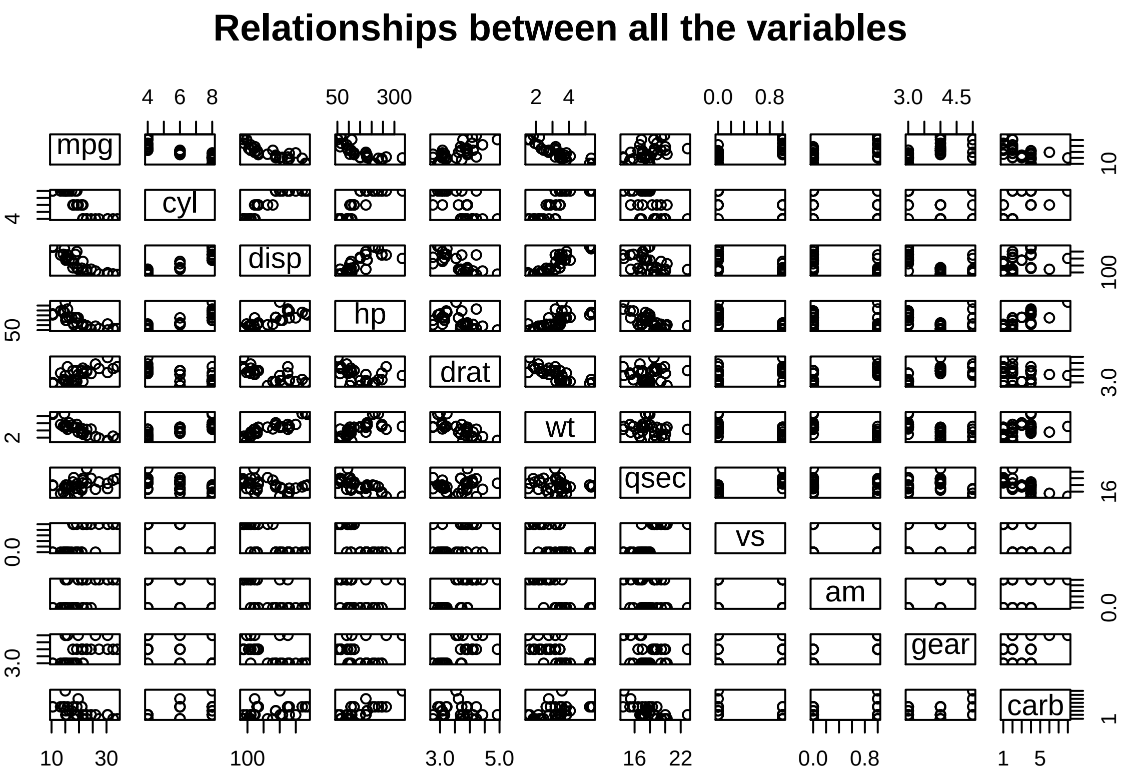


```
hTest
```

```
##
##  Two Sample t-test
##
## data:  mpg by am
## t = -4.1061, df = 30, p-value = 0.000285
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.84837  -3.64151
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

Figure 3: Relationship between all variables

```
#Chart shows relationship between all variables
pairs(mpg ~ ., data = mtcars, main="Relationships between all the variables")
```



Relationships between all the variables

```r
cor(mtcars) ## Correlations between all variables
```

```
##             mpg        cyl       disp         hp        drat         wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
## qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
## vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157
## am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953
## gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
## carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
##             qsec         vs          am       gear        carb
## mpg   0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
## cyl  -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
## disp -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
## hp   -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
## drat  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
## wt   -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
## qsec  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
## vs    0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
## am   -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
## gear -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
## carb -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000
```

Figure 4: Linear Regression Output

```
summary(linReg) ##Output of regression model result
anova(linReg) ##Output of Analysis of Variance
```

```
##
## 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
##
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value   Pr(>F)
## am         1 405.15  405.15   16.86 0.000285 ***
## Residuals 30 720.90   24.03
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

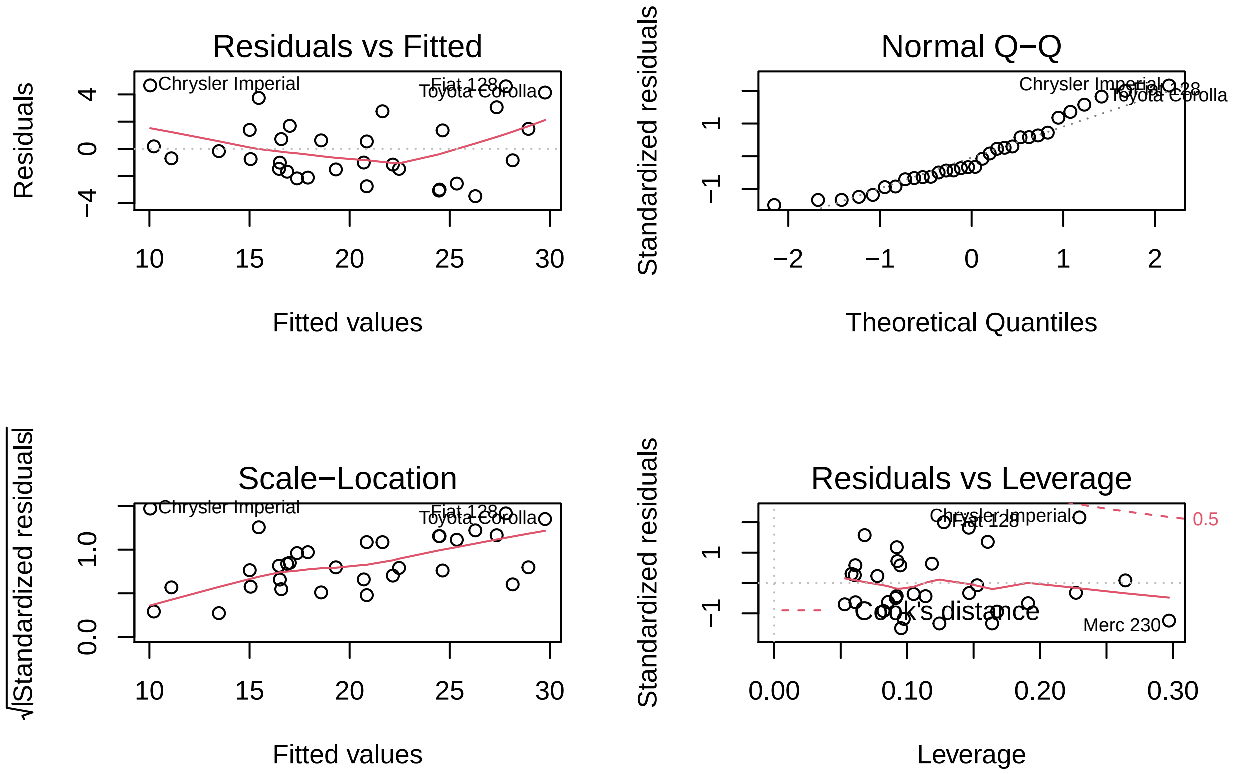
Figure 5: Multivariable Regression Output

```
summary(mulReg) ##Output of regression model result
anova(mulReg) ##Output of Analysis of Variance
vif(mulReg) ##Output of Variance Inflation Factor
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
##
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## cyl        1 817.71  817.71 116.4245 5.034e-10 ***
## disp       1  37.59   37.59   5.3526  0.030911 *
## hp         1   9.37    9.37   1.3342  0.261031
## drat       1  16.47   16.47   2.3446  0.140644
## wt         1  77.48   77.48  11.0309  0.003244 **
## qsec       1   3.95    3.95   0.5623  0.461656
## vs         1   0.13    0.13   0.0185  0.893173
## am         1  14.47   14.47   2.0608  0.165858
## gear       1   0.97    0.97   0.1384  0.713653
## carb       1   0.41    0.41   0.0579  0.812179
## Residuals 21 147.49    7.02
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##       cyl      disp        hp      drat        wt      qsec        vs        am
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873  4.648487
##      gear      carb
##  5.357452  7.908747
```

Figure 6: Stepwise Regression Output

```r
par(mfrow = c(2,2)) ##Multiple grapths into two by two plot
plot(stepReg) ##Plot stepwise regression model
```



```
summary(stepReg) ##Output of regression model result
anova(stepReg) ##Output of Analysis of Variance
vif(stepReg) ##Output of Variance Inflation Factor
```

```
##
## 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 6.95e-06 ***
## 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: 1.21e-11
##
## Analysis of Variance Table
```

```
##
## Response: mpg
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## wt         1 847.73  847.73 140.2143 2.038e-12 ***
## qsec       1  82.86   82.86  13.7048 0.0009286 ***
## am         1  26.18   26.18   4.3298 0.0467155 *
## Residuals 28 169.29    6.05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##       wt     qsec       am
## 2.482952 1.364339 2.541437
```

Stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some prespecified criterion. Usually, this takes the form of a sequence of F-tests or t-tests, but other techniques are possible, such as adjusted R2, Akaike information criterion, Bayesian information criterion, Mallows's Cp, PRESS, or false discovery rate.

The main approaches are:

Forward selection, which involves starting with no variables in the model, testing the addition of each variable using a chosen model fit criterion, adding the variable (if any) whose inclusion gives the most statistically significant improvement of the fit, and repeating this process until none improves the model to a statistically significant extent.

Backward elimination, which involves starting with all candidate variables, testing the deletion of each variable using a chosen model fit criterion, deleting the variable (if any) whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically insignificant loss of fit.

Bidirectional elimination, a combination of the above, testing at each step for variables to be included or excluded. Source: wikipedia

---

The platform specification used:

Spec	Description
OS	Windows 10 Pro - 64 bit
CPU	AMD Ryzen 5 - 3400G
RAM	16GB DDR4 3000MHz
Storage	500GB SSD - M.2 NVMe (PCIe)
Tool	RStudio