Motor Trend: Automatic vs manual transmission comparative study

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:
- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

Data	Description
mpg	Miles per US gallon
cyl	Number of cylinders
disp	Displacement (cubic inches)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (lb / 1000)
qsec	1 / 4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

1.Loading prerequisites

```
suppressMessages(
    {
        if(!require(manipulate)){
            install.packages("manipulate")
        }
        if(!require(GGally)){
            install.packages("GGally")
        }
        if(!require(lmtest)){
            install.packages("lmtest")
        }
        if(!require(dplyr)){
            install.packages("dplyr")
        }
        if(!require(ggplot2)){
            install.packages("ggplot2")
        }
        library(GGally)
        library(manipulate)
        library(lmtest)
        library(dplyr)
        library(ggplot2)
    }
)
```

1.1. Libraries

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

1.1. 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
```

2. Exploratory Analysis

Variables

```
names(mtcars)
```

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

Understanding variable types

```
apply(mtcars,2,class)
```

```
##       mpg       cyl      disp        hp      drat        wt      qsec        vs
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##        am      gear      carb
## "numeric" "numeric" "numeric"
```

Fixing the types

```
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
```

Summarizing each variables

```
summary(mtcars)
```

```
##       mpg        cyl         disp             hp             drat
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930
##        wt             qsec       vs             am     gear   carb
##  Min.   :1.513   Min.   :14.50   0:18   Automatic:19   3:15   1: 7
##  1st Qu.:2.581   1st Qu.:16.89   1:14   Manual   :13   4:12   2:10
##  Median :3.325   Median :17.71                         5: 5   3: 3
##  Mean   :3.217   Mean   :17.85                                4:10
##  3rd Qu.:3.610   3rd Qu.:18.90                                6: 1
##  Max.   :5.424   Max.   :22.90                                8: 1
```

3. Regression modelling

Fitting a model with all the variables

```
mdl_all = glm(mpg~., family = "gaussian", data = mtcars)
summary(mdl_all)
```

```
##
## Call:
## glm(formula = mpg ~ ., family = "gaussian", data = mtcars)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5087  -1.3584  -0.0948   0.7745   4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## amManual     1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 8.026845)
```

```
##
##     Null deviance: 1126.0  on 31  degrees of freedom
## Residual deviance:  120.4  on 15  degrees of freedom
## AIC: 169.22
##
## Number of Fisher Scoring iterations: 2
```

The above model tells us that the average **mpg** is 23.88

3.1 Model selection

- We would want to select a model that has larger adjusted and predicted R-squared values.
- In regression, p-values less than the significance level indicate that the term is statistically significant. We reduce the model by repeatedly removing parameters corresponding to coefficients that do not have significant effect on the model performance.

- Upon arriving at the manually identified "best" model, we will compare the model performance against an automated model selection procedure using step regression function. step()

- We will analyse whether there is significant effect on mpg when considering the transmission model using the t.test() function.

The wt, carb and disp parameters has least significant effects on the model performance, Hence we remove them to analyse the variation in the p-value of the other variables. If there is large improvements then it'd mean that there is no correlation between them.

```
mdl_test_1 = glm(mpg~.-wt-disp-carb, family = "gaussian", data = mtcars)
summary(mdl_test_1)
```

```
##
## Call:
## glm(formula = mpg ~ . - wt - disp - carb, family = "gaussian",
##     data = mtcars)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.4134  -1.3937   0.3857   1.6464   5.0131
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.75913   17.73868   1.001  0.32764
## cyl6        -1.03161    2.35733  -0.438  0.66593
## cyl8         3.19942    4.54838   0.703  0.48917
## hp          -0.06386    0.01874  -3.408  0.00252 **
## drat         1.74285    1.99326   0.874  0.39136
## qsec         0.05538    0.69880   0.079  0.93755
## vs1          3.66895    2.16897   1.692  0.10485
## amManual     4.42201    2.16990   2.038  0.05376 .
## gear4       -1.26917    2.39429  -0.530  0.60136
## gear5        2.19140    2.84390   0.771  0.44916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for gaussian family taken to be 7.39212)
##
##     Null deviance: 1126.05  on 31  degrees of freedom
## Residual deviance:  162.63  on 22  degrees of freedom
## AIC: 164.84
##
## Number of Fisher Scoring iterations: 2
```

Similary we remove qsec, gear, drat and vs

```
mdl_test_2 = glm(mpg~.-wt-disp-carb-qsec-gear-drat-vs, family = "gaussian", data = mtcars)
summary(mdl_test_2)
```

```
##
## Call:
## glm(formula = mpg ~ . - wt - disp - carb - qsec - gear - drat -
##     vs, family = "gaussian", data = mtcars)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -5.231  -1.535  -0.141   1.408   5.322
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.29590    1.42394  19.169  < 2e-16 ***
## cyl6        -3.92458    1.53751  -2.553  0.01666 *
## cyl8        -3.53341    2.50279  -1.412  0.16943
## hp          -0.04424    0.01458  -3.035  0.00527 **
## amManual     4.15786    1.25655   3.309  0.00266 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 7.303666)
##
##     Null deviance: 1126.0  on 31  degrees of freedom
## Residual deviance:  197.2  on 27  degrees of freedom
## AIC: 161
##
## Number of Fisher Scoring iterations: 2
```

Automated best model selection using step regression

```
mdl_best = step(mdl_all, direction = "backward")
```

```
summary(mdl_best)
```

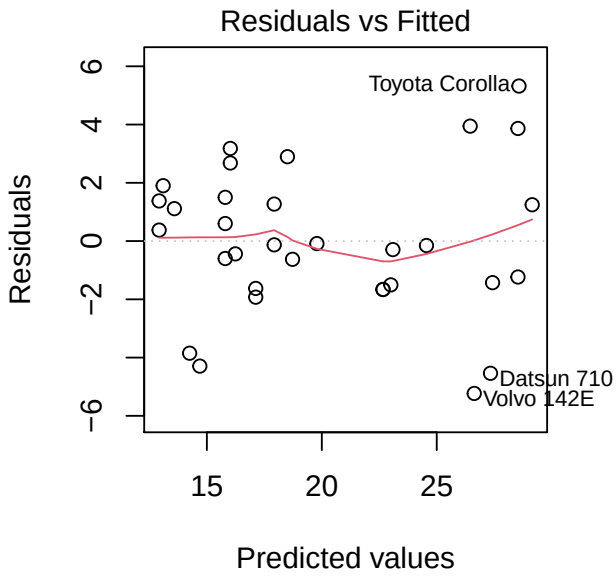
```
##
## Call:
## glm(formula = mpg ~ cyl + hp + wt + am, family = "gaussian",
##     data = mtcars)
##
```

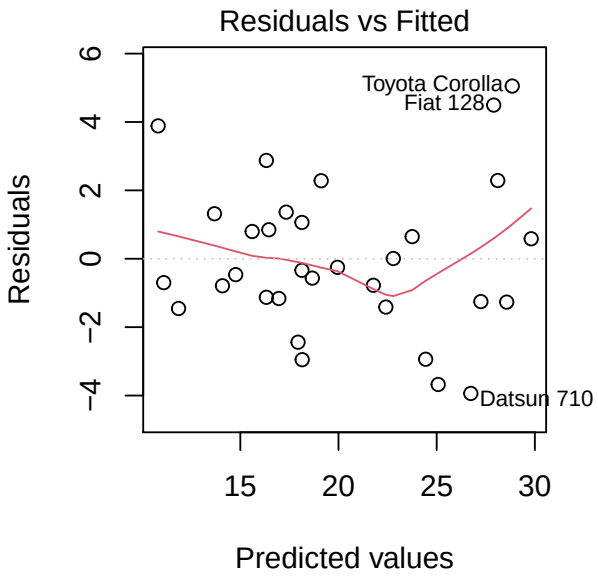
```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9387  -1.2560  -0.4013   1.1253   5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 5.808677)
##
##     Null deviance: 1126.05  on 31  degrees of freedom
## Residual deviance:  151.03  on 26  degrees of freedom
## AIC: 154.47
##
## Number of Fisher Scoring iterations: 2
```

The two models seem to be very similar with exception of just the wt parameter being omitted from the manually subsetted model.

3.2 Analyzing the models Analysing the residual plots manually fitted vs step regressed.

```
par(mfrow=c(1,2))
plot(mdl_test_2, which =1)
plot(mdl_best, which =1)
```

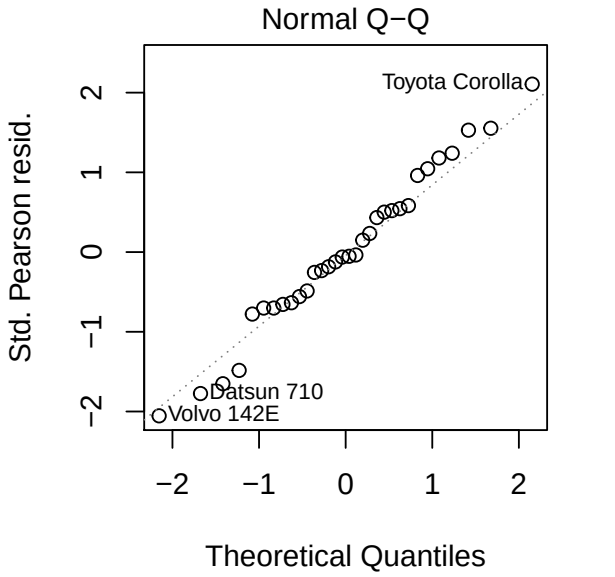


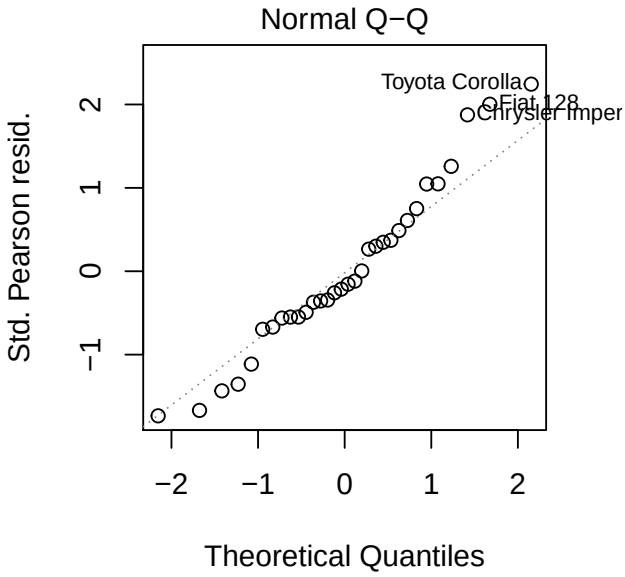


It is observed that the manually fitted model has lower standard error.

Asserting the normality of the residuals of each models

```
par(mfrow=c(1,2))
plot(mdl_test_2, which =2)
plot(mdl_best, which =2)
```





```r
t.test(mpg ~ am, data = mtcars)
```

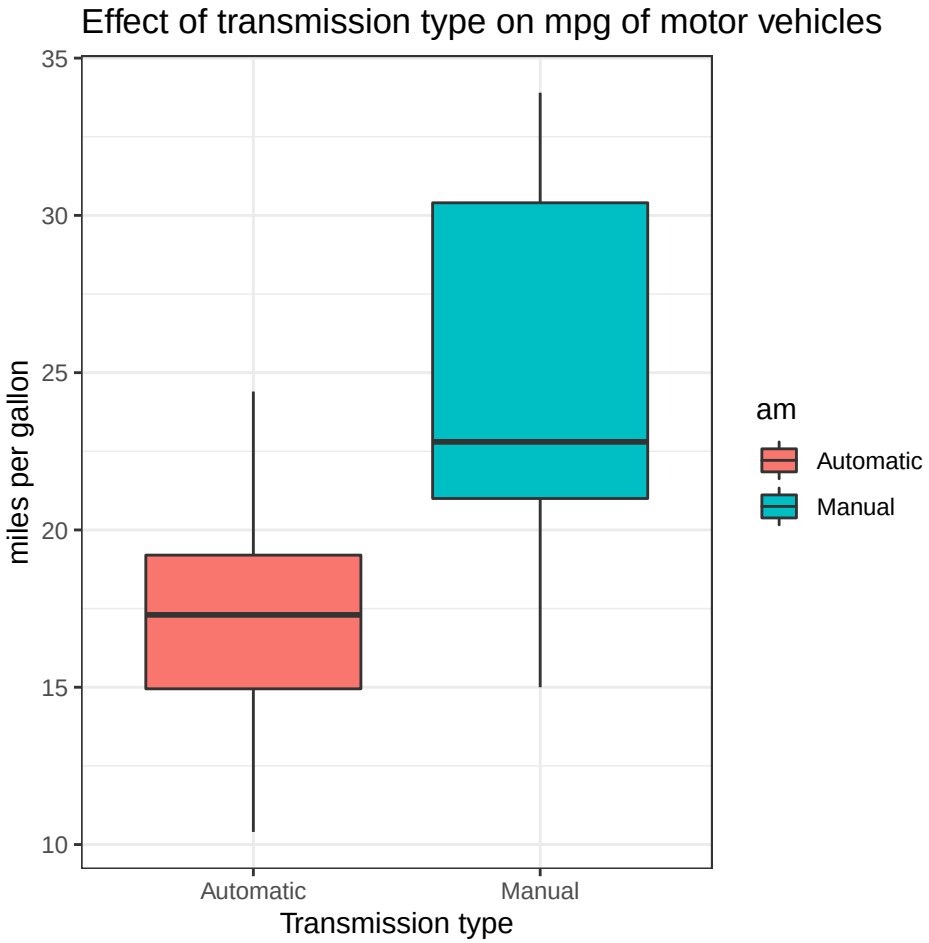
3.3 Using T-test to identify significance of transmission type on mpg

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

The above t.test with p-statistic <0.05 shows that there is significant effect of type of transmission on the mpg of a motor vehicle.

Visualizing the difference

```r
ggplot(
  mtcars,
  aes(
    x=am,
    y=mpg,
    fill = am
  )
) + geom_boxplot() +
  labs(x = "Transmission type", y = "miles per gallon") +
  ggtitle("Effect of transmission type on mpg of motor vehicles") +
  theme_bw()
```



4. Conclusion

Based on my analysis I was able to identify that there is significant effect of transmission type of the motor vehicle on its miles per gallon metric - that the manual transmission performes better than automatic type.

The rate of change of the conditional mean mpg with respect to am is about 4.1 conidering the manually subsetted model and 1.8 with the best subset selection using the step function.

Using the T-test we were able to infer that there is indeed a significance in transmission type on mpg visualized using boxplot shown above