Assignment: Regression Models Course Project

Motor Trend

Overview

Work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, MT are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). MT 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.

Requirements & setting

```
set.seed(42)
library(ggplot2)
```

Preprocessing

Let's load required data and transform certain variables into factors.

```
data(mtcars)
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am   <- factor(mtcars$am,labels=c("Automatic","Manual"))
```

Statistical Inference

T-test on the two subsets of mpg data: manual and automatic transmission assuming that the transmission data has a normal distribution and tests the null hypothesis that they come from the same distribution.

```
t.test(mtcars$mpg ~ mtcars$am)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg by mtcars$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
```

Based on the t-test results, we reject the null hypothesis that the mpg distributions for manual and automatic transmissions are the same.

Regression Analysis

First model (all to mpg)

First model includes all variables as predictors of mpg.

```
allModel <- lm(mpg ~ ., data = mtcars)
summary(allModel) # results hidden
```

Residual standard error - 2.833 and 15 degrees of freedom. Adjusted R-squared value - 0.779, which means that the model can explain about 78% of the variance of the MPG variable. None of the coefficients are significant at 0.05 significant level.

Best model

Then use first model in order to select significant predictors for the best model. The step function will perform this selection by calling lm repeatedly to build multiple regression models and select the best variables from them using both forward selection and backward elimination methods using AIC algorithm. This ensures that we have included useful variables while omitting ones that do not contribute significantly to predicting mpg.

```
bestModel <- step(allModel, direction = "both") #result hidden
```

```
summary(bestModel)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## 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
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

The best model obtained from the above computations shows that variables, cyl, wt and hp as confounders and am as the independent variable. Details of the model are depicted below.

Summory: Adjusted R-squared value - 0.84, which means that the model can explain about 84% (which is the maximum obtained considering all combinations of variables) of the variance of the MPG variable.

Residual Plots

The residual plots (in the appendix) of our regression model along with computation of regression diagnostics for our liner model give the following results:

1. The points in the Residuals vs. Fitted plot are randomly scattered on the plot that verifies the independence condition;
2. The Normal Q-Q plot consists of the points which mostly fall on the line indicating that the residuals are normally distributed;
3. The Scale-Location plot consists of points scattered in a constant band pattern, indicating constant variance.

Model comparison

Now let's compare the simple model with only am as the predictor variable and the best model which we obtained above containing confounder variables.

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

```
anova(simpleModel, bestModel)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the above results, the p-value obtained is highly significant and we reject the null hypothesis that the confounder variables cyl, hp and wt don't contribute to the accuracy of the model.

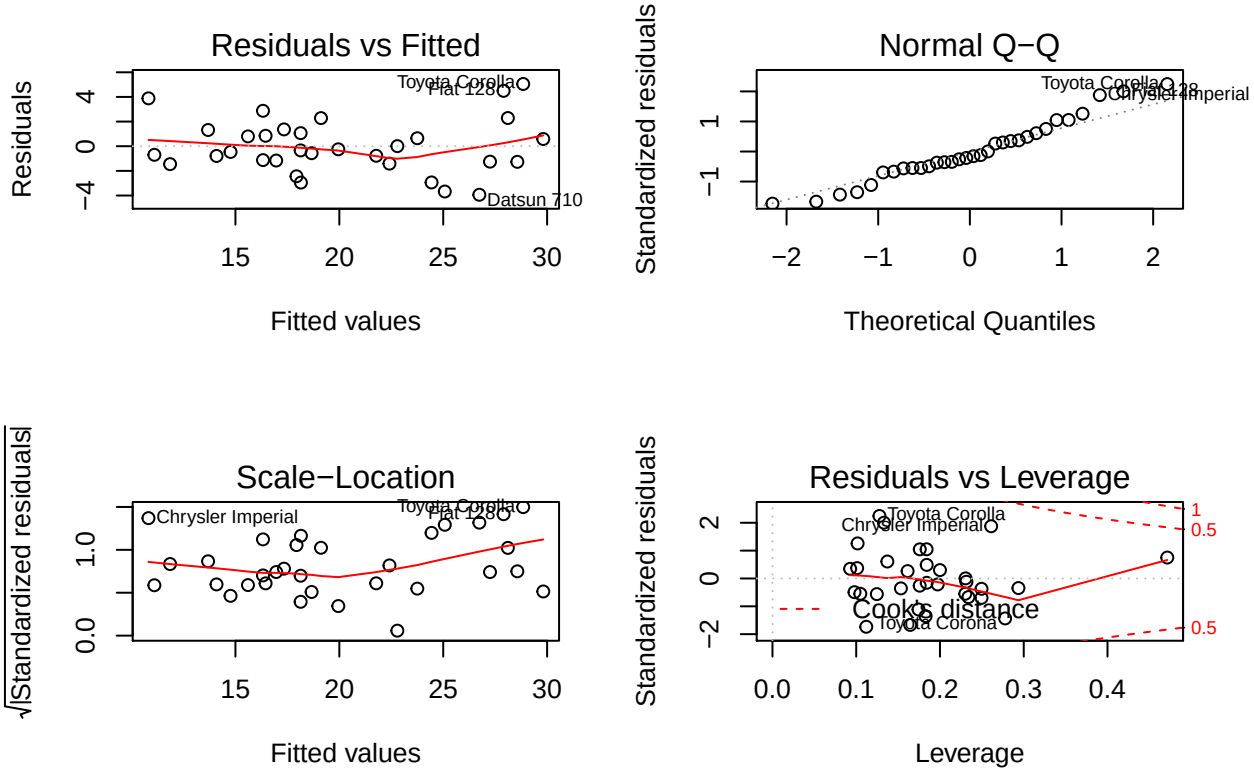
Conclusion

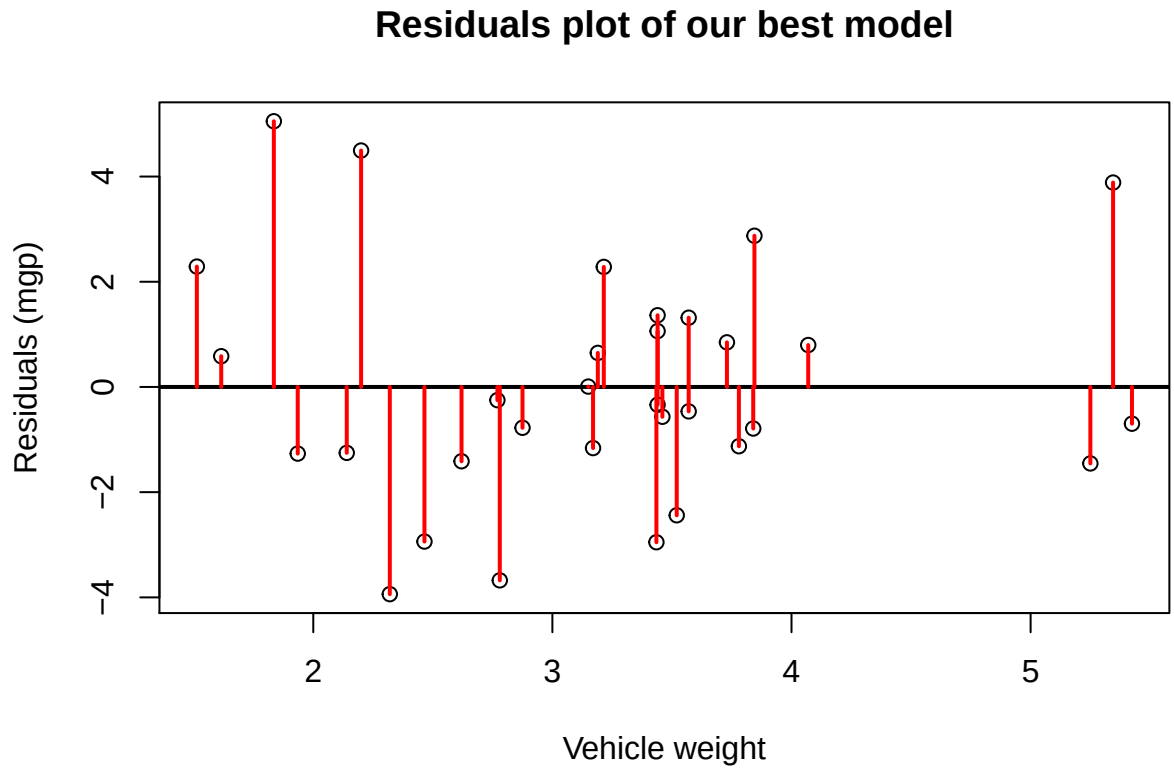
Based on the observations from our best model, we can conclude the following,

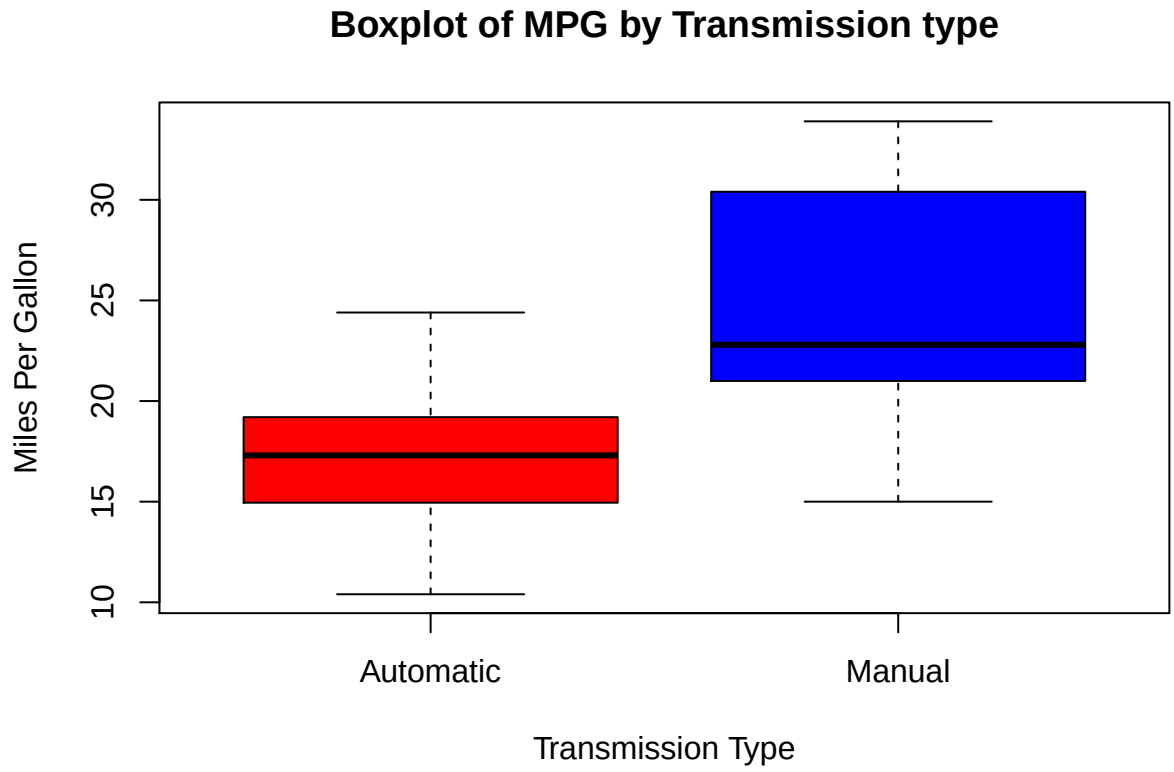
- cars with Manual transmission get more miles per gallon mpg compared to cars with Automatic transmission. (1.8 adjusted by hp, cyl, and wt);
- mpg will decrease by 2.5 (adjusted by hp, cyl, and am) for every 1000 lb increase in wt;
- mpg decreases negligibly with increase of hp;
- if number of cylinders, cyl increases from 4 to 6 and 8, mpg will decrease by a factor of 3 and 2.2 respectively (adjusted by hp, wt, and am).

Extra plots - appendix

Residual plots







Pair plots for mtcars dataset