Regression Models Course Project

Motor Trend Car Analysis

Executive Summary

Motor Trend magazine, looking at a data set of a collection of cars, is particulary 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:

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

Data Processing

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).

A data frame with 32 observations on 11 variables.

1. mpg Miles/(US) gallon
2. cyl Number of cylinders
3. disp Displacement (cu.in.)
4. hp Gross horsepower
5. drat Rear axle ratio
6. wt Weight (1000 lbs)
7. qsec 1/4 mile time

8. vs V/S
9. am Transmission (0 = automatic, 1 = manual)
10. gear Number of forward gears
11. carb Number of carburetors

Answering the questions

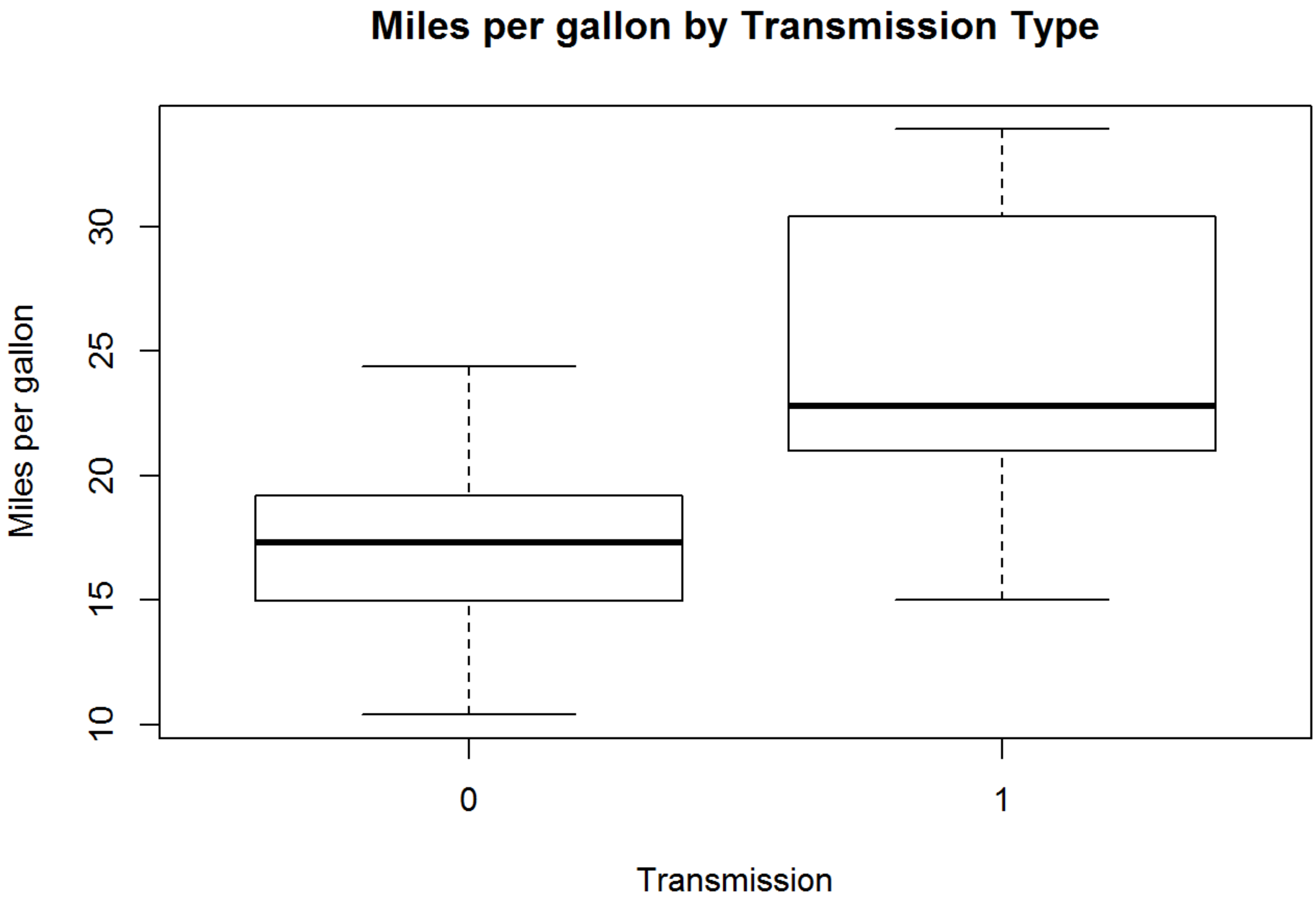
```
aggregate(mpg~am, data = mtcars, mean)
```

```
##   am      mpg
## 1  0 17.14737
## 2  1 24.39231
```

The mean transmission for manual is 7.24mpg higher than automatic.

Now let’s try to plot it:

```
boxplot(mpg ~ am, data = mtcars, xlab = "Transmission", ylab = "Miles per gallon", main="Miles per gallon by Transmission Type")
```



Now, let’s test whether this difference in mean values is significant:

```
auto <- mtcars[mtcars$am == 0,]
manual <- mtcars[mtcars$am == 1,]
```

```
t.test(auto$mpg, manual$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  auto$mpg and manual$mpg
## 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 of x mean of y
##  17.14737  24.39231
```

p-value=0.001374. That means that null hypeothesis that difference is not significant is hardly probable.

Now, let's build basic linear regression:

```
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
```

We see that automatic transmission runs at 17.147 mpg while manual transmission is 7.245 mpg more.

However, our R2 is only 0.36 so let's try to add more variables into model:

```
mvfit <- lm(mpg~am + wt + hp + cyl, data = mtcars)
summary(mvfit)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + hp + cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4765 -1.8471 -0.5544  1.2758  5.6608
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.14654    3.10478  11.642 4.94e-12 ***
## am           1.47805    1.44115   1.026   0.3142
## wt          -2.60648    0.91984  -2.834   0.0086 **
## hp          -0.02495    0.01365  -1.828   0.0786 .
```

```
## cyl         -0.74516    0.58279  -1.279   0.2119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.509 on 27 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.8267
## F-statistic: 37.96 on 4 and 27 DF,  p-value: 1.025e-10
```

We see that multi-variable model explains 84.9% of variance.

It may be concluded that on average, manual transmissions have 1.478 more mpg than automatic.

Let's test two models with anova test:

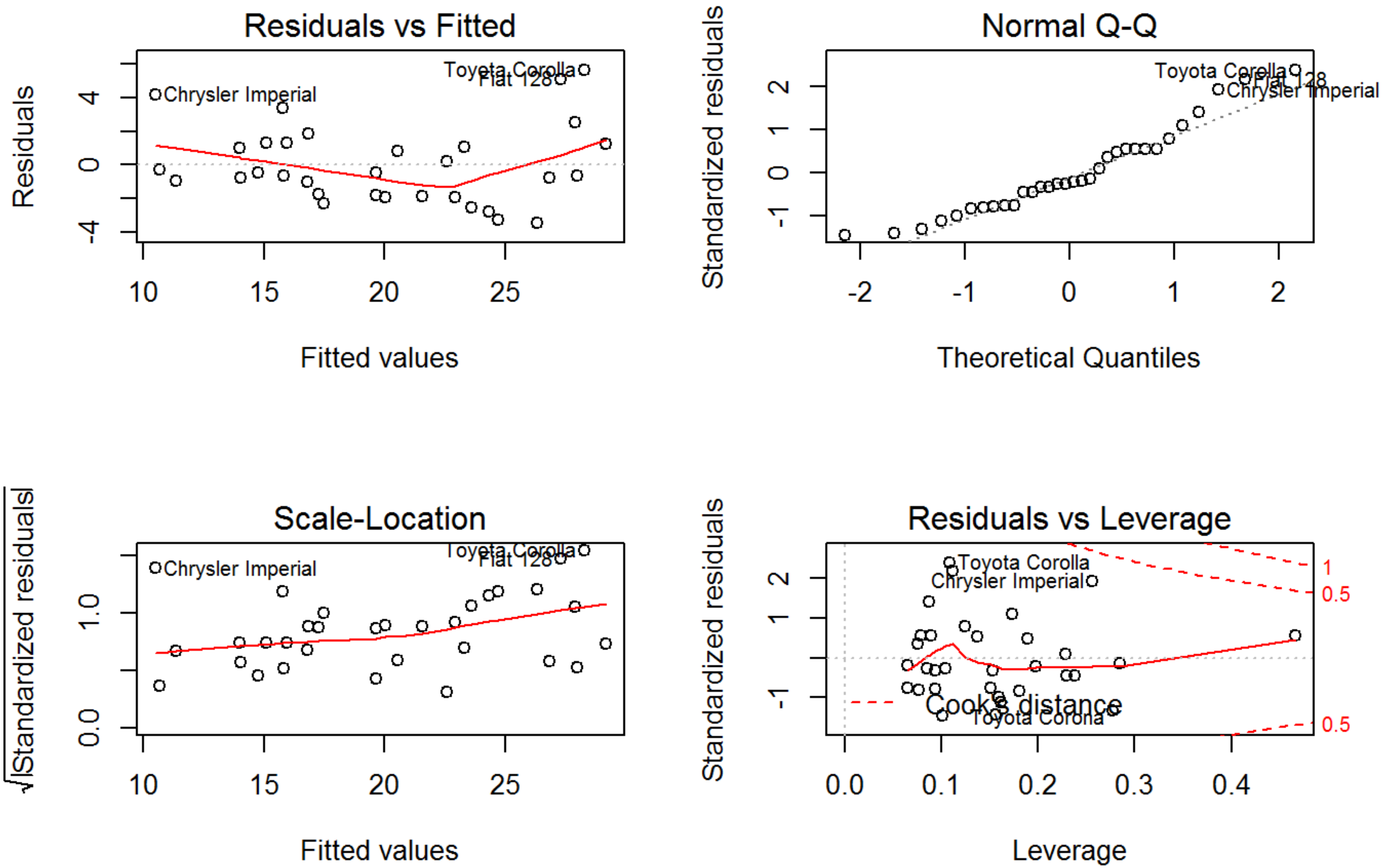
```
anova(fit,mvfit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + hp + cyl
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.9
## 2     27 170.0  3     550.9 29.166 1.274e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the low p-value we see that new model is appropriate.

Appendix

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



Residuals vs Fitted and Scale-Location plots show no pattern. Normal Q-Q plot indicates that Residuals approximately follow a Normal distributions.