Regression Models Course Project

Conor Breen

15 September 2015

Regression models analysis of mtcars dataset

NOTE: Produced in RMarkdown and Knitr via MSWord. .Rmd file and R Script available at https://github.com/cbreen66/Regression-Models

Executive summary

Analysing the mtcars dataset, the objective was to answer the questions "Is an automatic or manual transmission better for MPG?" and "Quantify the MPG difference between automatic and manual transmissions".

Two methods were employed. Firstly, exploratory data analysis examined initial relationships between MPG and transmission type. Regression modelling where causality is assumed was used to fully explore the relationship between transmission type, MPG and the other dataset variables.

Key findings were that cars with a manual transmission have get 24.4 miles per gallon on average compared to 17.1 miles per gallon for cars with an automatic transmission. Thus, manual transmission cars are more fuel efficient on average when no other variables are taken into account.

The most parsimonious regression model was found to be one with transmission type, car weight and quarter mile time in seconds (qsec). Holding weight and qsec constant, manual transmission cars get 2.94 miles per gallon more on average than automatic transmission cars. However, both weight and qsec have moderating effects on this relationship.

Introduction

This project involved exploratory analysis and regression modelling of the mtcars dataset which measures fuel consumption (in miles per gallon) and nine other car design features and performance measurements. Please note all figures are in the appendices.

The transmission type variable (am) was recoded where 1="manual" and 0="automatic" and converted to a factor.

```
data(mtcars)
mtcars$am <- ifelse(mtcars$am == 1, c("manual"), c("automatic"))
mtcars$am <- as.factor(mtcars$am)
```

Exploratory analysis

A t-test shows that manual cars perform better than automatic cars, getting 24.4 mpg compared to 17.1 mpg for automatic cars (see boxplot). With a p-value of 0.001374, we can safely reject the null hypothesis that the MPG of the automatic and manual transmissions are from the same population (assuming normal distribution).

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

##
##  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
```

We can note also that an additional t-test confirms that automatic transmission cars have a significantly greater mean weight than manual cars.

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

##
##  Welch Two Sample t-test
##
## data:  wt by am
## t = 5.4939, df = 29.234, p-value = 6.272e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8525632 1.8632262
## sample estimates:
## mean in group automatic    mean in group manual
##                3.768895                2.411000
```

Regression modelling

A linear regression model was tested which included mpg as the outcome and all 10 other variables in the mtcars dataset as predictors. A summary of this model shows an adjusted r-squared of 0.8066, indicating that the model can explain 81% of the variance in the dataset. However, the p-values indicate no significant explanatory effect of any of the variables. Furthermore, an analysis of variance inflation factors indicates significant multicollinearity.

```
Allin <- lm(mpg ~ ., data=mtcars)
summary(Allin)
```

```
##
## 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
## ammanual     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
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.2.2
```

```
vif(Allin)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873
##        am      gear      carb
##  4.648487  5.357452  7.908747
```

A backward stepwise approach was used to identify the most parsimonious model of statistically significant regressors using AIC as a means for model selection.

```
Model1 <- step(Allin, k=log(nrow(mtcars)))
```

The selected model with mpg as outcome has weight, qsec and transmission type as predictors.

A nested model testing method shows that all three predictors are significant.

```
fit1 <- lm(mpg ~ wt, data = mtcars)
fit2 <- update(fit1, mpg ~ wt + qsec)
```

```r
fit3 <- update(fit2, mpg ~ wt + qsec + am)
anova(fit1, fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + qsec
## Model 3: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 278.32
## 2     29 195.46  1    82.858 13.7048 0.0009286 ***
## 3     28 169.29  1    26.178  4.3298 0.0467155 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Measurement of the adjusted r-squared value shows that this model explains 83% of the variance.

```r
summary(Model1)
```

```
##
## 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 ***
## ammanual      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
```

Holding weight and qsec constant, manual transmission cars get 2.94 miles per gallon more on average than automatic transmission cars. However, both weight and qsec have moderating effects on this relationship. As weight increases, car performance in terms of miles per gallon decreases. As qsec increases, car performance also increases.

Diagnostic testing

Summary of variance inflation factors indicates the best fit regression model shows no significant multicollinearity. The mean of the residuals is close to zero.

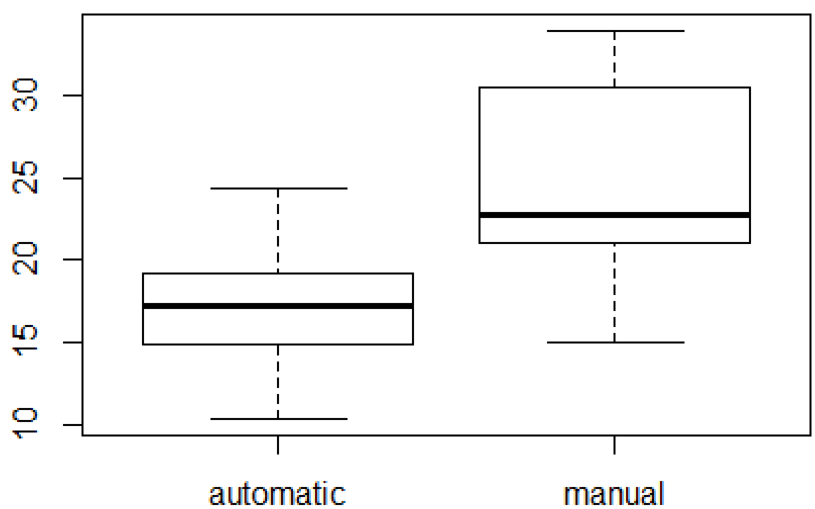
```r
vif(Model1)

##       wt     qsec       am
## 2.482952 1.364339 2.541437

mean(Model1$residuals)

## [1] 5.20417e-17
```

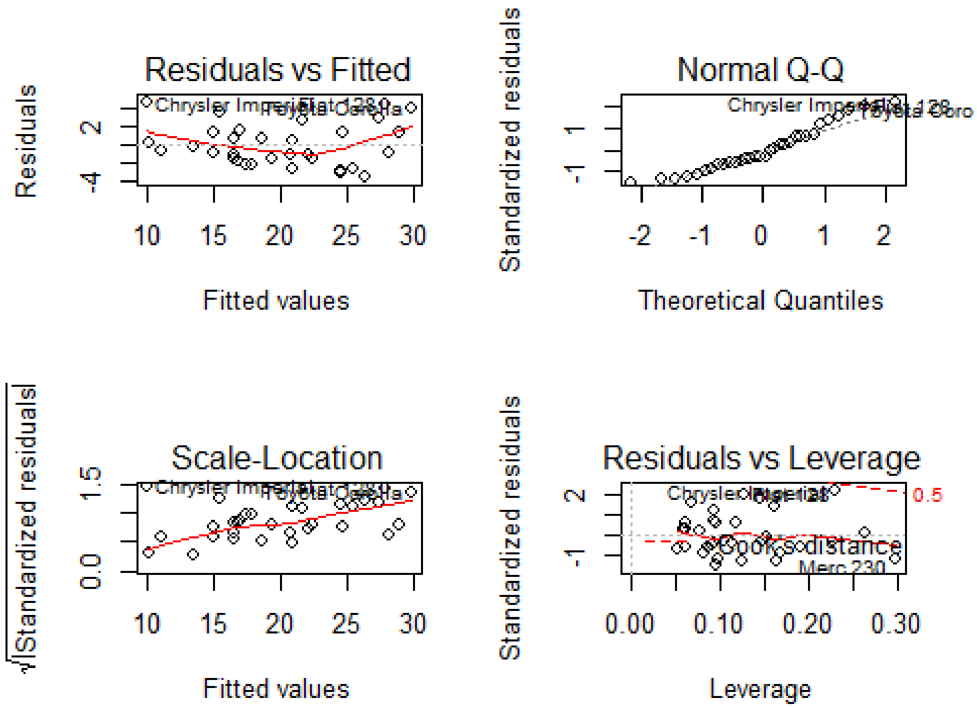
Four diagnostic plots were generated (see Appendix) and these show no systematic relationship between fitted values and residuals or standardised residuals. The residuals show a broadly normal distribution with some outliers.

Conclusion

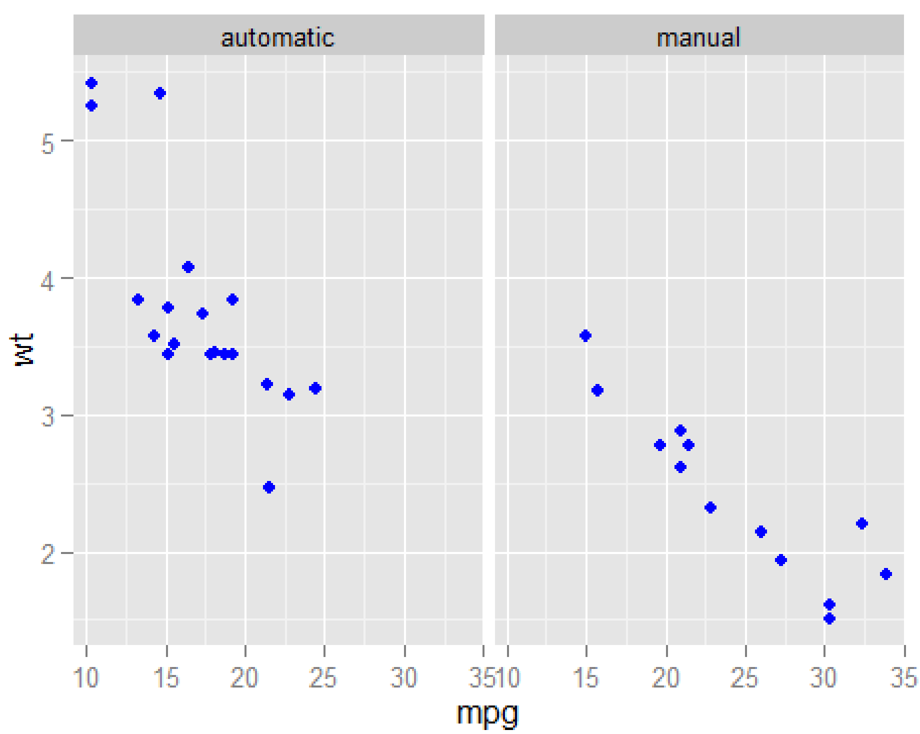
Using the small mtcars dataset, we can conclude that manual transmission cars have a better performance in terms of miles per gallon than automatics. However, weight and quarter mile time in seconds have moderating effects on this relationship. The lighter automatic transmission cars perform as well as the heavier manual transmission cars.

Appendix

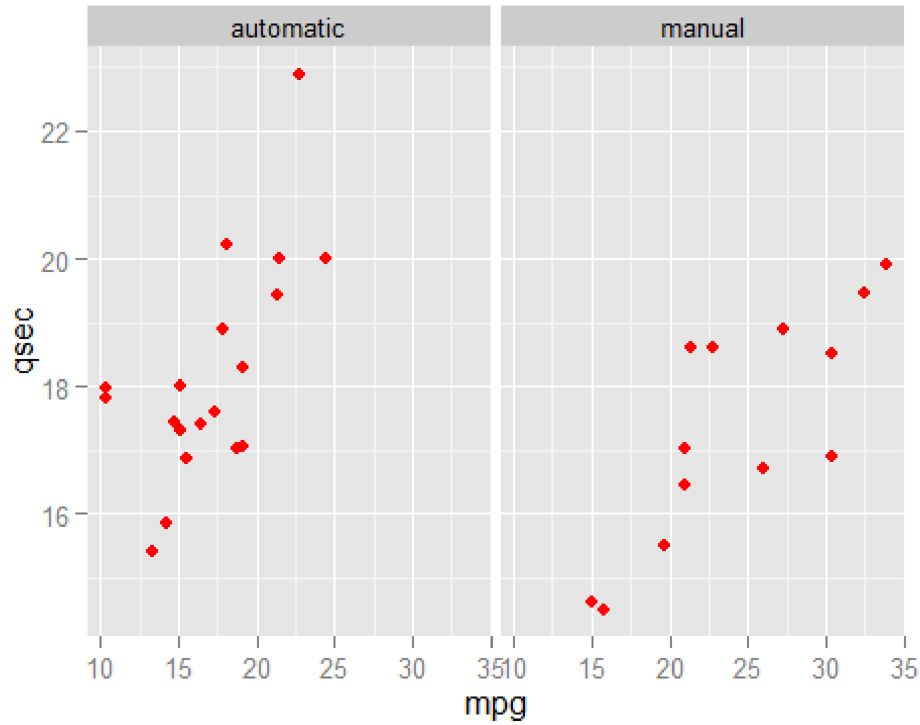




Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
Merc 230
0.5



automatic
manual
wt
mpg



automatic
manual
qsec
22
20
18
16
10
15
20
25
30
35
mpg