Regression Analysis

Michael Bristow

10 December 2017

R Markdown

Regression Models - Peer Assesement 1

Issue/Problem

Take the mtcars data set and write up an analysis to answer their question using regression models and exploratory data analyses.

Your report must be:

Written as a PDF printout of a compiled (using knitr) R markdown document.

Brief. Roughly the equivalent of 2 pages or less for the main text.

Supporting figures in an appendix can be included up to 5 total pages including the 2 for the main report. The appendix can only include figures. Include a first paragraph executive summary.

Executuve Summary

This report looks at historic data and trys to find drivers for MPG. It will look at trwo main items

- Is an automatic or manual transmission better for MPG

- Quantify the MPG difference between automatic and manual transmissions

Data Processing/Analysis

First load the data

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.3
```

```
data(mtcars)
xx <- mtcars
#convert "am"to a factor
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <-c("AT", "MT")

summary(mtcars)
```

```
##       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           am    
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000   AT:19  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000   MT:13  
##  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          
##       gear            carb      
##  Min.   :3.000   Min.   :1.000  
##  1st Qu.:3.000   1st Qu.:2.000  
##  Median :4.000   Median :2.000  
##  Mean   :3.688   Mean   :2.812  
##  3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :5.000   Max.   :8.000  
```

```r
head(mtcars)
```

```
##                    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 MT    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0 MT    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1 MT    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1 AT    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0 AT    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1 AT    3    1
```

Apply regsubsets to find best variable selection

```r
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 3.4.3
```

```r
reg.best<- regsubsets(mpg~.,mtcars,nvmax=5)
summary(reg.best)
```

```
## Subset selection object
## Call: regsubsets.formula(mpg ~ ., mtcars, nvmax = 5)
## 10 Variables  (and intercept)
##      Forced in Forced out
## cyl      FALSE      FALSE
## disp     FALSE      FALSE
## hp       FALSE      FALSE
## drat     FALSE      FALSE
## wt       FALSE      FALSE
## qsec     FALSE      FALSE
## vs       FALSE      FALSE
## amMT     FALSE      FALSE
## gear     FALSE      FALSE
## carb     FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          cyl disp hp  drat wt  qsec vs  amMT gear carb
## 1  ( 1 ) " " " "  " " " "  "*" " "  " " " "  " "  " " 
## 2  ( 1 ) "*" " "  " " " "  "*" " "  " " " "  " "  " " 
## 3  ( 1 ) " " " "  " " " "  "*" "*"  " " "*"  " "  " " 
## 4  ( 1 ) " " " "  "*" " "  "*" "*"  " " "*"  " "  " "
## 5  ( 1 ) " " "*"  "*" " "  "*" "*"  " " "*"  " "  " "
```

```
reg.summary<- summary(reg.best)
names(reg.summary)
```

```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

Look at the best r squared

```
reg.summary$rsq
```

```
## [1] 0.7528328 0.8302274 0.8496636 0.8578510 0.8637377
```

As can be seen, “weight” is the biggest “driver” of mpg followed by number of cylinders.

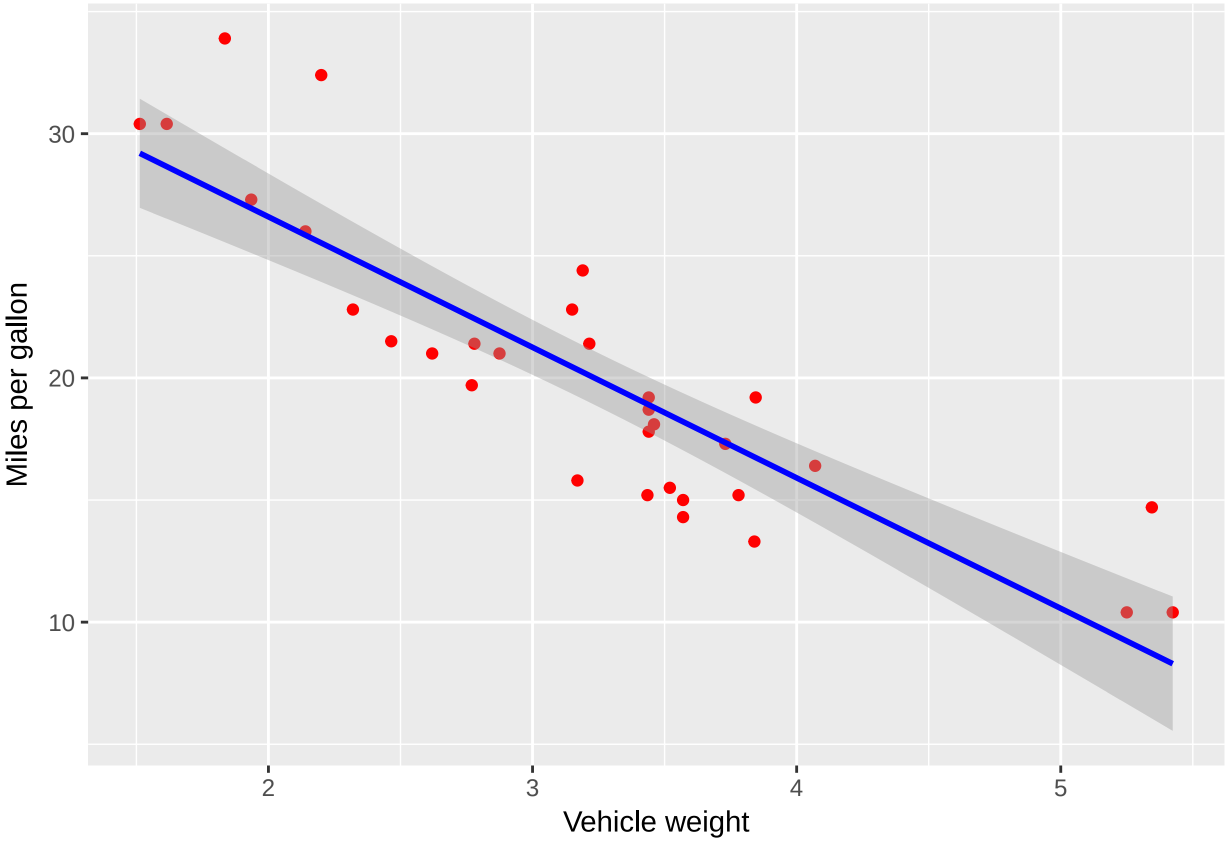
Our model should include

- weight
- cylinders

By adding more variaqbles, the r-squared only increases by small amounts and we need to be wary of overfitting

Plot weight against mpg and show regression line

```
ggplot(data = xx, aes(x = wt, y = mpg)) +
  geom_point(color='red') +
  geom_smooth(method = "lm", color = "blue") +
  labs(x = "Vehicle weight") +
  labs(y = "Miles per gallon")
```



As can be seen there is a direct correlation between the vehicle weight and it mpg

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

```
##
## Call:
## lm(formula = mpg ~ wt + cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2893 -1.5512 -0.4684  1.5743  6.1004
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.6863     1.7150  23.141  < 2e-16 ***
## wt           -3.1910     0.7569  -4.216 0.000222 ***
## cyl          -1.5078     0.4147  -3.636 0.001064 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.568 on 29 degrees of freedom
## Multiple R-squared:  0.8302, Adjusted R-squared:  0.8185
## F-statistic: 70.91 on 2 and 29 DF,  p-value: 6.809e-12
```

We can see that each extra ton in weight changes MPG by -3.1909721 when we use the full model using weight and cylinders

Effect of Transmission type on MPG

Question is, which has better MPG - Manual or Automatic We do this by running a regression using ONLY transmission type

```
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 ***
## amMT           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
```

It shows that on average

- a car has 17.147 mpg with automatic transmission, and if it is manual transmission, 7.245 mpg is increased.
- has the Residual standard error as 4.902 on 30 degrees of freedom.
- the Adjusted R-squared value is 0.3385, which means that the model can explain about 34% of the variance of the MPG variable.
- The low Adjusted R-squared value also indicates that we need to add other variables to the model.

Conclusion

Here are the conclusions fromn the analysis

- The primary driver of MPG is the cars weight
- The "optimal" model is to inlcude weight and cylinders
- By adding more variables, the r-quared value only increases marginally and we are in danger of overfitting the model
- Transmission typoe also has an efefct on MPG
- Having a manual transmission increases MPG by 7.245