Regression Models Course Project

nkvasg

Sunday, March 22, 2015

Executive Summary

Motor Trend, a magazine about the automobile industry is interested in exploring the relationship between a set of variables and miles per gallon (MPG). The report will explore data from the mtcars dataset to answer the 2 questions below:

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

The data from the dataset mtcars 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).

This report will use exploratory data analysis and regressional models to find the answers to the two questions above.

The t-test shows that the mean MPG of manual transmitted cars is approximately ~ 7 more than that of auto cars, with no other car design feature taken into account.

Several linear regression models were fitted to the data and the model with highest Adjusted R-squared value was selected. There is an interaction between car weight and transmission type on MPG. The model suggested that cars lower in weight with a manual transmission and those higher in weight with an automatic transmission will have higher MPG values.

Exploring Data

```
data(mtcars)
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  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
```

Cleaning and Preparing Data

```
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
levels(mtcars$am) <- c("Auto", "Manual")
```

Graphics

We will use some data visualization first to perform exploratory data analysis.

Boxplot of the variable "mpg" between Auto and Manual cars is presented as Figure 1 in Appendix. Also included a plot of the relationships between al variables of the dataset in Figure 2.

The scatter plot indicates there might be an interaction between "wt" and "am" variables.

Statistical Inference

We use t-test here to test if the mpg means are different, assuming that the milleage data has a normal distribution

```
t_test <- t.test(mtcars$mpg ~ mtcars$am)
t_test
```

```
##
##  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 Auto mean in group Manual
##             17.14737             24.39231
```

Regression Analysis

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

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## 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
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

This model has the Residual SE of 2.833 on 15 df. And the Adjusted R-squared value is 0.779, which means that the model can explain about 78% of the variance of the MPG variable. However, none of the coefficients are significant at 0.05 significant level.

We will therefore use backward selection to select some statistically significant variables.

```
stepmodel <- step(allmodel, k=log(nrow(mtcars)))
```

```
## Start:  AIC=101.32
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS     AIC
## - carb  5   13.5989 134.00  87.417
## - gear  2    3.9729 124.38  95.428
## - cyl   2   10.9314 131.33  97.170
## - am    1    1.1420 121.55  98.157
## - qsec  1    1.2413 121.64  98.183
## - drat  1    1.8208 122.22  98.335
## - vs    1    3.6299 124.03  98.806
## - disp  1    9.9672 130.37 100.400
## <none>              120.40 101.321
## - wt    1   25.5541 145.96 104.014
## - hp    1   25.6715 146.07 104.040
```

```
##
## Step:  AIC=87.42
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  2    5.0215 139.02 81.662
## - cyl   2   12.5642 146.57 83.353
## - disp  1    0.9934 135.00 84.187
## - drat  1    1.1854 135.19 84.233
## - vs    1    3.6763 137.68 84.817
## - qsec  1    5.2634 139.26 85.184
## - am    1   11.9255 145.93 86.679
## <none>              134.00 87.417
## - wt    1   19.7963 153.80 88.360
## - hp    1   22.7935 156.79 88.978
##
## Step:  AIC=81.66
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - cyl   2   10.4247 149.45 77.045
## - drat  1    0.9672 139.99 78.418
## - disp  1    1.5483 140.57 78.551
## - vs    1    2.1829 141.21 78.695
## - qsec  1    3.6324 142.66 79.022
## <none>              139.02 81.662
## - am    1   16.5665 155.59 81.799
## - hp    1   18.1768 157.20 82.129
## - wt    1   31.1896 170.21 84.674
##
## Step:  AIC=77.04
## mpg ~ disp + hp + drat + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - vs    1     0.645 150.09 73.717
## - drat  1     2.869 152.32 74.187
## - disp  1     9.111 158.56 75.473
## - qsec  1    12.573 162.02 76.164
## - hp    1    13.929 163.38 76.431
## <none>              149.45 77.045
## - am    1    20.457 169.91 77.684
## - wt    1    60.936 210.38 84.523
##
## Step:  AIC=73.72
## mpg ~ disp + hp + drat + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1     3.345 153.44 70.956
## - disp  1     8.545 158.64 72.023
## - hp    1    13.285 163.38 72.965
```

```
## <none>              150.09 73.717
## - am    1    20.036 170.13 74.261
## - qsec  1    25.574 175.67 75.286
## - wt    1    67.572 217.66 82.146
##
## Step:  AIC=70.96
## mpg ~ disp + hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1     6.629 160.07 68.844
## - hp    1    12.572 166.01 70.011
## <none>              153.44 70.956
## - qsec  1    26.470 179.91 72.583
## - am    1    32.198 185.63 73.586
## - wt    1    69.043 222.48 79.380
##
## Step:  AIC=68.84
## mpg ~ hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - hp    1     9.219 169.29 67.170
## <none>              160.07 68.844
## - qsec  1    20.225 180.29 69.186
## - am    1    25.993 186.06 70.193
## - wt    1    78.494 238.56 78.147
##
## Step:  AIC=67.17
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              169.29 67.170
## - am    1    26.178 195.46 68.306
## - qsec  1   109.034 278.32 79.614
## - wt    1   183.347 352.63 87.187
```

```
summary(stepmodel)
```

```
##
## 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
```

This model is "mpg ~ wt + qsec + am". It has the Residual SE of 2.459 on 28 df. And the Adjusted R-squared value is 0.8336, which means that the model can explain about 83% of the variance of the MPG variable. All the coefficients are significant at 0.05 significant level.

Because there is an interaction between "wt" and "am"

```
intmodel <- lm(mpg ~ wt + qsec + am + wt:am, data=mtcars)
summary(intmodel)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am + wt:am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.110893
## wt            -2.937      0.666  -4.409 0.000149 ***
## qsec           1.017      0.252   4.035 0.000403 ***
## amManual      14.079      3.435   4.099 0.000341 ***
## wt:amManual   -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

This model has the Residual SE of 2.084 on 27 df. And the Adjusted R-squared value is 0.8804, which means that the model can explain about 88% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level. This is the best model so far.

Next, we fit the simple model with MPG as the outcome variable and Transmission as the predictor variable.

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

```
##
## 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 ***
## amManual       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 approximately 17 mpg with automatic transmission, and if it is manual transmission, ~ 7 mpg is increased. This model has the Residual SE of 4.902 on 30 df. And 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.

We select the model with the interactive term between weight and transmission type as it gives the highest Adjusted R-squared value.

Residuals Analysis and Diagnostic

The Residual vs Fitted plot shows no consistent pattern. The Normal Q-Q plot indicates that the residuals are normally distributed. The Scale-Location plot has the points randomly distributed. We can compute and plot the levarage of each point. Last figure in the Appendix. The plot indicates that no significant outlier detected.

We also perform Dfbetas, the measure of how much an observation has effected the estimate of a regression coefficient:

```
influence <- dfbetas(intmodel)
summary(influence)
```

```
##   (Intercept)             wt                  qsec
##  Min.   :-0.488522   Min.   :-0.6772453   Min.   :-0.353637
##  1st Qu.:-0.088311   1st Qu.:-0.0384492   1st Qu.:-0.094013
##  Median :-0.003069   Median : 0.0045787   Median :-0.015935
##  Mean   : 0.001302   Mean   : 0.0001358   Mean   :-0.001941
##  3rd Qu.: 0.093811   3rd Qu.: 0.0390178   3rd Qu.: 0.075670
##  Max.   : 0.402406   Max.   : 0.9769875   Max.   : 0.536079
##     amManual          wt:amManual
##  Min.   :-0.398133   Min.   :-0.486475
```

```
##  1st Qu.:-0.090726   1st Qu.:-0.055569
##  Median : 0.012645   Median :-0.009068
##  Mean   :-0.001017   Mean   : 0.002509
##  3rd Qu.: 0.061173   3rd Qu.: 0.029015
##  Max.   : 0.608409   Max.   : 0.530727
```

We find the influential observations by selecting the ones with a dfbeta > 1 in magnitude.

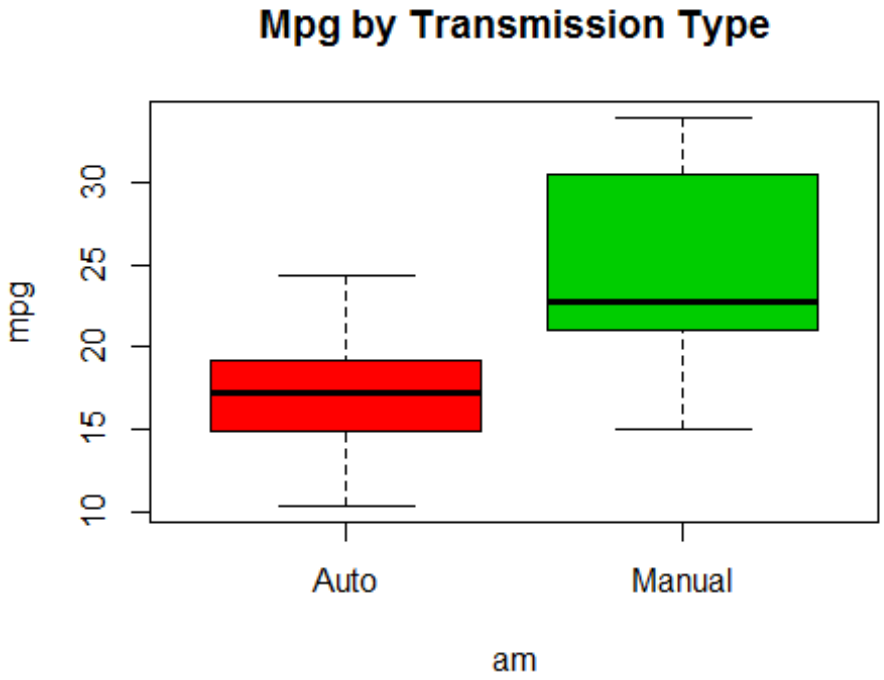
```
influence[sum(abs(influence)>1)]
```

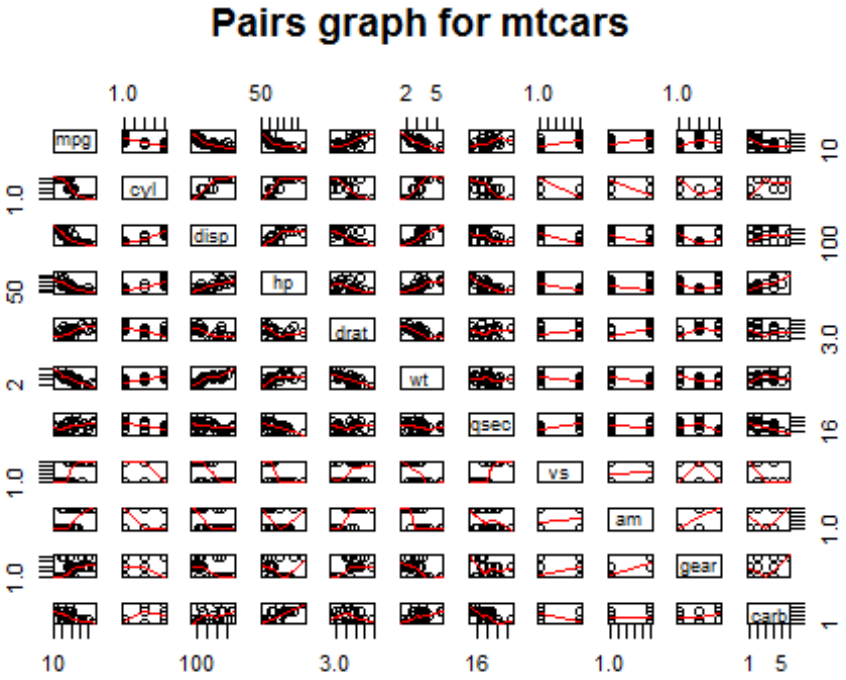
```
## numeric(0)
```

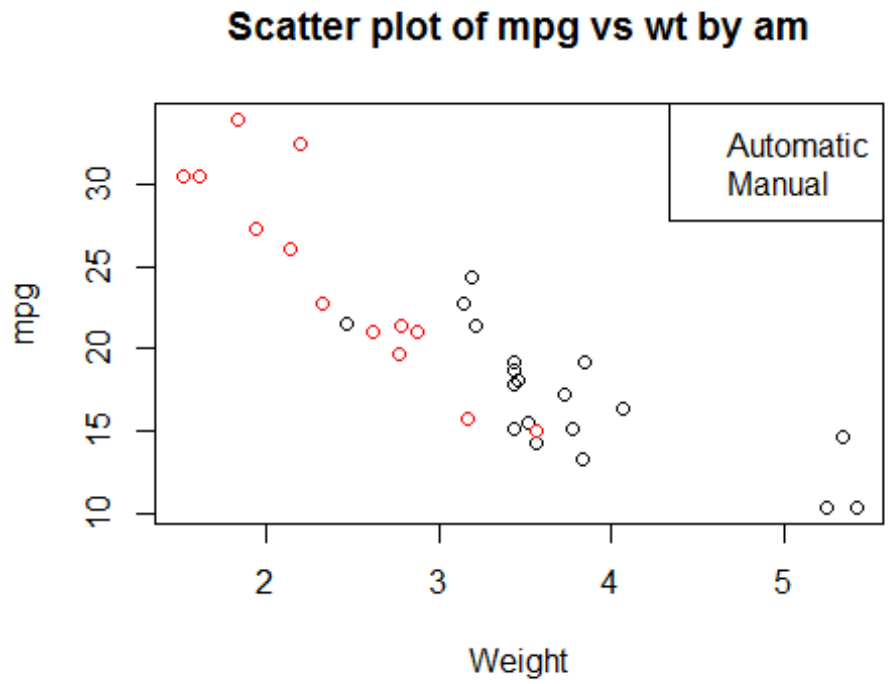
Conclusion

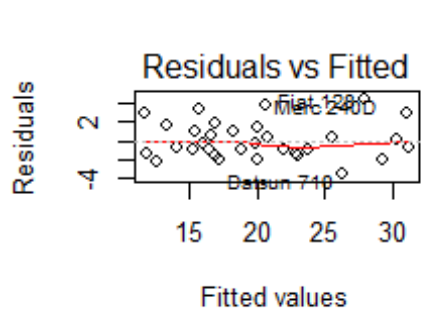
The results from the above analyses indicates that the model selected (intmodel), with an interactive term between car weight and transmission type is good to answer the question.

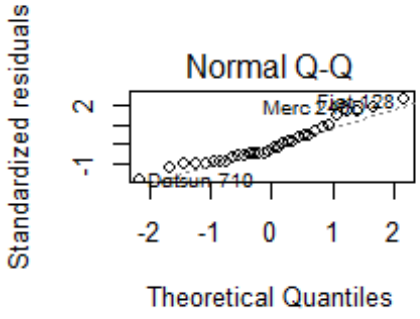
Appendix

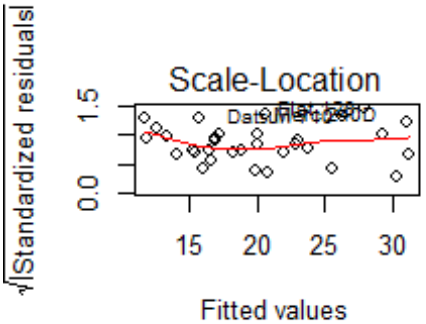

Mpg by Transmission Type
mpg
10
15
20
25
30
Auto
Manual
am

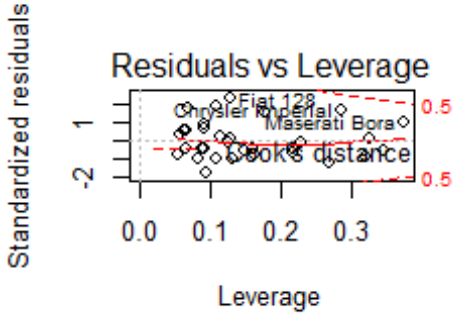

Pairs graph for mtcars
mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb

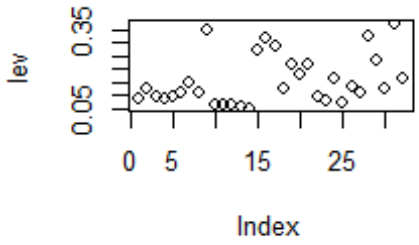

Scatter plot of mpg vs wt by am
Automatic
Manual
mpg
10
15
20
25
30
2
3
4
5
Weight


Residuals vs Fitted
Residuals
Fiat 128
Merc 240D
Datsun 710
-4
2
15
20
25
30
Fitted values


Normal Q-Q
Standardized residuals
Fiat 128
Datsun 710
-1
2
-2
-1
0
1
2
Theoretical Quantiles


Scale-Location
√|Standardized residuals|
Datsun 710
0.0
1.5
15
20
25
30
Fitted values


Residuals vs Leverage
Standardized residuals
Fiat 128
Chrysler Imperial
Maserati Bora
Cook's distance
0.5
0.5
-2
1
0.0
0.1
0.2
0.3
Leverage


lev
0.05
0.35
0
5
15
25
Index