Motor Trend MPG Data Analysis created with knitr

Executive Summary:

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

Load Data

```
> # Load Data
> library(ggplot2)
> data(mtcars)
> head(mtcars, n=3)
               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
> dim(mtcars)
[1] 32 11
> mtcars$cyl <- as.factor(mtcars$cyl)
> mtcars$vs <- as.factor(mtcars$vs)
> mtcars$am <- factor(mtcars$am)
> mtcars$gear <- factor(mtcars$gear)
> mtcars$carb <- factor(mtcars$carb)
> attach(mtcars)
The following objects are masked from mtcars (pos = 3):

    am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt

The following object is masked from package:ggplot2:

    mpg
```

Exploratory Analysis

```
> # T-test transmission type and MPG
> testResults <- t.test(mpg ~ am)
> testResults$p.value
[1] 0.001373638
```

Comment: Since P-value is less than 0.05, we reject the null hypothesis which says that the difference between transmission types is 0.

```
> testResults$estimate
mean in group 0 mean in group 1
       17.14737        24.39231
```

Comment: The difference estimate between the two transmissions is 7.24494 (24.39231 – 17.14737).

```
> # Regrssion analysis
> fullModelFit <- lm(mpg ~ ., data = mtcars)
> summary(fullModelFit)

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
am1          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

> summary(fullModelFit)$coeff
               Estimate  Std. Error     t value   Pr(>|t|)
(Intercept) 23.87913244 20.06582026  1.19004018 0.25252548
cyl6        -2.64869528  3.04089041 -0.87102622 0.39746642
cyl8        -0.33616298  7.15953951 -0.04695316 0.96317000
disp         0.03554632  0.03189920  1.11433290 0.28267339
hp          -0.07050683  0.03942556 -1.78835344 0.09393155
drat         1.18283018  2.48348458  0.47627845 0.64073922
```

```
wt          -4.52977584  2.53874584 -1.78425732 0.09461859
qsec         0.36784482  0.93539569  0.39325050 0.69966720
vs1          1.93085054  2.87125777  0.67247551 0.51150791
am1          1.21211570  3.21354514  0.37718957 0.71131573
gear4        1.11435494  3.79951726  0.29328856 0.77332027
gear5        2.52839599  3.73635801  0.67670068 0.50889747
carb2       -0.97935432  2.31797446 -0.42250436 0.67865093
carb3        2.99963875  4.29354611  0.69863900 0.49546781
carb4        1.09142288  4.44961992  0.24528452 0.80956031
carb6        4.47756921  6.38406242  0.70136677 0.49381268
carb8        7.25041126  8.36056638  0.86721532 0.39948495
```

Comment: None of the coefficients have P-values less than 0.05. We do not have strong evidence to conclude which variable is statistically significant.

So we have to use backward selection to determine which variables are statistically significant.

```
> # Backward selection
> stepFit <- step(fullModelFit)
Start:  AIC=76.4
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq    RSS    AIC
- carb  5   13.5989 134.00 69.828
- gear  2    3.9729 124.38 73.442
- am    1    1.1420 121.55 74.705
- qsec  1    1.2413 121.64 74.732
- drat  1    1.8208 122.22 74.884
- cyl   2   10.9314 131.33 75.184
- vs    1    3.6299 124.03 75.354
<none>              120.40 76.403
- disp  1    9.9672 130.37 76.948
- wt    1   25.5541 145.96 80.562
- hp    1   25.6715 146.07 80.588

Step:  AIC=69.83
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear

       Df Sum of Sq    RSS    AIC
- gear  2    5.0215 139.02 67.005
- disp  1    0.9934 135.00 68.064
- drat  1    1.1854 135.19 68.110
- vs    1    3.6763 137.68 68.694
- cyl   2   12.5642 146.57 68.696
```

```
- qsec  1    5.2634 139.26 69.061
<none>              134.00 69.828
- am    1   11.9255 145.93 70.556
- wt    1   19.7963 153.80 72.237
- hp    1   22.7935 156.79 72.855

Step:  AIC=67
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am

       Df Sum of Sq    RSS    AIC
- drat  1    0.9672 139.99 65.227
- cyl   2   10.4247 149.45 65.319
- disp  1    1.5483 140.57 65.359
- vs    1    2.1829 141.21 65.503
- qsec  1    3.6324 142.66 65.830
<none>              139.02 67.005
- am    1   16.5665 155.59 68.608
- hp    1   18.1768 157.20 68.937
- wt    1   31.1896 170.21 71.482

Step:  AIC=65.23
mpg ~ cyl + disp + hp + wt + qsec + vs + am

       Df Sum of Sq    RSS    AIC
- disp  1    1.2474 141.24 63.511
- vs    1    2.3403 142.33 63.757
- cyl   2   12.3267 152.32 63.927
- qsec  1    3.1000 143.09 63.928
<none>              139.99 65.227
- hp    1   17.7382 157.73 67.044
- am    1   19.4660 159.46 67.393
- wt    1   30.7151 170.71 69.574

Step:  AIC=63.51
mpg ~ cyl + hp + wt + qsec + vs + am

       Df Sum of Sq    RSS    AIC
- qsec  1     2.442 143.68 62.059
- vs    1     2.744 143.98 62.126
- cyl   2    18.580 159.82 63.466
<none>              141.24 63.511
- hp    1    18.184 159.42 65.386
- am    1    18.885 160.12 65.527
- wt    1    39.645 180.88 69.428
```

```
Step:  AIC=62.06
mpg ~ cyl + hp + wt + vs + am

       Df Sum of Sq    RSS    AIC
- vs    1     7.346 151.03 61.655
<none>              143.68 62.059
- cyl   2    25.284 168.96 63.246
- am    1    16.443 160.12 63.527
- hp    1    36.344 180.02 67.275
- wt    1    41.088 184.77 68.108

Step:  AIC=61.65
mpg ~ cyl + hp + wt + am

       Df Sum of Sq    RSS    AIC
<none>              151.03 61.655
- am    1     9.752 160.78 61.657
- cyl   2    29.265 180.29 63.323
- hp    1    31.943 182.97 65.794
- wt    1    46.173 197.20 68.191
> summary(stepFit)

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 **
am1          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

> summary(stepFit)$coeff
```

```
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
am1          1.80921138 1.39630450  1.295714 2.064597e-01
```

Comment: The new model has 4 variables (cylinders, horsepower, weight, transmission). The R-squared value of 0.8659 confirms that this model explains about 87% of the variance in MPG. The p-values also are statistically significantly because they have a p-value less than 0.05. The coefficients conclude that increasing the number of cylinders from 4 to 6 with decrease the MPG by 3.03. Further increasing the cylinders to 8 with decrease the MPG by 2.16. Increasing the horsepower is decreases MPG 3.21 for every 100 horsepower. Weight decreases the MPG by 2.5 for each 1000 lbs increase. A Manual transmission improves the MPG by 1.81.

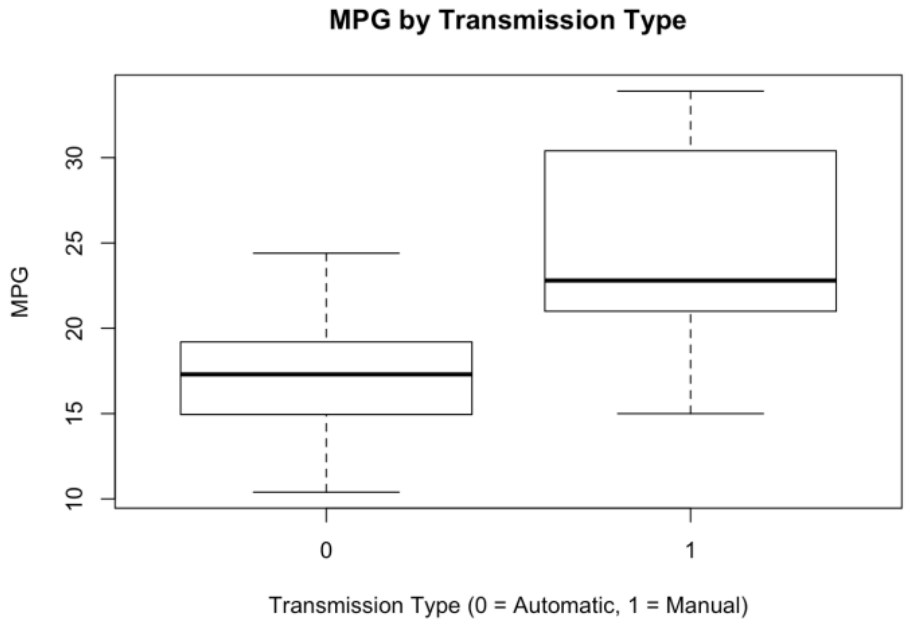
Residuals & Diagnostics

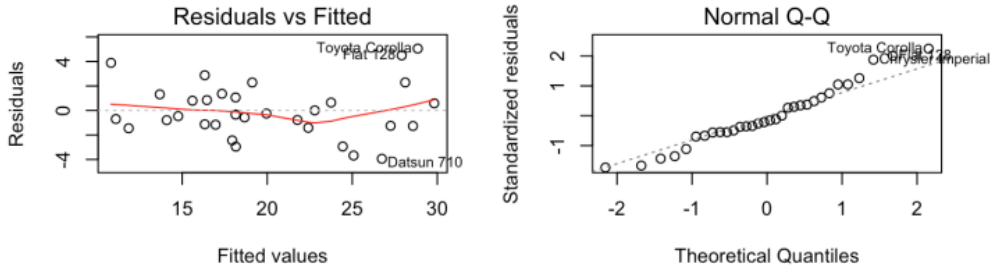
```
> # Residuals & Diagnostics
> sum((abs(dfbetas(stepFit)))>1)
[1] 0
```

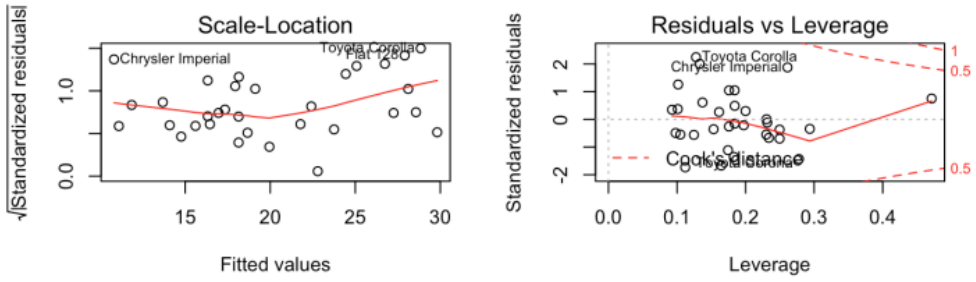
Conclusion

There are differences in MPG based on transmission types. Manual one has a slight MPG boost. In addition, weight, horsepower, number of cylinders are statistically significant in determing MPG.

Appendix


MPG by Transmission Type
MPG
10
15
20
25
30
0
1
Transmission Type (0 = Automatic, 1 = Manual)


Residuals vs Fitted
Residuals
Toyota Corolla
Fiat 128
Datsun 710
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles


Scale-Location
√|Standardized residuals|
Chrysler Imperial
Toyota Corolla
Fiat 128
Fitted values
Residuals vs Leverage
Standardized residuals
Toyota Corolla
Chrysler Imperial
Cook's distance
Leverage