Regmods-016 Course Project

Duncan Turnbull

21 December 2014

Executive Summary

This project seeks to answer two key questions around fuel consumption based on car data from the 1974 Motor Trend US magazine:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

The data provided has 32 samples with 11 variables. The data is reasonably normally distributed. Reviewing correlations between the variables suggested that weight and the number of cylinders were other variables of interest

The relationships were modelled using ordinary least squares. Initially it appeared as if a there was a strong relationship between transmission type and miles per gallon. However cnce consideration for weight and engine cylinders were included that relationship no longer appeared to be significant.

The primary factors driving miles per gallon are the car weight and the number of cylinders.

Given the available data it is not possible to reliably calculate or confirm there is a difference in miles per gallon between automatic and manual transmission at the 95% confidence level.

Examining the data

The observations for mpg and weight appear relatively normally distributed. The transmission is either automatic or manual, and there are either 4,6, or 8 cylinders (Appendix A)

Initial examination appears to show (Appendix B) that the mean Manual Transmission provides much better performance than the mean Automatic performance.

Identify variables of interest

By correlating the data (Appendix C) mpg correlates most strongly to weight (0.87) and then cylinders (0.85). It also correlates to displacement(0.85) and horsepower(0.78). The correlation to transmission (0.60) is relatively weaker, implying it is not the best predictor of miles per gallon.

Cylinders Horsepower Displacement

Looking at cylinders, horsepower & displacement they are all reasonably strongly correlated with themselves. They are all related to the engine size and so the model only needs one variable to represent that effect. Cylinders are the most strongly correlated with MPG and also they are a factor so Cylinders was chosen out of these three.

Modelling tests

Initial model

The initial model evaluated is miles per gallon(mpg) vs transmission (am). This appears to show an improvement of 7 mpg for manual transmission over automatic. The transmission as a predictor is significant at over 99%, however this maybe because it is the only predictor. We will look further. We notice the r.squared only explains 36% of the variance

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

```
## [1] "mpg~wt+cyl: r.squared: 0.830"
```

Transmission confidence levels

The 95% confidence levels (Appendix D) model of this model of transmission against mpg range across 0 suggesting the factors in this model are not significant.

Adding Weight to the model

Weight had the highest correlation to mpg. Adding weight to the model changes the impact such that transmission is no longer a statistically significant event (p=.988) Appendix ???

```
##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 37.32155131  3.0546385 12.21799285 5.843477e-13
## factor(am)1 -0.02361522  1.5456453 -0.01527855 9.879146e-01
## wt          -5.35281145  0.7882438 -6.79080719 1.867415e-07
```

Adding Cylinders

Weight had the highest correlation to mpg. If we add weight to the model

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 34.522443  2.6031842 13.261621 7.694408e-14
## factor(am)1  2.567035  1.2914280  1.987749 5.635445e-02
## cyl         -2.500958  0.3608282 -6.931159 1.284560e-07
```

Adding Weight + Cylinder

Adding both weight and cylinders to the model

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 39.4179334  2.6414573 14.9227979 7.424998e-15
## factor(am)1  0.1764932  1.3044515  0.1353007 8.933421e-01
## wt          -3.1251422  0.9108827 -3.4308942 1.885894e-03
## cyl         -1.5102457  0.4222792 -3.5764148 1.291605e-03
```

```
## mpg~am: r.squared: 0.360
## mpg~am+wt: r.squared: 0.753
```

```
## mpg~am+cyl: r.squared: 0.759
```

```
## mpg~am+wt+cyl: r.squared: 0.830
```

```
## mpg~wt+cyl: r.squared: 0.830
```

```
fit<- lm(mpg ~ factor(am) ,mtcars)
fit2 <-  update(fit, mpg ~ am + wt )
fit3 <-  update(fit, mpg ~ am + wt +cyl )
fit4 <-  update(fit, mpg ~ am + wt +cyl+hp )
anova(fit,fit2, fit3,fit4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + cyl
## Model 4: mpg ~ am + wt + cyl + hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 70.2925  5.39e-09 ***
## 3     28 191.05  1     87.27 13.8611 0.0009165 ***
## 4     27 170.00  1     21.05  3.3432 0.0785534 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The lack of significance at 95% level helps confirm that hp is not a useful variate to add to the model

But running the model the other way

```
fit<- lm(mpg ~  wt ,mtcars)
fit2 <-  update(fit, mpg ~ wt+ cyl )
fit3 <-  update(fit, mpg ~ wt+ cyl + am  )
fit4 <-  update(fit, mpg ~ wt+ cyl + am +hp )
anova(fit,fit2, fit3,fit4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + cyl
## Model 3: mpg ~ wt + cyl + am
## Model 4: mpg ~ wt + cyl + am + hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 278.32
## 2     29 191.17  1    87.150 13.8416 0.0009227 ***
## 3     28 191.05  1     0.125  0.0198 0.8890354
## 4     27 170.00  1    21.049  3.3432 0.0785534 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Look at multi variate model, rule out anything check VIF qsec (1/4 mile time) seems unlikely, maybe drat (rear axle ratio) too
Other variables probably affect fuel consumption similarly whether manual or auto i.e. weight, cylinders, displacement, horsepower, V/S number of carburetors The number of forward gears will probably influence the choice

.Final models

The final model for mpg does not require transmission. This as an extra factor will be biasing the results

```
summary( lm(mpg ~ am * wt ,mtcars) )$coef
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 31.416055  3.0201093 10.402291 4.001043e-11
## am          14.878423  4.2640422  3.489277 1.621034e-03
## wt          -3.785908  0.7856478 -4.818836 4.551182e-05
## am:wt       -5.298360  1.4446993 -3.667449 1.017148e-03
```

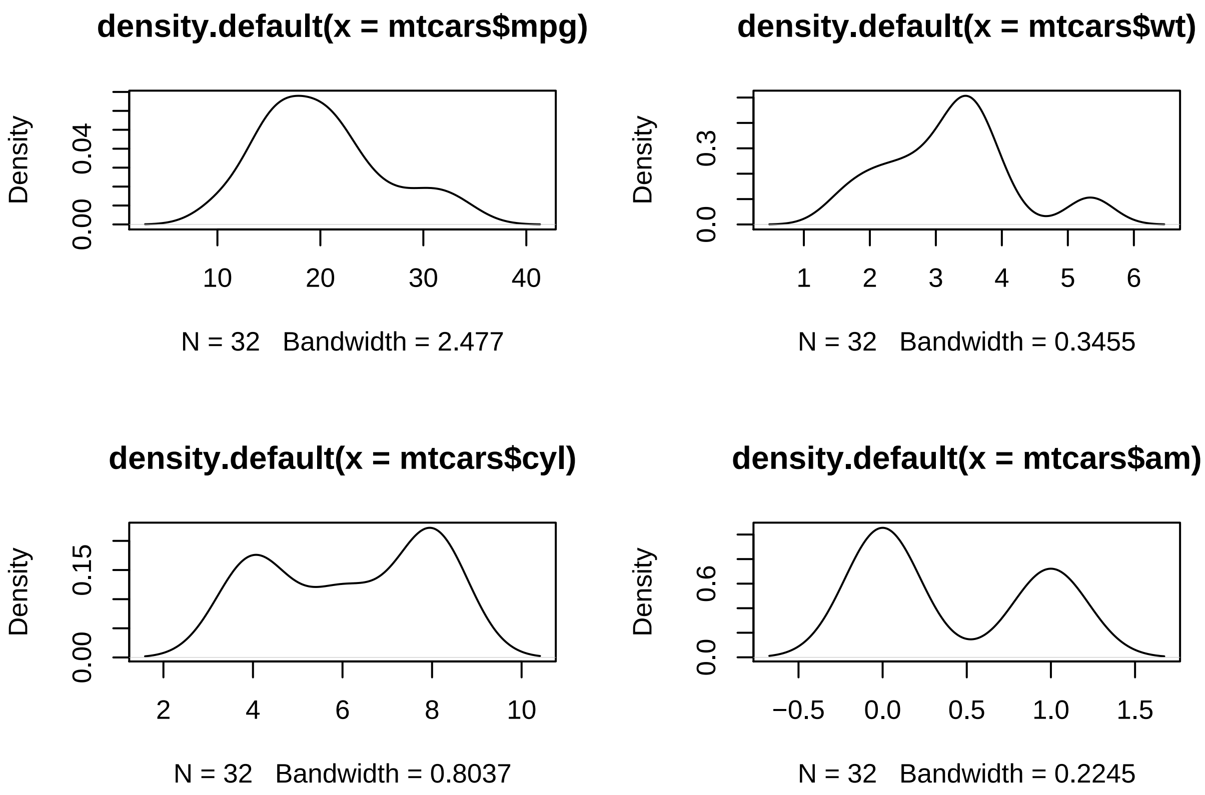
Final Outcome

Exploratory Data analyses Fit multiple models and detail strategy for model selection Did the student answer the questions of interest or detail why the question(s) is (are) not answerable? Did the student do a residual plot and some diagnostics? Did the student quantify the uncertainty in their conclusions and/or perform an inference correctly? Was the report brief (about 2 pages long) for the main body of the report and no longer than 5 with supporting appendix of figures? Did the report include an executive summary? Was the report done in Rmd (knitr)?

Appendix

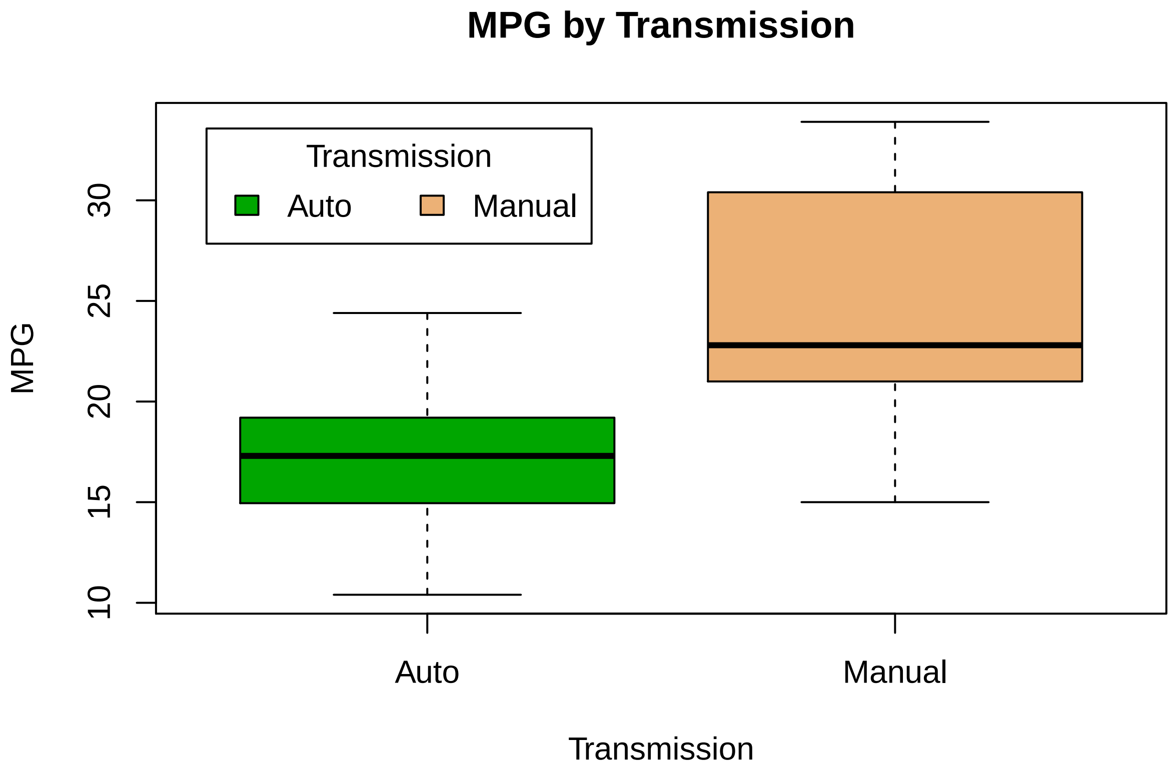
A: Density / distribution of interesting variable

```
par(mfrow=c(2,2)) ; plot(density(mtcars$mpg));plot(density(mtcars$wt));
plot(density(mtcars$cyl));plot(density(mtcars$am))
```



B: MPG by transmission

From the boxplot below it appears as if there are clear distinctions between the MPG for different transmission types. However there could be some confounding variables.



C: Correlations

The correlation table shows strong correlation to weight (wt), number of cylinders (cyl), size of engines (disp), gross horsepower (hp), however hp, disp and cyl are all strongly correlated with each other suggesting only one factor is required. For our purposes we choose cyl, the number of cylinders.

```
##             mpg        cyl       disp         hp       drat         wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.6999381  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.7102139  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.4487591  0.6587479
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.7124406  1.0000000
##            qsec         vs         am       gear       carb
## mpg   0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
## cyl  -0.5912421 -0.8108118 -0.5226070 -0.4926866  0.5269883
## disp -0.4336979 -0.7104159 -0.5912270 -0.5555692  0.3949769
## hp   -0.7082234 -0.7230967 -0.2432043 -0.1257043  0.7498125
## wt   -0.1747159 -0.5549157 -0.6924953 -0.5832870  0.4276059
```

D: How useful is transmission?

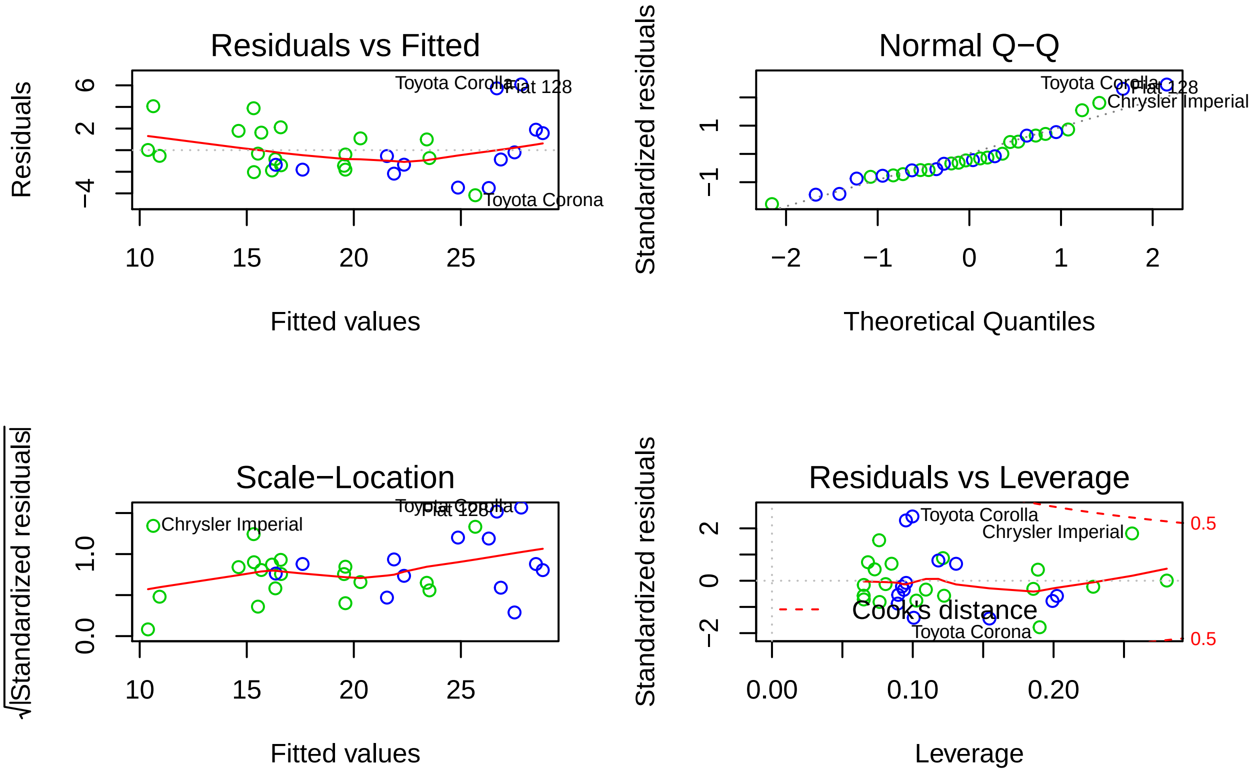
Looking at the confidence levels of a basic model through the origin shows that the levels are primarily either side of 0, meaning they could be 0. This doesn't suggest this is a significant variable.

```
fit <- lm(mpg~am ,mtcars)
confint(fit, level = 0.95)
```

```
##                 2.5 %   97.5 %
## (Intercept) 14.85062 19.44411
## am           3.64151 10.84837
```

E:Residuals

The residual plots appear relatively normally distributed and show any observations with undue influence.



Residuals vs Fitted
Residuals
−4
2
6
Toyota Corolla
Fiat 128
Toyota Corona
10
15
20
25
Fitted values
Normal Q−Q
Standardized residuals
−1
1
Toyota Corolla
Fiat 128
Chrysler Imperial
−2
−1
0
1
2
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
0.0
1.0
Chrysler Imperial
Toyota Corolla
Fiat 128
10
15
20
25
Fitted values
Residuals vs Leverage
Standardized residuals
−2
0
2
Toyota Corolla
Chrysler Imperial
Cook's distance
Toyota Corona
0.5
0.5
0.00
0.10
0.20
Leverage