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

Initial appearances were of a significant relationship between transmission type and miles per gallon. After further analysis, weight, engine cylinders and quarter mile time were also included. While cylinders appeared significant, the final model excluded them

The primary factors driving miles per gallon were transmission type, weight, and number of cylinders.

The conclusion was that a manual transmission car is better for miles per gallon performance for **cars upto a weight of approximately 2.8 tonnes**.

Quantifying the answer

The final model was `mpg ~ wt * am + cyl` which has an adjusted R-squared of 0.86, a p-value of 6.6e-12 and the intercept and all coefficients are significant at the .01 significance level.

The following table gives example predictions based on transmission type, car weight and number of cylinder with an upper and lower limits based on a confidence interval of 95%

The change in performance around the 2.8 tonne weight can be seen.

```
##   Weight  Trans Cylinders Predicted Lower CI Upper CI
## 1    1.6   Auto         4      25.8     22.5     29.0
## 2    1.6 Manual         4      31.0     28.7     33.2
## 3    2.2   Auto         4      24.3     21.8     26.9
## 4    2.2 Manual         4      27.1     25.6     28.5
## 5    2.8   Auto         6      20.6     18.8     22.3
## 6    2.8 Manual         6      20.7     19.2     22.3
## 7    3.4   Auto         8      16.8     15.1     18.4
## 8    3.4 Manual         8      14.4     11.9     17.0
```

The Process

Examining the data

The observations for mpg and weight(wt) 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 available predictor of mpg. Notably qsec did not appear strongly correlated to mpg. Cylinders, horsepower & displacement were reasonably strongly correlated with themselves. They all appear related to the engine size and so only one variable was included to represent that effect. Cylinders was the strongest correlation to MPG and a factor (4,6,8) so Cylinders was chosen out of these three.

Modelling tests

Initial model

The initial model evaluated was mpg vs transmission (am). This appeared to show a 7 mpg advantage for manual transmission over automatic. am is significant at over 99%, however this maybe because it is the only predictor. The R-squared only explains 36% of the variance. Both coefficients appear significant.

```
##              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~factor(am): r.squared: 0.360"
```

Adding Weight & Cylinders to the model

Weight had the highest correlation to mpg. If we add weight to the model it changes the impact such that the manual transmission coefficient is no longer statistically significant (p=.988). R-squared has improved noticeably (0.75) . More investigation is required. Adding cylinders looks similar to adding weight, although the manual transmission impact is improving slightly. Adding both weight and cylinders to the model improves it again. Below is a summary of the R-squared including a model excluding transmission. At this stage transmission doesn't look like it is a significant predictor.

```
## mpg ~ am + wt: r.squared: 0.753
## mpg ~ am + cyl: r.squared: 0.759
## mpg ~ am + wt + cyl: r.squared: 0.830
```

Step analysis & Anova

In order to get an alternate view the available factors were run through the step function (Appendix D). This resulted in a model that included the time over a quarter mile. The recommended model from the step function is: `mpg ~ wt + qsec`

It seems like transmission maybe more of an interaction than a stand alone variable. An Anova on some likely models was run. (Appendix E). This suggested two models for final analysis, `mpg ~ am + wt + cyl` & `mpg ~ am * wt + qsec`

Final model

The models from the step function and the the anova all had transmission coefficients that were not significant which prevented quantisation of the mpg differences. By manual playing around the final model was selected. **`mpg ~ am * wt + cyl`** qsec is no longer a predictor. Transmission interacts with weight, as the weight increases the mpg advantage of manual cars decreases.

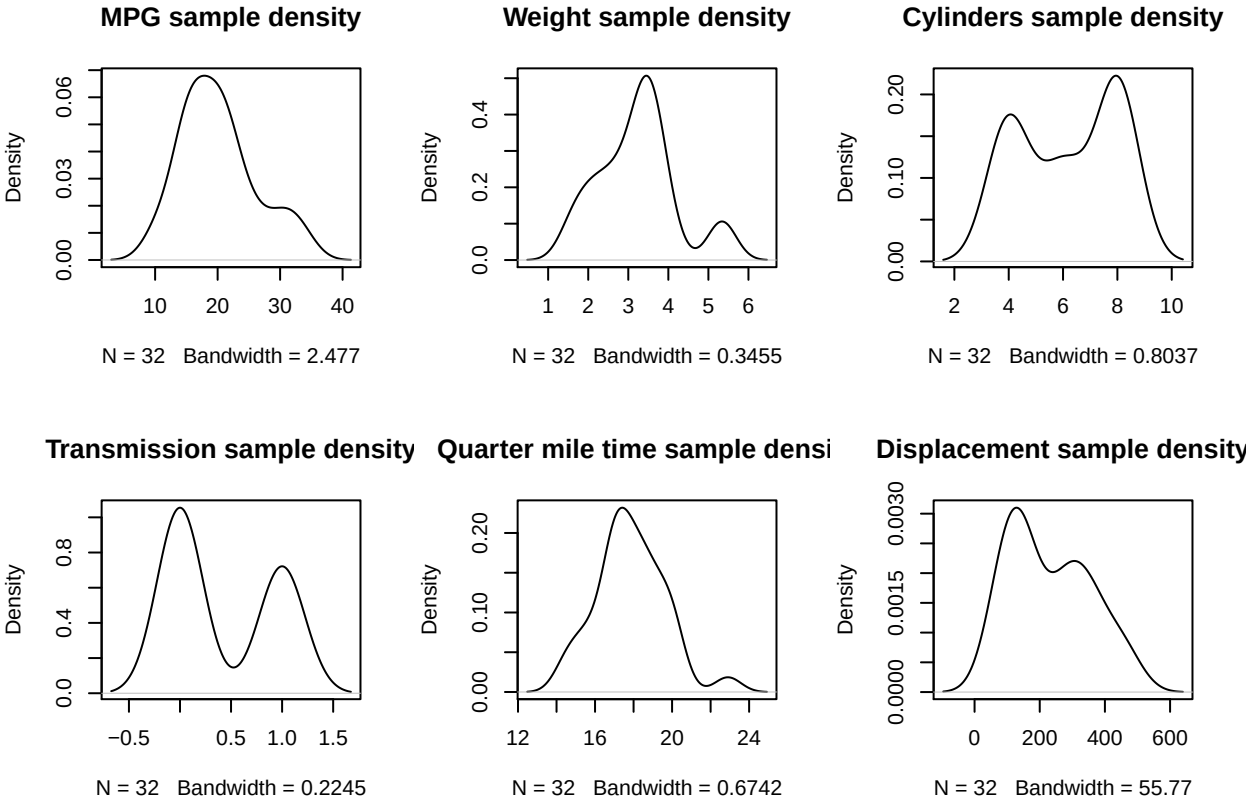
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    34.28      2.796   12.26 1.52e-12
## am             11.94      3.845    3.10 4.44e-03
## wt             -2.37      0.824   -2.87 7.82e-03
## cyl            -1.18      0.380   -3.11 4.42e-03
## am:wt          -4.20      1.312   -3.20 3.50e-03
```

```
## [1] "mpg ~ am * wt + cyl: r.squared: 0.877"
```

To verify this a residual plot *Appendix G) was done and the residuals exhibited no signs of undue influence.

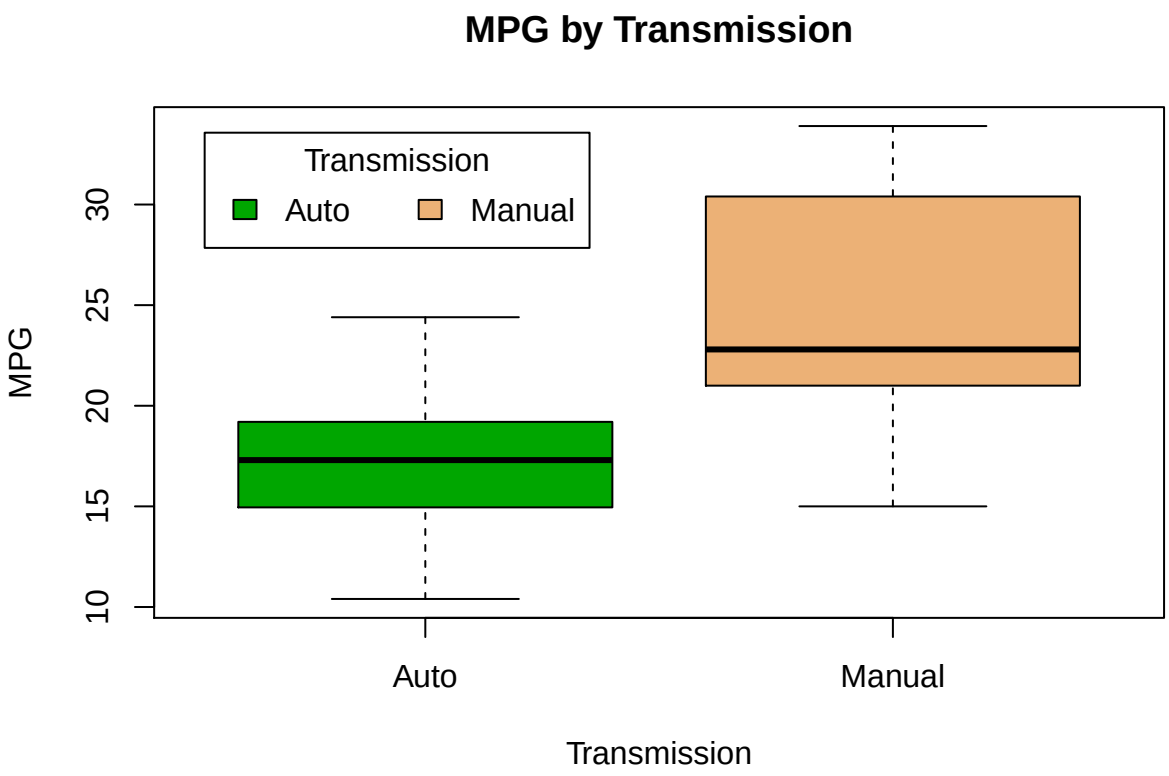
Appendix

A: Density / distribution of variables of interest



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  qsec    vs    am  gear  carb
## mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
## cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
## disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
## hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
## wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
```

D: Stepwise determination of likely models

Using a stepwise test to break down the model suggests mpg~wt+am+qsec is the best fit. The final result is shown

```
stepoutput <- step(lm( mpg~., mtcars))
```

```
## Step: AIC=61.31
## mpg ~ wt + qsec + am
##
## Df Sum of Sq RSS AIC
## <none> 169.29 61.307
## - am 1 26.178 195.46 63.908
## - qsec 1 109.034 278.32 75.217
## - wt 1 183.347 352.63 82.790
```

E: Anova of interactions

Using Anova to finalise the best interaction model

```
fit1 <- lm( mpg~am, mtcars)
fit2 <- update( fit1, mpg~am + wt)
fit3 <- update( fit1, mpg~am + wt + cyl)
fit4 <- update( fit1, mpg~am + wt + qsec)
fit5 <- update( fit1, mpg~am * wt )
fit6 <- update( fit1, mpg~am * wt + qsec)
fit7 <- update( fit1, mpg~am * wt + cyl)
anova( fit1, fit2, fit3, fit4, fit5, fit6 ,fit7)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + cyl
## Model 4: mpg ~ am + wt + qsec
## Model 5: mpg ~ am + wt + am:wt
## Model 6: mpg ~ am + wt + qsec + am:wt
## Model 7: mpg ~ am + wt + cyl + am:wt
##   Res.Df RSS Df Sum of Sq     F  Pr(>F)
## 1     30 721
## 2     29 278  1       443 101.9 1.2e-10 ***
## 3     28 191  1        87  20.1 0.00012 ***
## 4     28 169  0        22
## 5     28 188  0       -19
## 6     27 117  1        71  16.3 0.00040 ***
## 7     27 139  0       -21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F: Comparison of 3 best fit models

The Step function put forward `mpg ~ am + wt + qsec` while from the Anova analysis `mpg ~ am + wt + cyl` looked equally as strong but also the weight interaction model `mpg ~ am * wt + qsec` looks strong.

The model `mpg ~ am * wt + cyl` has the highest adjusted R-squared of 0.88, its p-value is the lowest 7.2e-13 of the three and all coefficients are significant to the 0.01 significance level or higher

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

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.62      6.960    1.38 1.78e-01
## am              2.94      1.411    2.08 4.67e-02
## wt             -3.92      0.711   -5.51 6.95e-06
## qsec            1.23      0.289    4.25 2.16e-04
```

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

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.418      2.641  14.923 7.42e-15
## am             0.176      1.304   0.135 8.93e-01
## wt            -3.125      0.911  -3.431 1.89e-03
## cyl           -1.510      0.422  -3.576 1.29e-03
```

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

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.72      5.899    1.65 0.110893
## am             14.08      3.435    4.10 0.000341
## wt             -2.94      0.666   -4.41 0.000149
## qsec            1.02      0.252    4.04 0.000403
## am:wt          -4.14      1.197   -3.46 0.001809
```

However none of these have significance in all their coefficients. This would make quantifying the differences in mpg subject to much more uncertainty.

Our last model we derive by just trying and seeing.

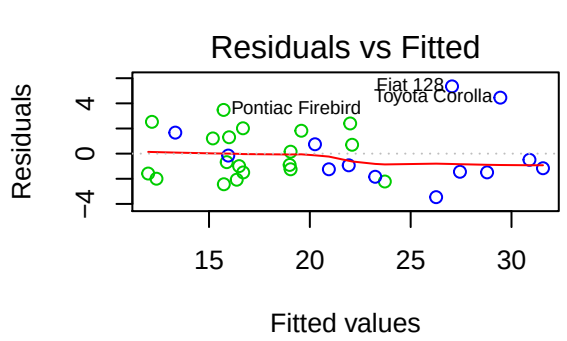
```
summary( lm( mpg~am * wt + cyl, mtcars) )
```

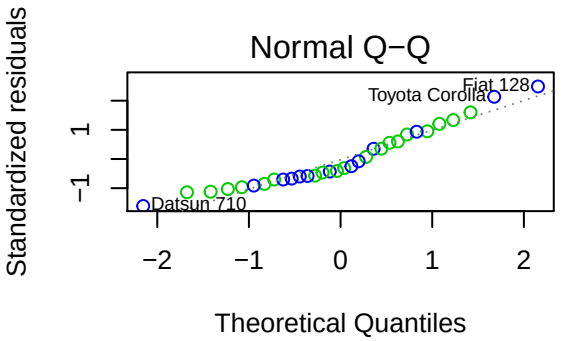
```
##
## Call:
## lm(formula = mpg ~ am * wt + cyl, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.462 -1.491 -0.788  1.396  5.350
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   34.283      2.796   12.26  1.5e-12 ***
## am            11.939      3.845    3.10   0.0044 **
## wt            -2.369      0.824   -2.87   0.0078 **
## cyl           -1.181      0.380   -3.11   0.0044 **
## am:wt         -4.197      1.312   -3.20   0.0035 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.26 on 27 degrees of freedom
## Multiple R-squared:  0.877,  Adjusted R-squared:  0.859
## F-statistic: 48.1 on 4 and 27 DF,  p-value: 6.64e-12
```

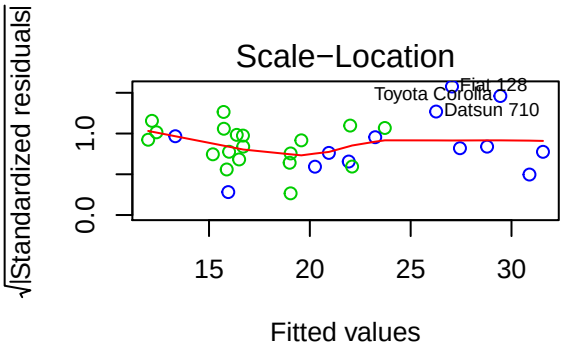
This model has signifiant coefficients, a relatively high adjusted R-squared, and a very low p-value. This is the one we are going for.

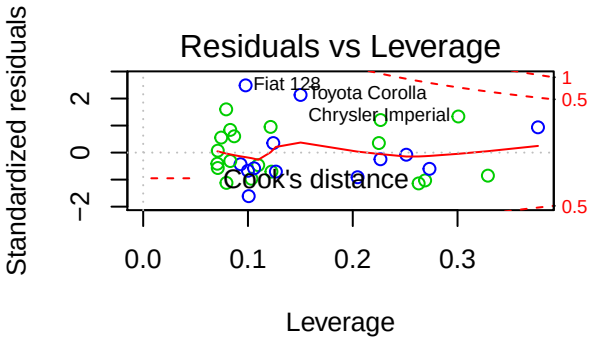
G:Residuals

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


Residuals vs Fitted
Residuals
Fitted values
Fiat 128
Toyota Corolla
Pontiac Firebird


Normal Q−Q
Standardized residuals
Theoretical Quantiles
Fiat 128
Toyota Corolla
Datsun 710


Scale−Location
√|Standardized residuals|
Fitted values
Fiat 128
Toyota Corolla
Datsun 710


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
Fiat 128
Toyota Corolla
Chrysler Imperial