Regression Models Project

Lachlan Glascott May 2020

Executive Summary

Motor Trend is a magazine about the automobile industry. Motor Trend is interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions: 1. Is an automatic or manual transmission better for MPG 2. Quantify the MPG difference between automatic and manual transmissions

Manual transmission is better for MPG than automatic transmission. It is estimated that manual transmissions are 2.9 MPG higher than automatic, holding weight and speed constant.

Is automatic or manual transmission better for MPG?

```
## # A tibble: 2 x 2
##   am     average_mpg
##   <chr>        <dbl>
## 1 Auto          17.1
## 2 Manual        24.4
```

```
fit <- lm(mpg ~ am,data = model_data)
summary(fit)$coeff
```

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

Manual cars have a mean mpg of 24.39 compared to an average mpg of 17.15 for automatic. From the t-test the p-value is less than 0.5 so at a 95% confidence level we we can reject the null hypothesis that there is no difference in mpg, and conclude with that manual cars have a higher mpg than automatic.

However, there may be other characteristics of cars which affect mpg which are correlated with the transmission type.

Exploratory Analysis

The scatter plots in the Appendix show that the am variable is highly correlated (postively or negatively) with several other variables. It may be these variables, such as the weight of the car, which are impacting on the mpg.

This can be tested using multivariate regression analysis to quantify the mpg difference between automatic and manual transmissions controlling for these other factors.

Modelling

A simple linear regression model shows the postive relationship between manual cars and mpg, and explains 36% of the variation. We now fit a multivariate model with all of the variables in the dataset to see how this impacts on the relationship.

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## amManual     2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     21 147.49  9     573.4 9.0711 1.779e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After including all of the variables in the dataset (refer to Appendix), the model estimates that miles per gallon for manual cars is 2.5 higher than automatic cars, holding all other variables constant. However, the p-value higher than 0.5 and as such we fail to reject the null hypothesis. In addition, there are no statsitically significant variables in the model which may be caused by overfitting by including correlated variables.

The analysis of variance inflation factors shows that there are statistically significant differences in the models and omore than just the am variable should be included.

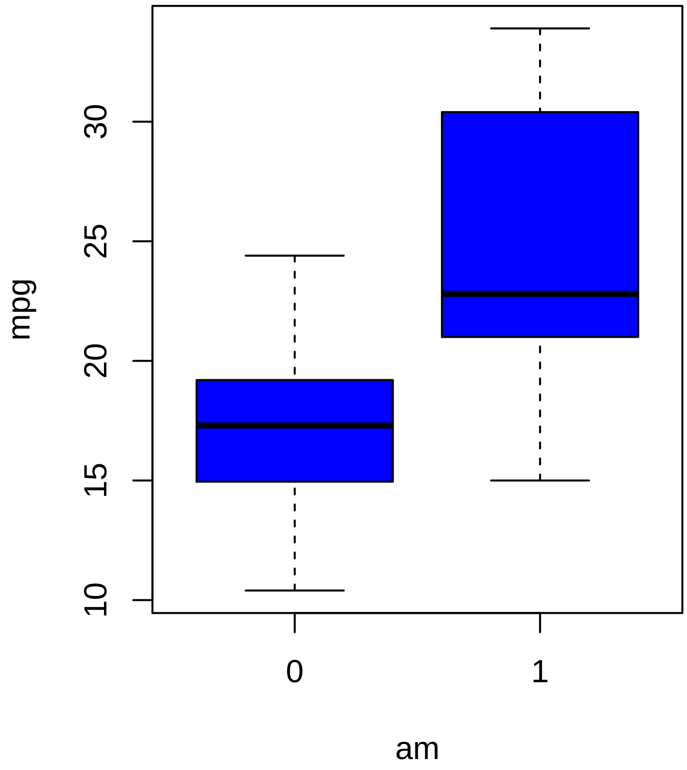
Based on the correlation plots and results from the multivariate regression model, the following model is fit.

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## amManual     2.935837  1.4109045  2.080819 4.671551e-02
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
```

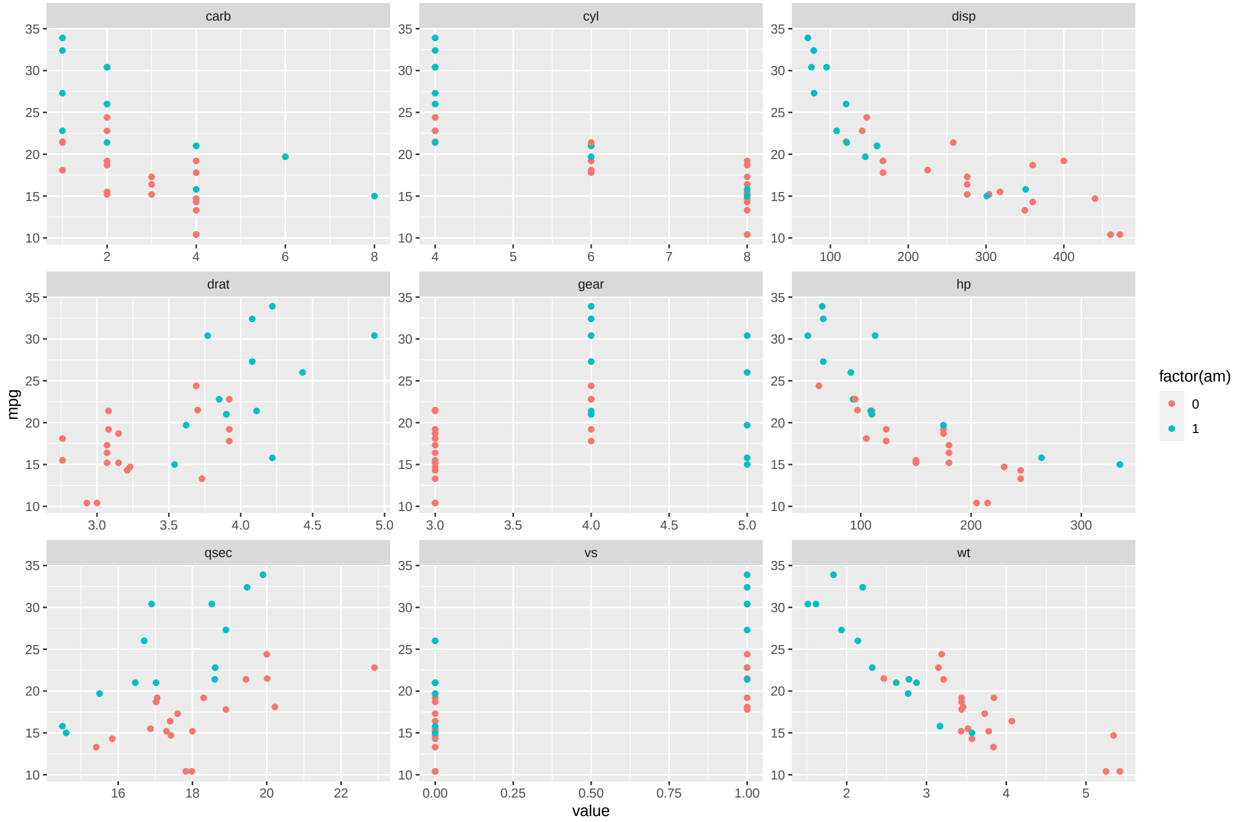
It is estimated that the MPG difference between manual transmission and an automatic transmission is 2.9, holding weight and quarter mile time constant.

Appendix

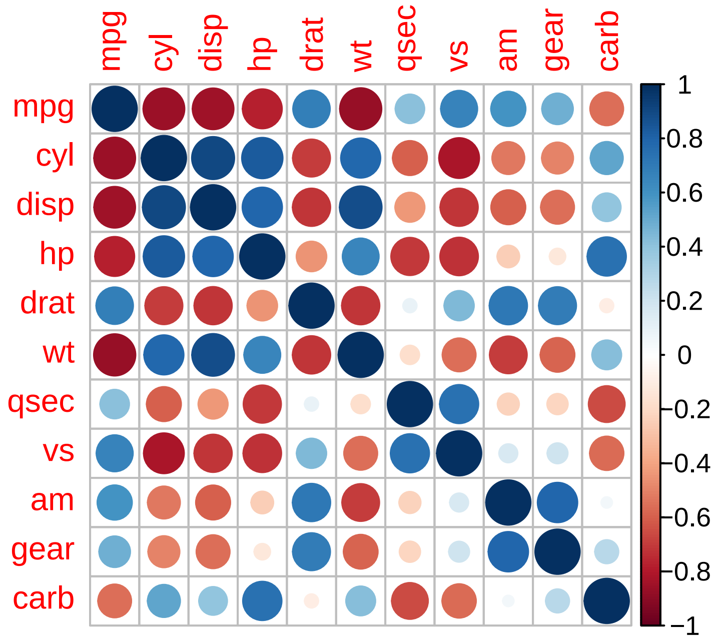
Boxplot mpg vs transmission type



MPG Scatter Plots



Correlation plot mtcars variables



Model summaries

```r
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = model_data)
##
## 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
```

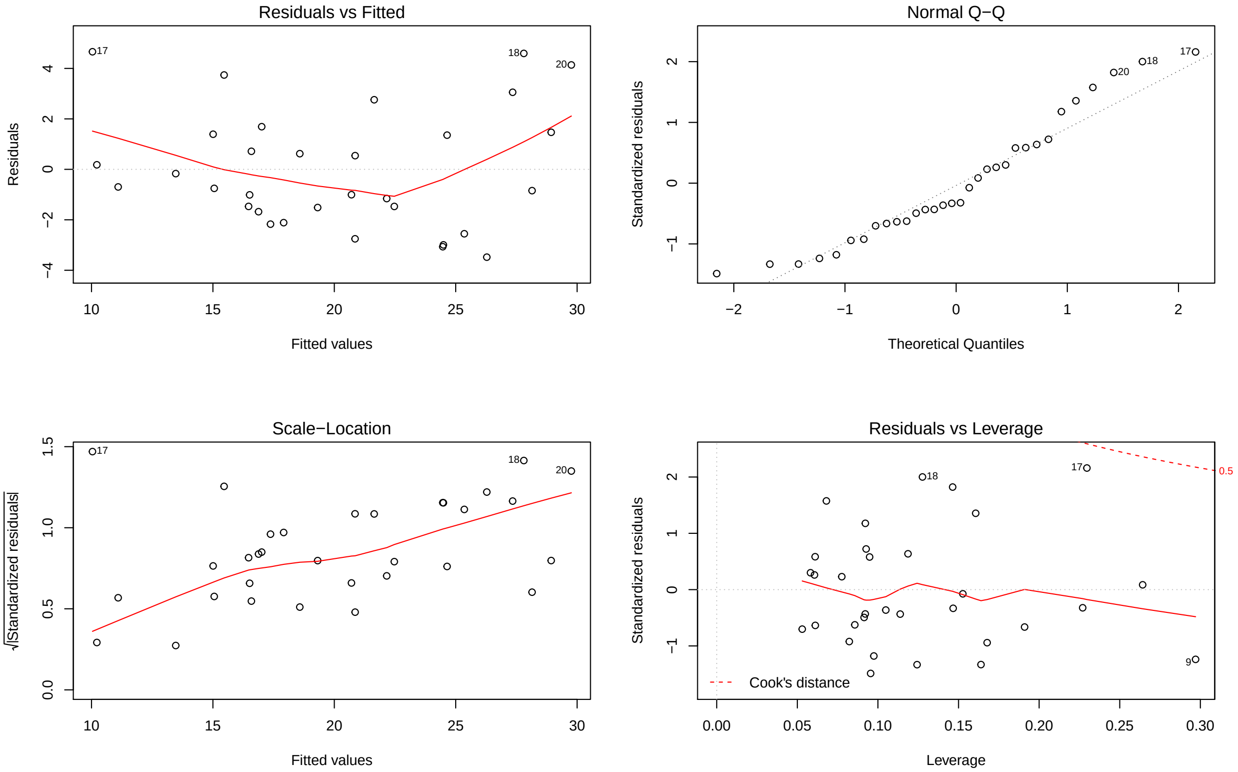
```r
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = model_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## amManual     2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```r
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec, data = model_data)
##
## 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
## amManual      2.9358     1.4109   2.081 0.046716 *
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## 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
```

Final model residual analysis


Residuals vs Fitted
Residuals
Fitted values
10
15
20
25
30
−4
−2
0
2
4
17
18
20
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−2
−1
0
1
2
−1
0
1
2
17
18
20
Scale−Location
√|Standardized residuals|
Fitted values
10
15
20
25
30
0.0
0.5
1.0
1.5
17
18
20
Residuals vs Leverage
Standardized residuals
Leverage
0.00
0.05
0.10
0.15
0.20
0.25
0.30
−1
0
1
2
0.5
Cook's distance
18
17
9