Analysis of the Affect of Transmission Type on Miles per Gallon in the mtcars Dataset

Executive Summary

Through an examination of the mtcars dataset, we have determined that the type of transmition in a car has a statistically significant affect on the mileage of the car. Specifically, stadard transmission cars have better gas mileage than automatic transmition cars. We discuss the results in detail below, including our atempt to separate confounding factors and determine the existence of outliers in the data.

Introduction

Our approach to the dataset follows three main paths:

1. Examine the direct relationship between transmission type and miles per gallon (mpg).
2. Examine the full data set to see if there is a parsimonious model representing mpg that includes transmission type.
3. Examine the model to see how outliers affect the model.

Linear Model Comparing Transmission Type to Mpg

Comparing transmission type to mpg directly does not present many difficulties and there are good reasons in this comparison to see a relationship. Figure 1 shows the histograms of vehicle mpg grouped by transmission type.

Looking at Figure 1 we can see that manual transmission vehicles have a higher mean mpg than automatics, but with a higher variance. This is confirmed by:

```
     [,1]             [,2]             [,3]          [,4]
[1,] "automatic mean" "automatic sdev" "manual mean" "manual sedv"
[2,] "17.1474"        "3.834"          "24.3923"     "6.1665"
```

So with a very basic analysis, we can see that there is a mean improvement of ~7 mpg going from manual to automatic transmission. We can expand this analysis with a linear model as follows:

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

```
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   17.147      1.125  15.247 1.134e-15
am             7.245      1.764   4.106 2.850e-04
```

```
[1] "Adjusted R^2:  0.338459"
```

The 2D linear model above echoes the results of looking at the classes individually. in addition we can see that the p-values are significant to greater than 0.99. Looking at the adjusted R^2 however, we can see that only about 34% of the variance is accounted for in this model.

From the R^2 number, the question arises: does the mtcars data imply a better model that includes transmission type?

A Parsimonious Model for Mpg

We can start this analysis by simply including all the mtcars columns in the linear model and then pulling out the columns that do not contribute or that are mixed with the other columns in the model.

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

Looking Table 1, we see that almost none of the p values are significant (we will look at 'wt' later) in this model, so lets look at the variance inflation to see if we can discover some independent columns for modeling mpg.

```
vif(full_fit)
```

```
   cyl   disp     hp   drat     wt   qsec     vs     am   gear   carb 
15.374 21.620  9.832  3.375 15.165  7.528  4.966  4.648  5.357  7.909 
```

The vif is very interesting as it shows that there are many confounding variable groups. By trying models of varying combinations of these columns, we can settle on the model `lm(mpg ~ am + carb + vs`. looking at the statistics for this model:

```
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   19.517     1.6091  12.130 1.156e-12
am             6.798     1.1015   6.172 1.155e-06
carb          -1.431     0.4081  -3.506 1.553e-03
vs             4.196     1.3246   3.168 3.696e-03
```

```
[1] "Adjusted R^2:  0.758473"
```

Our experimentally derived linear model now accounts for 76% of the variance using only three columns. also, looking at the vif for this model shows that these columns are relatively independent of each other:

```
   am  carb    vs 
1.067 1.535 1.575 
```

There are better models of mpg in this data that do not include transmission type (column 'am'). Probably the best parsimonious model is `lm(mpg~wt, data=mtcars)`.

This model is significant and accounts for 74% of the variance with a single variable. This seems pretty obvious since weight would logically be a big factor in determining gas milage.

Outliers

Within the scope of the model, we want to be sure that certain points aren't overly influencing the model. Table 2 below shows the PRESS residual for the model. For the PRESS residuals we can see that there are three models which stand out in the model:

- Datsun 710
- Volvo 142E
- Ford Pantera L

In fact, these are the extreme data points shown in the model's residual plots see Figure 3 below). These points do not seem to cause much distortion in the data set according to the residual plots.

Conclusion

In conclusion we find that transmission type is a significant factor in the determination of mileage given the mtcars dataset. Though there are other models than the ones that include transmission type, we can construct an effective model with transmission tyype as a major component.

Appendix

Table 1: Coefficients from a linear model of mpg using all mtcars data

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788  0.6573  0.51812
cyl         -0.11144    1.04502 -0.1066  0.91609
disp         0.01334    0.01786  0.7468  0.46349
hp          -0.02148    0.02177 -0.9868  0.33496
drat         0.78711    1.63537  0.4813  0.63528
wt          -3.71530    1.89441 -1.9612  0.06325
qsec         0.82104    0.73084  1.1234  0.27394
vs           0.31776    2.10451  0.1510  0.88142
am           2.52023    2.05665  1.2254  0.23399
gear         0.65541    1.49326  0.4389  0.66521
carb        -0.19942    0.82875 -0.2406  0.81218
```

```
[1] "Adjusted R^2:  0.806642"
```

Table 2: PRESS Residuals for the model `lm(mpg ~ am + carb + vs, data=mtcars)`

```
          Mazda RX4       Mazda RX4 Wag          Datsun 710
            0.46069             0.46069            -7.18033
     Hornet 4 Drive   Hornet Sportabout             Valiant
           -0.99152             2.29624            -4.69979
         Duster 360           Merc 240D            Merc 230
            0.54751             3.98232             2.18668
           Merc 280           Merc 280C          Merc 450SE
            1.55241            -0.24377             1.26893
         Merc 450SL         Merc 450SLC  Cadillac Fleetwood
            2.24091            -0.02705            -3.67467
Lincoln Continental   Chrysler Imperial            Fiat 128
           -3.67467             0.98056             3.79543
        Honda Civic      Toyota Corolla       Toyota Corona
            3.07245             5.51039            -0.87914
   Dodge Challenger         AMC Javelin          Camaro Z28
           -1.29836            -1.63536            -0.53510
   Pontiac Firebird           Fiat X1-9       Porsche 914-2
            2.85790            -2.03544             3.13241
       Lotus Europa      Ford Pantera L        Ferrari Dino
            3.07245            -5.41408             2.44473
      Maserati Bora          Volvo 142E
            0.22804            -6.98110
```

Figure 1: Histograms For Mpg on Manual and Automatic Transmission Vehicles in the mtcars Dataset.

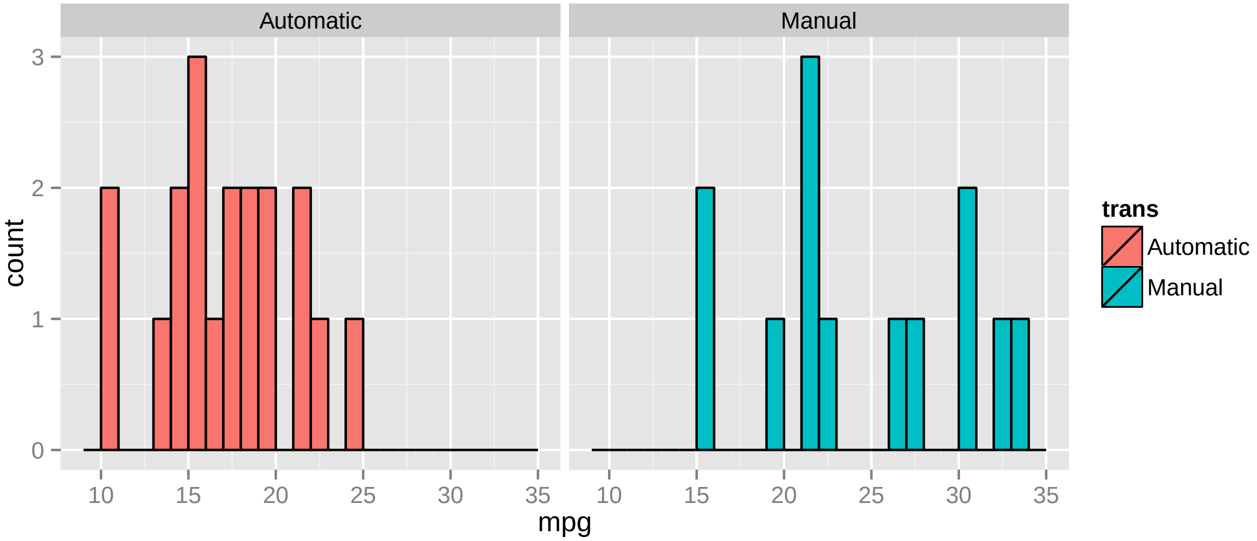


Figure 2: residual plots for the linear model `lm(mpg ~ am)`.

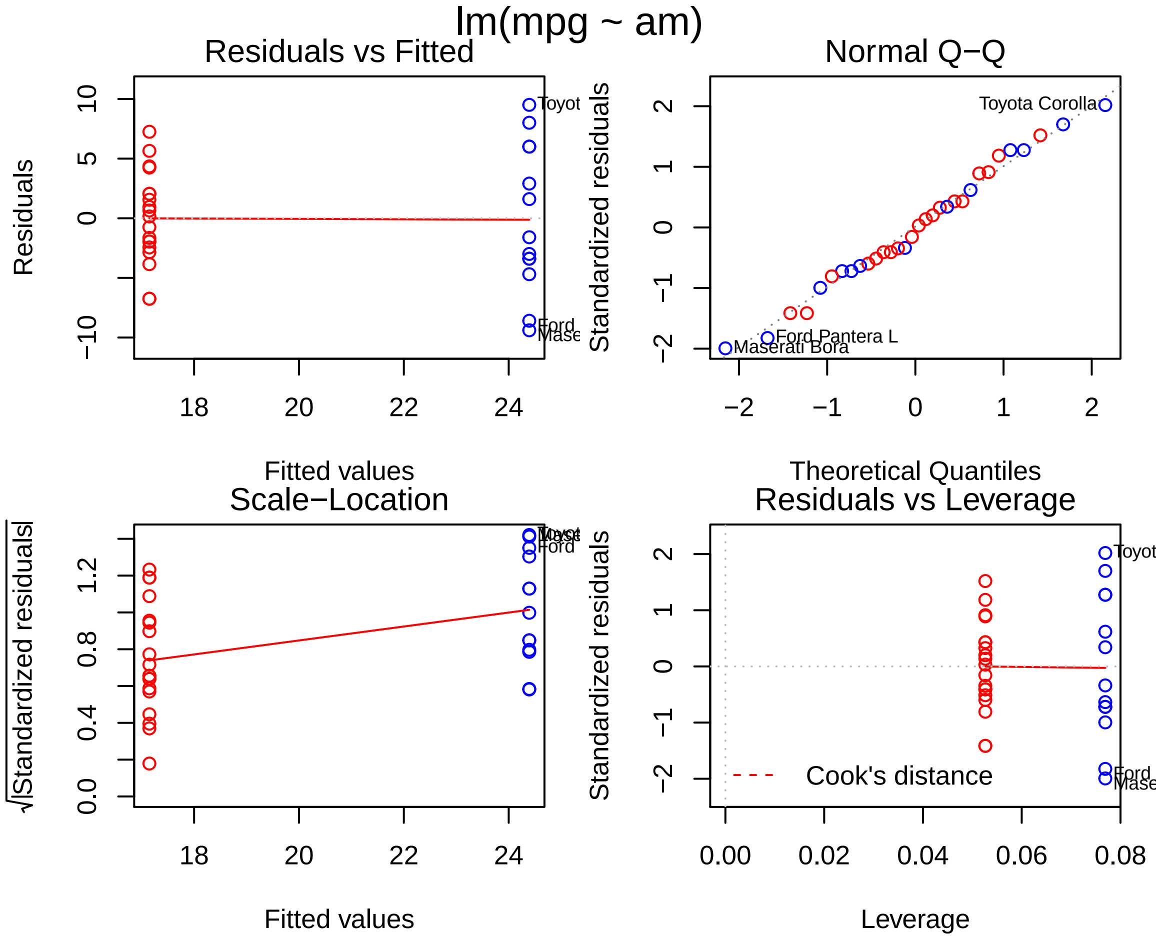
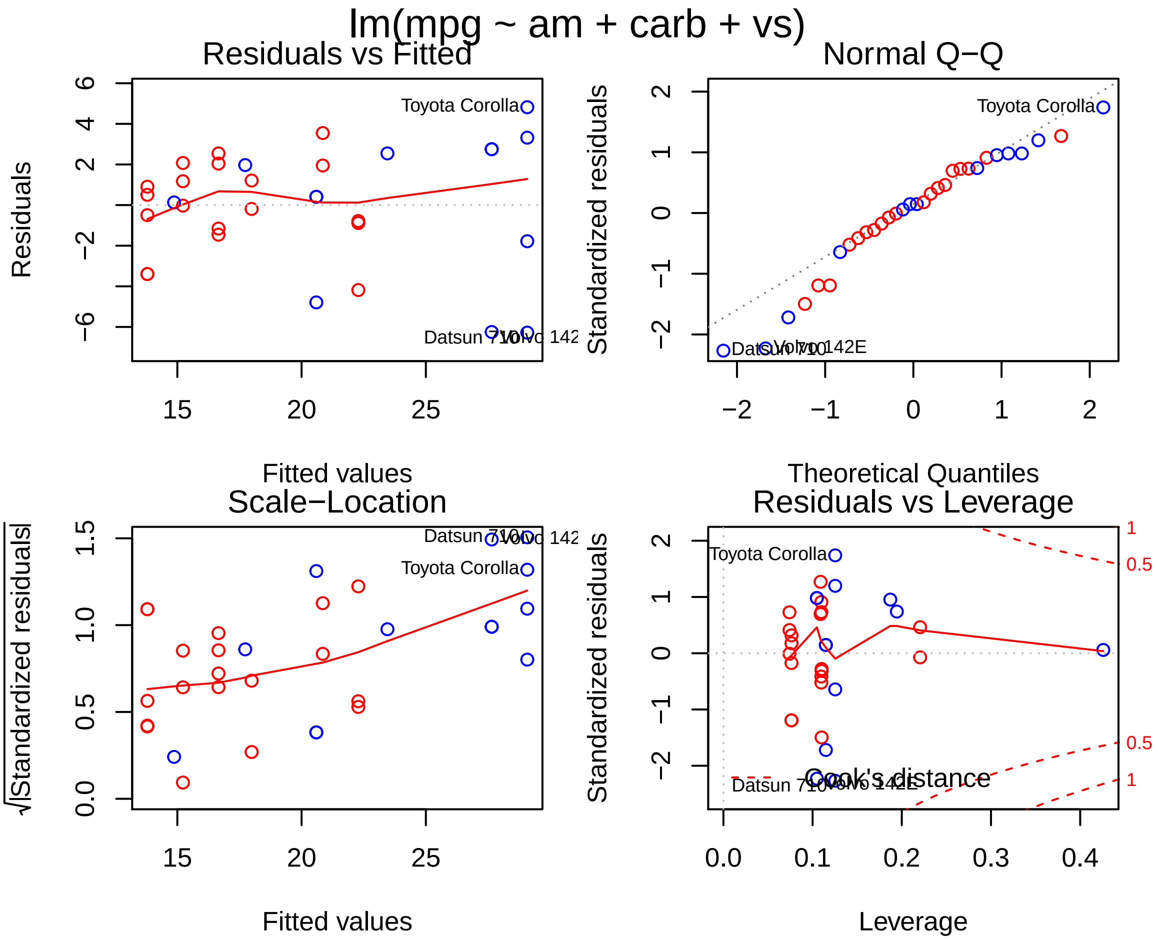


Figure 3: Residual plots for the linear model `lm(mpg ~ am + carb + vs, data=mtcars)`


lm(mpg ~ am + carb + vs)
Residuals vs Fitted
Residuals
−6
−2
2
4
6
Toyota Corolla
Datsun 710
Volvo 142
15
20
25
Fitted values
Normal Q−Q
Standardized residuals
−2
−1
0
1
2
Toyota Corolla
Datsun 710
Volvo 142E
−2
−1
0
1
2
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
Datsun 710
Volvo 142
Toyota Corolla
15
20
25
Fitted values
Residuals vs Leverage
Standardized residuals
−2
−1
0
1
2
Toyota Corolla
Datsun 710
Volvo 142E
Cook's distance
1
0.5
0.5
1
0.0
0.1
0.2
0.3
0.4
Leverage