Miles per Gallon (MPG) on Manual vs Automatic Transmission Vehicles

Frank D. Evans - Johns Hopkins Data Science, Regression Modeling

Executive Summary

The key questions of interest concern the car mileage (mpg) and type of transmission (0:automatic, 1:manual). Considering the relationship only between those factors shows a strong relationship–though one with a low probability of being statistically significant given the size of the data. When the other key attributes of the car are held constant, the relationship drops strongly to the transmission type–and other factors about the car show a high probability of being more strongly related to the mileage.

Exploratory Analysis

Load the data set, and required libraries.

```
library(datasets); library(MASS); data(mtcars)
```

Based on data documentation, it can safely be assumed that transmission type is a discrete factor dimension, while mpg is measured on a bounded, yet continuous scale that will consist exclusively of positive values. To verify assumptions and get a starting point of scale of the data, we will call a summary of the mpg attribute separated by the transmission factor.

```
by(mtcars$mpg, mtcars$am, summary)
```

```
## mtcars$am: 0
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    10.4    15.0    17.3    17.1    19.2    24.4 
## -------------------------------------------------------- 
## mtcars$am: 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    15.0    21.0    22.8    24.4    30.4    33.9 
```

To get a sense of scale, we’ll examine the number of records by factor for the transmission variable.

```
c(sum(mtcars$am == 0), sum(mtcars$am == 1))
```

```
## [1] 19 13
```

Regression Analysis

Single Variable

We first want to test a relationship directly between the variables of interest. A single variable regression seems to imply a considerable relationship.

```
fit_1var <- lm(mtcars$mpg ~ mtcars$am)
summary(fit_1var)$coefficients
```

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

The coefficient of 7.2 would imply that a manual transmission should expect an mpg increase of 7.2 miles per gallon. The t value and residual variation implies the model is a poor fit, confirmed by the plor of the residuals (Appendix: Figure 1).

Multi-Variable Analysis To find a better fitting model a generalized linear model is fit against the principle variables. Dimensional variables are fed into the model as a factor to generate dummy variables, and we use a stepwise method to consider alternate multi-variate regression models for better candidates.

```
step_fit <- glm(mpg ~ as.factor(am) + as.factor(cyl) + as.factor(gear) + disp + hp + drat + wt,
 data = mtcars)
model_fit <- stepAIC(step_fit, direction = 'both'); model_fit$anova
```

The results of an ANOVA test from the stepwise fit algorithm suggest a final model based on the transmission type compunded with the number of cylinders in the car’s engine (cyl), the horsepower of the car (hp) and the weight of the car (wt). From an intuition perspective, this makes probable sense.

```
final_model <- glm(mpg ~ as.factor(am) + as.factor(cyl) + hp + wt, data = mtcars)
summary(final_model)$coefficients; summary(final_model$residuals)
```

```
##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     33.70832    2.60489 12.9404 7.733e-13
## as.factor(am)1   1.80921    1.39630  1.2957 2.065e-01
## as.factor(cyl)6 -3.03134    1.40728 -2.1540 4.068e-02
## as.factor(cyl)8 -2.16368    2.28425 -0.9472 3.523e-01
## hp              -0.03211    0.01369 -2.3450 2.693e-02
## wt              -2.49683    0.88559 -2.8194 9.081e-03
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -3.940  -1.260  -0.401   0.000   1.130   5.050
```

The results of the sugested model suggest a much lower relationship, with a manual transmission attributable to a 1.8 mile per gallon lift when controlled for key corroborating factors. However, the high p-value of 0.21 for the transmission factor means that there is a relatively low probability that the relationship as observed in this data is significant.

More significant factors are found among the engine size, horsepower, and the weight of the car. Behind this factor, is a likely significant factor that the presence of a 6 cylindar engine as compared to a 4 cylindar baseline attributes a reduction of mpg by 3.0 miles per gallon.

A major corroborating factor surrounding statistical signifigance in this case is the high ratio of the total number of available variables (11) to the number of observations (32). Examination of the model fit and residual fit (Appendix: Figure 4) show that given the data and variables for analysis, incorporating the additional variable results in a significantly better explantory and predictive model.

Appendix

Figure 1: Single Variable Regression Plot and Residuals

```
par(mfrow = c(1,2))
plot(mtcars$am, mtcars$mpg, main = '1 Variable Regression: MPG vs Trans')
abline(fit_1var)
plot(mtcars$am, resid(fit_1var), main = 'Residuals: 1 Variable Exploratory')
abline(h = 0)
```

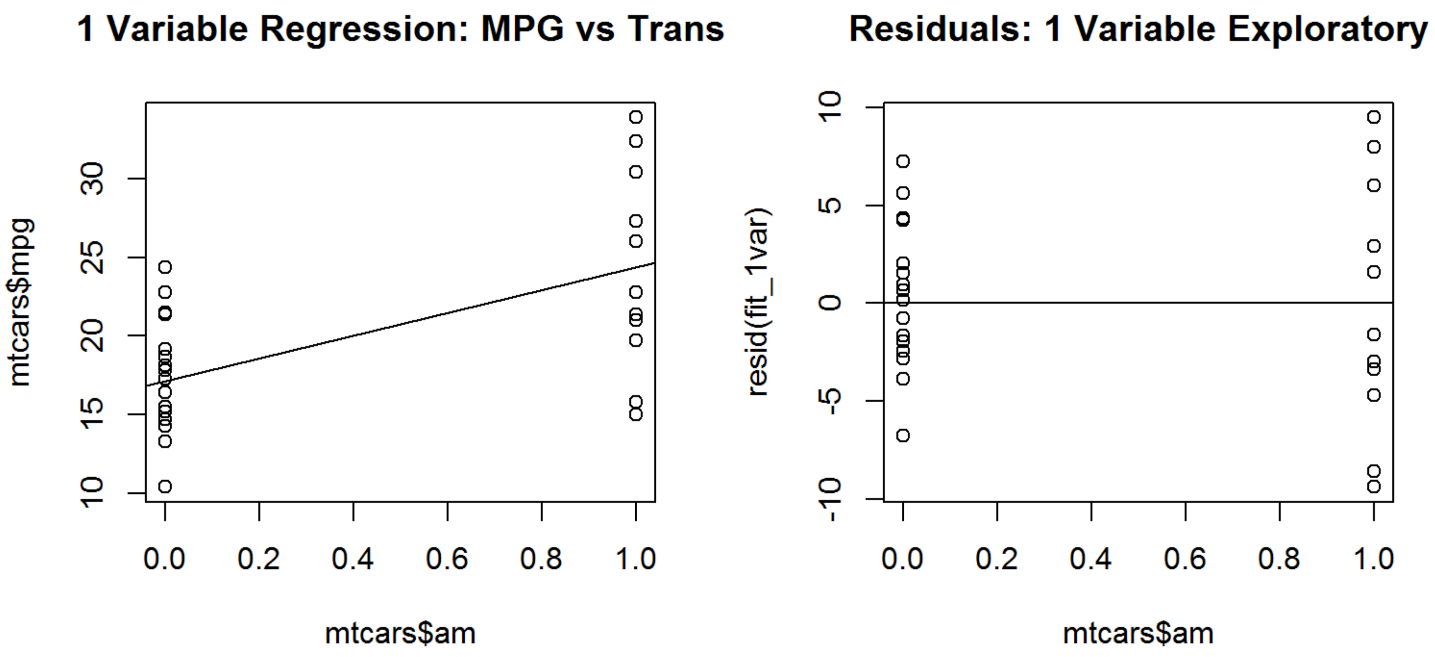


Figure 2: Single Variable Regression - Full Model Fit Plots

```
layout(matrix(c(1,2,3,4),2,2))
plot(fit_1var)
```

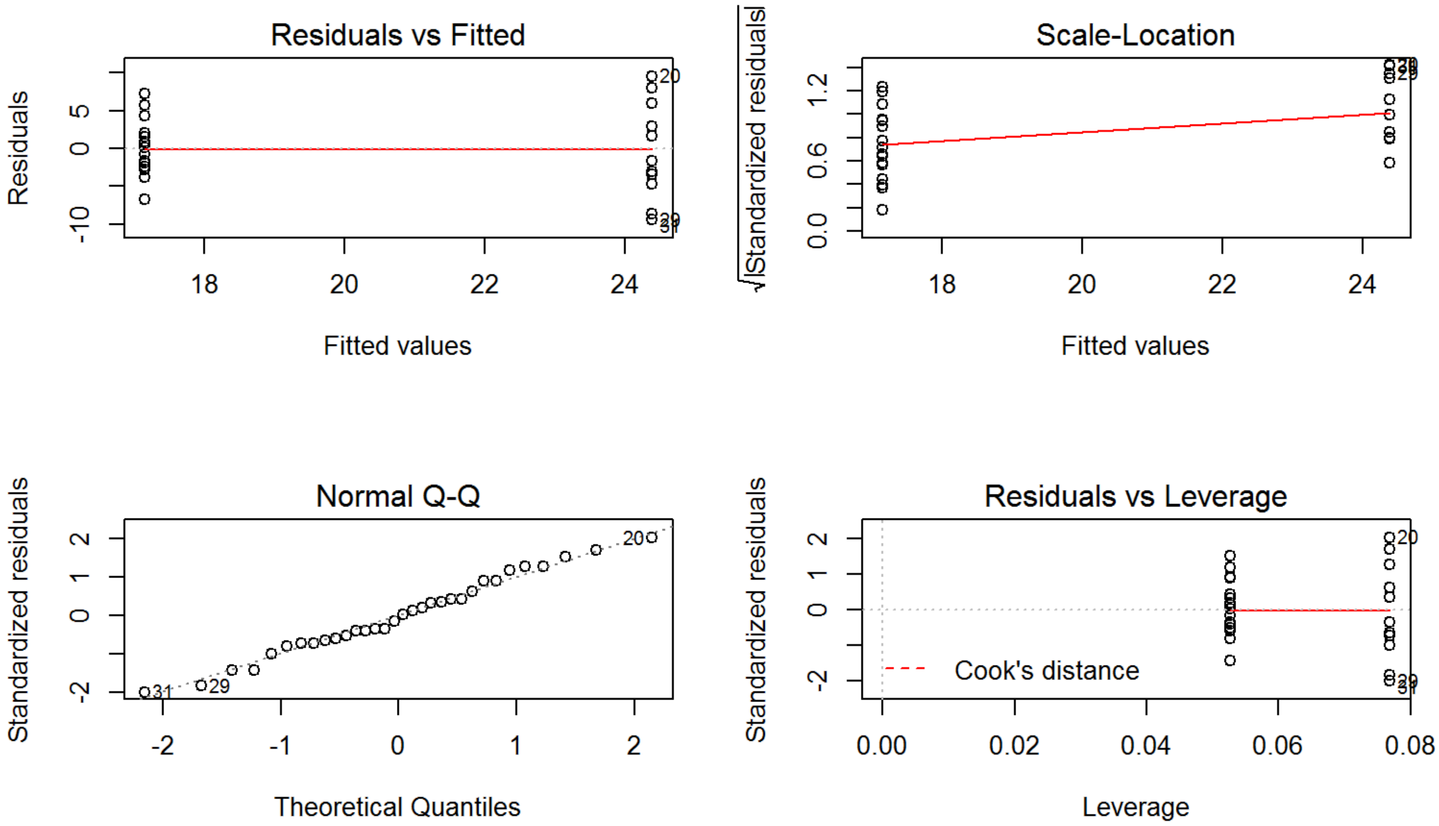


Figure 3: Pairs Plot for MTCars Attributes

```
test_vars <- c('mpg','cyl','disp','hp','wt','gear','carb')
pairs(x = mtcars[,test_vars],
      panel = panel.smooth,
      main = 'MTCars Attributes')
```

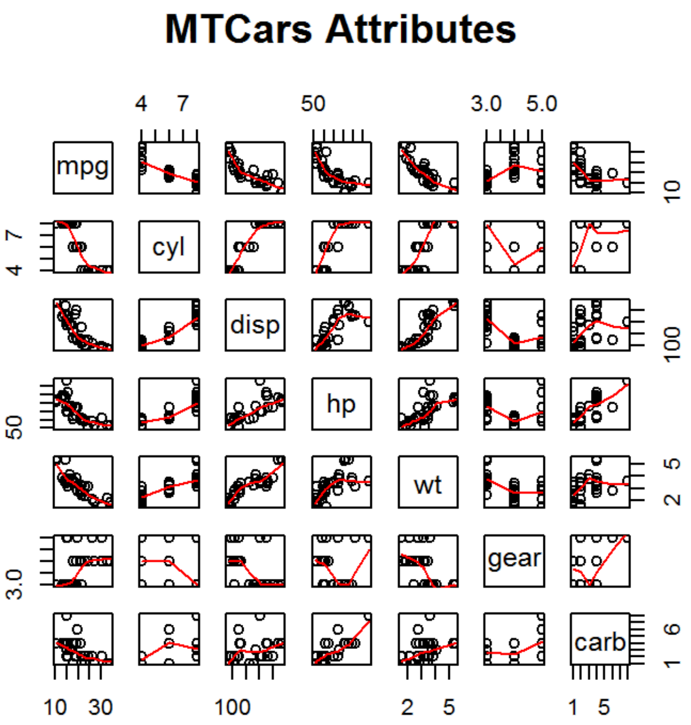
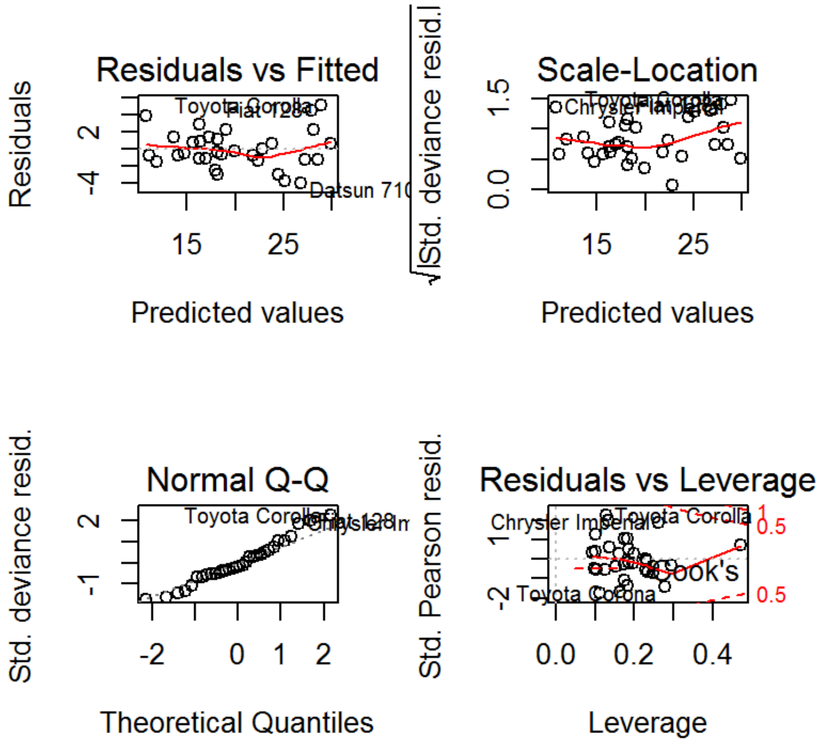


Figure 4: Final Model Regression - Full Model Fit Plots

```
layout(matrix(c(1,2,3,4),2,2))
plot(final_model)
```


Residuals vs Fitted
Residuals
2
-4
Toyota Corolla
Fiat 128
Datsun 710
15 25
Predicted values
Scale-Location
√|Std. deviance resid.|
1.5
0.0
Chrysler Imperial
Toyota Corolla
Fiat 128
15 25
Predicted values
Normal Q-Q
Std. deviance resid.
2
-1
Toyota Corolla
Fiat 128
Chrysler Im
-2 0 1 2
Theoretical Quantiles
Residuals vs Leverage
Std. Pearson resid.
1
-2
Chrysler Imperial
Toyota Corolla
Toyota Corona
Cook's
1
0.5
0.5
0.0 0.2 0.4
Leverage