$MotorTrend - VehicleAnalysis$

Abstract

In this report for Motor Trend Magazine, regression analysis reveiled changes in Miles per Gallon (MPG) with various features and makes of 32 automobiles. In a simplistic model, automatic vs manual transmission types showed that the manual transmission is 7.25 MPG better than a automatic transmission. With cylinder, displacement, horsepower, and weight as part of the equation, a multi-parameter regression model indicated that the manual transmission is 1.81 MPG better than the automatic transmission while the goodness of fit reached ~87%.

Analysis software, output files and data are at gitHub location (https://github.com/dvarney/Regression_Models)

Exploratory Data Analysis

The features of the data file, *mtcars*, for 32 vehicles with 11 characteristics is described in Table 1. Vehicles from the Honda Civic to the Lotus Europa and Maserati Bora were reviewed. Surprisingly, a Cadillac and Lincoln automobile were rated poorly, with the Lotus near the top. Lotus' fiberglass body probably contributed to the surprising performance for this expensive storts car. See Appendix for additional Tables and Figures.

To investivate the initial data, mpg vs automatic and manual, along with a density, Figure 1 & 2, were generated and inspected for insight. Definitions of the variables are listed below. For more details, see Table 1.

- 1 [mpg] Miles/(US) gallon
- 2 [cyl] Number of cylinders
- 3 [disp] Engine displacement (cu.in.)
- 4 [hp] Gross horsepower
- 5 [drat] Rear axle ratio
- 6 [wt] Weight (lb/1000)
- 7 [qsec] 1/4 mile time
- 8 [vs] V/S (unknown)
- 9 [am] Transmission (0 = automatic, 1 = manual)
- 10 [gear] Number of forward gears
- 11 [carb] Number of carburetors

Methods

The dataset was converted from numeric values to factor values, and subsequently MPG against five other variables. The p-values shows that the cylinder count, displacement in cubic centimeters (cc) and weight were significant predictors for MPG as the combined outcome.

The calculation of variance-inflation, which is used for linear and generalized linear models, as in this report, showed that four factors, cylinder, displacement, horsepower and weight were highly corelated with each other. See Table 3.

Linear Models

Multivariate models were used to analyze multiple parameters by adding extra variables to a single variable model. Considering the p-values and variability inflation, cylinders, displacement, weight and horsepower were selected for the regression model. The *p-value* is the probability of obtaining a value as numerically large as or larger that the observed *t*-tests, which are based on the assumption that the data is *normally distributed* (data builds a bell-curve shaped). Each variable was added to the model, along with the transmission type. A analysis of variance using the ANOVA function, showed the degree of freedom and p-values of each model, Table 1-3.

```
anova(fit1,fit2,fit3,fit4,fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + disp
## Model 4: mpg ~ am + cyl + disp + wt
## Model 5: mpg ~ am + cyl + disp + wt + hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 264.50  2    456.40 37.9300 2.678e-08 ***
## 3     27 230.46  1     34.04  5.6572  0.025339 *
## 4     26 182.87  1     47.59  7.9102  0.009429 **
## 5     25 150.41  1     32.46  5.3954  0.028621 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fit1)$coefficients[1:2,] #single variable model
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am1          7.244939   1.764422  4.106127 2.850207e-04
```

```
summary(fit5)$coefficients[1:2,] #multivariate model
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.864276   2.695416 12.563656 2.668321e-12
## am1          1.806099   1.421079  1.270935 2.154510e-01
```

Discussion

In the multivariate model, the manual transmission is 1.81 mpg better than automatic transmission. R-squared is a commonly used measure of the overall fit of the regression model, which in this instance, there is an increased of *goodness of fit* from 36% to ~87%. The residuals of fit5 are shown in Figure 3.

```
c(summary(fit1)$r.squared, summary(fit5)$r.squared)
```

```
## [1] 0.3597989 0.8664276
```

Appendix

Table 1, Summary of the *mtcars* dataset

```
##                      mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4           21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710          22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout   18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
```

Table 2, Analysis of Variance

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## cyl          2  824.8   412.4  51.377 1.94e-07 ***
## disp         1   57.6    57.6   7.181   0.0171 *  
## hp           1   18.5    18.5   2.305   0.1497    
## drat         1   11.9    11.9   1.484   0.2419    
## wt           1   55.8    55.8   6.950   0.0187 *  
## qsec         1    1.5     1.5   0.190   0.6692    
## vs           1    0.3     0.3   0.038   0.8488    
## am           1   16.6    16.6   2.064   0.1714    
## gear         2    5.0     2.5   0.313   0.7361    
## carb         5   13.6     2.7   0.339   0.8814    
## Residuals   15  120.4     8.0                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 3, Variance-inflation

```
##           GVIF Df GVIF^(1/(2*Df))
## disp  60.36569  1        7.769536
## hp    28.21958  1        5.312210
## wt    23.83083  1        4.881683
## cyl  128.12096  2        3.364380
```

Figure 1, Number of cars within bin size (MPG + 0.5) for MPG per Transmission

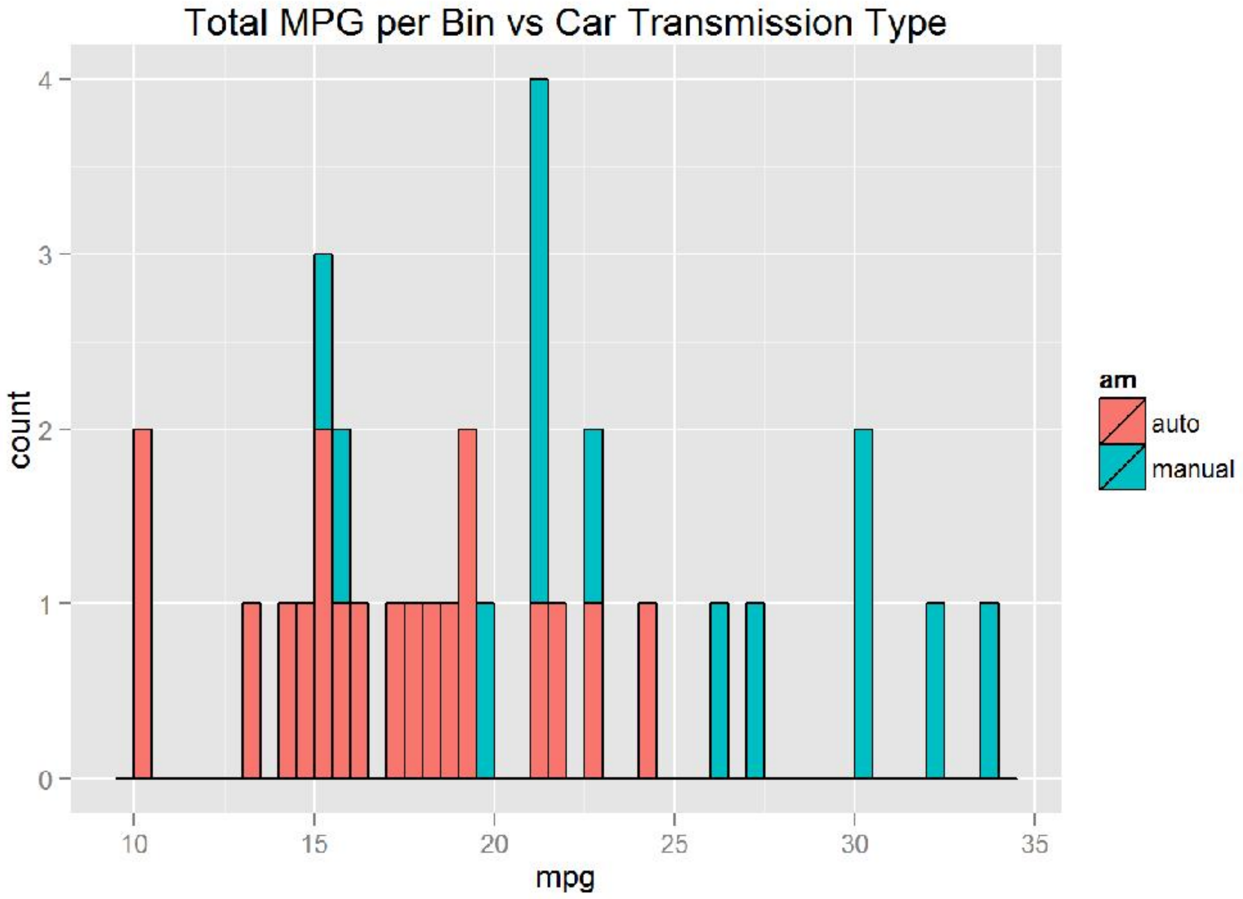


Figure 2, MPG vs Transmission

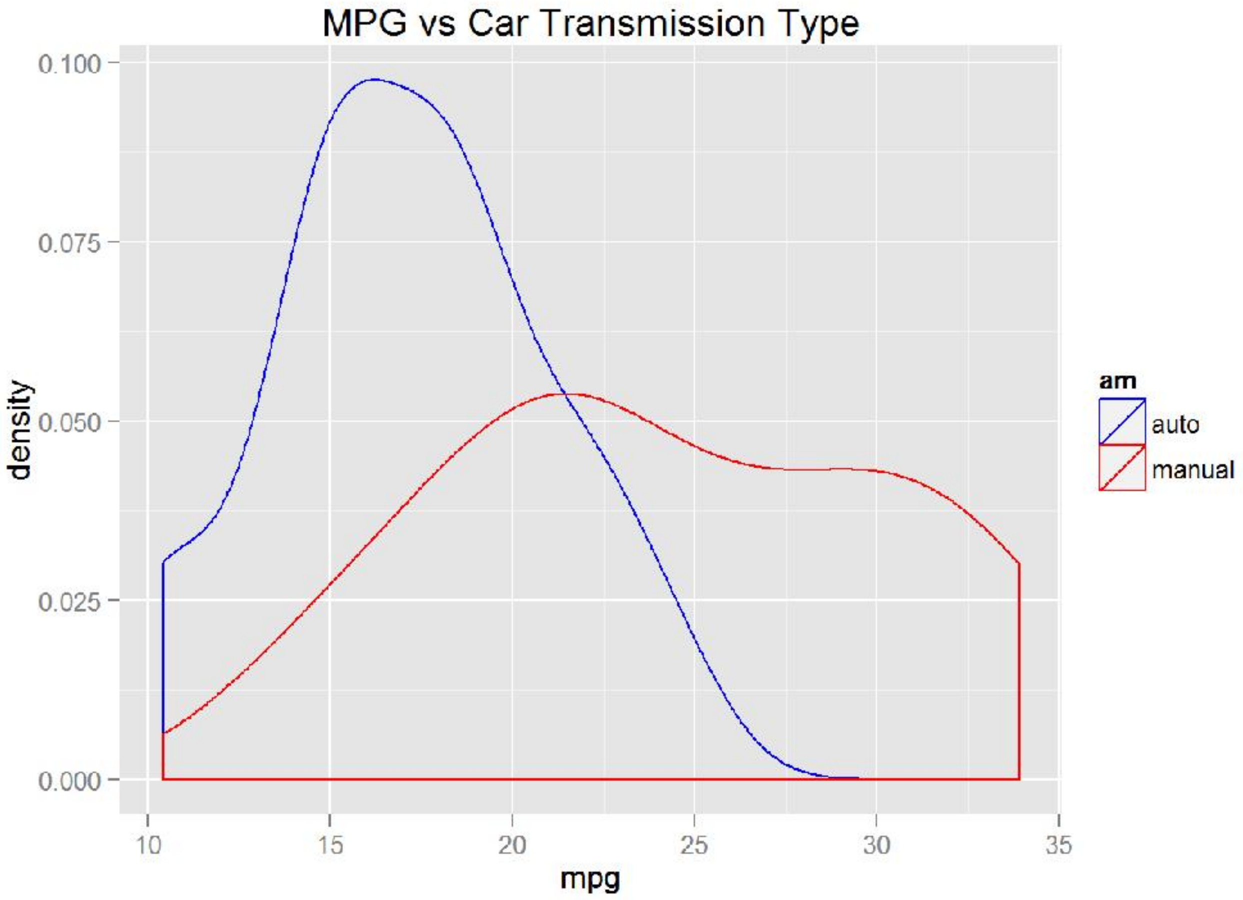
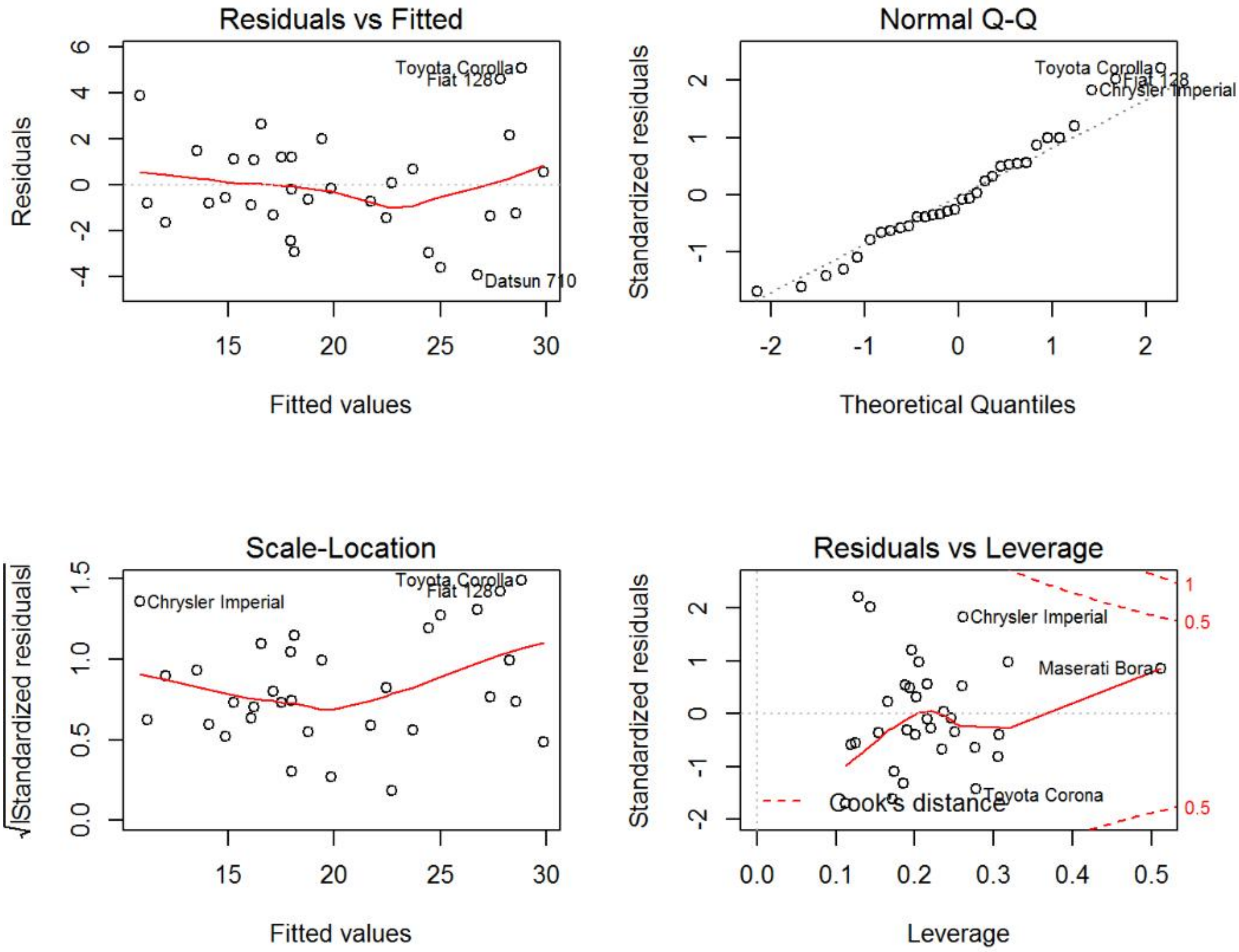


Figure 3, Residual Plots

```
par(mfrow=c(2,2))
plot(fit5)
```


Residuals vs Fitted
Residuals
-4
-2
0
2
4
6
15
20
25
30
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q-Q
Standardized residuals
-1
0
1
2
-2
-1
0
1
2
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
15
20
25
30
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
-2
-1
0
1
2
0.0
0.1
0.2
0.3
0.4
0.5
Leverage
Chrysler Imperial
Maserati Bora
Toyota Corona
Cook's distance
1
0.5
0.5