Regression Models: Motor Trend Car Road Tests - Effects of Transmission on MPG

Jhon H. Guzman

8/04/2017

Contents

1. Executive Summary

The analysis attempts to find out the relationship between miles-per-gallon (Mpg) and other variables in the mtcars data set. In particular, to determine whether an automatic or manual transmission is better for Mpg, and quantifies the Mpg difference. The Analysis focuses on inference with a simple linear regression model and a multiple regression model. Both models support the conclusion that the cars in this study with manual transmissions have on average significantly higher MPG's than cars with automatic transmissions. This conclusion holds whether we consider the relationship between `Mpg` and transmission type `am` alone or transmission type together with `weight` and `qsec` as aditional predictors. In the multiple regression model selected as best fit model, the MPG difference is `2.9358` Mpg in Manual transmission than those with automatic transmission.

2. Exploratory Data Analysis

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

2.1 Reading data

load our dataset from the library *datasets* in R

We first read in data from from the library *datasets* in R. The data is the data frame `mtcars (Motor Trend Car Road Tests)` Initially we do read the header data.

```
data(mtcars)
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

The variables of dataset are as follow:

- `mpg` as a measure of Miles/(US) gallon
- `cyl` as a measure of Number of cylinders
- `disp` as a measure of Displacement (cu.in.)
- `hp` as a measure of Gross horsepower
- `drat` as a measure of Rear axle ratio
- `wt` as a measure of Weight (1000 lbs)
- `qsec` as a measure of 1/4 mile time
- `vs` as a measure of V/S
- `am` as a measure of Transmission (0 = automatic, 1 = manual)
- `gear` as a measure of Number of forward gears
- `carb` as a measure of Number of carburetors

After reading, we check (there are `32`) observations and `11` variables in this dataset. 19 observations are for automatic transmission cars and 13 for manual transmission cars. There are no variables which allow us to compare manual vs automatic transmission directly.

2.2 Identify interested data

According to the box plot in *Figure 1*, manual transmission (1) yields a higher Mpg and there is another interesting variable `weight` thats shows in the box plot (weight ~ am) that cars with manual transmission have less weigth than those with automatic transmission.

The above suggests that Mpg performance depends not only on the type of transimision `am` but also on additional variables that must be included in the model like `wt` and `qsec`.

In *Figure 2*, a pair graph is used to generate scatterplots to show the relationship between best fit model variables. *Figure 3* shows the Residual best fit model mtcars in dataset.

3. Regression Analysis

First build different regression models using different variables in the model and find the best model fit using `step` R function, then perform analysis of residuals with the selected model.

3.1 Basic Model `mpg ~ am`

The basic model includes `mpg` as intercept and the interest variable `am` as predictor.

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

With the basic model `lm(formula = mpg ~ am, data = mt)`, a car with an automatic transmission has an average of `17.147` Mpg, and manual transmission increases mpg by `7.245`. However, this model has an Adjusted R-Squared of `0.3597989`, which means the model can only explain about `35.97%` of the variance of the MPG.

3.2 Stepwise Model

With stepwise model selection we use `step` R function that runs the linear models multiple times, build RMs, and select the best variables using forward selection and backward elimination methods with the AIC algorithm.

The result of `step` analysis shows that the best fit model is `lm(formula = mpg ~ wt + qsec + am, data = mt)` consists of three variables `wt` and `qsec` as confounders and variable `am` as independent variable. The Adjusted R-Squared value is `0.8496636`, which means the model can explain about `84.96%` of the variance of the MPG.

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

Then, the regression equation is *mpg = 9.618 -3.917 wt + 1.226 qsec + 2.9358 am.*

3.3 Selected Model Examination

The best fit model is `lm(mpg ~ wt + qsec + am, data = mt)`

Analysis of Variance Table (ANOVA)

The model include for the analysis are as follow:

`Model 1` Model with *(mpg ~ wt)*
`Model 2` Best fit model from step function *(mpg ~ wt + qsec + am)*

We are confident that `Model 2` is significantly better than `Model 1`, since the p-value analysis for the `am` coefficient in Model 2 is `0.4432` and the p-value for the best fit model of `1.55e-09` from Analysis of Variance Table is highly significant, we reject the null hypothesis and conclude that the three confounders `wt`, `qsec` and `am` contribute significantly to the model.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Coefficients

The best fit regression model shows that the coefficient for Manual transmission type is `2.9 Mpg` extra compared to automatic transmission.

```
## (Intercept)          wt        qsec          am
##    9.617781   -3.916504    1.225886    2.935837
```

Residuals and Diagnostics

In the residual plots (*Figure 3.*) we can verify the independence condition by looking at the Residuals vs. Fitted plot and verifying the randomness of the scatter of points but there is a bit of a curve, so that it departs slightly from normality. The residuals for the Merc 230, and Chrysler Imperial are called out because they exert potential influence as shown in *Figure 4.*.

The potential leverages coefficient are

```
##       Maserati Bora  Cadillac Fleetwood   Chrysler Imperial
##           0.1909815           0.2270069           0.2296338
## Lincoln Continental            Merc 230
##           0.2642151           0.2970422
```

The potential influencers in coefficent values are

```
## [1] 0.4765680 0.4968861 0.5481272 0.5626418 1.0938422
```

QQ Plot verify the normality condition for the regression residuals as the points in the plot mostly fall on the normal line.

The points in the Scale-Location are scattered in a constant pattern which verifies the constant variance condition. Using the Residuals vs. Leverage plot we can visually identify some outliers.

4. Conclusions

- The resulting best model `mpg ~ wt + qsec + am` is actually dependant on the transmission `am`, but also weight `wt` and 1/4 mile time `qsec`.
- Manual transmission is better for Mpg, holding all other parameters constant (wt, qsec), will increase an average of `2.94` more MPG than those with automatic transmission.
- The best fit model `mpg ~ wt + qsec + am` conclude that (with a $p < 0.05$ confidence) cars with manual transmission have `2.9358` more miles per gallon than automatic transmissions. The model seems clean with a $p < 0.05$ and a Multiple R-squared can explain about `84.97%` of the variance of the MPG.

Appendix Figures.

Figure 1. Boxplot of Mpg & Weight by transmission type

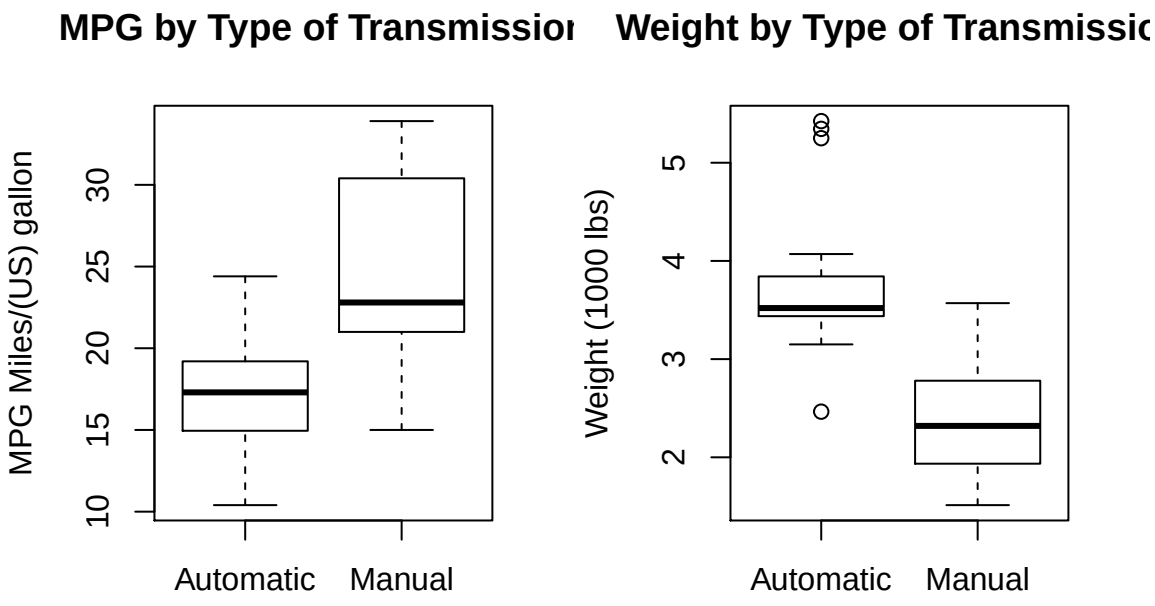


Figure 2. Pair plot mtcars

This is a pair plot for selected variables correlations associated with the best fit model with variables `mpg`, `wt`, `qsec` and `am`

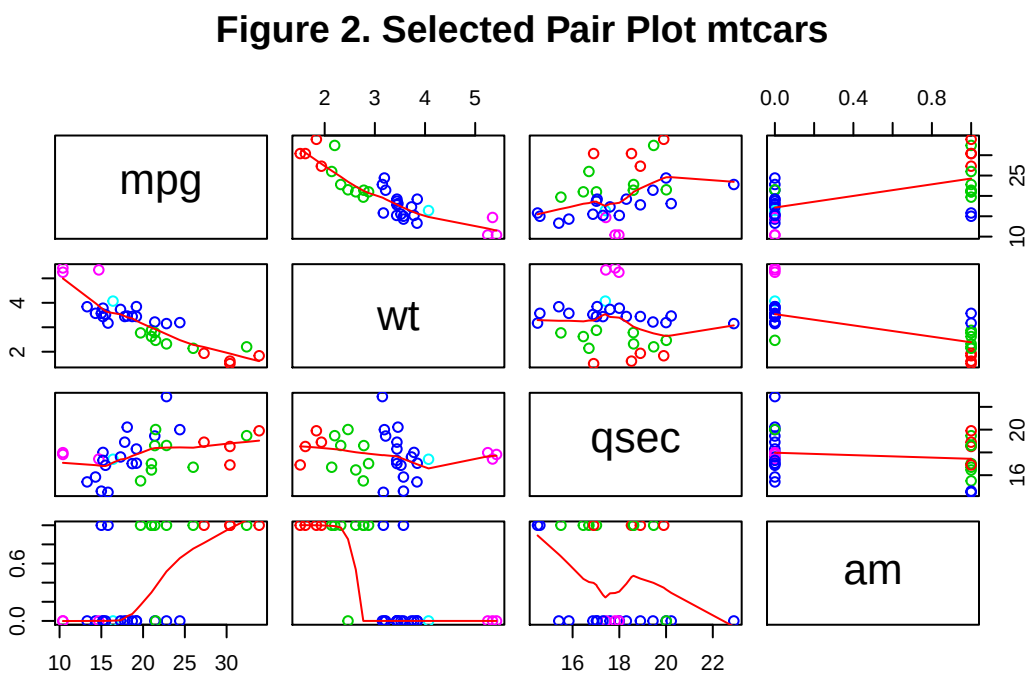


Figure 3. Residual best fit model mtcars

This is a pair plot for selected variables correlations associated with the best fit model with variables `mpg`, `wt`, `qsec` and `am`

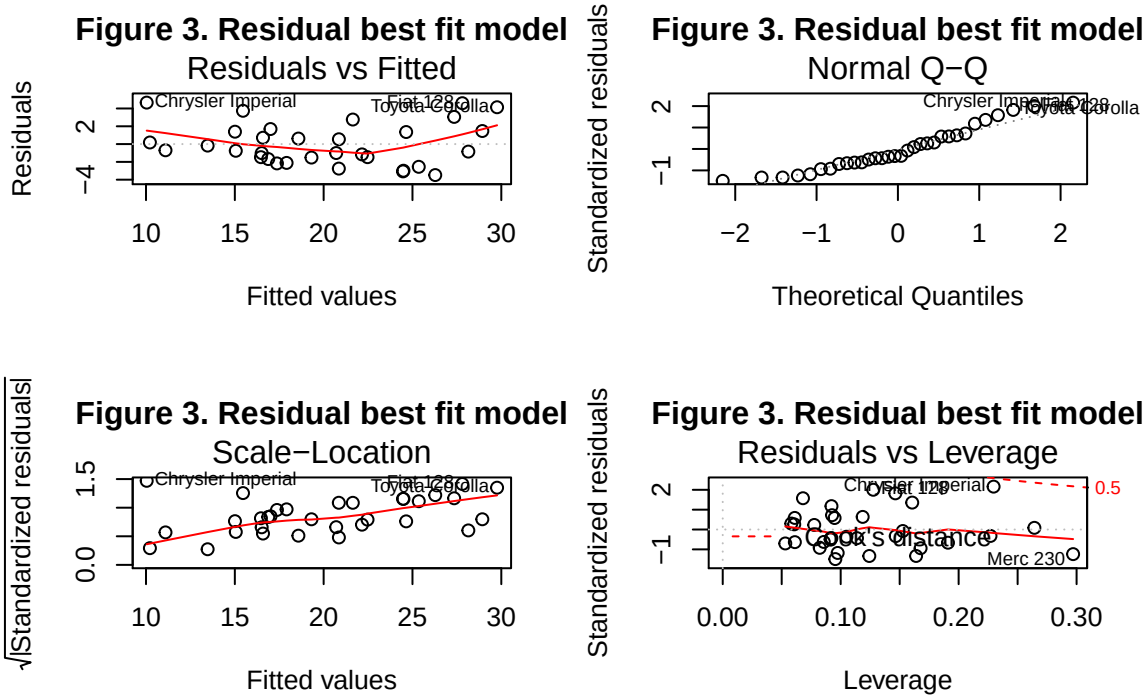
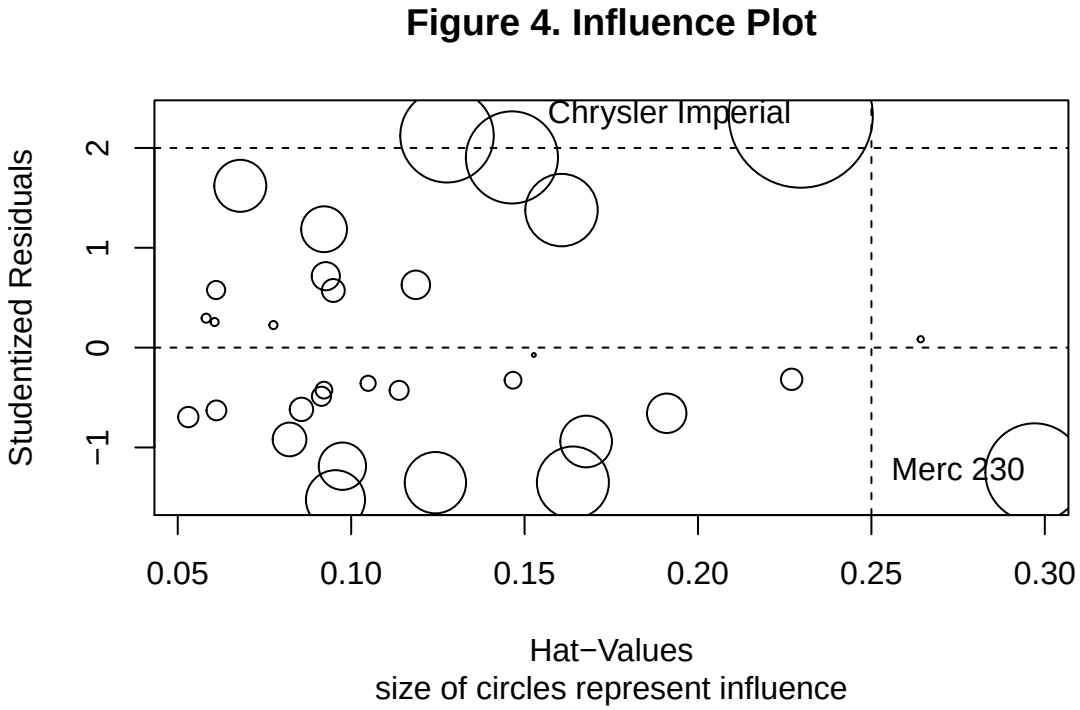


Figure 4. Influence Plot

This is a plot for influence hat values associated with the best fit model with variables `mpg`, `wt`, `qsec` and `am`




```
##                     StudRes       Hat     CookD
## Merc 230          -1.251106 0.2970422 0.1620827
## Chrysler Imperial  2.323119 0.2296338 0.3475974
```