Mtcars regression model

Dmitry S.

November 22, 2015

Contents

Executive summary

In ths project we are interested in the following two questions: "Is an automatic or manual transmission better for MPG (Mile per gallon)?" "Quantify the MPG difference between automatic and manual transmissions" I investigate that the average MPG for cars with automatic transmissions is 17.15 MPG, and the average MPG for cars with manual transmissions is 24.39 MPG. As we can see H0 hypotesis that there is not difference between averedge "Manual" MPG and average "Automatic" MPG must be regected.

The average MPG for cars with automatic transmissions is 17.15 MPG, and the average MPG for cars with manual transmissions is 24.39 MPG. In the simple model, the mean MPG difference is 7.245 MPG (the MPG of manual transmitted cars is at the average 7.245 MPG more than MPG of automatic transmitted cars); In the multiple regression model, the MPG difference is 2.9358 MPG (the MPG of manual transmitted cars is at the average 2.9358 MPG more than MPG of automatic transmitted cars.)if all other variables are constant. So manual transmission is better for MPG (Mile per gallon).

R-squared ~ 85% or 85% of the mpg variation is explained by the multiple regression model.

Exploratory Analysis

Summary mpg distribution sttistics for the cars with the automatic transmission

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.95   17.30   17.15   19.20   24.40
```

Summary mpg distribution sttistics for the cars with the manual transmission

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   21.00   22.80   24.39   30.40   33.90
```

As we see average mpg is different for different type of transmission.

Simple linear regression model

Let's check a simple linear regression (mpg ~ am)

```
## (Intercept)          am
##   17.147368    7.244939
```

Check the H0 - null hipotesis (that there is not difference between averedge "Manual" MPG and average "Automatic")

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

The p-value is 0.00137, we reject our null hypothesis (that there isn't difference between MPG for Automatic and Manual mpg), MPG for the automatic and manual transmissions is different. The MPG of manual transmitted cars is at the average 7.245 MPG higher than MPG of automatic transmitted cars.

Multiple regression model

Only 36% of the mpg variation (R-squared) is explained by the simple linear model, so we need to the multiple regression model

Model selection strategies: in this case I use backward selection.

Backward elimination begins with the largest model and eliminates variables one- by-one until we are satisfied that all remaining variables are important to the model. Forward selection starts with no variables included in the model, then it adds in variables according to their importance until no other important variables are found. When we care about understanding which variables are statistically significant pre- dictors of the response, or if there is interest in producing a simpler model at the potential cost of a little prediction accuracy, then the p-value approach is preferred. In backward elimination, we would identify the predictor corresponding to the largest p-value. If the p-value is above the significance level, usually $\alpha = 0{,}05$, then we would drop that variable, refit the model, and repeat the process. If the largest p-value is less than $\alpha = 0.05$, then we would not eliminate any predictors and the current model would be our best-fitting model.

Lagest model:

```
## (Intercept)         cyl        disp          hp        drat          wt
## 12.30337416 -0.11144048  0.01333524 -0.02148212  0.78711097 -3.71530393
##        qsec          vs          am        gear        carb
##  0.82104075  0.31776281  2.52022689  0.65541302 -0.19941925
```

Optimized model with significant variables

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

Conclusion: If all other variables are constant the MPG of manual transmitted cars is at the average 2.9358 MPG higher than MPG of automatic transmitted cars. R-squared ~ 85% or 85% of the mpg variation is explained by the multiple regression model.

Residuals and model diagnostics

The data points with the most leverage in the fit can be found by looking at the hatvalues() and those that influence the model coefficients the most are given by the dfbetas() function.

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

##    Toyota Corolla     Toyota Corona          Fiat 128 Chrysler Imperial
##         0.3174637         0.4050410         0.4765680         0.5626418
```

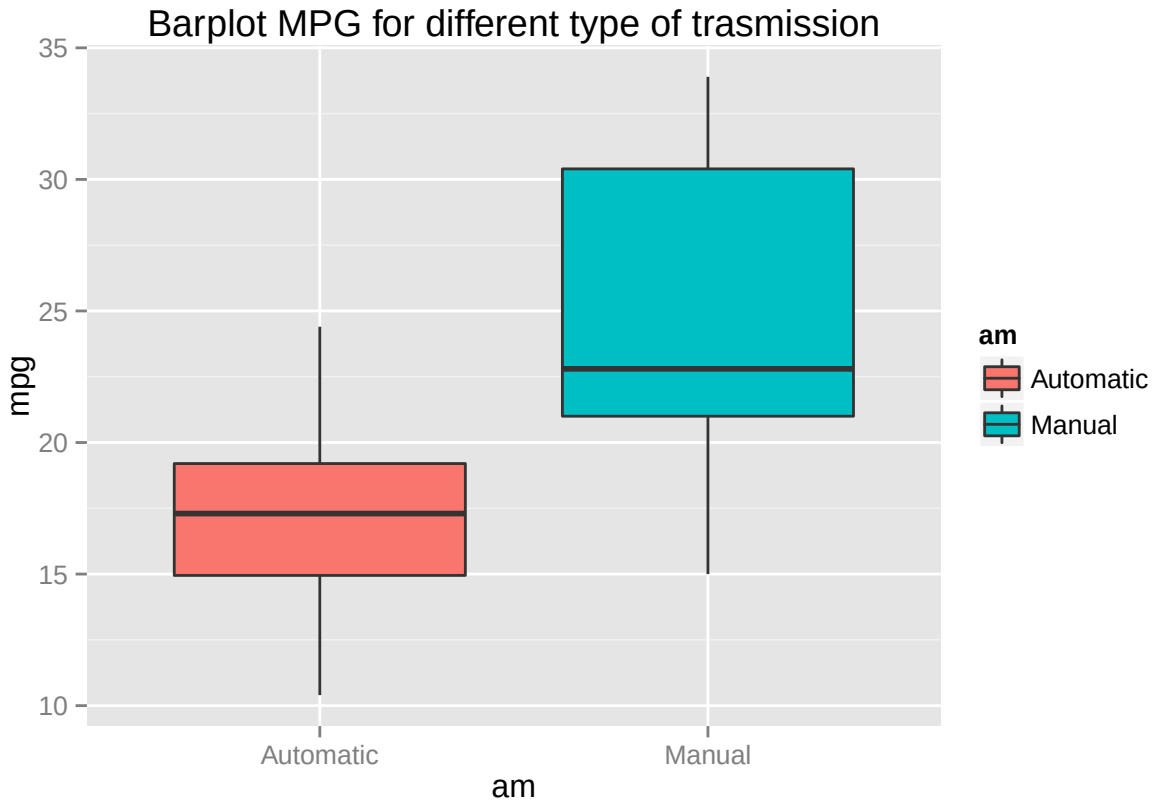
All residuals plots are in Appendix
The points in the Residuals vs. Fitted plot are randomly scattered on the plot that verifies the independence condition.

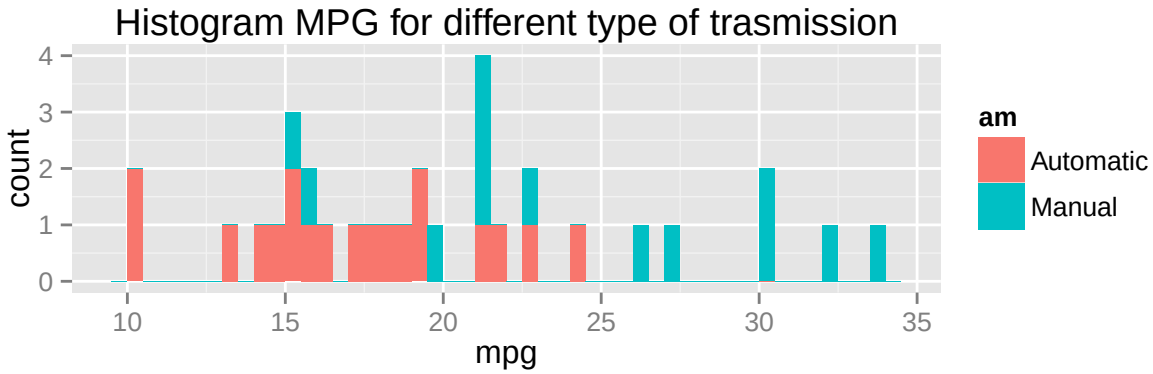
- The Normal Q-Q plot consists of the points which mostly fall on the line indicating that the residuals are normally distributed.
- The Scale-Location plot consists of points scattered in a constant band pattern, indicating constant variance.
- There are some distinct points of interest (outliers or leverage points) in the top right of the plots that may indicate values of increased leverage of outliers.

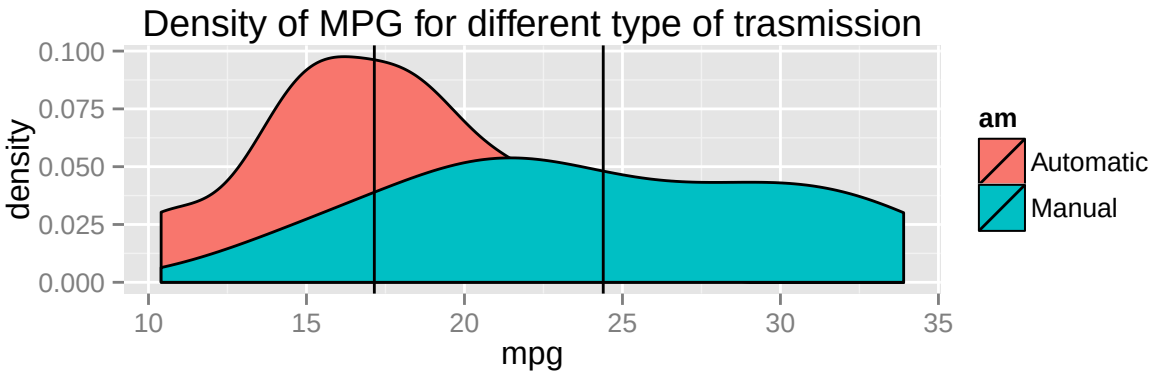
Looking at the above results, we notice that our analysis was correct, these are the same cars as mentioned in the residual plots.

Appendix

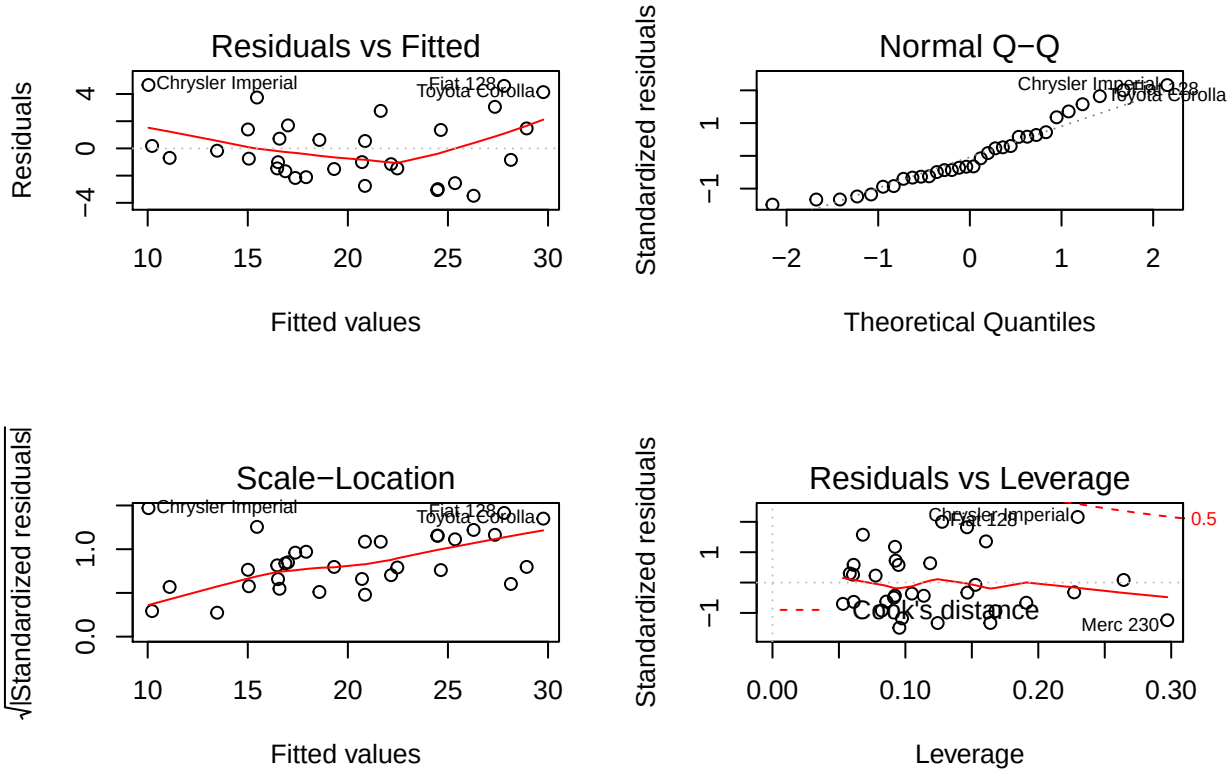
Summary statistics visualization




Histogram MPG for different type of trasmission
count
mpg
am
Automatic
Manual


Density of MPG for different type of trasmission
density
mpg
am
Automatic
Manual

Model Residuals



Variables correlation