Regression Models - Motor Trend

Rajesh Ekkaladevi

Mar 09, 2020

Executive Summary

The goal of the exercise is to review **mtcars** dataset and answer the following questions:

1. Is an automatic or manual transmission better for MPG.
2. Quantify the MPG difference between automatic and manual transmissions.

First, we will take a look at the input data and perform exploratory data analysis to understand the data better. Later we will fit linear regression models on the features and perform regression diagnostics. Finally come up with the best fit model along with assumptions made.

Data Processing

The **mtcars** dataset has 32 observation and 11 variables. None of the variables are missing data (having Null, NA or empty). Some of the variables are indeed factors so, the dataframe is converting into factors overwriting the original dataset. Refer to Appendix Section 1 for more details.

Exploratory Data Analysis

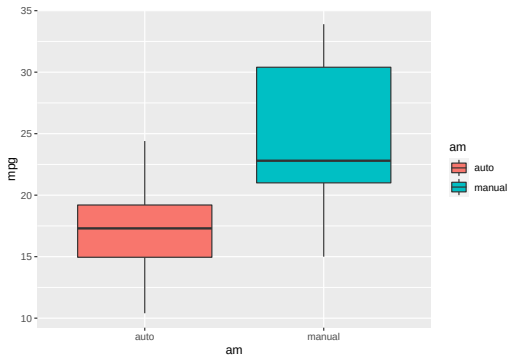
The exploratory data analysis with pair plot (refer to Appendix Section 2) shows how mpg is corrGelated with the other variables. Noteable observations are:
1. **am** with manual has more mpg compared to auto.
2. **cyl** with lower values (4, 6) have higher mpg compared to higher values (8) and the difference between manual vs auto transmission is higher at lower values with manual tranmission achieving higher mpg.
3. **gear** with lower values show higher difference between manual vs auto with manual trasmission achiving higher mpg.
4. **wt** shows higher correlation with manual (0.909) compared to auto transmission (0.768).
5. **qsec** shows higher correlation with manual (0.802) compared to auto transmission (0.657).

Regression Modeling

fit 1: lm(mpg ~ am, mtcars)

Single variate model between outcome (mpg) and predictor (am).

Based on the fit1 coefficients (refer to Appendix Section 3) the manual transmission motor is 7 miles efficients than auto. The below box diagram shows it as well.



In the given dataset **mtcars** there are many other variables provided that would impact the mileage. The above simple model could be biased if we ignore other variables that may correlate with **am** or it may be an underfit model.

fit 2: lm(mpg ~ ., mtcars)

Multi variate model between outcome (mpg) and using all predictors (.).

Based on the fit2 coefficients (refer to Appendix Section 4) the residual error is smaller than fit 1 but the F-static is lower (measure of significance of group of predictors) and also the P values are not significant either.

fit 3: step(lm(mpg ~ ., mtcars), trace=FALSE)

Multi variate model after using **step()** function that would provide the best model based on minimum variance and high degrees of freedom.

Based on the fit3 coefficients (refer to Appendix Section 5) the residual error is smaller than fit 1 and fit 2. The F-static is higher than any of the previous models and also the P values are significant for all the predictors (wt, qsec, am).

Looking at the diagnostic plot in Section 5 the residuals appear slightly off from normal distribution (the 95% confidence internval gray band). There is one data point which has high leverage according to cooks distance.

The fit 3 (mpg ~ cyl + wt + qsec + am) is the best model considered so far.

Conclusion

Answering to the questions:

1. Is an automatic or manual transmission better for MPG.
   All the models (fit1, fit2, fit3) show that, holding all other variables constant, manual transmission will increase mpg.
2. Quantify the MPG difference between automatic and manual transmissions.
   It is harder to answer this question as based on the models reviewed fit1, fit2, fit3 the mpg of manual is varying from 7.245, 1.212 to 1.809 miles respectively more than auto transmission. As the best fit model (fit 3) has higher F-static and significant P values for each of the predictor and also the adjusted R^2 (variance covered by the model) is very high (0.84) so, we can say that the manual transmission provides 1.809 miles more than auto.

The residual vs fitted model show the residuals are not normal (not symmetric to the horizontal line). Also, as there are only 32 observations which is very low number (the effect size is small) so, I cannot conclude that the model will fit all future observations.

Appendix

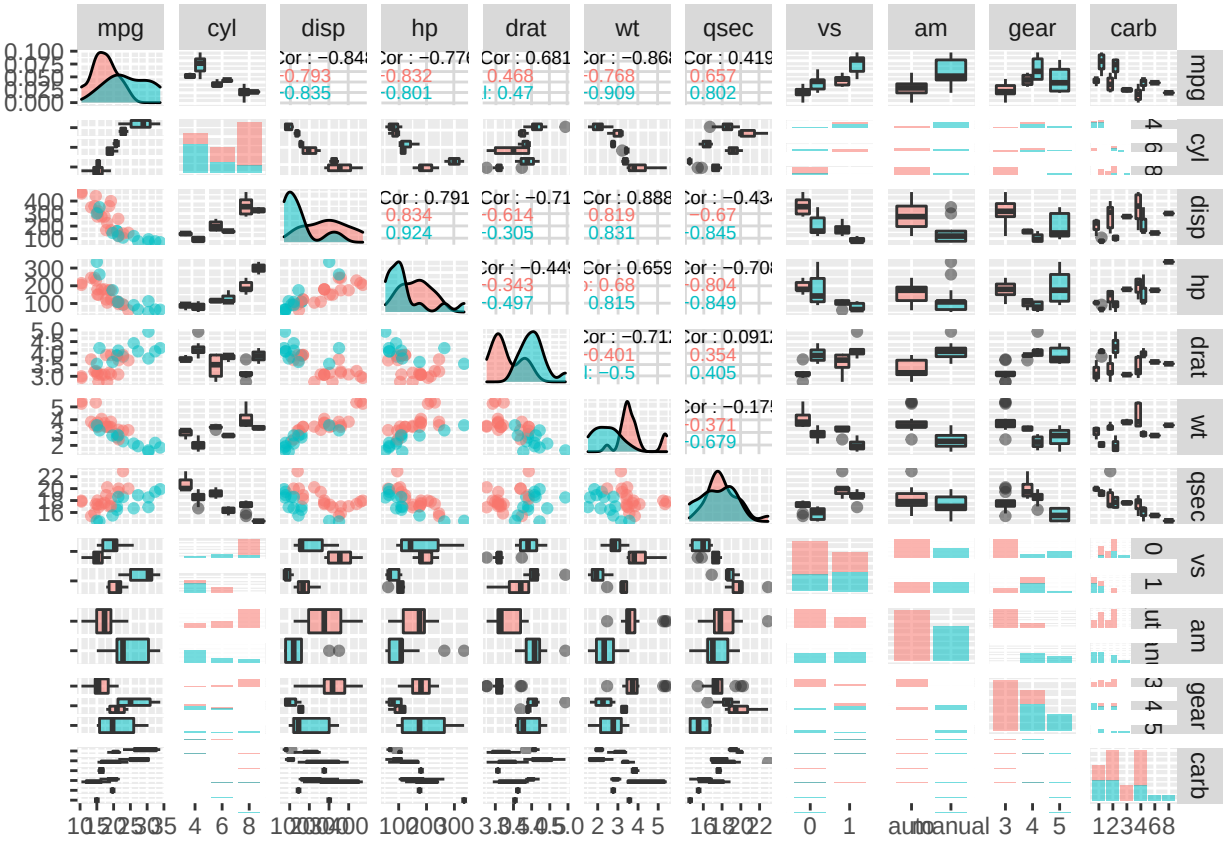
Section 1

Basic review of mtcars dataset:

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "auto","manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

Section 2

Exploratory data analysis between all variables of mtcars dataset:



Section 3

fit1 summary:

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## ammanual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

Section 4

fit2 summary:

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## ammanual     1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
```

```
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

Section 5

fit3 summary:

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## ammanual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

Diagnostic plots on multi variate best fit model: