Regression Models Course Project

Arpit Chaudhary

June 11, 2017

Contents

Executive Summary

This analysis tries to establish a relationship between MPG(Miles per gallons) with other variables present in the dataset mtcars(available by default with R). Specifically the analysis focusses on finding a relationship between MPG and transmission(Automatic/Manual) of the car.

- Manual cars have higher mpg for cars with weight less than 3400 lbs.
- Automatic cars have higher mpg for cars with weight greater than 3400 lbs.

Exact difference in mpg for similar(same weight and qsec) manual vs automatic car can be found using the equation below:

MPG_automatic - MPG_manual = 14.079 - 4.141*wt

Processing Data

Changing the "am" (transmission) variable to factor for better view in plots and so that the variable can directly be used for linear model fitting as factor(categorical) variable and not continuous.

```
myMtCars <- mtcars; myMtCars$am <- factor(myMtCars$am);
levels(myMtCars$am) <- c("Automatic", "Manual")
```

Fitting a linear model

```
fitAm <- lm(mpg ~ am, data = myMtCars)
bestFit <- step(lm(mpg ~ ., data = myMtCars), trace = 0, direction = "both")
improvedFit <- lm(mpg ~ wt + qsec + am*wt, data = myMtCars)
anova(fitAm, bestFit, improvedFit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
## Model 3: mpg ~ wt + qsec + am * wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
```

```
## 2     28 169.29  2    551.61 63.497 6.195e-11 ***
## 3     27 117.28  1     52.01 11.974  0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using step function to find the best linear model for the data provided. It can be seen from the boxplot that automatic cars in the dataset have higher weights than most manual cars. Thus adding interaction of am and wt in the model and testing all these models using anova as they are nested.

Interpretation of Coefficients

For the project we are using a linear model where mpg(Miles per Gallon) is the dependent variable while weight, time taken to cover quarter mile(measure of acceleration) are the predictors.

- (Intercept) : Estimated mpg value when weight of the the car is 0, qsec is 0 and the car is automatic. This is a base case which would not be seen practically in the real world.
- wt : Change in mpg value with a unit increase in weight of an automatic car keeping qsec constant.
- qsec : Change in mpg value for a unit increase in qsec value keeping all other variables(wt and transmission) constant.
- amManual : Difference in mpg value, when a the car under is consideration has manual transmission compared to a car with automatic transmission, keeping all other variables(wt and qsec) as constant.
- wt:amManual : Change in mpg value with a unit increase in weight of an manual car keeping qsec constant.

Conclusion

The model named improvedFit is statistically significant and can be used to predict the mpg for a car based on weight, acceleration and transmission. Adjusted R-square value for the fit is 0.88 which means that the model explains most of the variations in the mpg based on predictor variables used. Also p-value for all the coefficients of predictors and the overall model is very low which means that the model is statistically significant.

Summary of fit canbe found in appendix. The linear equation to predict mpg is as follows:

```
mpg = 9.723 - 2.937*wt + 1.017*qsec +  14.079*amManual - 4.141*amManual*wt
```

Is an automatic or manual transmission better for MPG?

It can be seen that manual cars have better mpg than the automatic cars for cars with lower weight while automatic cars have better mpg in the higher weight segment.

Quantifying the MPG difference between automatic and manual transmissions.

MPG_automatic - MPG_manual = 14.079 - 4.141*wt

where MPG_automatic is mpg for automatic car and MPG_manual is for a manual car, with both the cars having same wt and qsec values. At wt = 14.079/4.141 = 3.3999 (1000lbs) difference in mpg for similar(same weight and qsec) manual and automatic car is 0.

- Manual cars have higher mpg for cars with weight less than 3400(approx) lbs.
- Automatic cars have higher mpg for cars with weight greater than 3400(approx) lbs.

Appendix

For the Grader

- Did the student interpret the coefficients correctly?
  - Mentioned under "Interpretation of Coefficients"
- Did the student do some exploratory data analyses?
  - Available in Appendix
- Did the student fit multiple models and detail their strategy for model selection?
  - Under "Fitting a linear model"
- Did the student answer the questions of interest or detail why the question(s) is (are) not answerable?
  - Can be found under Conclusion
- Did the student do a residual plot and some diagnostics?
  - Available in Appendix
- Did the student quantify the uncertainty in their conclusions and/or perform an inference correctly?
  - I feel the inference is correct and model statistically significant. Although it would be better if the data available was randomized based on transmission as this would provide more trustable results. Current data have more cylinders/weight/displacement for automatic cars. This may have biased the analysis towards manual cars having higher mpg.
- Was the report brief (about 2 pages long) for the main body of the report and no longer than 5 with supporting appendix of figures?
  - Yes
- Did the report include an executive summary?
  - Yes, first heading
- Was the report done in Rmd (knitr)?
  - Yes

Exploratory Analysis

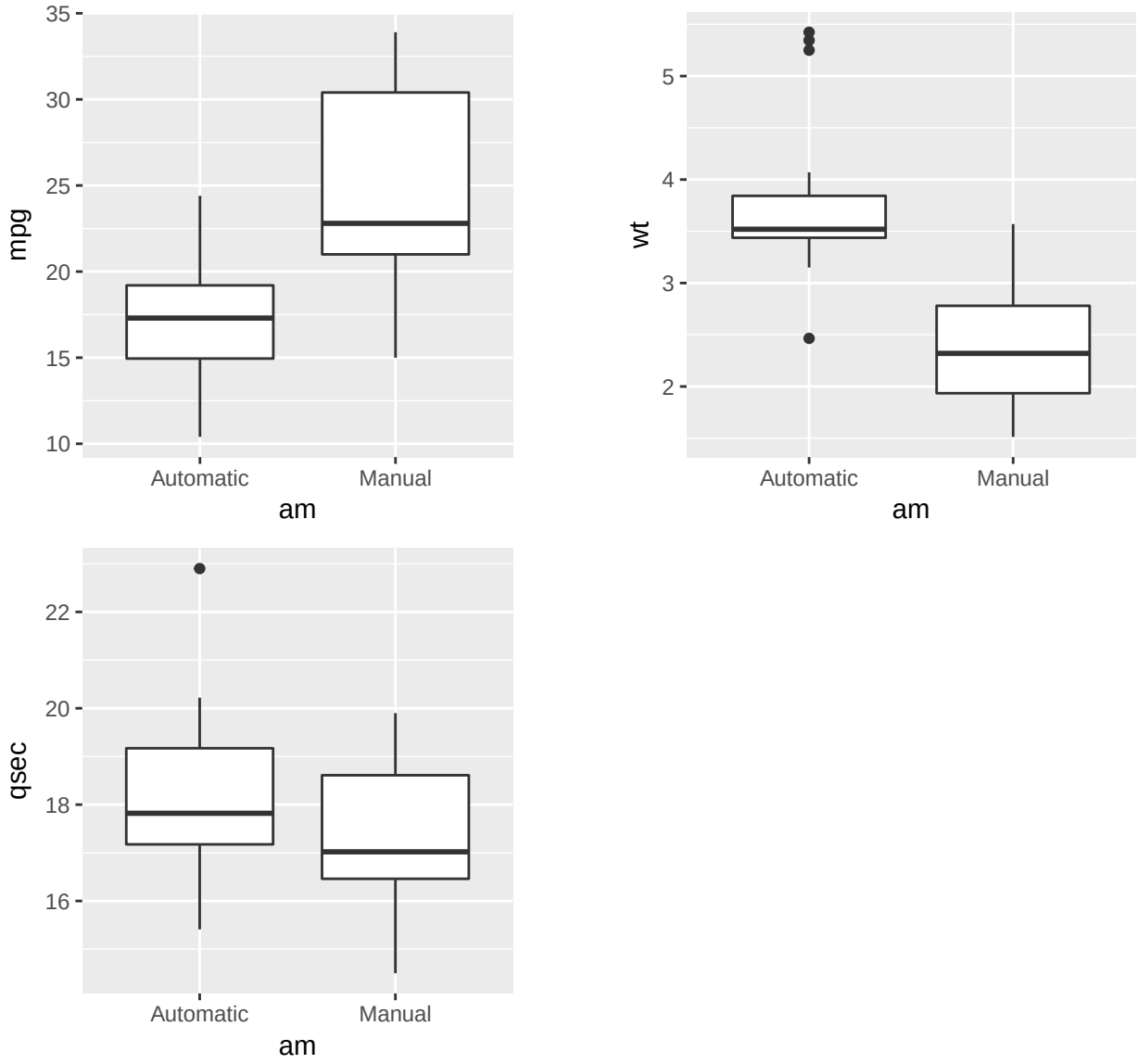
T-test for correlation between mpg and am variable.

```
t.test(mpg~am, data = mtcars)$p.value
```

```
## [1] 0.001373638
```

The above model takes into account just am to predict mpg but this model can be further improved by adding more variables which is done in the report.

Boxplots to show relation of mpg with other variables used in regression model



Summary of fit

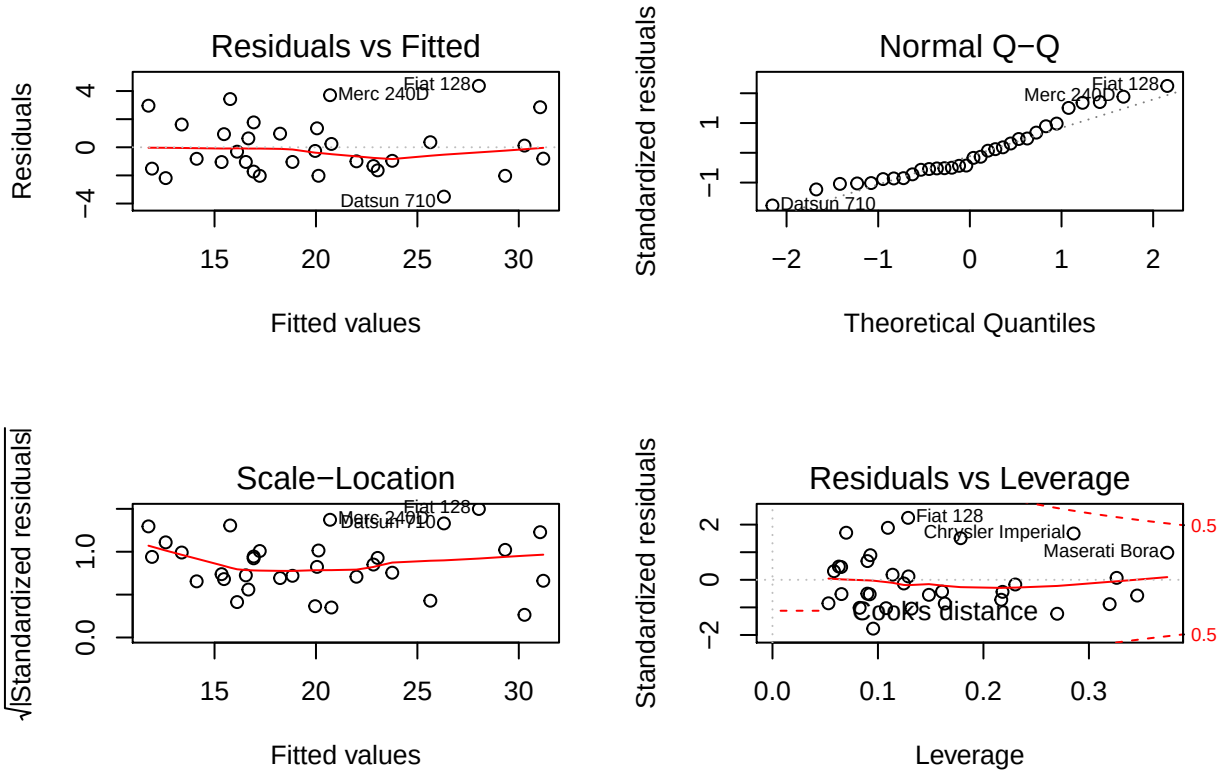
```
summary(improvedFit)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am * wt, data = myMtCars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    9.723      5.899   1.648 0.110893
## wt            -2.937      0.666  -4.409 0.000149 ***
## qsec           1.017      0.252   4.035 0.000403 ***
## amManual      14.079      3.435   4.099 0.000341 ***
## wt:amManual   -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

Residual plot and other diagnostics

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



- There is no pattern seen in the distribution of residuals.
- They appear normally distributed.
- heteroscedasticity is not seen.

Residuals follow the assumptions for a linear regression.