Regression Models - Assignment

H.Harvey

24 January 2016

Executive summary

In this report, we analyze the (mtcars) data set and explore the relationship between a set of variables and miles per gallon (MPG). We use exploratory data analysis and regression models to explore how automatic (am = 0) and manual (am = 1) transmissions features affect the MPG feature and answer the following question:

```
Q1-"Is an automatic or manual transmission better for MPG"
Q2-"Quantify the MPG difference between automatic and manual transmissions"
```

A t-test shows that a difference in performance exist between automatic and manual transmission cars in favor of the manual transmission. A linear regression model suggested that manual transmission (am=1) improves the MPG value once other variables are controlled. However, adding an interaction between terms suggested that the improvement in MPG is dependent on the weight (wt) of the car and the 1/4 mile time(qsec) with its transmission type.

Completer exploratory, data analysis, nested models and other instruction in Appendix.

Manual Modeling (Automated regression model validation in Appendix)

In order to explore the relationship between MPG and all other variables, a linear model was generated to identify the most influencial predictor. The weight (wt), horse power (hp), 1/4 mile time (qsec) and the transmission type (am) were selected based on their lower p-value. To complement the model, the cylinder (cyl) was included on this nested model for additional validation. (models in Appendix)

```
# Initial model including all variables & further nested model
fit_0 <- lm(mpg ~.,data=mtcars)
```

```
# anova_model <- anova(fit_1,fit_2,fit_3,fit_4,fit_5)
tidy(anova_model)
```

```
##   res.df      rss df     sumsq statistic      p.value
## 1     30 278.3219 NA        NA        NA           NA
## 2     29 195.4636  1 82.858306 14.386948 0.0008414121
## 3     28 186.0593  1  9.404334  1.632904 0.2130404042
## 4     27 160.0665  1 25.992837  4.513218 0.0437011550
## 5     25 143.9817  2 16.084730  1.396421 0.2661264879
```

The nested model demonstrates that the inclusion of the (hp) (model 3) and (cyl) model (5) predictor doesn't improve the model. Hence, the final model includes (wt), (qsec) and (am). At equal weight and acceleration time, manual transmission seems to have a higher MPG.

```
# fit_6 <- lm(mpg ~ wt + qsec + factor(am), data = mtcars)
tidy(fit_6)
```

```
##          term  estimate std.error statistic      p.value
## 1 (Intercept)  9.617781 6.9595930  1.381946 1.779152e-01
## 2          wt -3.916504 0.7112016 -5.506882 6.952711e-06
## 3        qsec  1.225886 0.2886696  4.246676 2.161737e-04
## 4 factor(am)1  2.935837 1.4109045  2.080819 4.671551e-02
```

The coeff. for manual (am=1) indicates an increase of 2.936 MPG above of the automatic coeff. (9.618) with weight and the 1/4 mile time held constant. The model also suggests that an increase of 1T pounds in weight (wt) leads to a 3.917 reduction in MPG, while a slower acceleration time of 1 second (qsec) increases MPG by 1.226.

A permutation of interaction between predictor (made possible due to the limited numbers of predictor(3)) of the final model shows that there is an interaction between the type of transmission and the wt, which is leading to a better predicting model.(models in Appendix)

```
# fit_7 <- lm(formula = mpg ~ wt + qsec + factor(am) + wt:factor(am) , data = mtcars)
summary(fit_7)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + factor(am) + wt:factor(am), data = mtcars)
##
## 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 ***
## factor(am)1      14.079      3.435   4.099 0.000341 ***
## wt:factor(am)1   -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
```

Conclusion

We can infer that light manual transmission cars listed in the mtcars dataset shows a slightly better MPG than automatic transmission car. However, we cannot conclude that manual transmission cars have better MPG in general. In fact, the interaction between (wt) and the transmission type (am) suggest that the MPG gain in changing from automatic to manual would decreases (-4.141 above automatic coeff. -2.937) when the weight (wt) increase. Therefore, we could predict that light cars with a slow acceleration time would have better MPG.

Even thought the glmulti (automated modeling pck) produced a better model, the author decided to maintain the (fit_7) model for ease of explanation/understanding.

Appendix

The package glmulti and broom must be installed first Exploratory data analysis

```
#Mean and SD can be calculated for each model Automatic and Manual
model_mean_sd <- mtcars %>% select(am,mpg) %>% group_by(am) %>% summarise(mean(mpg),sd(mpg))
model_mean_sd
```

```
## Source: local data frame [2 x 3]
##
##   am mean(mpg)  sd(mpg)
## 1  0  17.14737 3.833966
## 2  1  24.39231 6.166504
```

H0: Automatique & Manual mpg are = 0 H1: Automatique & Manual mpg are difference <> 0

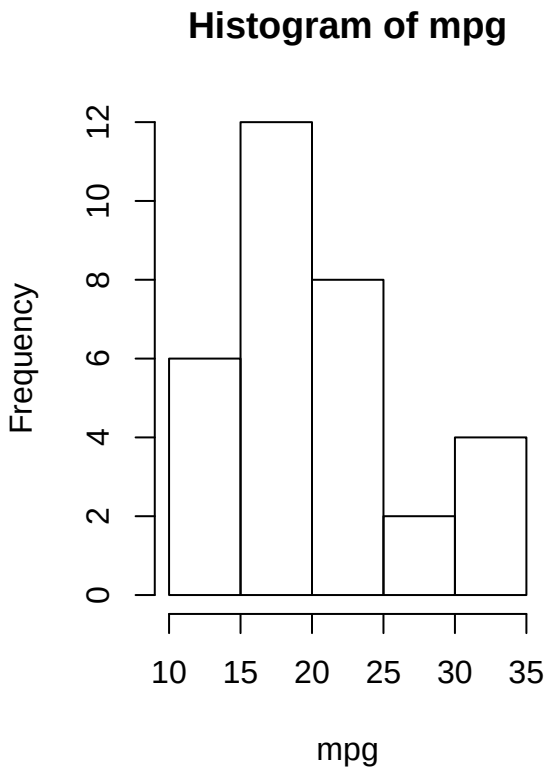
```
trans_aut <- mtcars %>% filter(am==0) %>% select(mpg)
trans_man <- mtcars %>% filter(am==1) %>% select(mpg)
```

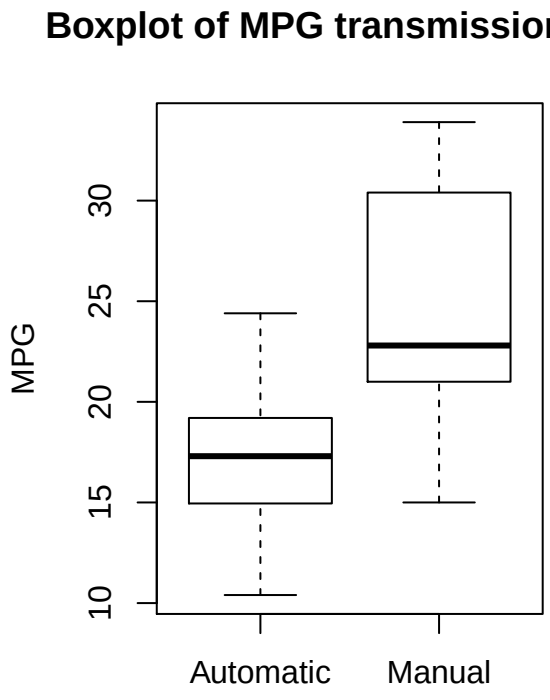
```
t.test(trans_aut,trans_man, alternative = 'two.sided')$p.value
```

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

We observe that manual transmission has a mean of 24.39 +/- 6.17 MPG while Automatic has a mean of 17.4 +/- 3.83 MPG. Where 0 = automatic and 1 = manual. Since the p-value is 0.00137, we reject our null hypothesis, which indicate that the automatic and manual transmissions are from different group.

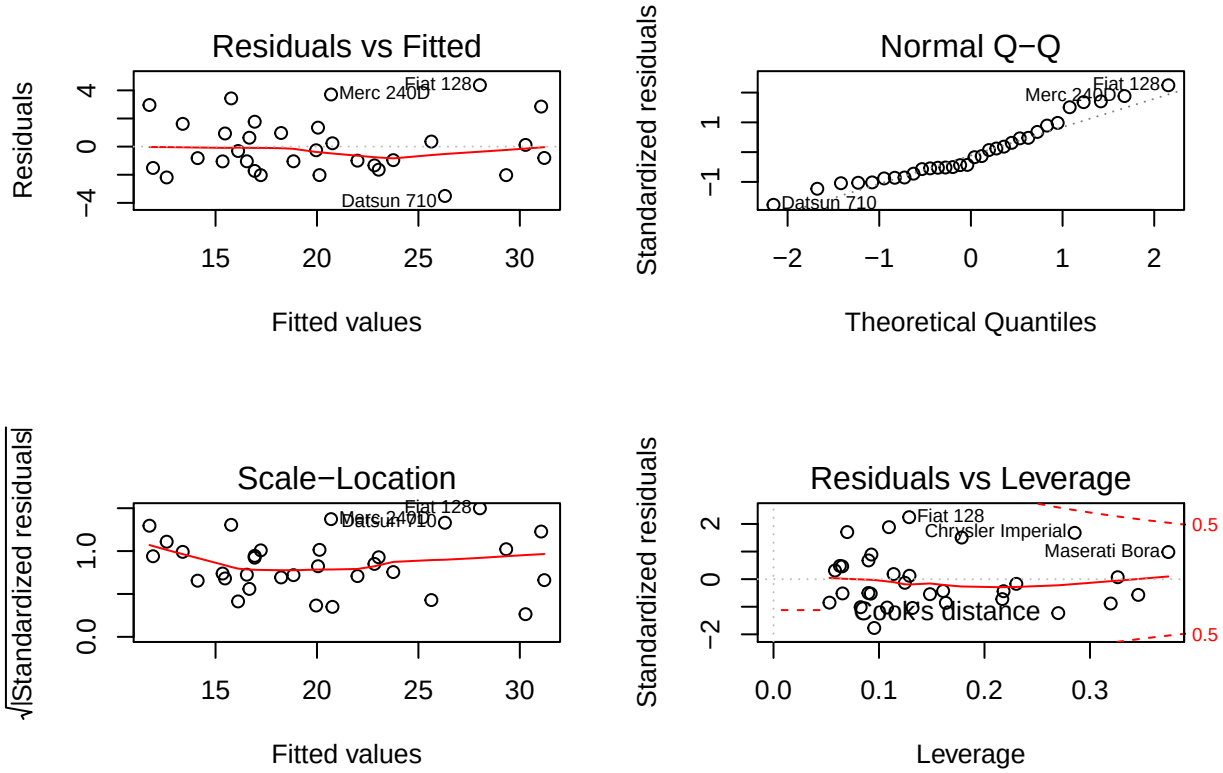
```
par(mfrow = c(1,2)); hist(mpg); boxplot(mpg~am, names = c("Automatic", "Manual"),ylab="MPG", main="Boxpl
```





Residual

```
par(mfrow = c(2,2))
plot(fit_7)
```



We can observe that there seems to be a few outliers that are skewing the data.

Automated Modeling

The utilisation of the glmulti package (automated modeling package) allowed to validate the above manual process. We first modeled the complete set of predictor without interaction between the predictor to keep the model simple. The number of models have been limited to 1024 (10exp2). This saves CPU time.

fit_auto <- glmulti(mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb, fitfunction = lm, level = 1, method = "h", confsetsize = 1024, report = FALSE, plotty = FALSE)

```
summary(fit_auto@objects[[1]])
```

```
##
## Call:
## fitfunc(formula = as.formula(x), data = data)
##
## 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
```

This result support the initial linear model.

Automated Modeling (interaction discovery)

With the selected predictor (wt),(am),(qsec), a complete interaction model was generated. Level=2 is the exhaustive method where all permutation are calcultated.

fit_auto_final <- glmulti(mpg ~ am + wt + qsec , fitfunction = lm, level = 2, method = "h", report = FALSE, plotty = FALSE)

```
summary(fit_auto_final@objects[[1]])
```

```
##
## Call:
## fitfunc(formula = as.formula(x), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5377 -1.4422 -0.5678  1.1146  4.1990
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.77584    4.68678   0.166 0.869753
## am          13.88153    3.37765   4.110 0.000331 ***
## qsec         1.52787    0.23781   6.425 6.96e-07 ***
## am:wt       -4.10098    1.17665  -3.485 0.001696 **
## qsec:wt     -0.16766    0.03661  -4.580 9.41e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.05 on 27 degrees of freedom
## Multiple R-squared:  0.8992, Adjusted R-squared:  0.8843
## F-statistic: 60.21 on 4 and 27 DF,  p-value: 4.634e-13
```

- fit_1 <- update(fit_0, mpg ~ wt, data = mtcars)
- fit_2 <- update(fit_1, mpg ~ wt + qsec , data = mtcars)
- fit_3 <- update(fit_2, mpg ~ wt + qsec + hp , data = mtcars)
- fit_4 <- update(fit_3, mpg ~ wt + qsec + hp + factor(am), data = mtcars)
- fit_5 <- update(fit_4, mpg ~ wt + qsec + hp + factor(am) + factor(cyl), data = mtcars)
- fit_7 <- lm(formula = mpg ~ wt + qsec + factor(am) + wt:factor(am) , data = mtcars)
- fit_8 <- lm(formula = mpg ~ wt + qsec + factor(am) + qsec:factor(am) , data = mtcars)
- fit_9 <- lm(formula = mpg ~ wt + qsec + factor(am) + wt:factor(am) + qsec:factor(am) , data = mtcars)