Executive Summary, Regression Models

ALo

Feb - 2015

Introduction

This document describe the analysis to complete the final project for the course "Regression Models". It consists of a model to answer the questions:

1 - Is an automatic or manual transmission better for MPG

2 - Quantify the MPG difference between automatic and manual transmissions

My analyis is broken down in 5 stages: *1 Data analysis* 2 Inference *3 Regression Models* 4 Best Model fit *5 Conclusions

Note: Some code is shown throuhgout the exec summary to assist understanding the info required to generate the results under each section. Plots/Full computaions are found in Appendix.

Data Analysis

Description: mtcars -> "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).

First, I analyse the date set 'mtcars' to explore the relationship between all the variables andmiles per gallon (MPG) using criteria: a) A Box plot to compare mpg vs transmission type

b) Determining the variables are close correlated by using calculations assisted by a "pairs" graph (refer to Appendix fig 2)

Observations: Data Analysis

a) In a first instance, Figure 1 'plot boxplot' shows that manual transmission yields higher values of MPG than the automatic transmission.

b) However, the calculations show that the are higher correlation with the variables wt, disp, cyl and hp.Figure 2 - pair shows such correaltion and supports argument 1. Further analysis to these variables must be then performed to answer the questions.

- Boxplot mpg vs am

```
p<-ggplot(data = mtcars, aes(x = interaction(am), y = mpg)) + geom_boxplot()
```

- Thi code determines close correlated variables.

```
mpgcor <- abs(data.frame(cor(mtcars))[1, ]);mpgcor1 <- order(mpgcor, decreasing = TRUE);
mpgcor2<- mpgcor1[1:5]; names(mtcars)[mpgcor2][1:5]
```

- Please Refer to 'pair plot' correlation of variables

```
dta <- mtcars; dta.r <- abs(cor(dta)), dta.col <- dmat.color(dta.r),
dta.o <- order.single(dta.r), cpairs(dta, dta.o, panel.colors=dta.col,
gap=.5, main="Variables Ordered by Correlation" )
```

Inference

As further investigation is required, I'm using a t-test to provide a better insight about mpg and the transmission relation. The result of this analysis is always given in terms of a null hypothesis, being H0 and H1 the hypothesis.Two outcomes can be obtained with this analysis "Reject H0 in favour of H1" or "Do not reject H0". If the result shows a "Do not reject H0", this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence against H0 in favor of H1. Rejecting the null hypothesis then, suggests that the alternative hypothesis may be true. Applying this concept to mpg and am

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

Observations: Inference

As shows in table above p-value = .00137 This establishes that the null hypothesis is rejected and an alternative hypothesis exists. This test assumes these variables are each normally distributed when testing the null hypothesis.By default, this performs a two-sided test and assumes unequal variances. The p-value of the test is very low therefore rejecting the null hypothes. suggesting that the alternative hypothesis exists. The automatic and manual transmissions are from different populations.

Regression Analysis

This analyis includes 4 models ass described below:

1) m_0 'mg vs am'

```
model_0 <- lm(mpg~am, data=mtcars)
```

My first guess is to test directly the linear model between the variables mpg and am and prove the need to explore the variables found in the previous steps.

m_0 yields a Residual standard error: 4.902 on 30 degrees of freedom with 174.147 mpg for manual transmission versus and increased 7.245 mpg for an automatic transmission. Although those are good figures I noticed that R-squared yield a value of 0.3598, which is rather low demonstrating the need to analyse more variables.

2) m_1 is full model,

```
model_1 <-  lm(mpg~., data=mtcars)
```

As opposed to model_0, this model tests mpg as an outcome and all the other variables as regressors. This model yields a Residual standard error of 2.65 on 21 degree degrees of freedom. And the Adjusted R- squared: 0.8066, which means that the model accuracy is 80% of the variance of the MPG variable. However, these coefficients are not significant at level of 0.05.

3) m_2 executes the Stepwise Model Selection. This process is to select significant predictors to determine the final, best model.

The step function performs such selection by calling lm repeatedly. It selects the best variables to use in predicting mpg with the Akaike information criterion, which implements both a forward and a backward elimination. This ensures that we have included useful variables while omitting ones that do not contribute significantly to predicting mpg.

```
model_2 <- step(model_1, direction = "both")
```

The outcome of m_2 reveals the following formula "mpg ~ wt + qsec + am". The Residual standard error as 2.459 on 28 degrees of freedom. And the Adjusted R-squared value is 0.8336, which means that the model can explain about 83% of the variance of the MPG variable. The coefficients are at 0.05 significant level.

In order to understand better the relation between the variables revealed in the previous step, we plot that model as follow:

```
ggplot(mtcars, aes(x=wt, y=mpg, group=am, color=am, height=3, width=3)) +
geom_point() + scale_colour_discrete(labels=c("Automatic", "Manual"))+xlab("weight")+
ggtitle("MPG vs Wt-Transmission")
```

This graph reveals that the interaction between "wt" and "am" is quite significant. It can be assumed that automatic cars may be heavier than manual cars. If we adjust model to include such interaction we create m_3.

4) m_3 factoring mg and am.

```
model_3 <-lm(mpg ~ wt + qsec + am + wt:am, data=mtcars)
```

This model yields a Residual standard error as 2.084 on 27 degrees of freedom. The Adjusted R-squared =0.8804. This is 88% accuracy. The coefficients are significant at 0.05. This model has yield the best results up to this point

The Analysis of variance (anova) results very to cpmpare models, as follows.

```
anova(model_0, model_2, model_3,model_1 ),confint(model_3)
```

The investigation yield model_2 and model_3 as best candidates. Although model_2 has better P value, model_3 has the highest adjusted R-square qualifying as my preferred choice.

Conclusions

The results yield a result such that: 1 - Question 1 -"When wt and qsec remain constant, cars in the dataset mtcars with manual transmission have an increased MPG (miles per gallon) on average than cars with automatic transmission". 2 - Question2 - Being 14.079 + (-4.141)*wt A manual car wighted at 1000 lbs has 9.938 more MPG than an automatic car with the same weight.

Other observations - When the weight increases, there is a decrease of the mpg of (around) f Thus, starting, with cars more than 3400 lbs, the automatic cars should be the choice of preference. - From the Residuals vs Fitted plot we see that the residuals are randomly scattered and thus verify the independece condition. Any pattern would indicate underfitting. - The Normal Q-Q plot shows points that fall on or close to the line, indicating the residuals are approximately normally distributed. - The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed - The Residuals vs. Leverage reveals ther are no outliers are present, since all values are within the 0.5 range.

```
summary(model_3)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am + wt: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 ***
## am            14.079      3.435   4.099 0.000341 ***
## wt:am         -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
```

Appendix

```
## Loading required package: cluster
```

Diagrams and Plots

Figure 1 mpg vs am

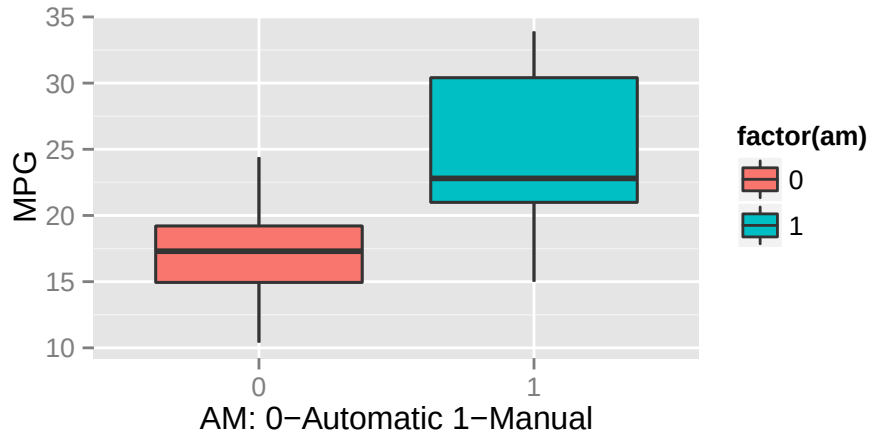


Figure 2 Pairs graph

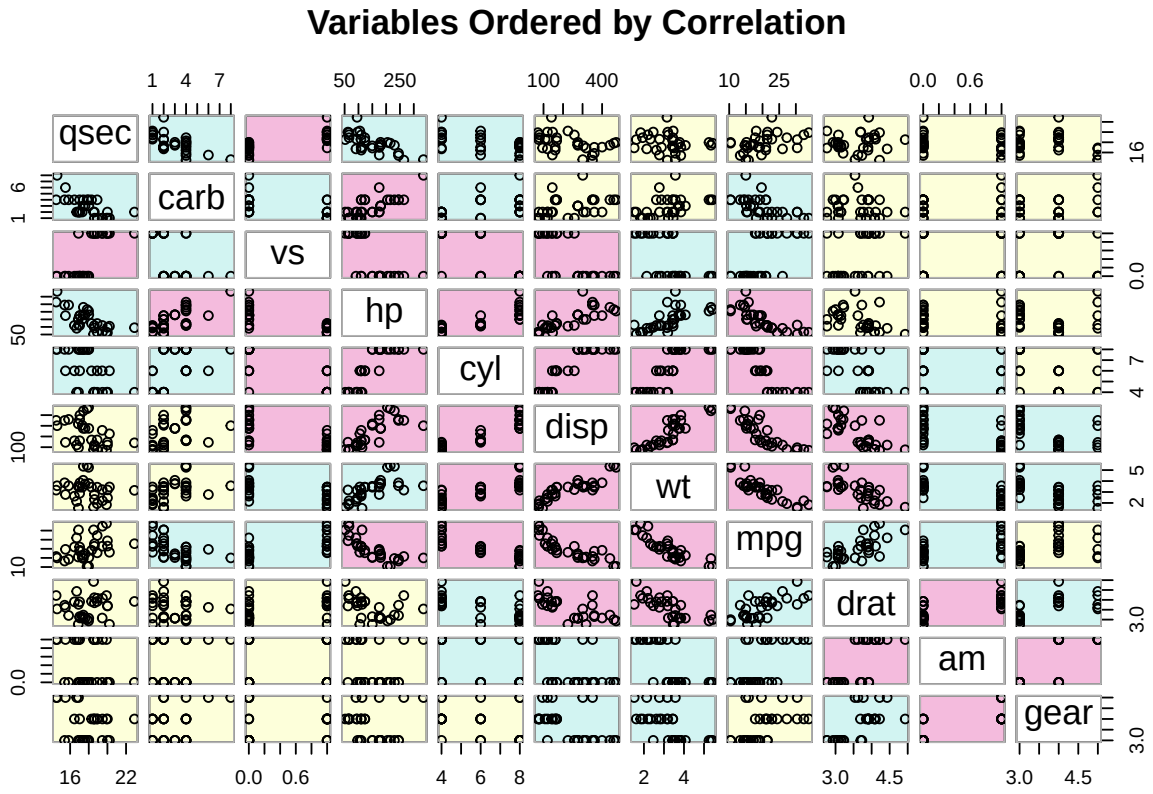


Figure 3 scatter plot mpg vs am

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

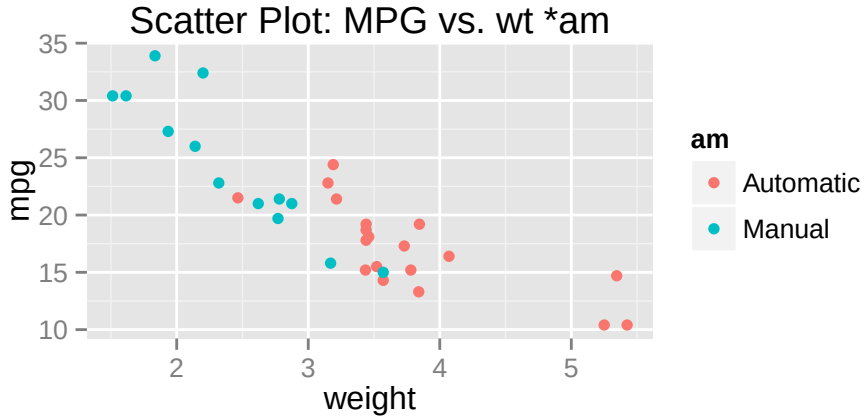
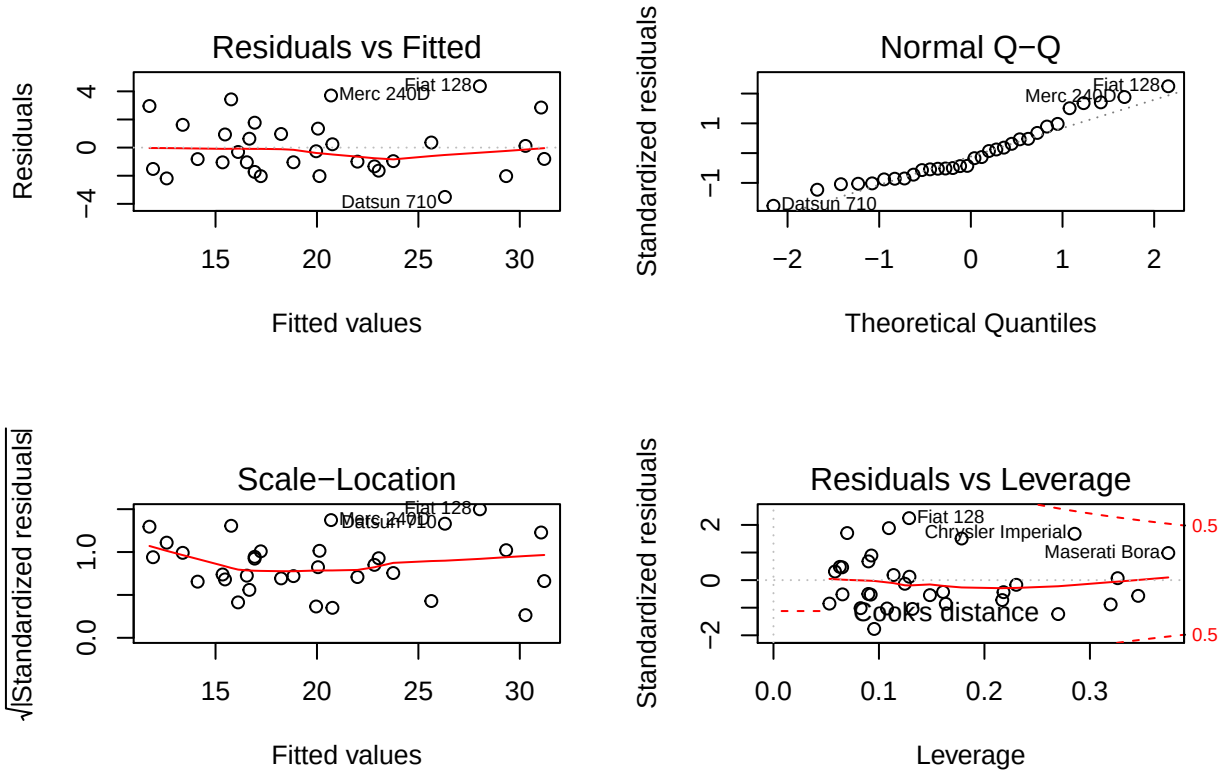


Figure 4 - Residuals



Additional Section showing computations

This section is additonal. Athough it's not required in the project I decided to inlcude it as additional appendix.

This code calculates and extract the relevant variables

```
data(mtcars)
mpgcor <- abs(data.frame(cor(mtcars))[1, ]);mpgcor1 <- order(mpgcor, decreasing = TRUE)
mpgcor2<- mpgcor1[1:5]; names(mtcars)[mpgcor2][1:5]
```

```
## [1] "mpg"  "wt"   "cyl"  "disp" "hp"
```

The following code computes the 'models: m_0, m_1, m_2, m_3'

1) m_0 tests 'mg vs am'

```
model_0 <- lm(mpg~am, data=mtcars)
summary(model_0)
```

```
##
```

```
## 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 ***
## am             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
```

2) m_1 is full model, which tests mpg as an outcome and other the variables as regressors

```
model_1 <-  lm(mpg~., data=mtcars)
summary(model_1)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

3) m_2 executes the Stepwise Model.

```r
model_2 <- step(model_1, direction = "both")
```

```r
summary(model_2)
```

```
##
## 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
```

4) m_3 factoring mg and am.

```r
model_3 <-lm(mpg ~ wt + qsec + am + wt:am, data=mtcars)
summary(model_3)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am + wt: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 ***
## am            14.079      3.435   4.099 0.000341 ***
## wt:am         -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
```