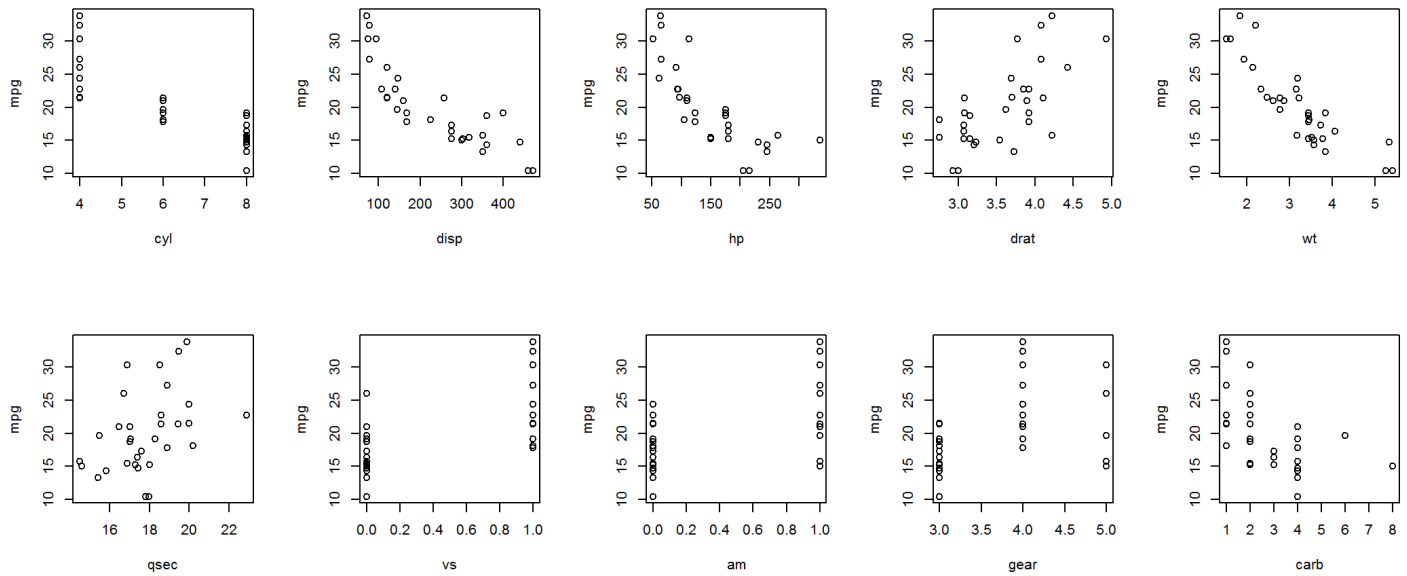
Coursera Regression Models Course Project

Note : The text is a bit less than 2 pages long. The tables and figures make it much longer but I have chosen to put them into the body of the text anyway. Otherwise it makes a 2 pages text really unpleasant to read.

Executive summary : In this report, I examine the effect of the type of transmission of a car on its fuel consumption efficiency measured in miles per gallon using linear regression modeling. I find that other features are more appropriate to explain fuel efficiency especially the weight and the 1/4 mile time. Adding the type of transmission to a linear model including those variables does not add significant information.

In this project I study the database “mtcars” provided in R base.This database contains observations of 11 features of 32 different cars. It does not contain any NULL value. My goal is to answer the following questions : is an automatic or manual transmission (variable “am”) better for Miles Per Gallon (variable “mpg”) ? The second part of the project is to quantify how different is the mpg between automatic and manual transmission.
As this project is done for a course about linear regression models, I will use this technique to answer the questions. I will proceed as follow : first I will do some explanatory analysis, then I will select the variables that explain the variance in mpg in order to control for their effect in my final model, then I will assess the effect of transmission on miles per gallon and answer the questions.

To apply linear modelling to mtcars dataset, I first check that the variables are linearly related to the outcome. For this purpose let’s first plot the scatterplots of each of the variable and the outcome.



I observe from the plot that variables cyl, vs, am, gear and carb are categorical while the others are numeric. Variables disp, hp and wt need reshaping as they are not scattered arount a straight line. Drat has a linear relationship with mpg even if it has a high variance, the relationship between qsec and mpg is linear. I can not really see any outlier.

I try two transformations of disp, hp and wt: sqrt() and log(). To check which is the best one, I look at their correlation coefficient with mpg.

```
##      Correlation coef Correlation with sqrt() Correlation with log()
## disp          -0.8476                 -0.8807                -0.9071
## hp            -0.7762                 -0.8194                -0.8488
## wt            -0.8677                 -0.8890                -0.9001
```

In term of the creation of a linear relationship, the sqrt() transformation dominates the no-transformation case and log() dominates sqrt() . I am to keep the log() of those variables.

Now that the variables are prepared for linear modeling, let's have a look at the correlation coefficient of the dependant variables with mpg.

```
##  disp    wt   cyl    hp  carb  qsec  gear    am    vs  drat   mpg
## -0.91 -0.90 -0.85 -0.85 -0.55  0.42  0.48  0.60  0.66  0.68  1.00
```

With respectively -0.91, -0.90, -0.85 and -0.85, variables disp, wt, cyl and hp are good candidates. Note that each of these variables are highly correlated with each other and have a negative correlation coefficient with mpg (see appendix, Table 1). I have reasons to believe that they explain the same thing. I am now looking for variables uncorrelated to wt,cyl,disp and hp that can explain the residual variance. Let's have a look at the correlation matrix between those 4 variables and the others.

```
##        cyl  disp    hp    wt
## mpg  -0.85 -0.91 -0.85 -0.90
## cyl   1.00  0.93  0.88  0.80
## disp  0.93  1.00  0.86  0.90
## hp    0.88  0.86  1.00  0.73
## drat -0.70 -0.76 -0.56 -0.73
## wt    0.80  0.90  0.73  1.00
## qsec -0.59 -0.45 -0.68 -0.18
## vs   -0.81 -0.73 -0.76 -0.56
## am   -0.52 -0.64 -0.35 -0.72
## gear -0.49 -0.55 -0.22 -0.58
## carb  0.53  0.44  0.70  0.44
```

I can see 3 interestings pairs of variables : wt and qsec, hp and gear, hp and am. Cyl and disp do not really have features uncorrelated with them. I decide not to include them in my final model.

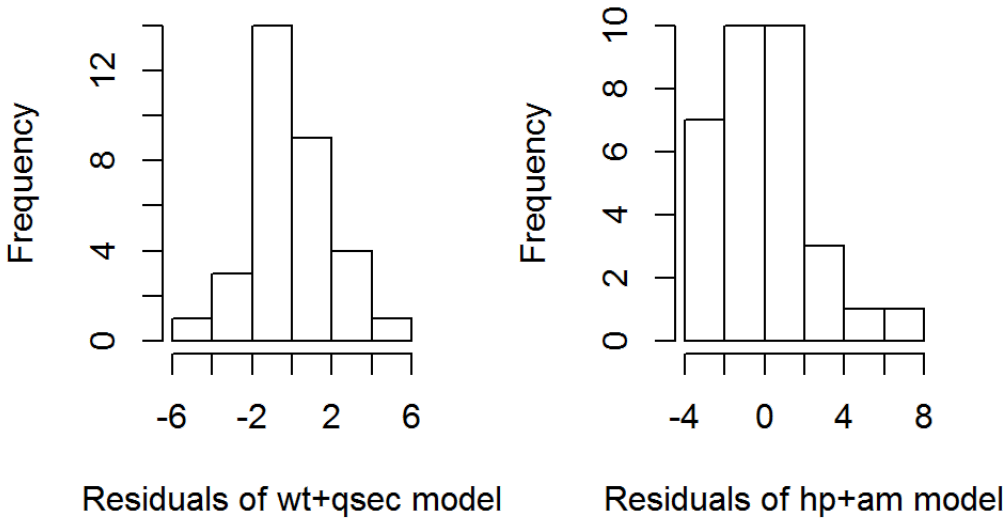
Since my final model includes wt and qsec I will explain my work with this pair of variables. I will then explain why I do not keep the hp+am model even if it includes am. The case hp+gear will not be discussed explicitly as the reasons I do not keep it are the same that for the hp+am case.

A regression model with only the wt variable have a R^2 of 0.81. This variable alone explains 81% of the variance with a high statistical significance. The p-value of the coefficient is 2.391e-12. In other words, there is only a 2.391e-12 probability that we can have picked such sample values given that the hypothesis that the true coefficient of wt is 0 is true. The R^2 with only qsec in the model is low and the qsec coef is significant (R^2 = 0.15 and p value = 0.0171). However, when qsec is added to wt, it explains more than one third of the residual variance once accounted for wt and it has a really low pvalue (0.000384). 87% of the variance is explained with this 2 variables model. That is efficient in term of model complexity (for details see appendix, LM1).

A linear model with explanatory variables hp and am has a lower R^2. In addition, as shown on the following plot, the residuals are skewed. Their correlation coefficient with mpg (0.42) is higher than for the wt+qsec model (0.35) . There is definitely something not good with this model and a further investigation

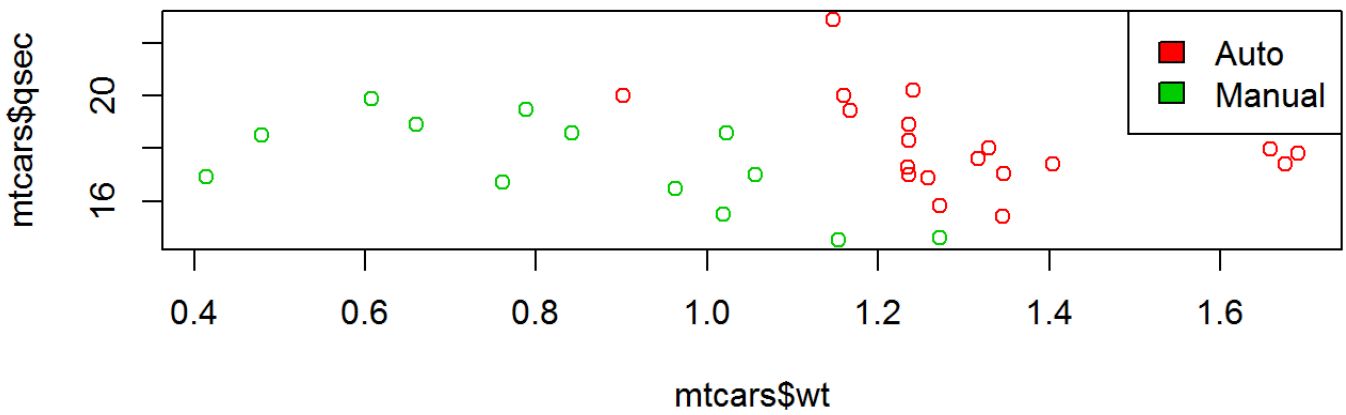
should be done to really discredit it.

f resid(lm(mpg ~ wt + qsecƆf resid(lm(mpg ~ hp + am,



What about variable am then ? Let's recap my strategy : I have found the variables that best explain the variance of mpg. The variable am is not among them. It is time to include it in the regression to see if it can explain some variance. It should not be surprising that the effect of am is small. Still, it can be statistically significant.

Including am is disapointing. It does not really add explanatory power to our model. Including it increases the R^2 by 1 percentage point (from 0.87 to 0.88) and its coefficient is not statistically significant. The following graph illustrates why adding am does not add much explanatory power.



You can almost separate cars with different transmission by a straight line when they are in the wt-qsec space. In other words, once we know the wt and qsec of a car, we can very accurately predict its transmission type. The converse is not true since am is one-dimensional and binary.

The coefficient of am in the final model must be interpreted this way : once taken into account the weight and qsec, a manual transmission on a car increases the number of miles travelled per gallon by 1.6 on average. The probability that we get this sample value for the coefficient while its true value is 0 is 0.24. I am not confident enough that this value of the parameter is not due to randomness to reject the hypothesis that am has no impact on mpg. Interpreting the intercept does not make sense in this model since it would involve the concept of a car with a weight equal to 0. The sign of qsec means that once

taken into account the weight and transmission type, adding one qsec unit increases mpg by 1.06 on average. The sign of the wt coefficient is less intuitive. Since it is a log transformation of the original variable, it means that a X% increase in weight leads to a X*(-14.20) change in mpg. A higher weight reduces the number of miles traveled per galon ; that makes sense.

Appendix

Table 1

```
round(cor(mtcars[,c(2,3,4,6)]),2)
```

```
##       cyl disp   hp   wt
## cyl  1.00 0.93 0.88 0.80
## disp 0.93 1.00 0.86 0.90
## hp   0.88 0.86 1.00 0.73
## wt   0.80 0.90 0.73 1.00
```

LM1

```
##
## Call:
## lm(formula = mpg ~ wt + qsec, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.073 -1.388 -0.437  0.749  5.469
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   22.297      4.460    5.00  2.5e-05 ***
## wt           -16.178      1.252  -12.92  1.5e-13 ***
## qsec           0.893      0.222    4.02  0.00038 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.18 on 29 degrees of freedom
## Multiple R-squared:  0.878,  Adjusted R-squared:  0.87
## F-statistic:  104 on 2 and 29 DF,  p-value: 5.66e-14
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.348 -1.417 -0.464  0.959  4.912
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    16.46       6.55    2.51  0.01804 *
## wt            -14.20       2.06   -6.90  1.7e-07 ***
## qsec            1.06       0.26    4.07  0.00035 ***
## am              1.60       1.32    1.21  0.23749
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.16 on 28 degrees of freedom
## Multiple R-squared:  0.884,  Adjusted R-squared:  0.872
## F-statistic: 71.1 on 3 and 28 DF,  p-value: 3.26e-13
```