Linear Models - Assignment

This report provides the reasoning, methods and results of the analysis between the variables miles per gallon and type of transmission of the R databank "mtcars". To reach the results, it was performed a linear regression between the variables of the bank, specially the one that interest us the most - type of transmission - and the response variable - miles per gallon. Codes, images can be found in the later appendices. It was found that there is a significant difference between the two types of transmissions when it comes to the ratio miles per gallon.

Miles per gallon is a quantitative variable, that has a domain between zero and plus infinity. Theoretically, the linear model premisse, that the errors follow a normal distribution, implies that the response variable should follow a normal distribution, but this premisse can be relaxed depending on the variable distribution. It can be seen that the MPG has the following measures of location and distribution:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.42   19.20   20.09   22.80   33.90

## Variance: 36.3241

## Standard Deviation: 6.026948
```

For the model, initially, it was performed a regression with all the variables contained in the bank. The reasoning behind this approach comes from the fact that other variables can mislead the conclusions. For example, suppose that the most powerfull cars consume more gallons of gas per mile. If all of these cars have an automatic transmission, than, probably, the procedure 'lm' will result that automatic cars consume more, but that doesn't mean that it is due to the transmission. Therefore, it was performed a regression with all the variables to observe the variation in the response when all the other variables are held constant. Only with that, we will be able to state whicht type of transmission produces a highter rate of miles per gallon. Performing this regression, we can observe that the coeficients are not statistically significant:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.097055  18.666824   0.594   0.5583
## cyl         -0.301014   1.026604  -0.293   0.7721
## disp         0.004086   0.015191   0.269   0.7905
## drat         0.927430   1.628212   0.570   0.5747
## wt          -3.265846   1.837757  -1.777   0.0894 .
## qsec         0.899225   0.726106   1.238   0.2286
## vs          -0.247732   2.023787  -0.122   0.9037
## am           2.423864   2.053110   1.181   0.2504
## gear         0.523594   1.486390   0.352   0.7280
## carb        -0.630106   0.704109  -0.895   0.3805
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.649 on 22 degrees of freedom
```

```
## Multiple R-squared:  0.8629, Adjusted R-squared:  0.8069
## F-statistic: 15.39 on 9 and 22 DF,  p-value: 1.446e-07
```

Despite the Results from the 'lm' procedure for the individual coeficients, we have statistical significance to say that at least one of the predictors has some linear association with the variable mpg (F test p-value = 1.446e-07). With that, we can procede the analisys to find the best model.
To determine the best model, it was used a stepwise backward procedure that takes the full model, with all the variables, and delete the ones that do not present any statistical significance to the response. Using these procedure, the final model is:

mpg = 9.617781 - 3.916504 * Weigth + 1.225886 * QuarterMileTime + 2.935837 Type of Transmission

```
## 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
```

The final model contains three variables: Car Weight, Type of transmission and quarter mile time. Two of those are quantitative, car weight and quarter mile time, and one is a dumy variable, type of transmission. We are particularly interested in the effect of the type of the transmission on the ratio miles per gallon.

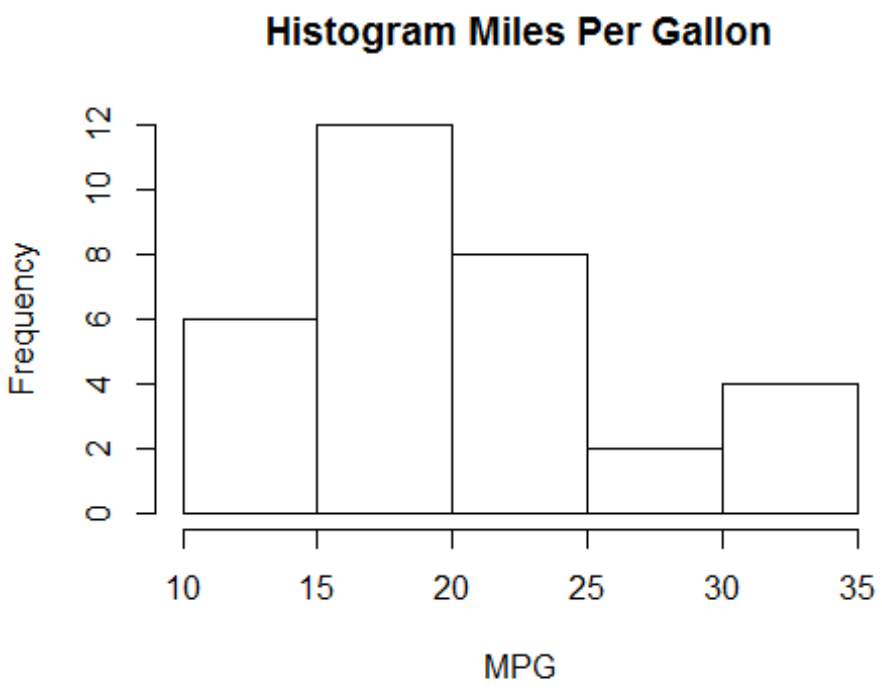
We can observe that the type of transmission has a significant effect on the miles per gallon ratio. If the car has an manual transmission (am =1), the expected miles per ration increases in 2.9358 when all the other variables in the model are held constant.In other words, cars with manual transmission are more economic. Other logical relation that can be drawn from the model is the fact that as the car gets heavier, smaller the variable miles per gallon will be. An analisys of the quality of the fitted model include checking all the premisses and the percent of the variance that's explained by it. The residuals of the fitted model attend the assumptions of the linear regression model that are homoscedasticity and normality. Homoscedasticity implies that the variance of the residuals,a random variable, is constant. It can be checked using the residual plots of the 'lm' procedure. QQ-Plots and Normality test can give us clues about the behaviour of the residuals. In the fitted model, the Shapiro-Wilks test (considering alfa = 0.05) didn't reject the hipothesis of normality. Remember that, an important assumption of the linear regression models is the normality of the error. Because the residuals can be though as the errors, the residuals must follow a normal distribution.
Observing the R-Squared of the model, it can be seen that the three variables included in the model explain a total of 0.8497 (R-Squared), but in a multiple regression the statustic that counts is the Adjusted R-Squared. In the model, the Adjusted R-Quared resulted in 0.8336 of the variance explained by the model.

Through the results of the analisys it was shown that the weight of the car, quarter mile time and type of transmission are the best variables to explain and predict the mile per gallon ration. These three variables produce a model that explain aproximadetly 83.36% of the variance of the response variable. All the assuption were met - Normality of the residuals and constant variance - and they were confirmed using normality test and Graphics. Manual transmission cars produced a better ratio for the response variable that cars with automatic transmission. In summary, cars with manual transmission are expected to have a 2.9358 greater miles per gallon ration than automatic cars.

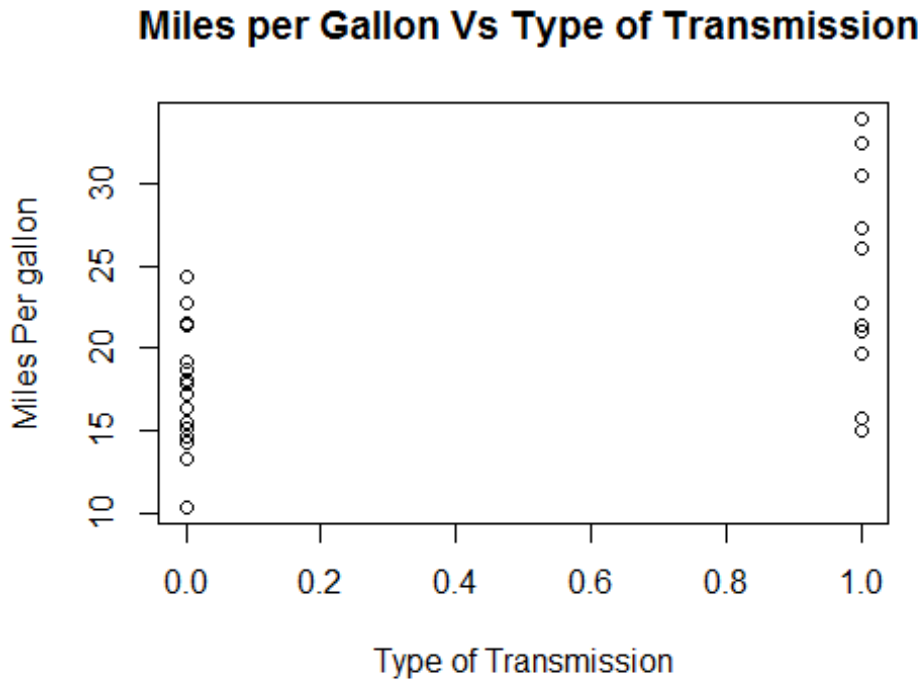
Final note:

Sorry for any mistake in the language. I'm not a native English speaker. Thanks for your time and good luck.

Appendix



MPG VS Type of transmission



Residuals for the Final Model and Backward Procedure