On the Impact of Transmission and Various Measurements on MPG

Overview

In this report, we perform simple analyses of the mtcars data in the R datasets package. We are interested in exploring the relationship between a set of variables and miles per gallon (MPG). In particular, we analyse the impact of transmission on fuel consumption and quantify the difference of the two transmission modes (automatic/manual) on MPG.

Exploring the Data

First, the dataset is made available for processing, transforming transmission and engine type into factors, and basic information is obtained:

```
library(datasets); data("mtcars"); # str(mtcars)
# trasmission and engine type are more natually encoded as factor variables
mtcars$am <- factor(mtcars$am)
mtcars$vs <- factor(mtcars$vs)
```

Let's first qualitatively observe the difference between transmission modes in terms of MPG. Here are common distribution values for the sample of MPG in the two groups:

```
summary(subset(mtcars, am == "0")$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.95   17.30   17.15   19.20   24.40
```

```
summary(subset(mtcars, am == "1")$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   21.00   22.80   24.39   30.40   33.90
```

with cars with automatic transmission being a lot less fuel effective than the manual transmission ones. This is confirmed by the plot in Fig.1 in the appendix.

We then look at the scatterplots between all possible pairs of variables, visually searching for correlations and other relationships among them. The scatterplots are reported in Fig.2 in the appendix, colored according to the trasmission type. Overall, all variables seem to have an influence on fuel consumption, but are also highly correlated between each other. Not surprisingly, MPG clearly decreases with the number of cylinders, displacement, horse power and weight, and slightly less with the number of carburetors. There seem to be a positive influence of rear axle ratio (drat) and straight engines compared to v engines (vs).

Data Analyses

Comparison of Automatic/Manual Transmission Cars

Even if there's overlap, the above analysis pointed to a difference in the distribution of MPG values of the two groups (automatic/manual). Here we verify normality of the mpg samples and verify that the average MPG value of automatic transmission cars is significantly smaller than the one of manual transmission cars:

```
sapply(list(subset(mtcars,am==0)$mpg,subset(mtcars,am==1)$mpg),
       function(elt) shapiro.test(elt)$p.value)
```

```
## [1] 0.8987358 0.5362729
```

```
test <- t.test(mpg ~ am, paired = FALSE, var.equal = FALSE,
               data = mtcars, alternative = "less")
c(test$p.value, test$conf.int[2])
```

```
## [1]  0.0006868192 -3.9132558169
```

The Shapiro Test fails to reject hypothesis of normality, and the t.test confirms the mean MPG of automatic cars is significantly smaller than the one of manual cars. With 95% probability, manual cars run on average at least around 4 miles longer than automatic cars. Note that the same results could have obtained by using a linear model of MPG vs transmission.

Linear Modeling

Besides trasmission, MPG of an engine depends on several interrelated factors which are measured by other variables in the dataset under consideration. In order to better understand the relationship between MPG and transmission, we try to model MPG as the outcome of a linear model using transmission and other variables as possible regressors.

Model Selection

Model selection occurs with a customised forward selection iterative strategy using likelihood ratio tests comparing nested models. Since the focus is on the effect of trasmission on MPG, the first model uses only the transmission binary variable *am* as regressor and the latter as outcome:

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

As reported in appendix A.2, coefficient estimates are deemed to be significant, with the intercept 17.15 representing the mean MPG for the automatic cars; the other coefficient, 7.24, is the estimated increase in the mean MPG we observe when looking at manual cars instead. This confirms the finding of the previous section. Nevertheless, the adjusted R-squared indicates a significant fraction of the outcome is not explained by trasmission alone.

When looking at adding other regressors, we must judiciously consider which variables to add based on knowledge of the system under observation, i.e. some variables are alternative representations of other variables, and there exist a strong correlation between subset of the observed variables. For instance, one must choose whether to consider either the number of cylinders or displacement, the latter being clearly a continuous representation of the former. The forward selection strategy we apply is to consider, at each step and for each of the remaining variables, its simultaneous effect on the variance inflation factor (VIF) of the other regressors and the increase in variance explained by the model. The algorithm selects at each step the variable which brings a combined minimum VIF and largest increase in adjusted R2:

```
# create a list of regressors to evaluate, do not consider qsec
regressors <- names(mtcars)
regressors <- regressors[regressors!="mpg" & regressors!="am" & regressors!="qsec"]
models <- list(am = mpg_am_fit)
formula <- "mpg ~ am"
while(TRUE) {
  min = 10000; bestr <- ''; bestm <- ''; bestf <- ''
  for (regr in regressors) { # loop over remaining regressors
    f <- paste(formula, "+", regr)
```

```
    fit <- do.call("lm", list(as.formula(f), data=as.name("mtcars")))
    s <- summary(fit)
    vifs <- vif(fit)[-length(vif(fit))]; # remove VIF on current regressor
    v <- c(vifs, 1-s$adj.r.squared) # to simulaneously and equally evaluate VIFs and R2
    if(mean(v) < min) { min <- mean(v); bestr <- regr; bestm <- fit; bestf <- f }
  }
  formula <- bestf
  models[[bestr]] <- bestm
  regressors <- regressors[regressors!=bestr] # remove chosen regressor
  if(!length(regressors)) { break }
}
```

We then consider nested likelihood ratio tests and choosing the model where the terms of its successive model are deemed to be not significant ($\alpha = 0.05$). Results are shown in Appendix A.2. The above strategy selects a very parsimonious model using only transmission and horse power as regressors. In order to obtain a more meaningful interpretation of the intercept, we recompute the selected model by shifting the hp values by their average. The confidence intervals of the estimates are again reported in Appendix A.3:

```
rescaled_best <- lm(mpg ~ am + I(hp-mean(hp)), data = mtcars)
coefs <- summary(rescaled_best)$coefficients
```

The advantage of having such a simple parsimonious model with one categorical regressor is that we obtain a useful and clear interpretation of the intercept and other model tersm:: the intercept can be interpreted as the expected MPG of the average HP car with automatic transmission, while the sum of the intercept and the transmission coefficient is the expected MPG of the average HP car with manual transmission.

Some diagnostic plots are shown in Appendix A.3. The Residuals vs Fitted and Normal QQ plots confirm the residuals are approximately normally distributed with zero mean and no clear correlation with the fitted values. The Residuals vs Leverage plot shows there might be some outliers with a certain influence which probably be taken into account in order to obtain a better fit.

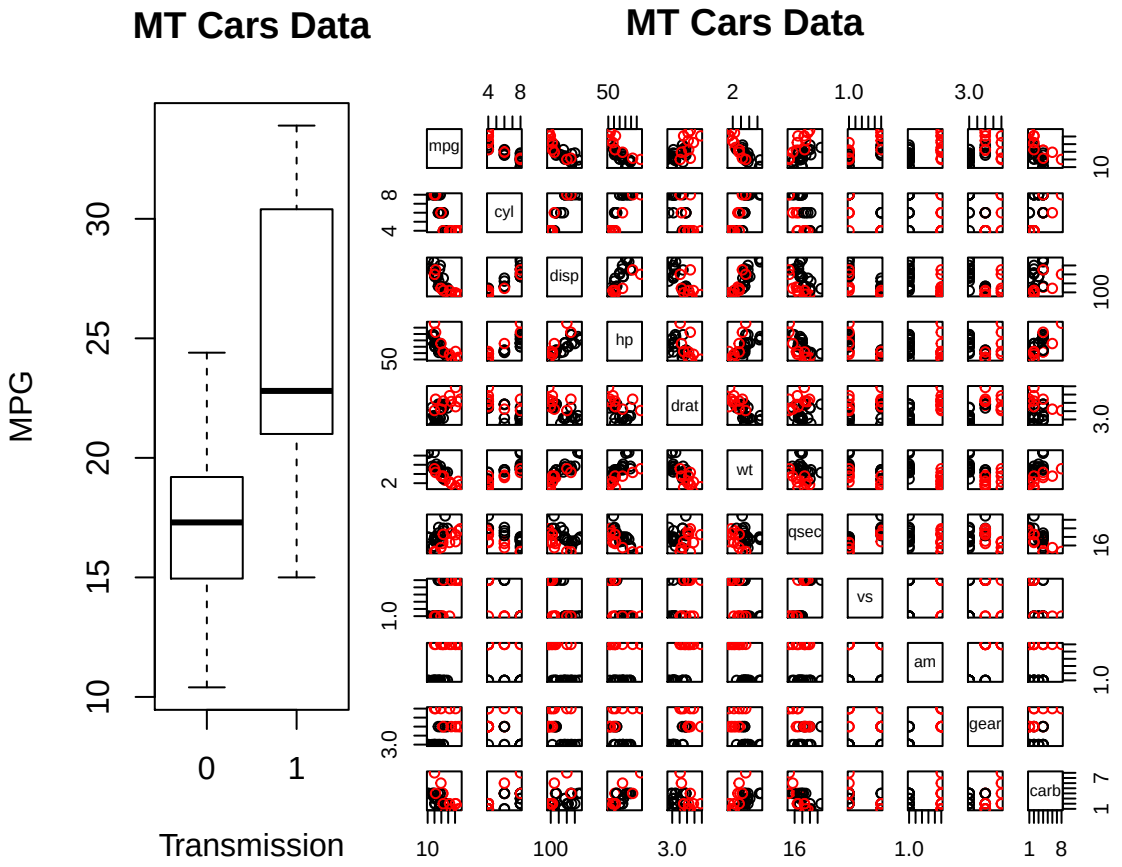
Summary

We've found manual cars to be on average significantly more fuel effective than automatic cars. A parsimonious model with transmission and hp has been found to be a good and significant fit of the available observations. Based on this model, with 95% confidence we quantify the expected MPG of the average HP car with automatic transmission to be around 17.95, the expected MPG of the average HP car with manual transmission to be close to 23.22, and the expected decrease in MPG per 1hp increase with transmission remaining constant to be around 0.06 MPGs.

Appendix

A.1 Exploratory Plots

```
par(mfrow=c(1,2))
with(mtcars, boxplot(mpg ~ am, main="MT Cars Data", xlab="Transmission", ylab="MPG"))
pairs(mtcars, main = "MT Cars Data", col = mtcars$am)
```



A.2 Evaluation of Nested Models for MPG

The following summarises the model of MPG with transmission type:

```
summary(mpg_am_fit)
```

```
##
## 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 ***
## am1            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
```

This is a summary of the likelihood ratio tests of various incremental models developed over the course of the

above analyses:

```
anova(models[[1]],models[[2]],models[[3]],models[[4]],
      models[[5]],models[[6]],models[[7]],models[[8]],models[[9]])
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + hp
## Model 3: mpg ~ am + hp + vs
## Model 4: mpg ~ am + hp + vs + drat
## Model 5: mpg ~ am + hp + vs + drat + wt
## Model 6: mpg ~ am + hp + vs + drat + wt + gear
## Model 7: mpg ~ am + hp + vs + drat + wt + gear + carb
## Model 8: mpg ~ am + hp + vs + drat + wt + gear + carb + cyl
## Model 9: mpg ~ am + hp + vs + drat + wt + gear + carb + cyl + disp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 245.44  1    475.46 66.8979 4.079e-08 ***
## 3     28 218.88  1     26.56  3.7370   0.06619 .
## 4     27 213.58  1      5.30  0.7456   0.39720
## 5     26 167.81  1     45.77  6.4401   0.01875 *
## 6     25 167.76  1      0.05  0.0075   0.93170
## 7     24 158.32  1      9.43  1.3274   0.26164
## 8     23 157.75  1      0.58  0.0810   0.77863
## 9     22 156.36  1      1.39  0.1953   0.66286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A.3 Summary of the Best Selected Model

The following reports a summary of the best selected model using the forward selection strategy outlined above:

```
summary(models[[2]])
```

```
##
## Call:
## lm(formula = mpg ~ am + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3843 -2.2642  0.1366  1.6968  5.8657
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.584914   1.425094  18.655  < 2e-16 ***
## am1          5.277085   1.079541   4.888 3.46e-05 ***
## hp          -0.058888   0.007857  -7.495 2.92e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.909 on 29 degrees of freedom
## Multiple R-squared:  0.782,  Adjusted R-squared:  0.767
## F-statistic: 52.02 on 2 and 29 DF,  p-value: 2.55e-10
```

These are the 95% confidence intervals of this model using transmission and the mean centered hp values:

```
coefs[1,1] + c(-1,1)*qt(.975, df = rescaled_best$df)*coefs[1,2]
```

```
## [1] 16.56447 19.32915
```

```
coefs[2,1] + c(-1,1)*qt(.975, df = rescaled_best$df)*coefs[2,2]
```

```
## [1] 3.069177 7.484994
```

```
coefs[3,1] + c(-1,1)*qt(.975, df = rescaled_best$df)*coefs[3,2]
```

```
## [1] -0.07495665 -0.04281896
```

And here are some diagnostic plots:

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