Investigating the relationship between transmission type and MPG

Roman Preobrazhensky

Preliminary remark: *I am not a native English speaker and I apologize for possible grammar mistakes in the following text. I hope you will be able to fully understand it.*

Also in my understanding it's rather impossible to perform decently thorough analysis with all the code chunks in just 2 pages, so I think what has been mentioned in the assignment is 2 pages of ***text only****, disregarding necessary code chunks and their results. I was trying to be as concise as I possibly could though.*

Summary

In this document I'm trying to investigate the MPG (miles per gallon) difference between cars with automatic and manual transmissions and answer these questions: "Is an automatic or manual transmission better for MPG" and to "Quantify the MPG difference between automatic and manual transmissions". First I'm going to look at the `mtcars` dataset and perform some exploratory analisys.Then I'll try to fit linear regression models with different sets of varibales. After picking the best model, there will be statistical exploration of that model including quantifying the uncertainty in conclusions, residuals and diagnostics.

Exploratory analysis

We're working with `mtcars` dataset which containes 32 observations of cars with 11 variables each. Some of these variables take discrete values and can be treated as factors, so I've converted them into factors. The code for this is in the appendix by the name "Code Chunk 1". Notice that I've renamed the dataset to `mtc`.

The next thing I did was create a pairsplot which contains regression lines in its upper half and in the lower half it has scatterplot cells colored according to the covariance of corresponding variables. You can see this plot in the appendix as "Figure 1". Amongst colored cells those cells with higher covariance have a color closer to red. We can see already that `mpg` and `wt` have highest covariance, so we will probably use 'wt' to predict `mpg`.

Regression models

We're forced by the questions to use the trasmission type varible, the `mpg`, as a regressor. It would be, of course, kind of silly to think that MPG is affected by the transmission only. And one can check this by comparing a model with only `am` as a regressor and a model with all of the variables, which can be done with ANOVA as shown in the appendix as "Table 1".

We see that there is statistically significant difference between these two models, which implies that at least one more variable can count as a predictor. Here I must tell why I'm not going to use all of them. First of all, simple experiment can show that model fitted with all of the variables will not produce one statistically significant coefficient. That is because we have a very small dataset that has only three times more observations than it has variables. So if we take one variable and fix all the remaining ones at some particular values, we will often obtain one or two observations (there are not many cars with most of characteristics matching). It is also sort of required, as well as I know, in statistical science to have at least 10 times more observations than considered variables. So, I will use 3 variables in my final linear model.

Now let's start adding other variables in our model one at a time (so that the overall number of regressors will be at the value of 2) and see, if it's making a significant decrease in RSS. We will do it with the `anova` function and grab only the p-values.

```
for(x in colnames(mtc[, -match(c("am", "mpg"), names(mtc))])){
        newfit <- lm(as.formula(paste("mpg ~ am + ", x)), mtc);
        print(paste0(x, ': ',anova(lm(mpg ~ am, mtc), newfit)[2,6]))}
```

```
## [1] "cyl: 8.01010927659694e-07"
## [1] "disp: 5.7475278734969e-07"
## [1] "hp: 2.92037483129984e-08"
## [1] "drat: 0.0106954829638931"
## [1] "wt: 1.86741503808364e-07"
## [1] "qsec: 6.27075924788626e-06"
## [1] "vs: 6.50096242484413e-06"
## [1] "gear: 0.0373294158717765"
## [1] "carb: 0.000526999272735578"
```

From this output we see that variables `wt` (weight) and `hp` (horsepower) are the most significant two. Let's compare their affect on RSS and consider an interaction term as well:

```
print(anova(lm(mpg ~ am + hp, mtc), lm(mpg ~ am*hp, mtc))$RSS)
```

```
## [1] 245.4393 245.4340
```

```
print(anova(lm(mpg ~ am + wt, mtc), lm(mpg ~ am*wt, mtc))$RSS)
```

```
## [1] 278.3197 188.0077
```

This tells us that we should definitely use `am*wt` for our predictions as it has the most significant effect on RSS. Now we must choose another variable, so let's add and compare a bunch of them:

```
for(x in colnames(mtc[, -match(c("am", "mpg", "wt"), names(mtc))])) {
        newfit <- lm(as.formula(paste("mpg ~ am*wt + ", x)), mtc);
        print(paste0(x, ': ',anova(lm(mpg ~ am, mtc), newfit)[2,6]))}
```

```
## [1] "cyl: 5.33182913468136e-09"
## [1] "disp: 4.42326209404355e-09"
## [1] "hp: 1.79900830186733e-09"
## [1] "drat: 3.06890704472072e-08"
## [1] "qsec: 8.8465759731758e-11"
## [1] "vs: 2.10526895179722e-09"
## [1] "gear: 2.02229249531001e-08"
## [1] "carb: 3.33740726897652e-06"
```

And the winner is `qsec` (which is basically a speed - the time of passing 1/4 of a mile in seconds) with the most significant effect. Let's take a look at out final model:

```
fit <- lm(mpg ~ am*wt + qsec, mtc)
summary(fit)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.723053  5.8990407  1.648243 0.1108925394
## am1         14.079428  3.4352512  4.098515 0.0003408693
## wt          -2.936531  0.6660253 -4.409038 0.0001488947
## qsec         1.016974  0.2520152  4.035366 0.0004030165
## am1:wt      -4.141376  1.1968119 -3.460340 0.0018085763
```

```r
summary(fit)$adj.r.squared
```

```
## [1] 0.8804219
```

We see that over 88% of variance is explained by the proposed model, and every coefficient is statistically significant (except for the intersept, but its significance are generally disregarded). Also from this model we can tell that there is an average 14.079 MPG increase while switching from `am = 0` (which is an automatic transmission) to `am = 1` (which is a manual transmission), so that manual seems to be better for MPG than automatic. To quantify how certain this is we can show the 95%-confidence intervals:

```r
confint.lm(fit, parm = 'am1')
```

```
##        2.5 %   97.5 %
## am1 7.030875 21.12798
```

So, we're 95% confident that cars with manual transmission have MPG larger by some value from 7.031 to 21.128 miles per gallon.

Residuals and diagnostics

The residual plot is shown in the appendix as Figure 2. There seems to be no pattern in the residuals, although they are slightly off the normal distribution. There are no significant outliers as well. Outliers can be tested with `outlierTest` function from `car` package, which performs a Bonferonni p-value test.

```r
library(car)
outlierTest(fit)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##          rstudent unadjusted p-value Bonferonni p
## Fiat 128 2.443362            0.02165       0.6928
```

This function tells us that there are no statistically significant outliers and just showing that the most distant observation has an adjusted P-value for the hypothesis of not being an outlier at 0.693.

Results

Summarising the above, the proposed model that considers interaction between regressors `am` and `wt` with `qsec` as a confounder explains more than 88% of variablility and suggests that cars with manual transmission are better for MPG and have an increase of it in interval of 7.031 to 21.128 with the average value of 14.079.

Appendix

Code Chunk 1. Code for converting necessary variables to factors

```
data("mtcars")
mtc <- mtcars
mtc[, c('am', 'cyl', 'vs', 'gear', 'carb')] <-
        lapply(mtcars[,c('am', 'cyl', 'vs', 'gear', 'carb')], factor)
```

Figure 1. Exploratory plot. This is a paired plot with regression lines and color heatmap for covariances

```
library(gclus)
panel.regression <- function (x, y, col = par("col"), bg = NA, pch = par("pch"),
                              cex = 1, col.regres = "red", ...)
{
        points(x, y, pch = pch, col = col, bg = bg, cex = cex)
        ok <- is.finite(x) & is.finite(y)
        if (any(ok))
                abline(stats::lm(y[ok] ~ x[ok]), col = col.regres, ...)
}
dta.r <- abs(cov(mtcars))
dta.col <- dmat.color(dta.r, colors = heat.colors(20, alpha = 0.9))
dta.o <- order.single(dta.r)
cpairs(mtcars, dta.o, panel.colors = dta.col, gap = 0.5, upper.panel = panel.regression)
```

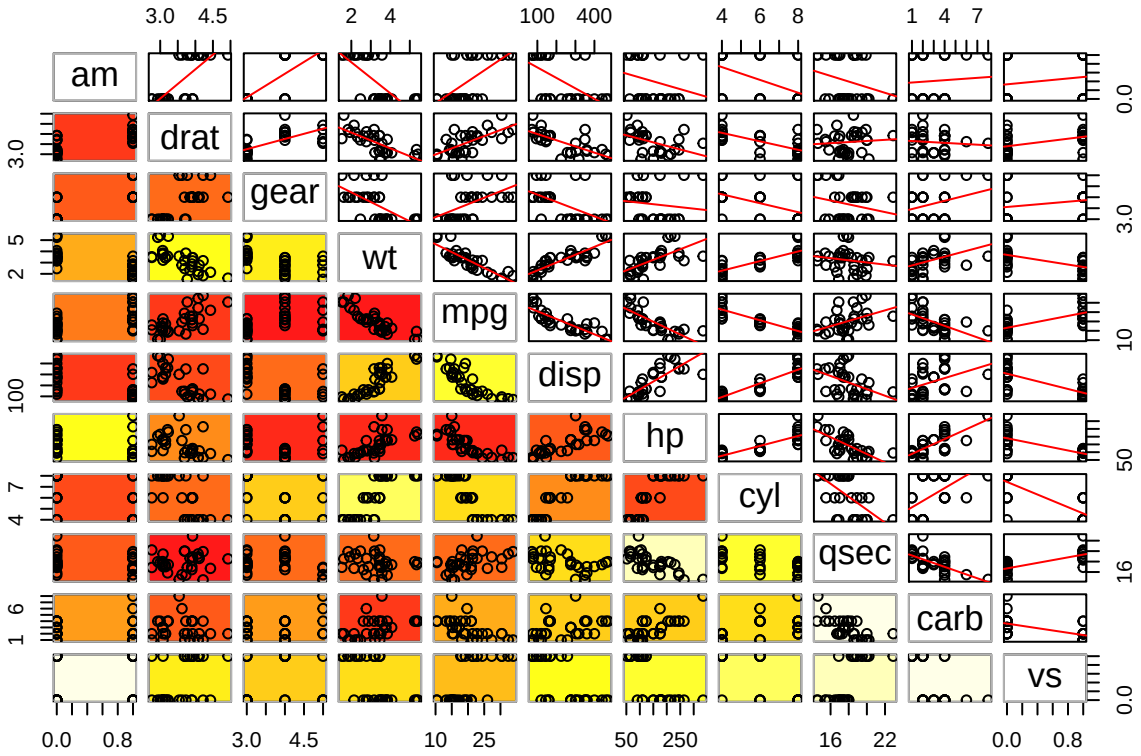

Table 1. Comparison of model with one variable and model with all the variables

```
anova(lm(mpg ~ am, mtc), lm(mpg ~ ., mtc))
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.9
## 2     15 120.4 15    600.49 4.9874 0.001759 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2. Residual plot for the final model

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