The transmission type effect on Miles per gallon in the *mtcars* R dataset - *Elena Civati*

Summary

In this analysis we investigated the relationship between transmission (automatic or manual) and fuel consumption (MPG, Miles per US gallon), using data for 32 car models from the 1974 *Motor Trend* US magazine.
Although there is a significant difference in mean MPG in the 2 groups (manual transmission cars being more fuel saving), further investigation shows that this relationship isn't that clear once we account for other aspects of automobile design.

Exploratory analysis

First, some of the variables in the dataset were modified to be treated more conveniently.

```
library(dplyr); data("mtcars")
cars<-mutate(mtcars, cyl=as.factor(cyl), vs=ifelse(vs==0, "Vshaped", "Straight"),
             am=ifelse(am==0, "Automatic", "Manual"))
```

In Plot 1, we can see the distribution of MPG with respect to the Transmission variable.
Performing a t test would also confirm the difference shown by the plot:

```
t.test(cars$mpg ~ cars$am)
```

```
##
##  Welch Two Sample t-test
##
## data:  cars$mpg by cars$am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

Nonetheless, other characteristics of the cars are highly correlated with our regressor of interest, thus being potential confounders; in Plot 2 we summarize these correlations.

Model selection

As the outcome is a continuos variable, we choose to use a linear model with the categorical transmission variable ("am") as the main regressor. From Plot 2, potential confounders are: cyl (number of cylinders), disp (engine displacement), drat (Rear exle ratio), wt (car weight) and gear (number of forward gears); some of them being highly correlated with each other, it wouldn't be appropriate to include them all in our model, as shown by the following variance inflation analysis:

```
library(car); fit<-lm(mpg ~ am+cyl+disp+drat+wt+gear, data=cars); vif(fit)
```

```
##           GVIF Df GVIF^(1/(2*Df))
## am    3.744802  1        1.935149
## cyl   7.774422  2        1.669810
## disp 12.958482  1        3.599789
## drat  3.266570  1        1.807365
## wt    6.641311  1        2.577074
## gear  3.153163  1        1.775715
```

We decided to discard “disp” and “wt” that lead to great inflation of variance. We then used ANOVA to decide which of the remaining variables will be used in the final model.

```
fit0<-lm(mpg ~ am, data=cars); fit1<-lm(mpg ~ am+cyl, data=cars)
fit2<-lm(mpg ~ am+cyl+drat, data=cars); fit3<-lm(mpg ~ am+cyl+drat+gear, data=cars)
anova(fit0, fit1, fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + drat
## Model 4: mpg ~ am + cyl + drat + gear
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 264.50  2    456.40 23.1569 1.678e-06 ***
## 3     27 264.32  1      0.17  0.0175    0.8958
## 4     26 256.22  1      8.11  0.8225    0.3728
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results show that “drat” and “gear” give no added value to the model. So, we’ll keep the transmission type as regressor, and just the number of cylinders as a confounder (as a factor, as the effect of adding a cylinder is probably not additive). This factor has 3 levels:

```
levels(cars$cyl)
```

```
## [1] "4" "6" "8"
```

Conclusions

```
round(summary(fit1)$coef, 5)
```

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 24.80185    1.32261 18.75213  0.00000
## amManual     2.55995    1.29758  1.97287  0.05846
## cyl6        -6.15612    1.53572 -4.00861  0.00041
## cyl8       -10.06756    1.45208 -6.93319  0.00000
```

The Intercept coefficient gives us a mean estimate of 24.802 mpg for a 4 cylinder car with automatic transmission.
The amManual coefficient suggest an average increase of 2.56 mpg for a car with a manual transmission, given a constant number of cylinders. However, its *p* value is 0.05846: quite low, but not enough to call this coefficient significant at a standard level of 0.05. In fact, its 95% confidence interval contains 0, so we can’t infer that manual transmission implies fuel saving.

```
confint(fit1)[2,]
```

```
##       2.5 %      97.5 %
## -0.09801611  5.21792352
```

Diagnostic

The residual plots (Plot 3) show that our model (or maybe the dataset) is affected by heteroskedasticity. The variance isn’t constant, but tends to be higher for the more extreme values of MPG.

Plots

```
boxplot(cars$mpg ~ cars$am, ylab="Miles/US gallon", xlab="Transmission",
        main="MPG by transmission type", col="lightblue")
```

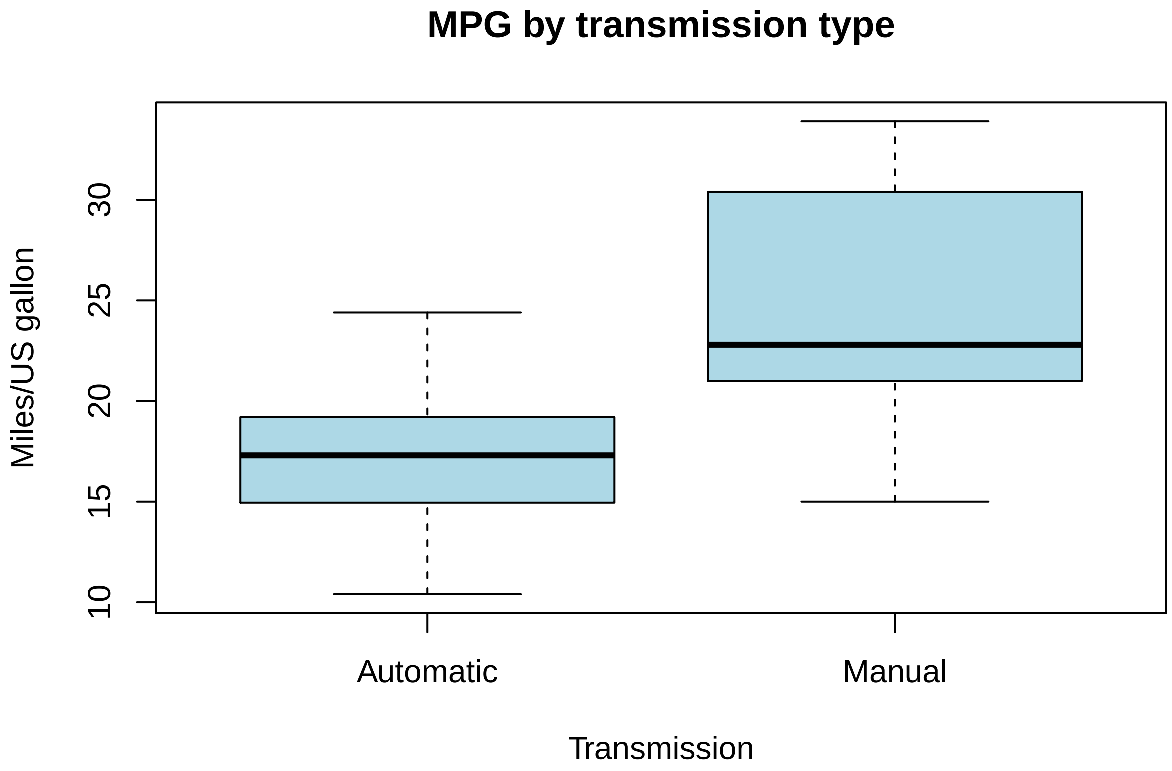

Figure 1: Plot 1

```
library(ggplot2)
library(GGally)
g<-ggpairs(cars)
g
```

```
par(mfrow=c(2,2))
plot(fit1)
```

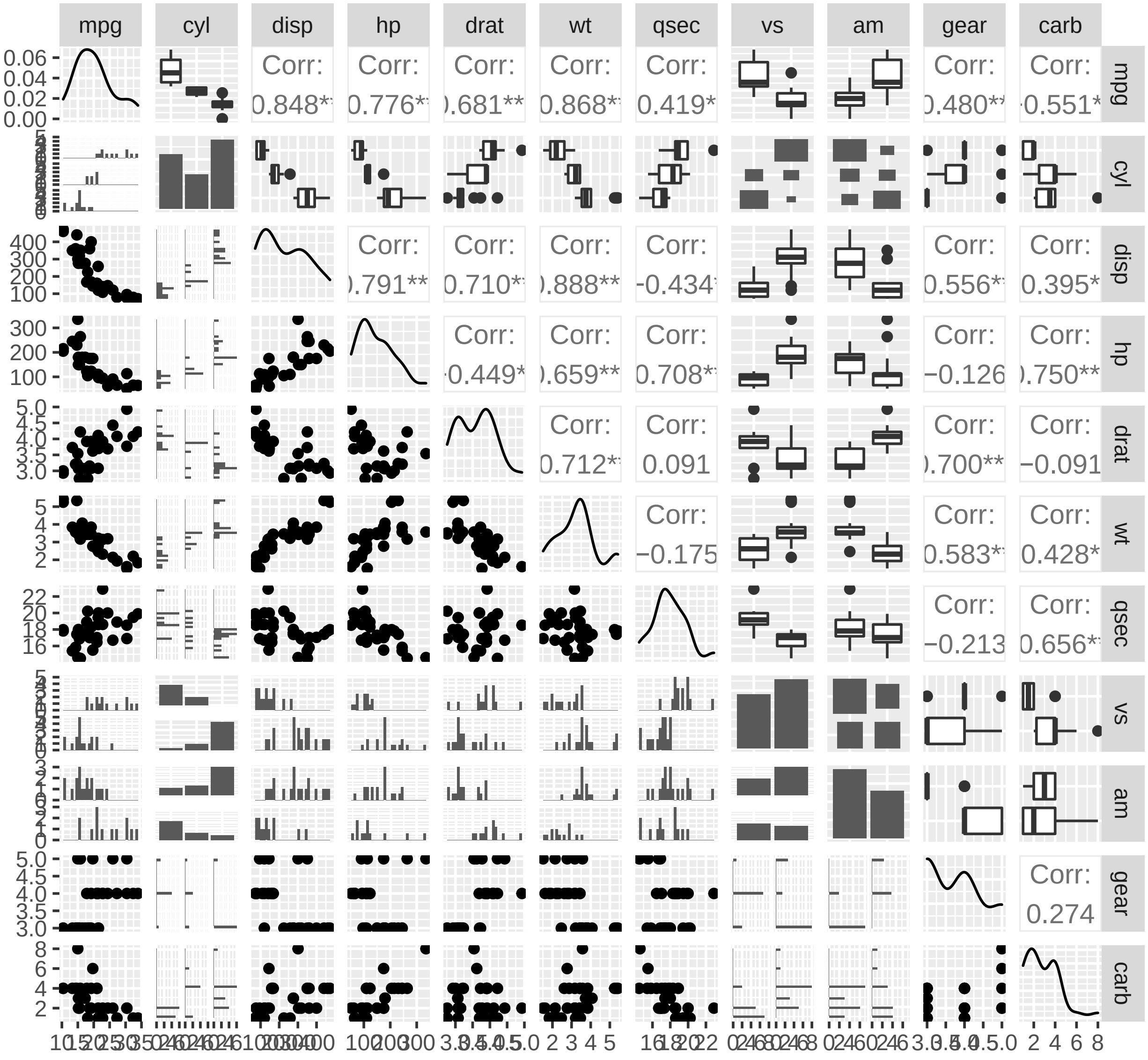

Figure 2: Plot 2

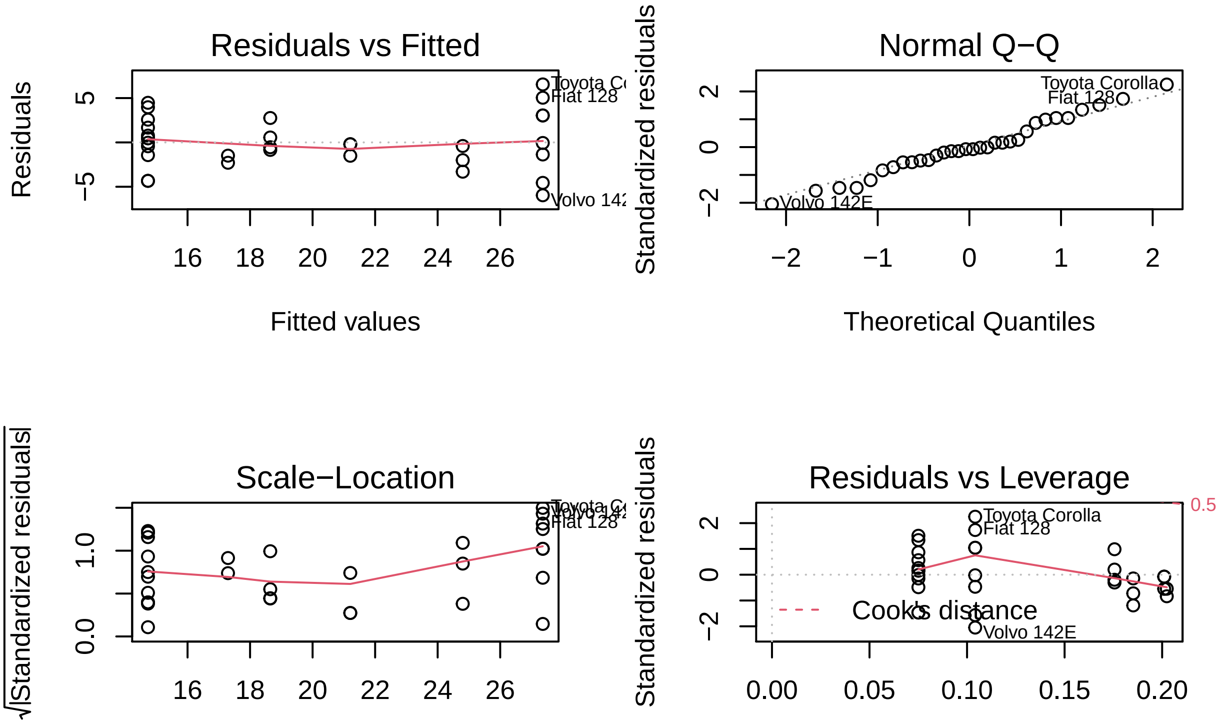

Figure 3: Plot 3