Relationship between variables and miles per gallon

Executive Summary

Task

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmission"

Exploration Data analysis

The first step of our analysis is to load the `mtcard` dataset and if we are interested we check its data structure by using `str()` and `?mtcard`.

```
data("mtcars")
str(mtcars)
```

Now we will focus on study the relation between the type of transmission and the number of miles per gallon, `am` (0 = automatic and 1 = manual) and `mpg` variables respectively.

We will plot some relations between the transmission `am` and miles per gallon `mpg` in order to identify it help us to identify any patterns. We will work with `ggplot2`, `GGally` and `lmtest` libraries so we need to load both.

Our first figure will show the values of MPG related to manual and automatic cars but first we will add factor and label it as automatic and manual transmission.

```
mtcars$trans <- factor(mtcars$am, labels = c("automatic", "manual"))
```

```
  plot1 <- ggplot(mtcars, aes(x=mpg, fill = factor(mtcars$trans))) +
  theme_bw(base_size = 8, base_family = "Helvetica") +
  geom_bar( binwidth = 1, col = 'black', position="dodge") +
  labs( fill = 'Transmission') +
  labs( title = "Fig 1. Miles per gallon", x = "MPG")
```

From the Fig 1. (attached to the appendix) we can identify that manual cars have a higher values on miles per gallon. Following the Fig 2. a boxplot for miles per gallon and transmission.

```
plot2 <- ggplot(mtcars, aes(x=factor(mtcars$trans), y=mpg,
                            fill = factor(mtcars$trans))) +
  theme_bw(base_size = 8, base_family = "Helvetica") + geom_boxplot(adjust = 1) +
  geom_jitter(size = 1) + labs( fill = 'Transmission') +
  labs( title = 'Fig 2. MPG for transmission', x = 'Transmission', y = 'MPG')
```

As you can see in both figures attached in the appendix, seems that there is some pattern between MPG and Transmission, anyway this patterns can not be explained just with this evidences. We have to check if it is related to other variables. In order to check this relations we can proceed by plot correlation heatmap which show us in which measure the variables are correlated.

```
heatmap(cor(mtcars[1:11]), main= "mtcars dataset heatmap correlation");
```

From the previous plot Fig 3. a heatmap showing correlation (also attached to the appendix), we notice that there is very little correlation between transmission and miles per gallon, variables `am` and `mpg` respectively.

Regression models - Simple model

Following we will give an estimation of the effect of transmission on miles per gallon. In order to do that we will build a linear model and compute the confidence interval.

```
simple_model <-lm(mpg ~ trans, data=mtcars)
model_coef <- summary(simple_model)$coefficients
model_coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## transmanual  7.244939   1.764422  4.106127 2.850207e-04
```

```
interval_simple <- model_coef[2,1] + c(-1,1) *
                              qt(0.975, df = simple_model$df) *
                              model_coef[2,2]
interval_simple
```

```
## [1]  3.64151 10.84837
```

From the above results we obtained a p-value: 2.8502074×10^{-4} and an interval of $[3.6415096, 10.848369]$. Considering the model and the results, compared both transmisions we can conclude that a manual transmission increases on average 7.2449393 miles per gallon over the automatic transmission and the 95% confidence interval is $[3.6415096, 10.848369]$

For complet the analysis in a simple model we can plot the residuals and check for heteroskedasticity that suggest the presence of more variables.

```
plot3 <- ggplot(mtcars, aes(x = trans, y = resid(simple_model), fill = trans)) +
  theme_bw(base_size = 8, base_family = "Helvetica") + geom_boxplot(adjust = 1) +
  geom_jitter(size = 1) +
  labs(title='Fig 4. Residuals - Simple Model',x='Transmission',y='Residuals')
```

From the result shown in the appendix we can observe a large variability for manual transmission and a possible heteroskedasticity case, presence of large variance in the model. Will be necessary to consider a multivariable model. We can ensure the presence of heteroskedasticity by checking studentized Breusch-Pagan test. For this purpose we will need to load `lmtest` library.

```
bptest(simple_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  simple_model
## BP = 5.0771, df = 1, p-value = 0.02424
```

The presence of low p-value 0.0242439 give us strong evidences of possible hidden variables so we will proceed with a multivariable model.

Regression models - Multivariable model

Following the multivariable study:

```r
multi_model <- lm(mpg ~ ., data = mtcars)
mmodel_coef <- summary(multi_model)$coefficients
mmodel_coef
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## am           2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

One approach in this multiple model is to perform a backwards elimination, start with all the predictors in the model and remove the variable with the higher P value. We will perform this process to removes the less significant variable till all values are smaller than an specific value in this case `0.05`

Following the iterative process:

```r
dat <- mtcars[, c("mpg", "wt", "qsec", "am")]
fit <- lm(mpg ~ . - 1, data = dat)
summary(fit)$coefficients

dat <- mtcars[,1:11]
dat <- dat[, names(dat) != "cyl"]; summary(lm(mpg ~ ., data = dat))$coefficients
dat <- dat[, names(dat) != "vs"]; summary(lm(mpg ~ ., data = dat))$coefficients
dat <- dat[, names(dat) != "carb"]; summary(lm(mpg ~ ., data = dat))$coefficients
dat <- dat[, names(dat) != "gear"]; summary(lm(mpg ~ ., data = dat))$coefficients
dat <- dat[, names(dat) != "drat"]; summary(lm(mpg ~ ., data = dat))$coefficients
dat <- dat[, names(dat) != "disp"]; summary(lm(mpg ~ ., data = dat))$coefficients
dat <- dat[, names(dat) != "hp"]; summary(lm(mpg ~ ., data = dat))$coefficients
```

```r
complex_model <- lm(mpg ~ . - 1, data = dat)
model_coef2 <- summary(complex_model)$coefficients
model_coef2
```

```
##       Estimate Std. Error   t value     Pr(>|t|)
## wt   -3.185455  0.4827586 -6.598442 3.128844e-07
## qsec  1.599823  0.1021276 15.664944 1.091522e-15
## am    4.299519  1.0241147  4.198279 2.329423e-04
```

As is shown above we removed the unrelated variables one by one, the final model just contains `wt`, `qsec` and `am`. We can conclude that the predicting model for the miles per gallon of a vehicle is `y = -3.185455wt + 1.599823qsec + 4.299519am`

```r
interval_complex <- model_coef2[2,1] + c(-1,1) *
                              qt(0.975, df = complex_model$df) *
                              model_coef2[2,2]
interval_complex
```

```
## [1] 1.390948 1.808697
```

The adjusted R-squared for the model is 0.9857902, which is satisfying. And the adjusted 95% confidence interval is $[1.3909482, 1.8086969]$. In apendix also included the residuals.

Conclusions

The study has shown that there is no correlation between the variables am and mpg that justify that we cannot answer this questions without consider other relevant variables as wt and qsec, weigh and 1/4 mile time for explain the mile per gallon consumption. The better aproximation to both answer is to evaluate the function `y = -3.185455wt + 1.599823qsec + 4.299519am`

Appendix

Fig 1. Miles per gallon

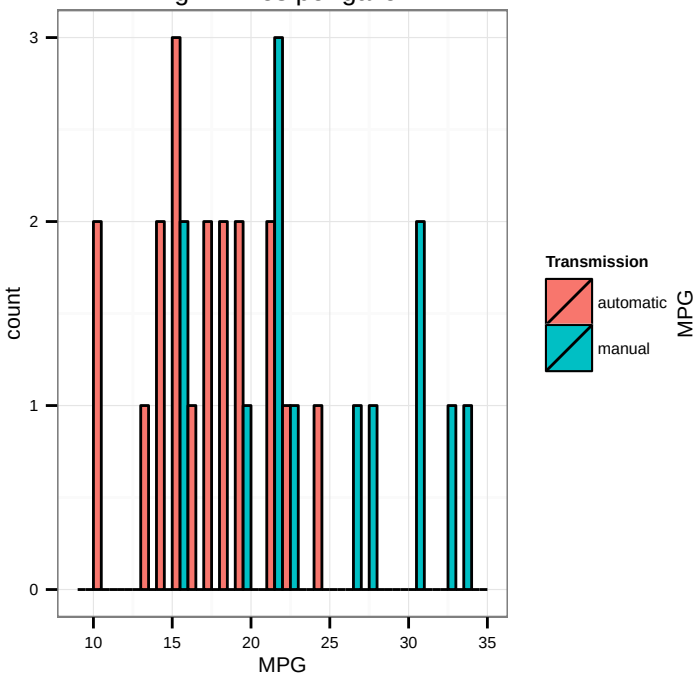
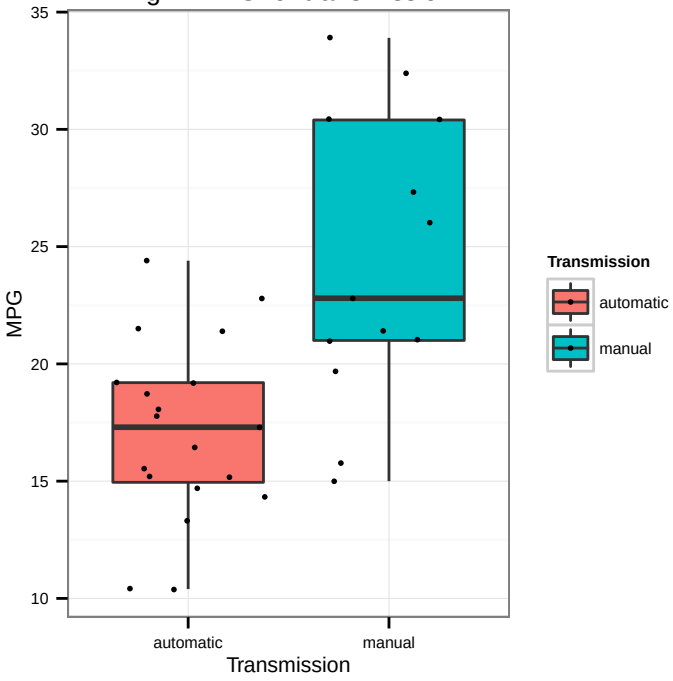


Fig 2. MPG for transmission



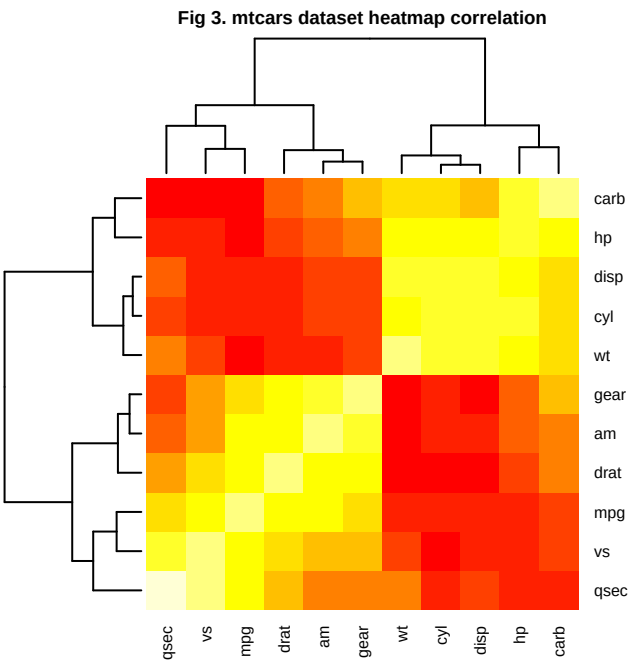


Fig 3. mtcars dataset heatmap correlation
carb
hp
disp
cyl
wt
gear
am
drat
mpg
vs
qsec
qsec
vs
mpg
drat
am
gear
wt
cyl
disp
hp
carb

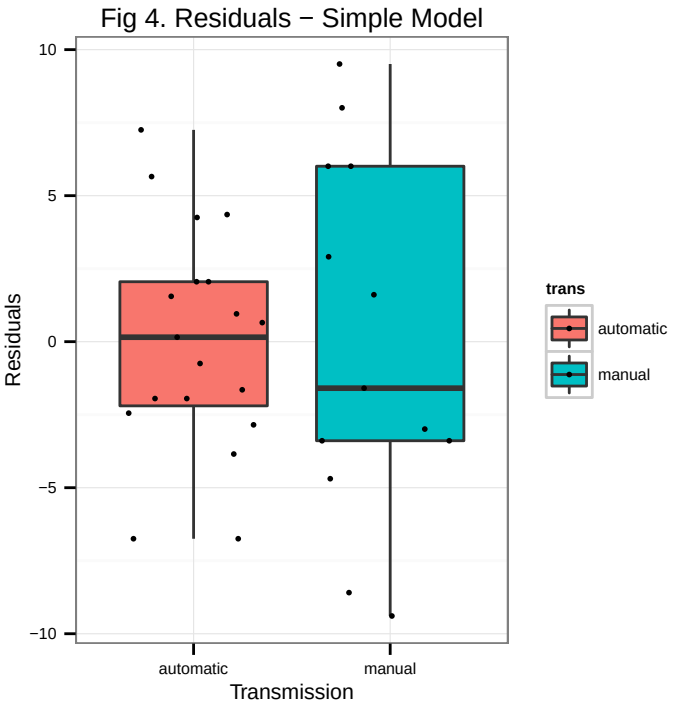
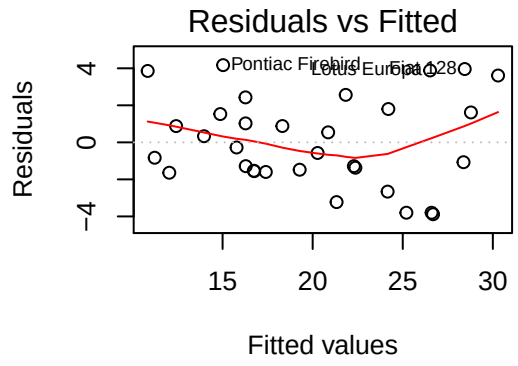
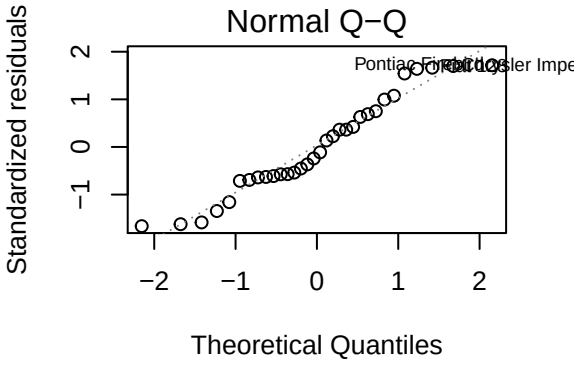
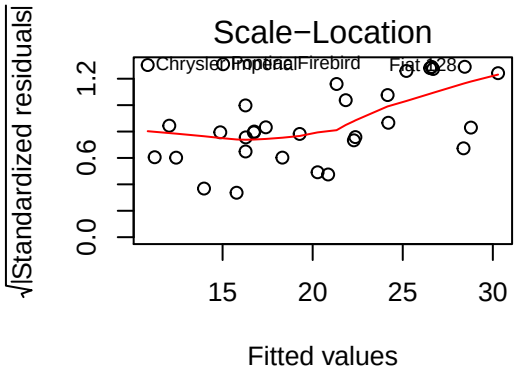
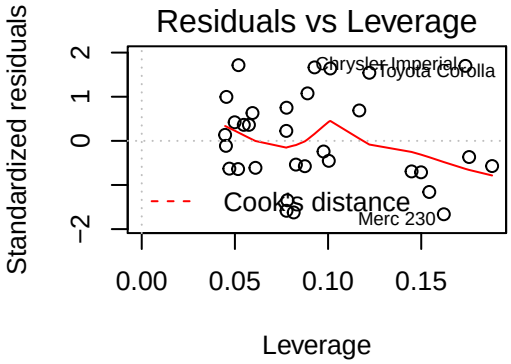


Fig 4. Residuals − Simple Model
Residuals
10
5
0
−5
−10
automatic
manual
Transmission
trans
automatic
manual


Residuals vs Fitted
Residuals
Pontiac Firebird
Lotus Europa
Fiat 128
4
0
−4
15
20
25
30
Fitted values


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−2
−1
0
1
2
Theoretical Quantiles


Scale−Location
√|Standardized residuals|
Chrysler Imperial
Pontiac Firebird
Fiat 128
1.2
0.6
0.0
15
20
25
30
Fitted values


Residuals vs Leverage
Standardized residuals
Chrysler Imperial
Toyota Corolla
Cook's distance
Merc 230
2
1
0
−2
0.00
0.05
0.10
0.15
Leverage