The influence of manual and automatic transmission on fuel consumption based on the dataset mtcars

Bruno

2023-08-15

Summary

We analyzed the dataset *mtcars*, which considers the fuel consumption and other car features for 32 automobiles, extracted from the 1974 Motor Trend US magazine.

Using a simple linear regression model, we obtained a linear relationship between the variables *mpg* and *am*, where *mpg* stands for fuel consumption (in miles per gallon) and *am* reffers to the car transmission type (automatic or manual). The increase in the mean fuel consumption of cars with manual transmission with respect to the mean for cars with automatic transmission was equal to 7.24, with a p-value of 0.00029. In this case, the 95% confidence interval for the slope of our line is $[3.64151, 10.84837]$. By inspecting models considering a larger set of variables, for the most significant plot, we obtained that the coefficient related to *am* is compatible with zero (p-value equal to 0.20). Therefore, we concluded that the influence of the transmission type on *mpg* is not statistically significant, and the most appropriate model should not include the variable *am*.

Introduction

The data used in this work was extracted from the 1974 Motor Trend US magazine and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

The data frame *mtcars* contains 32 observations on 11 variables. The variables with their description are shown as follows.

- **mpg**: Miles per (US) gallon
- **cyl**: Number of cylinders
- **disp**: Displacement (cu.in.)
- **hp**: Gross horsepower
- **drat**: Rear axle ratio
- **wt**: Weight (1000 lbs)
- **qsec**: 1/4 mile time
- **vs**: Engine (0 = V-shaped, 1 = straight)
- **am**: Transmission (0 = automatic, 1 = manual)
- **gear**: Number of forward gears
- **carb**: Number of carburetors

In this work, we used the following packages.

```r
library(tidyverse)
library(GGally)
```

Importing and cleaning data

First, let us start by importing the dataset *mtcars* and checking its variables.

```
df <- mtcars
str(df)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Through a brief inspection of the dataset, we can verify that there are no duplicated rows or missing values in it.

```
# Number of duplicated rows
sum(duplicated(df))
```

```
## [1] 0
```

```
# Number of rows with NAs
sum(!complete.cases(df))
```

```
## [1] 0
```

Let us convert the variables *cyl*, *vs*, *am*, *gear*, and *carb* as factor variables, where for the variable *vs* we label the value 0 as "V" and the value 1 as "S"; and for the variable *am*, we label 0 as "Automatic" and 1 as "Manual."

```
df$vs <- factor(df$vs, levels = c(0,1), labels = c("V", "S"))
df$am <- factor(df$am, levels = c(0,1), labels = c("Automatic", "Manual"))

fac_var <- c("cyl", "gear", "carb")
for (i in fac_var) df[[i]] <- factor(df[[i]])
```

```
summary(df)
```

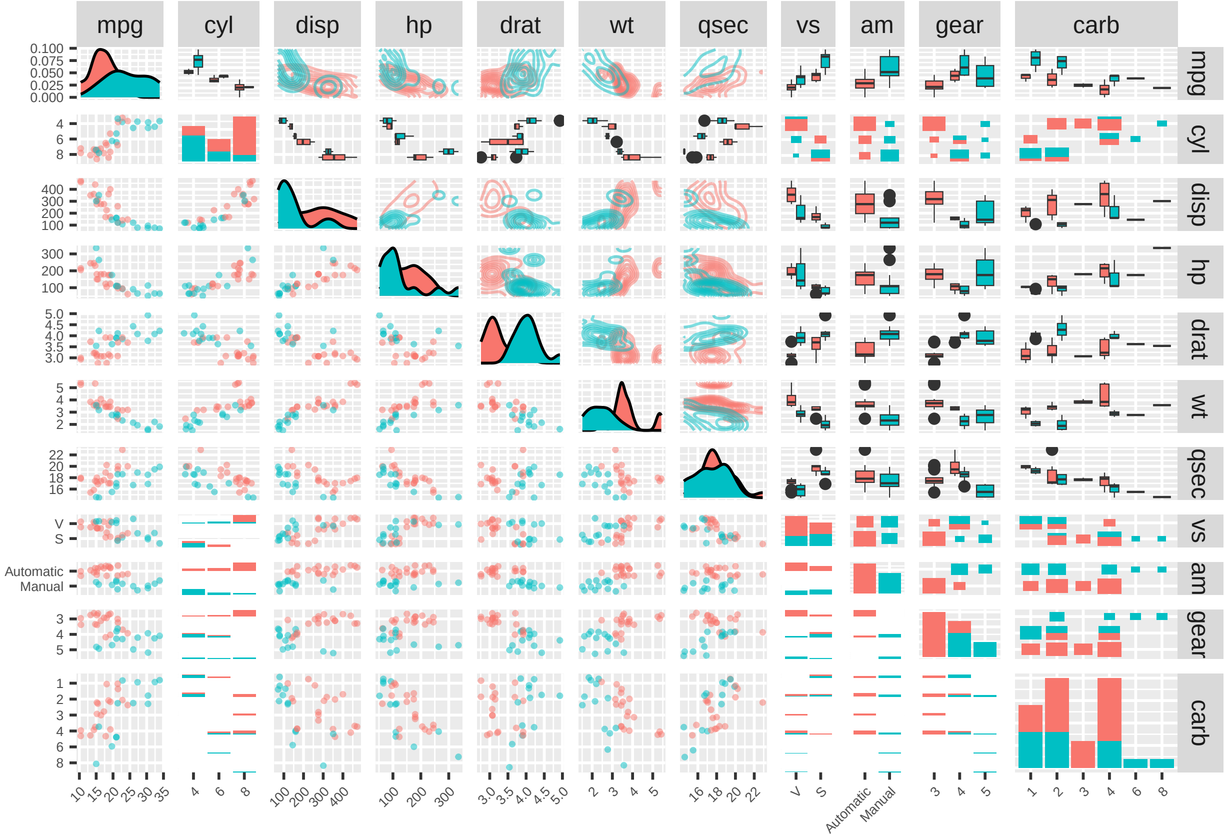
```
##       mpg        cyl         disp             hp             drat
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930
##        wt             qsec       vs             am     gear   carb
##  Min.   :1.513   Min.   :14.50   V:18   Automatic:19   3:15   1: 7
##  1st Qu.:2.581   1st Qu.:16.89   S:14   Manual   :13   4:12   2:10
##  Median :3.325   Median :17.71                         5: 5   3: 3
##  Mean   :3.217   Mean   :17.85                                4:10
##  3rd Qu.:3.610   3rd Qu.:18.90                                6: 1
##  Max.   :5.424   Max.   :22.90                                8: 1
```

Exploratory analysis

In the following, we explore the possible dependence between the variables involved. Since we are interested in answering questions about the dependence of *mpg* against *am*, we use different colors to highlight the data associated with cars with automatic or manual transmissions.

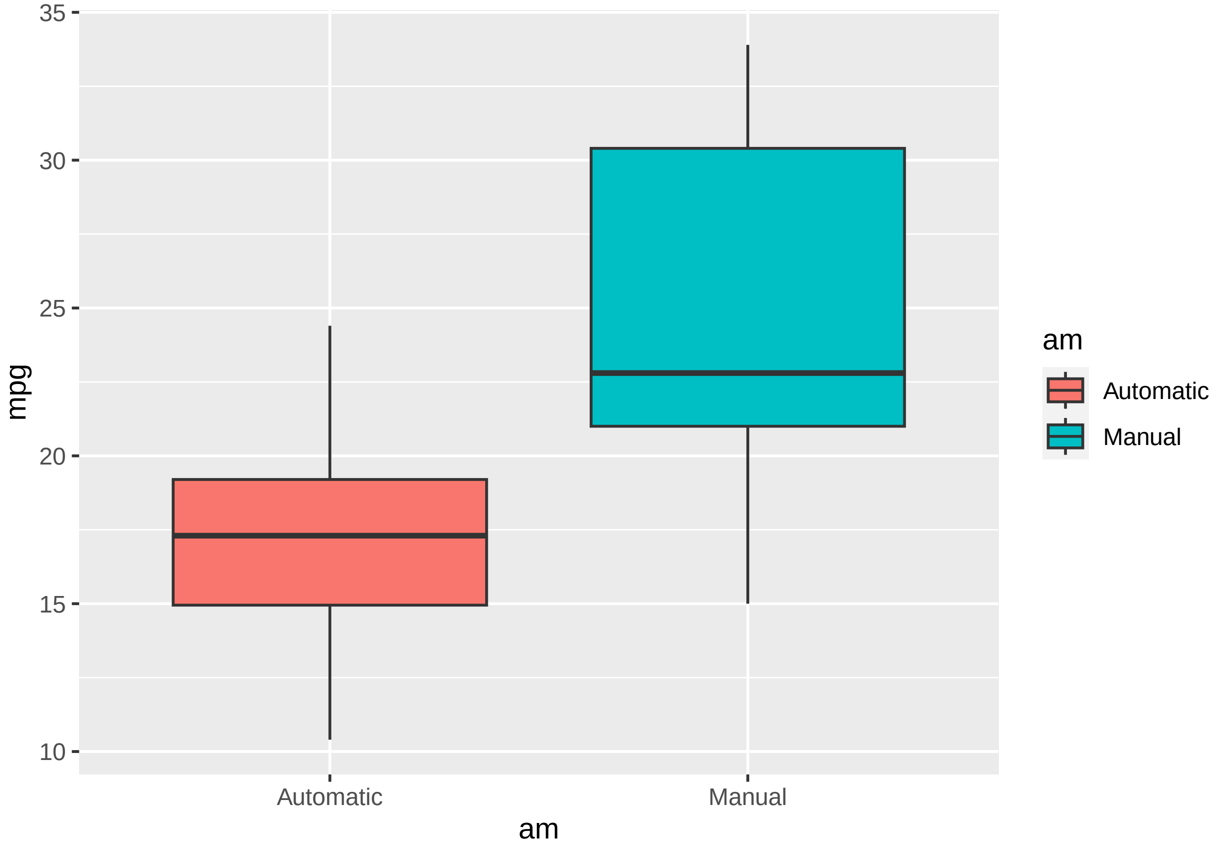
```
g <- ggpairs(
    df,
    mapping = aes(color = am),
    upper = list(continuous = wrap("density", alpha = 0.5),
                 combo = wrap("box_no_facet", lwd  = 0.2)),
    lower = list(continuous = wrap("points", alpha = 0.5, size = 0.5),
                 combo = wrap("dot_no_facet", alpha = 0.5, size = 0.5)),
    proportions = "auto"
) +
    theme(axis.text = element_text(size = 5),
          axis.text.x = element_text(angle = 45, hjust = 1))

print(g)
```



Let us explore the values of *mpg* across the variable *am*.

```
ggplot(data = df,
       mapping = aes(x = am, y = mpg, fill = am)) +
    geom_boxplot()
```



In the boxplot above, we can verify that according to our dataset, there should be a difference between the mean of *mpg* for cars with automatic and manual transmissions. It is important to highlight that our data was not randomized, and there is the possibility of the existence of confounders. Let us make a linear regression in order to discover the possible dependence of *mpg* against the other variables.

Linear regression

mpg vs am

First, let us start with a very simple model where we explore the dependence of *mpg* only against the variable *am*.

```r
fit1 <- lm(mpg ~ am, data = df)
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = df)
##
## 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 ***
## amManual       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
```

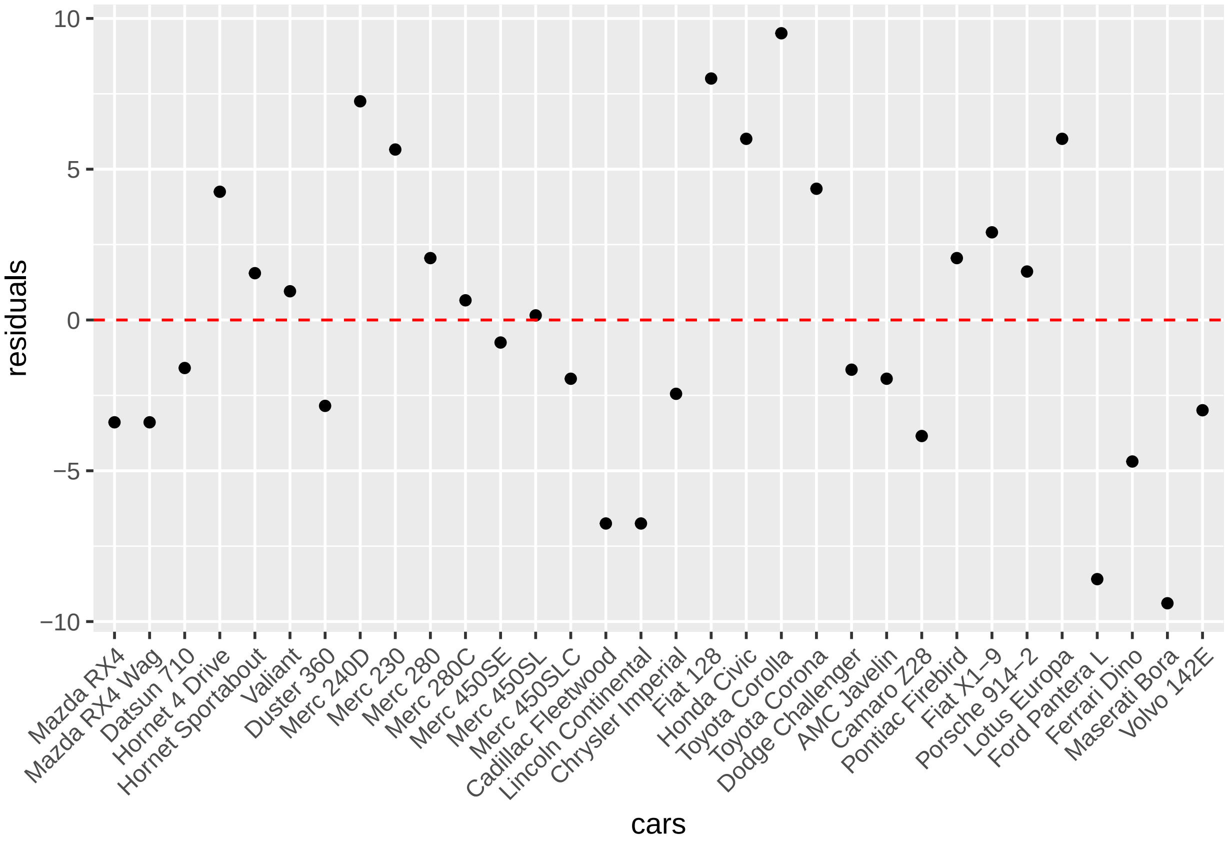
```
confint(fit1)
```

```
##                2.5 %   97.5 %
## (Intercept) 14.85062 19.44411
## amManual     3.64151 10.84837
```

According to the results shown above, there is a dependence between *mpg* and *am*, where the increase (using automatic transmission as a reference) on the mean mpg is equal to 7.24 with a p-value of 0.00029. In this case the 95% confidence interval for the slope of our line is $[3.64151, 10.84837]$

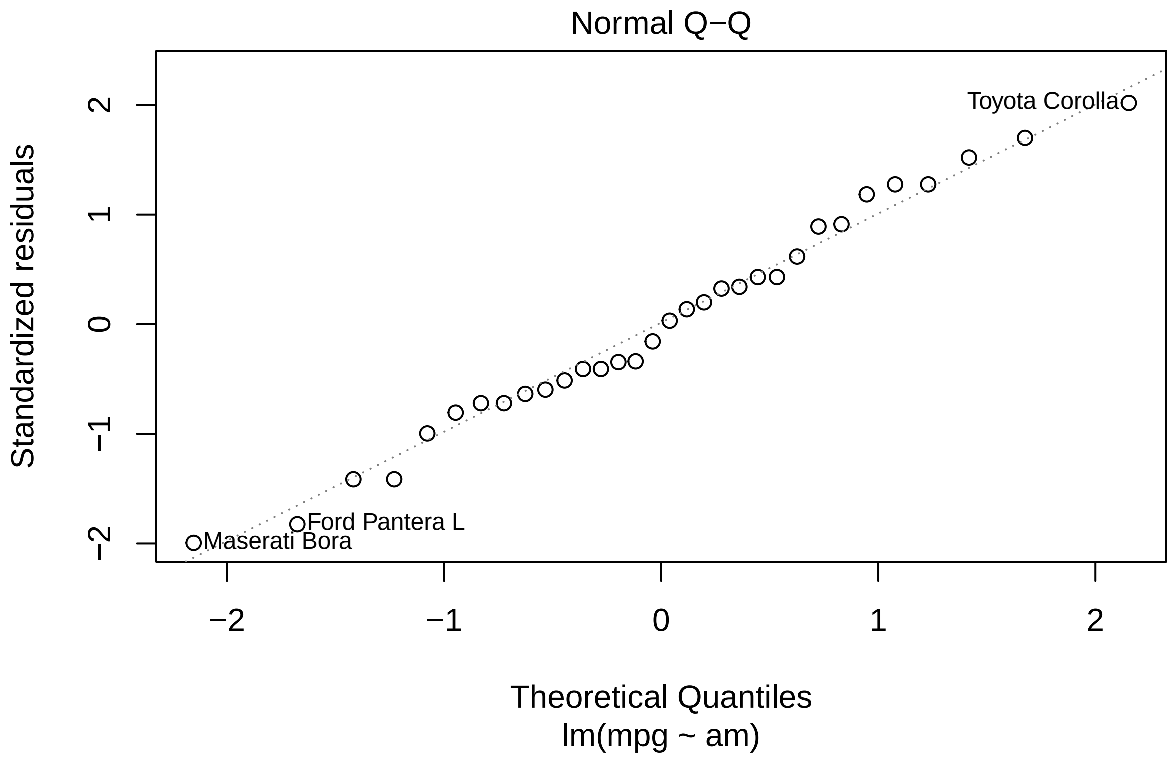
However, this result may be biased in case mpg truly depends on a larger set of variables. In order to address the quality of our fit, let us analyze the residual plot.

```
ggplot(
     data = data.frame(
         cars = factor(rownames(df), levels = rownames(df)),
         residuals = resid(fit1)
     ),
     mapping = aes(x = cars, y = residuals)
 ) +
     scale_x_discrete(guide = guide_axis(angle = 45)) +
     geom_point() +
     geom_hline(yintercept = 0,
                linetype = "dashed",
                color = "red")
```



Based on the plot above, we conclude that despite the fact that the coefficient associated with *am* being non-zero is statistically significant, the residuals do not seem to distribute randomly around zero.

```r
plot(fit1, which = 2)
```



Furthermore, by inspecting the QQ plot above, we see that the residuals do not follow a normal distribution. Therefore, we have to analyze models with a larger set of variables to obtain a better fit and check whether such a dependence still holds.

Model selection

Let us consider the series of nested models that include the dependence on the variable *am*:

- Model 1: mpg ~ am
- Model 2: mpg ~ am + wt
- Model 3: mpg ~ am + wt + cyl
- Model 4: mpg ~ am + wt + cyl + hp
- Model 5: mpg ~ am + wt + cyl + hp + gear

```
fit1 <- lm(mpg ~ am, data = df)
fit2 <- lm(mpg ~ am + wt, data = df)
fit3 <- lm(mpg ~ am + wt + cyl, data = df)
fit4 <- lm(mpg ~ am + wt + cyl + hp , data = df)
fit5 <- lm(mpg ~ am + wt + cyl + hp + gear, data = df)

 anova(fit1,fit2,fit3, fit4, fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + cyl
## Model 4: mpg ~ am + wt + cyl + hp
## Model 5: mpg ~ am + wt + cyl + hp + gear
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 70.9708 1.249e-08 ***
## 3     27 182.97  2     95.35  7.6452  0.002698 **
## 4     26 151.03  1     31.94  5.1223  0.032945 *
## 5     24 149.67  2      1.36  0.1091  0.897096
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to ANOVA, adding the variable *gear* to Model 4 gives us an F-statistic with a p-value greater than 0.05. Note that among the 5 models considered above, Model 4 is the most appropriate one. The values for the intercept and coefficients for fit4 are given as follows.

```
summary(fit4)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + cyl + hp, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## amManual     1.80921    1.39630   1.296  0.20646
## wt          -2.49683    0.88559  -2.819  0.00908 **
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

Although the F-statistic for Model 4 is statistically significant with a p-value equal to 4.87×10^{-11}, the hypothesis that the coefficient associated with *am* is non-zero cannot be rejected since the p-value for the t-test is equal to 0.20.

Conclusion

By using a simple linear regression model, we found a relationship between the variables *mpg* and *am*, where the increase (using automatic transmission as a reference) in the mean mpg was equal to 7.24 with a p-value of 0.00029. In this case, the 95% confidence interval for the slope of our line is $[3.64151, 10.84837]$. By inspecting the residuals of our simple linear regression, we considered a model which considers a larger set o variables. For the most significant plot, we obtained that the coefficient related to *am* is compatible with zero (p-value equal to 0.20). Therefore, we concluded that the influence of the transmission type on *mpg* is not statistically significant, and the most appropriate model should not include the variable *am*.