Øving 2

2022-04-21

Problem 1

```
library(ISLR)

Auto = subset(Auto, select = -name)  #Removing the name column form the dataset

str(Auto)
```

```
## 'data.frame':    392 obs. of  8 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : num  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : num  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : num  1 1 1 1 1 1 1 1 1 1 ...
```

We have 392 samples of 8 variables.

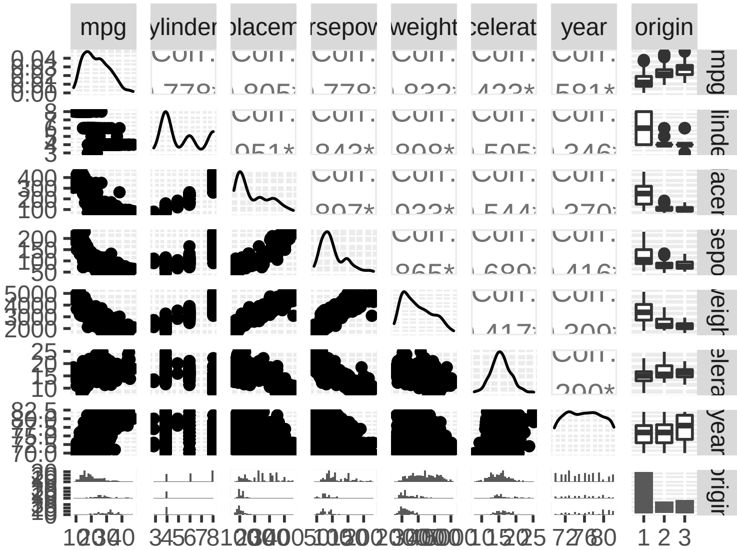
Because origin is not a quantitative variable, but a qualitative encoded with 1, 2, and 3, we need to let R know that these variables are not numerical values so we do not get wrong model fits. We use the factor() function.

```
Auto$origin = factor(Auto$origin)
```

a) Use the function ggpairs() from GGally to produce a scatterplot matrix which includes all of the variables in the data set,

```
library(GGally)

ggpairs(Auto)
```



b)

Compute the correlation matrix between the variables.

```r
ReducedAuto = Auto[, -8]    #Removing the 8th column (origin)

corr_matrix = cor(ReducedAuto)

corr_matrix
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
##              acceleration       year
## mpg             0.4233285  0.5805410
## cylinders      -0.5046834 -0.3456474
## displacement   -0.5438005 -0.3698552
## horsepower     -0.6891955 -0.4163615
## weight         -0.4168392 -0.3091199
## acceleration    1.0000000  0.2903161
## year            0.2903161  1.0000000
```

c) Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables (except name) as the predictors.

Comment on:

i) Is there a relationship between the predictors and the response?

ii) Is there evidence that the weight of a car influences mpg? Interpret the regression coefficient β_{weight}.
iii) What does the coefficient for the year variable suggest?

```r
model = lm(mpg ~ ., data = Auto)

summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0095 -2.0785 -0.0982  1.9856 13.3608
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.795e+01  4.677e+00  -3.839 0.000145 ***
## cylinders    -4.897e-01  3.212e-01  -1.524 0.128215
## displacement  2.398e-02  7.653e-03   3.133 0.001863 **
## horsepower   -1.818e-02  1.371e-02  -1.326 0.185488
## weight       -6.710e-03  6.551e-04 -10.243  < 2e-16 ***
## acceleration  7.910e-02  9.822e-02   0.805 0.421101
## year          7.770e-01  5.178e-02  15.005  < 2e-16 ***
## origin2       2.630e+00  5.664e-01   4.643 4.72e-06 ***
## origin3       2.853e+00  5.527e-01   5.162 3.93e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.307 on 383 degrees of freedom
## Multiple R-squared:  0.8242, Adjusted R-squared:  0.8205
## F-statistic: 224.5 on 8 and 383 DF,  p-value: < 2.2e-16
```

i) We observe that the p-value is quite small (2.2e-16), meaning that the probability of observing H_0, that there is no relationship between the predictors and the response is very small, meaning that there definitely is a relationship. The R^2 is also quite high, meaning our model fits well for the data.

ii) The p-value for weight is also very small (2e-16), meaning that this variable has a huge influence on the mpg. The $\beta_{weight} = -6.710 \cdot 10^{-3}$, meaning that for one unit increase in weight, we get a β_{weight} decrease in mpg. So, a car that weights 1000 kg more than another car, would drive 6.7 miles less far per gallon of fuel.

iii) The coefficient $\beta_{year} = 7.77 \cdot 10^{-1}$ suggest that for one unit increase in year, the mpg increases by 0.777. Newer models tend to be able to drive further per gallon of fuel than older models.

iv)