MAS202M Applied Data Analysis — Exercise 3.9

Kári Hlynsson

January 17, 2024

In this exercise I will be looking into the `Auto` data set, the same one as in Exercise 2.8. Start with the packages:

Let's start with some routine exploration of the data. First of all, the data types:

```
glimpse(Auto)
```

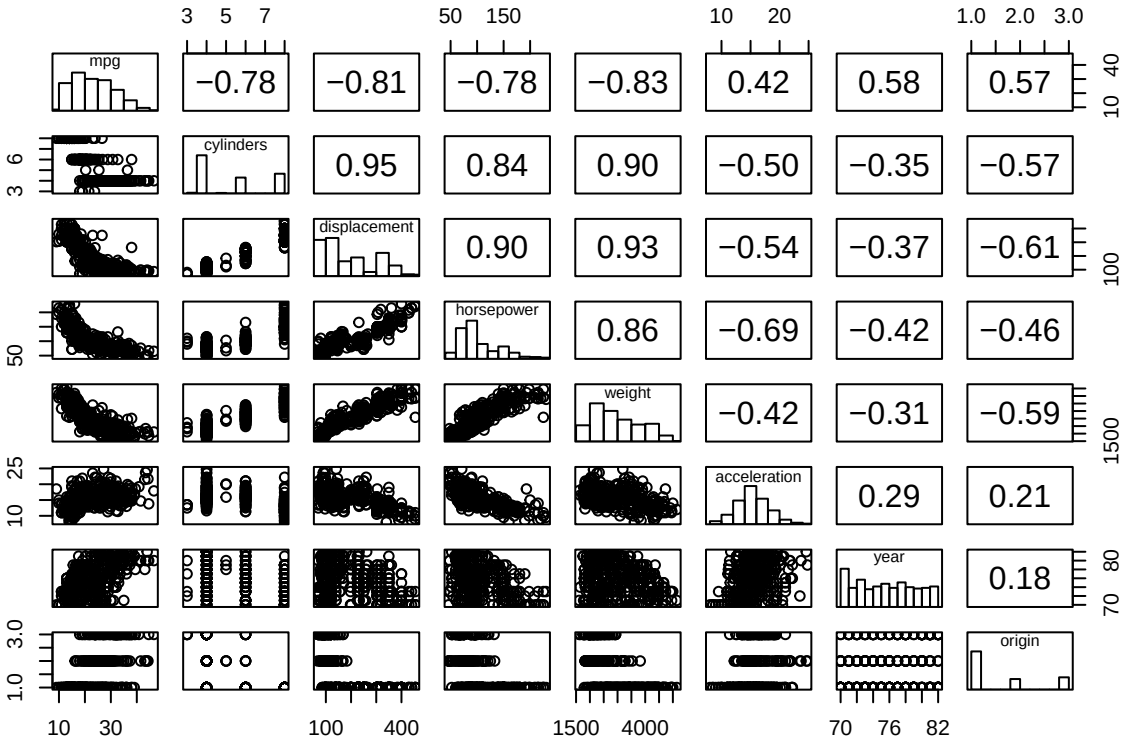
```
## Rows: 392
## Columns: 9
## $ mpg          <dbl> 18, 15, 18, 16, 17, 15, 14, 14, 14, 15, 15, 14, 15, 14, 2~
## $ cylinders    <int> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 4, 6, 6, 6, 4, ~
## $ displacement <dbl> 307, 350, 318, 304, 302, 429, 454, 440, 455, 390, 383, 34~
## $ horsepower   <int> 130, 165, 150, 150, 140, 198, 220, 215, 225, 190, 170, 16~
## $ weight       <int> 3504, 3693, 3436, 3433, 3449, 4341, 4354, 4312, 4425, 385~
## $ acceleration <dbl> 12.0, 11.5, 11.0, 12.0, 10.5, 10.0, 9.0, 8.5, 10.0, 8.5, ~
## $ year         <int> 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 7~
## $ origin       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 3, ~
## $ name         <fct> chevrolet chevelle malibu, buick skylark 320, plymouth sa~
```

The data is 392 observations of 9 variables. Let's create a scatterplot matrix of the variables in the data set:

Part (a)

Produce a scatterplot matrix which includes all of the variables in the data set.

```
pair_plot(Auto |> select(-name))
```



Part (b)

Correlation matrix:

```
cor(Auto |> select(-name))
```

```
##                     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
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

As I commented on previously, there are some quite strong correlations between some of the variables. For instance, `displacement` and `horsepower` have a correlation coefficient of 0.9329944.

Part (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. Use the `summary()` function to print the results. Comment on the output. For instance:

Start by fitting the model:

```
model_all <- lm(mpg ~ . - name, data = Auto)

summary(model_all)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

The linear model fitted to the data is specified by

$$\widehat{\text{mpg}} = -17.2184 - 0.4934(\text{cylinders}) + 0.0199(\text{displacement}) - 0.017(\text{horsepower}) - 0.0065(\text{weight}) + 0.0806(\text{acceleration}) + 0.7508(\text{year}) + 1.4261(\text{origin}) \tag{1}$$

Significant non-intercept coefficents include `displacement`, `weight`, `year` and `origin`. The fact that the rest are not significant may be due to due to collinearity between variables (e.g. `weight` and `displacement`) as seen earlier in the correlation matrix from part (b). Thus we can not conclude that the coefficients for these variables are zero, but we will certainly investigate to see whether accounting for variable interactions changes our findings.

i. Is there a relationship between the predictors and the response? Yes.

ii. Which predictors appear to have a statistically significant relationship to the response? See above.

iii. What does the coefficient for the year variable suggest? The coefficient for year is 0.750773, which suggests a positive relationship between `year` and `mpg`, i.e. more recent vehicles correlate with higher fuel consumption per mile driven. A unit increase in `year` where other predictors is kept constant causes an increase of 0.750773 in `mpg`.

Part (d)