Exercise 2

TMA4268 Statistical Learning V2023

Henrik Olaussen

i dag

#Problem 2

MSE: Mean Square Error. $MSE = 1/n \sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$, $\hat{f}(x_i)$ is the prediction given from the estimated $\hat{f}$ at the i´th value. This MSE uses the training data. Can think of it as training MSE. We do not care too much about the training MSE. We are interested in the accuracy of the prediction on some unseen test-data. Want our method to accuratly predict the future, not fit well with the past. We want the lowest test-MSE not lowest training-MSE. Moreover, there are no guarantee that the lowest training MSE will give the lowest test MSE. For example, a too flexible model will often fit the predictions badly, as the model is too closely related the training data, and hence fit predictions badly, as test data will differ from the training data, we need a trade-off.

Furthermore, a small variance (high bias) means that we have under-fitted the data. High bias means that the fitted model captures the true relationship badly, e.g. a straight line that is supposed to fit a logarithmic function. If the fitted line is very flexible and fits the training data points perfectly, we have low bias. This might cause overfitting in the test data. When we look at how the model fits the test data, the difference in fit is measured through variance. The model that does not capture the true relationship tend to get a smaller variance (sums of squares measured between the fit of the training data and the test data) to the test data than the flexible model (with high bias). Hence, the variance is smaller in the high bias case, and larger in the small bias case.

#Problem 3

dimentions of the data: 392 observations, and 9 covariates. Qualitative predictors: cylinders, origin and name. Quantitative: mpg, dipl, horsepower, weight, acc, year.

```
library(ISLR)
library(GGally)
#Auto
#?Auto

#the range of the predictors
range(Auto$mpg)
```

```
## [1]  9.0 46.6
```

```
range(Auto$origin)
```

```
## [1] 1 3
```

```r
#mean
mean(Auto$mpg)
```

```
## [1] 23.44592
```

```r
#standard deviation
sd(Auto$mpg)
```

```
## [1] 7.805007
```

```r
#reduced dataset
redAuto <- Auto[-c(10:85),] #remove data 10->85

ggpairs(redAuto[, c(1:7)]) + theme_minimal()
```

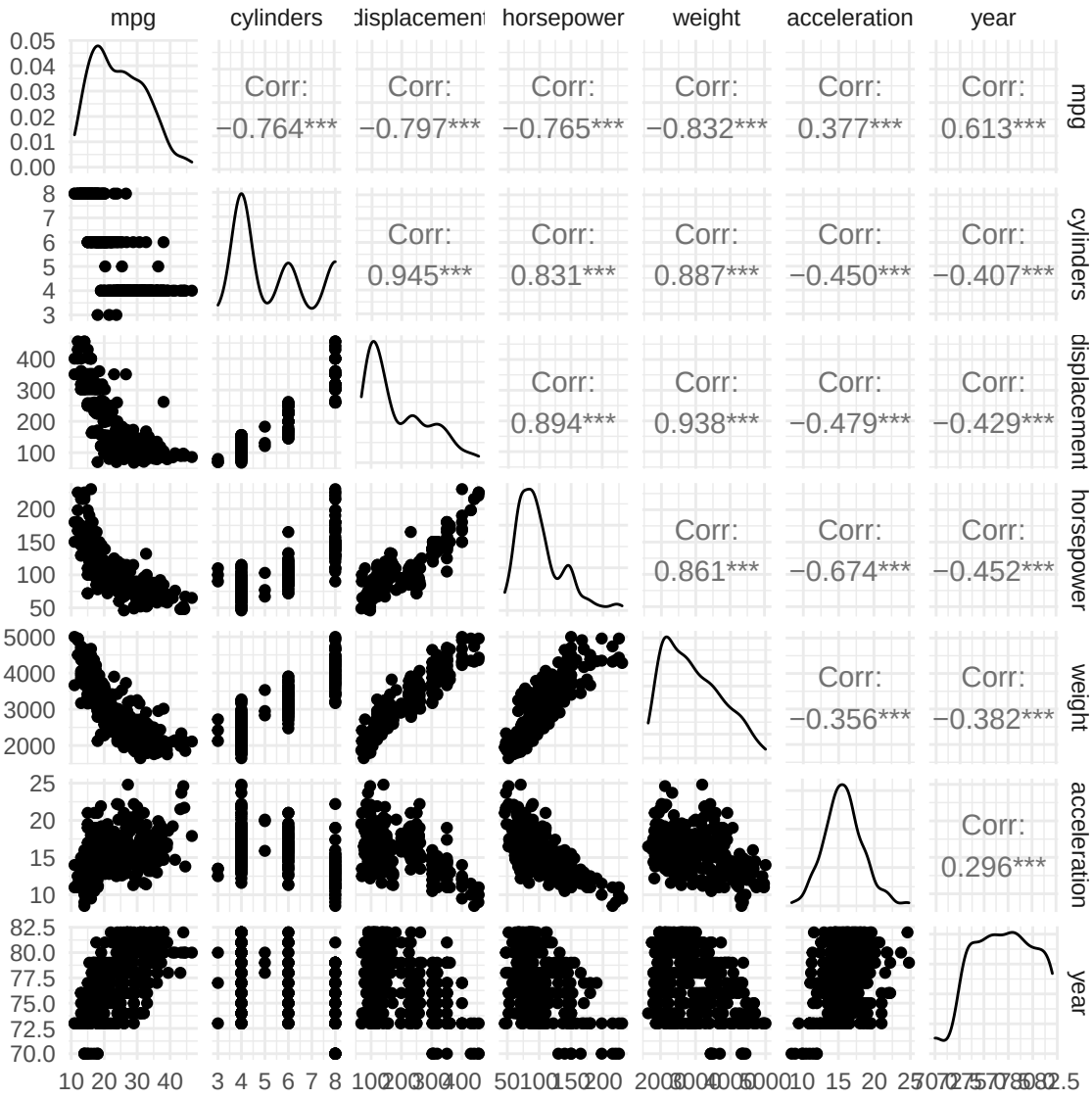

Figure 1: Pairs plot of the academic salary data set.