Statistical Computing, Exercise Sheet 4

Ulrike Grömping

16 November 2021

Contents

1 The data for the description tasks

The R package **sm** holds a data frame `geys3d` with data on the eruption behavior of the Old Faithful geyser: these are durations in minutes of a consecutive sequence of eruptions, waiting times until the *respective* eruption (`Waiting`) and until the *next* eruption (`Next.waiting`), also in minutes. (How are `Waiting` and `Next.waiting` related?)

```
require(sm)
```

```
## Loading required package: sm
```

```
## Package 'sm', version 2.2-5.6: type help(sm) for summary information
```

```
summary(geys3d);head(geys3d); tail(geys3d)
```

```
##     Waiting        Next.waiting       Duration
##  Min.   : 43.00   Min.   : 43.00   Min.   :0.830
##  1st Qu.: 59.00   1st Qu.: 59.00   1st Qu.:2.000
##  Median : 76.00   Median : 76.00   Median :4.000
##  Mean   : 72.29   Mean   : 72.29   Mean   :3.465
##  3rd Qu.: 83.00   3rd Qu.: 83.00   3rd Qu.:4.380
##  Max.   :108.00   Max.   :108.00   Max.   :5.450
```

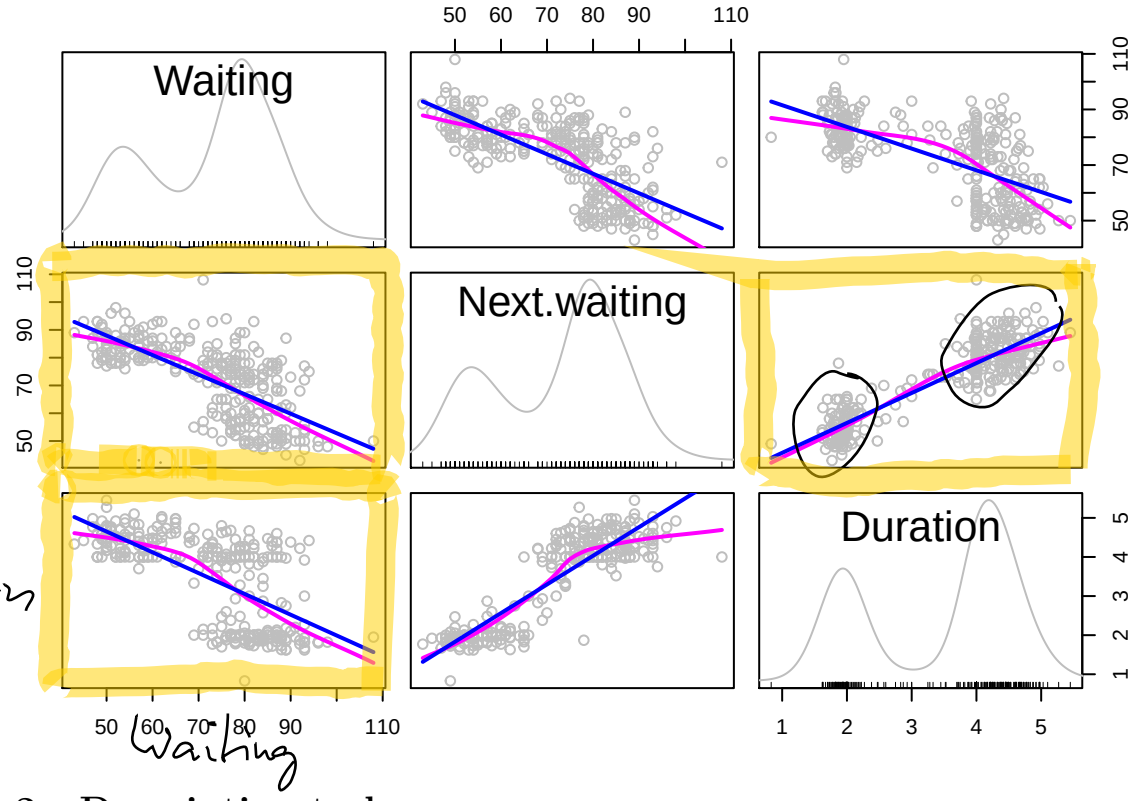
```
##   Waiting Next.waiting Duration
## 1      80           71     4.02
## 2      71           57     2.15
## 3      57           80     4.00
## 4      80           75     4.00
## 5      75           77     4.00
## 6      77           60     2.00
```

```
##     Waiting Next.waiting Duration
## 293      54           87     4.42
## 294      87           52     2.13
## 295      52           85     4.08
## 296      85           58     2.07
## 297      58           88     4.00
```

```
## 298       88            79      4.00
```

As a first shot for a "big picture", the function `scatterplotMatrix` from package **car** displays an advanced version of what the `plot` function would (see code chunk below). The blue line is the least squares line, the magenta line is the loess line based on local linear fits (i.e., nonparametric).

```
## default has everything in the same color
car::scatterplotMatrix(geys3d, col="grey",
                   regLine=list(method=lm, col="blue"),
                   smooth=list(smoother=loessLine, spread=FALSE,
                       lty.smooth=1, lwd.smooth=2,
                       col.smooth="magenta"))
```



Duration

Waiting

2 Description tasks

Please inspect the relation among all three pairs of variables of the `geys3d` data. Remember that X should be the predictor variable and Y the response, if two variables are not considered on equal footing. If there is a calendar time difference between two variables, then X should be chosen to precede Y.

(i) Describe in words, what you observe in the scatter plot matrix.
(ii) Calculate the matrix of Bravais-Pearson correlations. cor(geys3d)
(iii) Calculate the least squares line that describes the relation between the eruption duration (X) and the next waiting time (Y). Obtain it in three different ways:

- using function `lm`,
- manually in R, by using functions `mean`, `cov` and/or `var`.
- manually in R, by using only very basic functions like `mean` and `sum`.

(iv) Interpret the line obtained in (iii).
(v) What is the R^2 value for the least squares line obtained in (iii)? Obtain it in two different ways:

- from the `summary` applied to the linear model object,
- from the correlation matrix.

(vii) An eruption has lasted for four minutes. Based on the least squares line, predict the next waiting time.
(viii) Remember that the equation for the least squares line was obtained for both times measured in minutes. Can you modify the equation for the case that

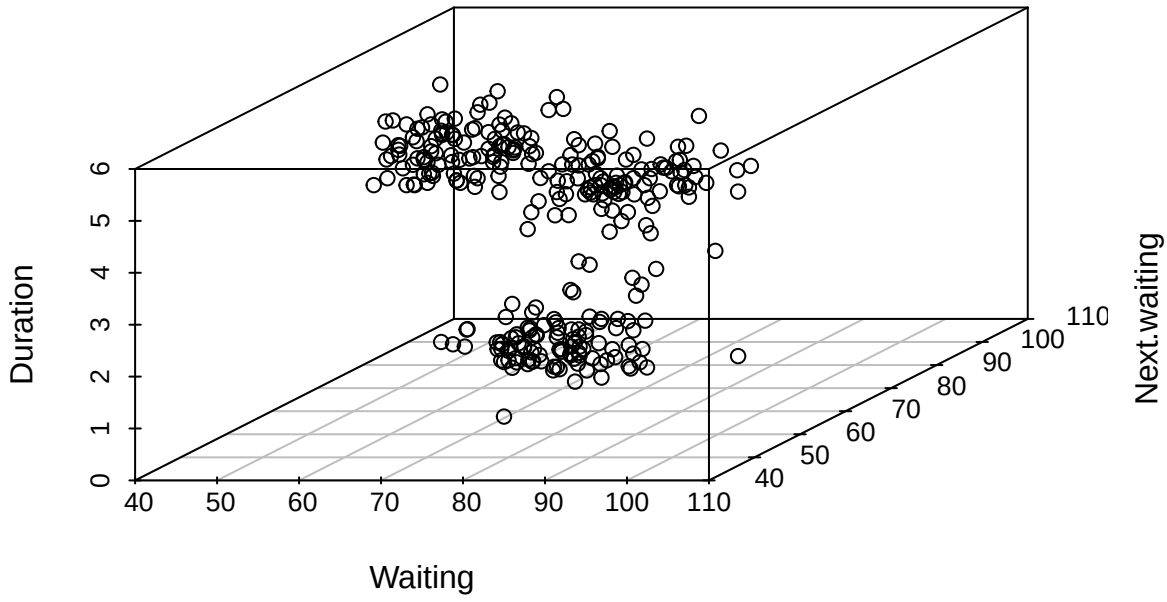
- both times are in seconds instead of minutes?
- next waiting time remains in minutes, but duration becomes seconds?

Note: these data are interesting in 3D, see the 3D scatter plot below (static in pdf, interactive in HTML).

```
if (params$html) require(rgl) else require(scatterplot3d)
```

```
## Loading required package: scatterplot3d
```

```
if (params$html) {
  plot3d(geys3d)
  rglwidget()} else scatterplot3d(geys3d)
```



3 Calculation tasks

These tasks are meant to be done manually, using a pocket calculator, where needed.

Consider the following small data set:

```
x:  8   3   -2  9   -3
y:  -7  -4  -2  -6  -1
```

(i) Calculate the means and variances of both variables.
(ii) Calculate the covariance, the Bravais-Pearson correlation and the coefficient of determination of the least squares line as a model for the data.
(iii) Calculate the least squares line.
(iv) Draw a scatterplot, and add the line to it.
(v) There is another correlation coefficient called the "Spearman correlation", which is obtained by calculating separate ranks of the x and y values and calculating the Bravais-Pearson correlation of those ranks. Spearman for this example: -0.9, obtained as Bravais-Pearson correlation of the ranks below:
rank Rx: 4 3 2 5 1
rank Ry: 1 3 4 2 5
The R command for the Spearman correlation is `cor(x, y, method="spearman")`. Think about what this coefficient measures. When does it become -1 or 1?