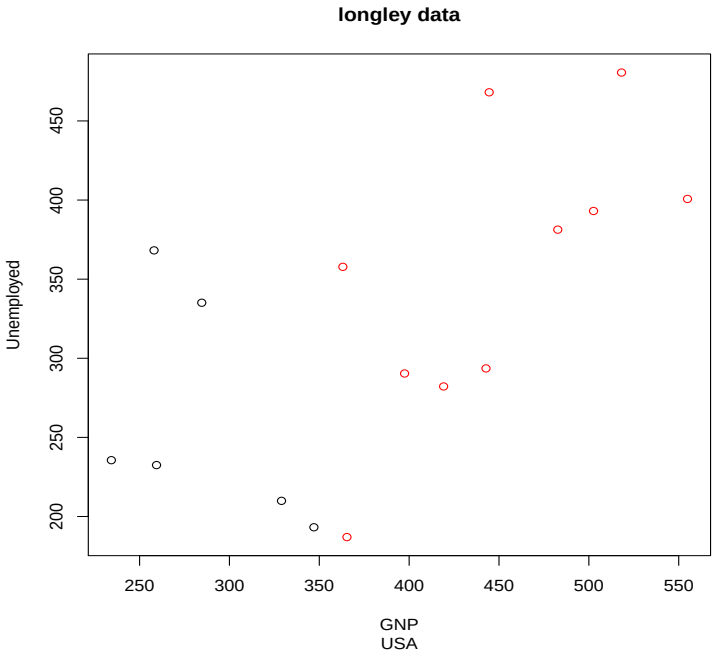
Exercise Series 2

1. Labelling of points

Let's visualize Longley's macroeconomic data set. This data set is part of a standard R package and can be loaded with the command `data(longley)`. Use `?longley` after loading the data to get more information about the data set.



a) Plot the number of unemployed (variable `Unemployed`) against GNP (variable `GNP`).

Hint:

- Remember that you can access a data frame's column values by using either `[,"name"]` or the `$name` trick: `longley[,"GNP"]` and `longley$GNP` both return the vector of GNP values.
- Use the arguments `main`, `xlab` and `ylab` of the `plot` function to modify the title and the axes labels of the graphic.

b) Color the points in red if their corresponding population (variable `Population`) is greater than 115.

Hint: To define the color of the points, first create a new categorical variable based on the evaluation of the population: `pop <- factor(longley$Population > 115)`. Now you can use this vector for the `col` argument of the `plot()` function.

Challenge: try to find an alternative way of defining the vector `pop` by using the function `ifelse()`.

c) Label each point with its year (variable `Year`).

Hint: Use the low-level plotting function `text()` in order to add labels to points (e.g. year values). Try to use the optional argument `pos` of the function `text()`.

2. Graphics - demo

Run the following exisiting examples to obtain a display of simple (1-dim, x-y), image-like (2-dim) and perspective capabilities with R. Try to follow the commands and find out what they do.

Hint: Press *Enter* in the R-window to step forward within the example.

a) `demo(graphics)`

b) `demo(image)`

c) `demo(persp)`

d) `library(scatterplot3d); example(scatterplot3d)`

e) `example(points)`

3. Test of paired samples: bituminous binder

Samples of bituminous binders, which are used in street construction, are examined in a lab. They are slowly cooled down and bended regularly. Cracks appear after reaching a material-dependent temperature. The temperature in the lab is then recorded as "break point after Fraass" for the binder under investigation. This examination was previously carried out manually, but it is now common use a machine. We now want to compare the two procedures.

The data set `bitumen.sav` contains the break points for six different binders, measured manually (`MAN`) or automatically (`AUT`), respectively.

The data set can be loaded via

```
t.url <- "http://stat.ethz.ch/Teaching/Datasets/NDK/bitumen.dat"
bitumen <- read.table(t.url, header = TRUE)
```

Calculate the differences between the break points of the two methods and store them in the variable `diff`. Then take a look at the data using a boxplot of `diff` and make a Q-Q plot of the variable `diff`. Test on the 5%-level whether the two methods `AUT` and `MAN` differ significantly with a t-Test and a Wilcoxon Test if their respective assumptions are met.

Hint: We first visualize the data in order to check the assumptions needed for the different tests. Generally, one decides which test to use according to the distribution of the data, but here we just apply all tests. We use a paired test because each binder has been measured twice: once manually and once automatically.

(Source: Data from O. Neubauer, EMPA, Dübendorf)

4. Chi-squared test for independence: environmental protection

A survey about environmental pollution inquired (amongst other things) whether one felt affected by pollutants (`Beeintr`). The possible answers were

- (1): "not at all affected"
- (2): "slightly affected"
- (3): "quite affected"
- (4): "very affected".

A potential question of investigation is whether the given answer was related to the level of education (`Schule`). The corresponding results can be found in the data set `umwelt` which can be loaded via:

```
url <- "http://stat.ethz.ch/Teaching/Datasets/WBL/umwelt.dat"
umwelt <- read.table(url, header = TRUE)
```

a) Display the data in a cross table.

b) Is the answer to the question if one felt affected by pollutants independent from the level of education?

Hint: Use the function `chisq.test(...,...)`.

c) Compare the expected counts with the observed counts using the residuals. Visualize the residuals using a mosaic plot.

Hint: Extract the expected counts from the fit

```
fit <- chisq.test(..., ...)
fit$...
```

To visualize the residuals:

```
require(vcd)
mosaic(~ ... + ..., shade = TRUE, data = ...)
```

5. Forbes' data on boiling points in the Alps

This data set contains 17 observations of boiling points of water and barometric pressures. Forbes wanted to estimate the air pressure from the boiling point in order to estimate the altitude.

The Forbes' data is included in the package `MASS`. The data frame `forbes` contains the variables `bp` (boiling point, in degrees Fahrenheit) and `pres` (barometric pressure, in inches of mercury). Use the following code to load the package `MASS` and copy the data frame `forbes` to a new data frame named `d.forbes`:

```
require(MASS)
d.forbes <- forbes
```

We will continue working with the data frame `d.forbes`.

a) Add the variable `logpres` $= \log_{10}($`pres`$)$ to the data frame `d.forbes`.

Hint: Use the base 10 logarithm function `log10()`. One way to add a new column to a data frame is

```
dataframe$name_of_new_variable <- newvariable
```

b) Plot the new variable `logpres` against `bp`. What do you notice?

c) Add a regression line to the scatter plot of `logpres` against `bp`. What do you observe?

Hint: Use the command

```
fit <- lm(... ~ ..., data = forbes)
abline(fit)
```

d) Fit the linear regression model

$$\texttt{logpres}_i = \beta_0 + \beta_1 \cdot \texttt{bp}_i + E_i,$$

where the random errors E_i are assumed to be i.i.d. normally-distributed with mean 0 and constant variance σ^2. Are the assumptions of the model fulfilled?

e) Test the null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_A : \beta_1 \neq 0$ at the 5%-level, where β_1 is the slope of the regression line.

Hint: The function `summary()` prints a nice summary of the regression output.

f) Exclude the 12^{th} observation from the data frame. Refit the model given in d). Compare the estimates $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\text{se}^{(\hat{\beta}_0)}$, $\text{se}^{(\hat{\beta}_1)}$ and $\widehat{\sigma}^2$ with the ones from part b). Do you have an explanation for the difference in the results?

Source: S. Weisberg, Applied Linear Regression, Wiley (1985), p. 3

6. Anscombe Data

This exercise is about data constructed by F.J. Anscombe. It contains four Y- and four X-Variables. We want study the four models

$$Y_i^{(k)} = \alpha + \beta \cdot X_i^{(k)} + E_i \quad \text{for } k = 1, ..., 4.$$

Read in the data using following command:

```
t.url <- "http://stat.ethz.ch/Teaching/Datasets/NDK/anscombe.dat"
d.anscombe <- read.table(t.url, header = TRUE)
```

a) Calculate the coefficients and corresponding standard errors for all four models. Also note the $\widehat{\sigma}^2$ and R^2 values. Create a table of these values.

Hint: You can store the output of `summary()` in a variable. Use `str()` on this new variable to learn how to access the desired values.

b) For all four models show the regression line in a scatter plot.

c) Discuss the results of parts a) and b).

Source: F.J. Anscombe, Graphs in statistical analysis, American Statistician 27, 17-21 (1973)