Statistical Geophysics
Exercises

Nikolaus Umlauf

2014/15

1. **Exercise**

   The file `Messina.txt` contains the historical sequence of earthquake events in a region around the Strait of Messina, a narrow passage between the eastern tip of Sicily and the southern tip of Calabria in the south of Italy. This region is highly seismic. The earthquake list provided in the file is limited to earthquakes of magnitude 4.5 or above, during the period 1700-1980. The time of occurrence is given in fractional year.

   (a) Read the data into R and prepare a `data.frame` with two columns, one for variable `time` and one for `magnitude`.

   (b) Make an analysis of the interarrival times (use function `diff()`) of the earthquakes for all events irrespective of magnitude and only for events with magnitude greater than 5 (you may use the function `subset()` to extract a subset of the data in which the magnitude is greater than 5).

   (c) For the subset and the whole data, carry out the following steps:
   - Define a variable for the interarrival times (the vector of differences of the variable `year`).
   - Use the function `hist()` to plot a histogram of the interarrival times.
   - Calculate $\hat{\lambda} = \frac{1}{\bar{x}}$.
   - Make use of the functions `dexp()` and `lines()` to superimpose the exponential density with parameter $\hat{\lambda}$ on the histogram.

   (d) Save your code in an R script that runs smoothly using function `source()`.

2. **Exercise**

The following table gives the amount of rain (in litres per square metre), measured at the volcano Merapi (Indonesia) between January 1st and January 20th, 1995.

rain	date	rain	date
2	1995/01/01	50	1995/01/11
9	1995/01/02	12	1995/01/12
18	1995/01/03	0	1995/01/13
2	1995/01/04	0	1995/01/14
23	1995/01/05	0	1995/01/15
42	1995/01/06	0	1995/01/16
11	1995/01/07	3	1995/01/17
13	1995/01/08	3	1995/01/18
40	1995/01/09	40	1995/01/19
12	1995/01/10	48	1995/01/20

(a) Determine the type of scale of the variable `rain`.

(b) Draw a histogram using the intervals [0,10),[10,20),[20,30),[30,40),[40,50].

(c) Read the data into R and do the histogram again with R.

(d) Now plot the empirical cumulative distribution function (with R) and determine graphically how large the percentage of days is on which it rained more than 35 litres per square meter?

(e) Calculate (by hand!) the values for the following measures of location and dispersion for the variable `rain`: mode, median, arithmetic mean, lower quartile, upper quartile, variance, standard deviation, and coefficient of variation.

(f) Use the results obtained in (e) to draw a boxplot of the empirical distribution of `rain`. Do the boxplot again in R.

(g) Calculate a 95% confidence interval for the expected value of rainfall, μ.

3. **Exercise**

Given the following random sample

$x_1 = 1.16$, $x_2 = 0.21$, $x_3 = 0.1$, $x_4 = 0.96$, $x_5 = 0.08$, $x_6 = 0.67$

of a gamma distributed random variable $X \sim Ga(\alpha, \beta)$. The density function of X is given by

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} \exp(-\beta x),$$

with $x > 0$, $\alpha > 0$ and $\beta > 0$. Derive the maximum likelihood estimator for β if $\alpha = 1.71$.

4. **Exercise**

The density function of a normally distributed random variable X is given by

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{\sigma^2 2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Derive the maximum likelihood estimator for μ.

5. **Exercise**

   Assume that X is a continuous random variable that is uniformly distributed on the interval $[a, b]$, denoted by $X \sim \mathcal{U}(a, b)$. The pdf of the continuous uniform distribution is

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} .$$

   (a) Find the expectation and variance of the continuous uniform distribution on $[a, b]$.

   (b) Use the software R to generate uniformly distributed random numbers on the interval $[4, 8]$ for different sample sizes ($n_1 = 100, n_2 = 1000, n_3 = 10000$). Calculate their mean and variance and compare the results to the results obtained in (a).

6. **Exercise**

   In this exercise we will investigate again some rain data collected at the volcano Merapi (Indonesia). The file `rain.dat` contains data on the amount of rain (in litres per square metre), measured at two stations (Jrakah and Kaliurang). Suppose the days on which the measurements were taken are different for both stations. The random variables

$$Y_A = \text{“Rain at station Jrakah”} \quad \text{and} \quad Y_B = \text{“Rain at station Kaliurang”}$$

   are independent and follow a normal distribution with equal variances:

$$Y_A \sim \mathcal{N}(\mu_A, \sigma^2), \quad Y_B \sim \mathcal{N}(\mu_B, \sigma^2).$$

   (a) We would like to test the research hypothesis that the amount of rain at station Jrakah is different from the amount of rain at Kaliurang ($\alpha = 0.05$). Formulate the null and alternative hypotheses for this test problem.

   (b) Calculate the test statistic and make a decision whether or not the null hypothesis can be rejected. Do the following:

      i. Read the data into R and determine means and empirical variances.

      ii. Calculate the pooled empirical variance and calculate the test statistic.

      iii. To which value do you have to compare the test statistic in order to get a decision?

      iv. Is there a faster way in R to carry out a Student's t-Test?

   (c) Carry out a test for the hypothesis that there is less rain at station Jrakah than at station Kaliurang ($\alpha = 0.05$).

   (d) Suppose the 20 days on which the measurements were taken are the same for both stations, that is, the samples are dependent. Carry out an appropriate test to check the assumption that there is a difference in the amount of rain at the two stations ($\alpha = 0.05$).

7. **Exercise**

In some applications of simple linear regression a model without an intercept is required (when the data are such that the line must go through the origin), that is, a model of the form

$$y_i = \beta_1 x_i + \epsilon_i \quad (i = 1, \ldots, n).$$

(a) Derive the least squares estimator for β_1.

(b) In the R package **gamair**, `data('hubble')` contains data from the Hubble space telescope on distances and velocities of 24 galaxies (see also `?hubble`). Fit the following simple linear regression model without intercept to the data:

$$\texttt{velocity} = \beta_1 \texttt{distance} + \epsilon. \tag{1}$$

This is essentially what astronomers call Hubble's Law and β_1 is known as Hubble's constant. We can use the estimated value of β_1 to find an approximate value for the age of the universe.

(c) Fit a quadratic regression model, i.e., a model of the form

$$\texttt{velocity} = \beta_1 \texttt{distance} + \beta_2 \texttt{distance}^2 + \epsilon \tag{2}$$

to the `hubble` data. Which model, (1) or (2), would you choose on the basis of the Akaike Information criterion (AIC) and Bayesian Information criterion (BIC)? The AIC and BIC are implemented in the R functions `AIC()` and `BIC()`, respectively.

8. **Exercise**

The file `consumption.txt` contains the weight in pounds (1st column) and the petrol consumption in miles per gallon (2nd column) of 32 vehicle variants.

(a) Read the data `consumption.txt` into R.

(b) Estimate with R the influence of the weight variable on petrol consumption and interpret the parameter estimates.

(c) Why does it make more sense to specify the consumption in gallons per mile than in miles per gallon? Fit the model and check if the transformation leads to a better fit. Try to get a graphical idea about the relationship of the variables.

(d) How do the parameter estimates and p-values change if one specifies the weight in kilogram and the consumption in litres per 100 km (1 kg = 2.2046 American pounds, 1 km = 0.6214 miles, 1 litre = 0.22 gallons)? Derive the general conversion formula for the LS estimates after the following linear transformations have been applied:

$$\begin{aligned} x_i &\rightarrow t_i = a_0 + a_1 x_i, \quad \text{with } a_1 \neq 0 \ , \\ y_i &\rightarrow u_i = b_0 + b_1 y_i, \quad \text{with } b_1 \neq 0 \ , \end{aligned}$$

and then calculate the new parameter estimates.

9. **Exercise**

Weather modification, or cloud seeding, is the treatment of individual clouds or storm systems with various inorganic and organic materials in the hope of achieving an increase in rainfall. In the R package **HSAUR2**, `data('clouds')` contains data collected in the summer of 1975 from an experiment, which was conducted in the area of Florida, investigating the use of massive amounts of silver iodide in cloud seeding to increase rainfall (see also `?clouds`). Use the R function `step()` to perform stepwise model selection by AIC. As the initial model for the stepwise search use

```
R> f <- rainfall ~ seeding * (sne + cloudcover + prewetness +
  echomotion) + time
R> b <- lm(f, data = clouds)
```

10. **Exercise**

(a) Setup an R script and name it `ZambiaNutrition.R`.

(b) Download the data on malnutrition in Zambia (e.g. using function `download.file()`) from:

`http://www.stat.uni-muenchen.de/~bayesx/tutorials/zambia.raw`

(c) Now, generate a `data.frame` with the covariates and names as specified in the following table:

Variable	Description
`stunting`	Standardized Z-score for stunting.
`mbmi`	Body mass index of the mother.
`agechild`	Age of the child in months.
`district`	District where the mother lives.
`memployment`	Mother's employment status with categories 'yes' and 'no'.
`meducation`	Mother's educational status with categories for no education or incomplete primary 'no', complete primary but incomplete secondary 'primary' and complete secondary or higher 'secondary'.
`urban`	Locality of the domicile with categories 'yes' and 'no'.
`gender`	Gender of the child with categories 'male' and 'female'.

Make sure that the labeling of the `factor` variables is correct, too.

(d) Save the data in `.rda` format (function `save()`) and name it `ZambiaNutrition.rda`.

(e) Carry out a first descriptive analysis on this data set, therefore examine the influence of all covariates on the response `stunting`. To improve visualization of the relationship of the response and the continuous covariates, split the range of the covariates in equidistant intervals (function `seq()`) and draw `boxplot`s of the observations of each interval (function `cut()` returns a `factor` for which function `boxplot()` can easily create the desired plot).

(f) From your findings in (e), estimate a linear regression model excluding the spatial information of variable `district` (function `lm()`) and interpret your results (just add the interpretation in your script as a little comment `##`).

(g) Make sure the script runs by calling `source('ZambiaNutrition.R')` and include a check if the data is already downloaded, so you don't need to do point (b) every time you run the script.