LINKÖPING UNIVERSITY

Computational Statistics 732A89 – Spring 2025
Computer Lab 5

Martin Andrae, Bayu Brahmantio, Frank Miller
Department of Computer and Information Science (IDA), Linköpings University

February 18, 2025

This computer laboratory is part of the examination for the Computational Statistics course. Create a group report (which is directly presentable, if you are a presenting group), on the solutions to the lab as a PDF file. Be concise and do not include unnecessary printouts and figures produced by the software and not required in the assignments.
All R code should be included as an appendix to your report.
A typical lab report should contain 2-4 pages of text plus some figures plus an appendix with codes. In the report, refer to all consulted sources and disclose all collaborations.
The report should be handed in via LISAM (or alternatively in case of problems by email) by **23:59 February 25, 2025** at the latest. Notice that there is a deadline for corrections 23:59 08 April 2025 and a final deadline of 23:59 29 April 2025 after which no submissions or corrections will be considered, and you will have to redo the missing labs next year. The seminar for this lab will take place **March 11, 2025**.
The report has to be written in English.

Question 1: Bootstrap for regression

The dataset `kresseertrag.dat` has three columns: the observation number $(1, \ldots, 81)$, the concentration of fertilizer used in %, the yield of cress in *mg*. The data should be analysed with a regression model, assuming that the errors ε are independent with equal variance. In the lecture, the straight-line regression was considered; but here, you should use a potentially better model.

a. Read in the dataset and fit a cubic regression model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$, where $x =$ concentration and $y =$ yield, using the `R`-function `lm`.

b. You might improve the model in a. by removing or adding (a) model term(s). For the chosen model, estimate the coefficients in the model together with their 95%-confidence intervals using the `R`-function `lm`. Create a plot for yield vs. concentration and add the estimated regression curve to the plot.

c. Derive a 95%-bootstrap confidence interval **for one of** the model parameters β_i based on the percentile method. Do not use a bootstrap package for this calculation; program the bootstrap on your own. Use at least 10000 bootstrap replicates. Plot a histogram with the bootstrap distribution for the parameter.

d. Derive now 95%-confidence intervals for the chosen parameter using the package `boot` with percentile and BCa method.

e. Compare the confidence intervals from b., c., and d. and comment on it. Do you judge that the intervals are similar or do you see relevant differences? What is your overall conclusion from these confidence interval results about that model parameter?

Question 2: Simulation of power curves

The Gumbel distribution (with scale parameter set to 1 and location parameter $\mu + c$ where $c = \log(\log(2)) \approx -0.36651$) has the distribution function

$$F(x) = \exp(-\exp(-(x - \mu - c))), \quad c = \log(\log(2)).$$

The median of a random variable with this distribution is μ.

Assume that $n = 13$ independent observations are made and they follow a Gumbel distribution with location $\mu + c$. To test the null hypothesis $H_0 : \mu = 0$ (median is 0) versus $H_a : \mu > 0$, one can use the Sign test. We want to investigate the power (=probability to reject H_0) of this test in this situation using a simulation study.

a. Generate Gumbel random variables using the Inverse Transformation Method.

b. Simulate the power of the Sign test for $\mu \in [0, 2]$ using an appropriate grid for μ-values on $[0, 2]$ and an appropriate number of repetitions for each μ (under consideration of the simulation's precision; provide some reasoning for the chosen number of repetitions). Plot the power curve (i.e., power versus μ). For the Sign test, you might use, e.g., the function `SIGN.test` in the package `BSDA`; or alternatively, the function `binom.test` might be used when applied to the number of positive observations.