Stockholms universitet

Computational Statistics
Advanced Course, 7.5 HEC, Spring 2022
Home Assignment 3 until March 18

Mahmood Ul Hassan and Frank Miller, Department of Statistics

March 11, 2022

You are supposed to solve these problems individually without any cooperation. Record your answers clearly. If you cite material from other sources, or use intellectual ideas or code from others, point out this clearly with stating the source. Please include your R computer code and relevant output in your solution. Send an email to `mahmood.ul-hassan@stat.su.se` with your solutions until the above deadline with a file name like `CompStat2022HA3March_[Your name].pdf` and extra file(s) with similar name with your computer code.

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*. You are supposed to fit a quadratic regression, $y = \beta_0 + \beta_1 x + \beta_2 x^2$, where x = concentration and y = yield.

a. Read in the dataset, fit a quadratic regression and estimate the three coefficients of the regression 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.

b. Derive a 95%-bootstrap confidence interval for the three model parameters based on the percentile method (which was used in the lecture). 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 β_2.

c. Compare the confidence intervals for β_2 from a. and b. and comment on it. Which of the confidence intervals do you recommend for β_2 and why?

d. Interpret the chosen confidence interval in c. What does this mean for the model choice when analysing the yield-data?

2 Simulation of power curves

Assume that $n = 21$ independent observations are made and they follow an $N(\mu, 1)$ distribution. To test the null hypothesis $H_0 : \mu = 0$ versus $H_a : \mu > 0$, the standard test is the one-sample t-test, but one can also use the sign test if one is uncertain about the distribution of the data. We want to compare the power of these two tests using a simulation study. We use a significance level $\alpha = 0.10$.

a. Simulate the power of the t-test and the sign test for $\mu = 0, 0.1, 0.2, \ldots, 1$ using an appropriate number of repetitions for each μ (under consideration of the simulation's precision). For the sign test, you might use e.g. the function `SIGN.test` in the package `BSDA`; or alternatively, `binom.test` might be used when applied to the number of positive observations.

b. Plot the power curves (power versus μ) for both tests using your results you have generated in a.

3 Sampling algorithms

Consider the following density with a triangle-shape (another triangle distribution than considered in Lecture 5):

$$f(x) = \begin{cases} 0 & \text{if } x < -1 \text{ or } x > 1, \\ x + 1 & \text{if } -1 \leq x \leq 0, \\ 1 - x & \text{if } 0 < x \leq 1. \end{cases}$$

We are interested in generating draws of a random variable X with this density.

a. Derive manually the distribution function F and the inverse F^{-1} of it and write down also steps of your computation. Program a random generator for X using the Inverse Transformation Method.

b. Choose an appropriate and simple envelope $e(x)$ for the density and program a random generator for X using Rejection Sampling.

c. In Lecture 5, another triangle distribution was generated using the Inverse Transformation Method, see page 8-9 of the lecture notes. Let Y be a random variable following this distribution. A random variable $-Y$ has a triangle distribution on the interval $[-1, 0]$. Program a random generator for X using Composition Sampling based on Y and $-Y$.

d. Sums or differences of two independent uniformly distributed variables can also have some triangle distribution. Choose two appropriate uniform distributions and program a generator for X based on a linear combination of them.

e. Check your random generators in each of a. to d. by generating 10000 random variables and plotting a histogram. Which of the four methods do you prefer if you had to generate samples of X?

4 Markov chain Monte Carlo integration

We want to generate multivariate random vectors $X = (X_1, X_2)$ which have the density

$$h(x_1, x_2) = c \cdot \exp\{-(x_1 - 1)^2 - (x_2 + 2)^2 + 2(x_1 - 1)(x_2 + 2) - |x_1| - |x_2|\}$$

for some constant $c > 0$. A sample should be produced with Markov chain Monte Carlo and the Metropolis algorithm.

a. Write an own program using a Metropolis algorithm to generate draws following h.

b. Suggest two different proposal distributions. Use these proposal distributions and generate 10000 observations. Generate histograms for X_1 and X_2 (the marginal distributions).

c. Choose one of the two proposal distributions in b. and argue why you prefer it. Based on the generated sample of 10000 observations, determine an estimate for $EX = E\,(X_1, X_2)^\top$ and for $P(X_1^2 + X_2^2 \leq 1)$.