Computer Lab 5
Computational Statistics

Linköpings Universitet, IDA, Statistik

2023 XII 01

Kurskod och namn:	732A90 Computational Statistics
Datum:	2023 XI 28—2023 XII 5 (lab session 1 XII 2023 SU24/25)
Delmomentsansvarig:	Krzysztof Bartoszek, Bayu Brahmantio, Héctor Rodriguez Déniz
Instruktioner:	This computer laboratory is part of the examination for the Computational Statistics course Create a group report, (that 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 into your report.** A typical lab report should 2-4 pages of text plus some amount of figures plus appendix with codes. In the report reference **ALL** consulted sources and disclose **ALL** collaborations. **Copying solutions from others, the Internet and other sources is NOT allowed.** The report should be handed in via LISAM (or alternatively in case of problems by e–mail) by **23:59 5 December 2023** at latest. Notice there is a deadline for corrections **23:59 31 January 2024** and a final deadline of **23:59 29 February 2024** after which no submissions nor corrections will be considered and you will have to redo the missing labs next year. The seminar for this lab will take place **12 December 2023, 10:15, KY27**. The report has to be written in English.

Question 1: Hypothesis testing

In 1970, the US Congress instituted a random selection process for the military draft. All 366 possible birth dates were placed in plastic capsules in a rotating drum and were selected one by one. The first date drawn from the drum received draft number one, the second date drawn received draft number two, etc. Then, eligible men were drafted in the order given by the draft number of their birth date. In a truly random lottery there should be no relationship between the date and the draft number. Your task is to investigate whether there can be doubts concerning the randomness of the selection of the draft numbers. The draft numbers (`Y=Draft_No`) sorted by day of year (`X=Day_of_year`) are given in the file `lottery.xls`. The data was originally published by the U.S. Government, and most conveniently made available online at `http://jse.amstat.org/jse_data_archive.htm` (see also Starr Norton (1997) Nonrandom Risk: The 1970 Draft Lottery, Journal of Statistics Education, 5:2, DOI: 10.1080/10691898.1997.11910534)
.

1. Create a scatterplot of Y versus X, are any patterns visible?

2. Fit a curve to the data. First fit an ordinary linear model and then fit and then one using `loess()`. Do these curves suggest that the lottery is random? Explore how the resulting estimated curves are encoded and whether it is possible to identify which parameters are responsible for non–randomness.

3. In order to check if the lottery is random, one can use various statistics. One such possibility is based on the expected responses. The fitted loess smoother provides an estimate $\hat{Y}$ as a function of X. It the lottery was random, we would expect $\hat{Y}$ to be a flat line, equalling the empirical mean of the observed responses, $\overline{Y}$. The statistic we will consider will be

$$S = \sum_{i=1}^{n} |\hat{Y}_i - \overline{Y}|.$$

   If S is not close to zero, then this indicates some trend in the data, and throws suspicion on the randomness of the lottery. Estimate S's distribution through a non–parametric bootstrap, taking $B = 2000$ bootstrap samples. Decide if the lottery looks random, what is the p–value of the observed value of S.

4. We will now want to investigate the power of our considered test. First based on the test statistic S, implement a function that tests the hypothesis
   H_0: Lottery is random
   versus
   H_1: Lottery is non–random.
   The function should return the value of S and its p–value, based on 2000 bootstrap samples.

5. Now we will try to make a rough estimate of the power of the test constructed in Step 4 by generating more and more biased samples:

   (a) Create a dataset of the same dimensions as the original data. Choose k, out of the 366, dates and assign them the end numbers of the lottery (i.e., they are not legible

for the draw). The remaining $366 - k$ dates should have random numbers assigned (from the set $\{1, \ldots, 366 - k\}$). The k dates should be chosen in two ways:

i. k consecutive dates,

ii. as blocks (randomly scattered) of $\lfloor k/3 \rfloor$ consecutive dates (this is of course for $k \geq 3$, and if k is not divisible by 3, then some blocks can be of length $\lfloor k/3 \rfloor + 1$).

(b) For each of the Plug the two new not–completely–random datasets from item 5a into the bootstrap test with $B = 2000$ and note whether it was rejected.

(c) Repeat Steps 5a–5b for $k = 1, \ldots,$ until you have observed a couple of rejections.

How good is your test statistic at rejecting the null hypothesis of a random lottery?

Question 2: Bootstrap, jackknife and confidence intervals

The data you are going to continue analyzing is the database of home prices in Albuquerque, 1993. The variables present are `Price`; `SqFt`: the area of a house; `FEATS`: number of features such as dishwasher, refrigerator and so on; `Taxes`: annual taxes paid for the house. Explore the file `prices1.xls`. The source of the original is the Data and Story Library (`https://dasl.datadescription.com/`) and it can be recovered from (`https://web.archive.org/web/20151022095618/http://lib.stat.cmu.edu/DASL/Datafiles/homedat.html`).

1. Create a scatter plot of `SqFt` versus `Price`. Fit a linear model to it—does a straight line seem like a good fit?

2. While the data do seem to follow a linear trend, a new sort of pattern seems to appear around 2000ft^2. Consider a new linear model

$$\texttt{Price} = b + a_1 \cdot \texttt{SqFt} + a_2 \cdot (\texttt{SqFt} - c)\mathbf{1}_{\texttt{SqFt}>c},$$

   where c is the area value where the model changes. You can determine c using an optimizer, e.g., `optim()`, with the residual sum of squares (RSS) as the value to be minimized. For each value of c, the objective function should estimate b, a_1, and a_2; then calculate (and return) the resulting RSS.

3. Using the bootstrap estimate the distribution of c. Determine the bootstrap bias–correction and the variance of c. Compute a 95% confidence interval for c using bootstrap percentile, bootstrap BCa, and first–order normal approximation
   (**Hint**: use `boot()`,`boot.ci()`,`plot.boot()`,`print.bootci()`)

4. Estimate the variance of c using the jackknife and compare it with the bootstrap estimate

5. Summarize the results of your investigation by comparing all of the confidence intervals with respect to their length and the location of c inside them.