Computer Lab 1
Computational Statistics

Linköpings Universitet, IDA, Statistik

2019/01/23

Kurskod och namn:	732A90 Computational Statistics
Datum:	2019/01/22—2019/01/31 (lab session 23 January 2019)
Delmomentsansvarig:	Krzysztof Bartoszek, Eric Herwin, Sara Johansson
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. The report should be handed in via LISAM (or alternatively in case of problems e–mailed to krzysztof.bartoszek@liu.se or Eric Herwin erihe068@student.liu.se or Sara Johansson sarjo775@student.liu.se), by **23:59 31 January 2019** at latest. Notice there is a final deadline of **23:59 1 April 2019** 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 **7 March 2019**. The report has to be written in English.

Question 1: Be careful when comparing

Consider the following two R code snippets

```
x1<-1/3;x2<-1/4
if (x1-x2==1/12){
print("Subtraction is correct")
}else{
print("Subtraction is wrong")
}
```

and

```
x1<-1;x2<-1/2
if (x1-x2==1/2){
print("Subtraction is correct")
}else{
print("Subtraction is wrong")
}
```

1. Check the results of the snippets. Comment what is going on.
2. If there are any problems, suggest improvements.

Question 2: Derivative

From the defintion of a derivative a popular way of computing it at a point x is to use a small ϵ and the formula

$$f'(x) = \frac{f(x+\epsilon) - f(x)}{\epsilon}.$$

1. Write your own R function to calculate the derivative of $f(x) = x$ in this way with $\epsilon = 10^{-15}$.
2. Evaluate your derivative function at $x = 1$ and $x = 100000$.
3. What values did you obtain? What are the true values? Explain the reasons behind the discovered differences.

Question 3: Variance

A known formula for estimating the variance based on a vector of n observations is

$$\mathrm{Var}(\vec{x}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$$

1. Write your own R function, `myvar`, to estimate the variance in this way.
2. Generate a vector $x = (x_1, \ldots, x_{10000})$ with 10000 random numbers with mean 10^8 and variance 1.
3. For each subset $X_i = \{x_1, \ldots, x_i\}$, $i = 1, \ldots, 10000$ compute the difference $Y_i = \texttt{myvar}(X_i) - \texttt{var}(X_i)$, where $\texttt{var}(X_i)$ is the standard variance estimation function in R. Plot the dependence Y_i on i. Draw conclusions from this plot. How well does your function work? Can you explain the behaviour?
4. How can you better implement a variance estimator? Find and implement a formula that will give the same results as `var()`?

Question 4: Linear Algebra

The Excel file "tecator.xls" contains the results of a study aimed to investigate whether a near–infrared absorbance spectrum and the levels of moisture and fat can be used to predict the protein content of samples of meat. For each meat sample the data consists of a 100 channel spectrum of absorbance records and the levels of moisture (water), fat and protein. The absorbance is -log10 of the transmittance measured by the spectrometer. The moisture, fat and protein are determined by analytic chemistry. The worksheet you need to use is "data" (or file "tecator.csv"). It contains data from 215 samples of finely chopped meat. The aim is to fit a linear regression model that could predict protein content as function of all other variables.

1. Import the data set to R
2. Optimal regression coefficients can be found by solving a system of the type $\mathbf{A}\vec{\beta} = \vec{b}$ where $\mathbf{A} = \mathbf{X}^T\mathbf{X}$ and $\vec{b} = \mathbf{X}^T\vec{y}$. Compute $\mathbf{A}$ and $\vec{b}$ for the given data set. The matrix $\mathbf{X}$ are the observations of the absorbance records, levels of moisture and fat, while $\vec{y}$ are the protein levels.
3. Try to solve $\mathbf{A}\vec{\beta} = \vec{b}$ with default solver `solve()`. What kind of result did you get? How can this result be explained?
4. Check the condition number of the matrix $\mathbf{A}$ (function `kappa()`) and consider how it is related to your conclusion in step 3.
5. Scale the data set and repeat steps 2–4. How has the result changed and why?