TMA4300 Computer Intensive Statistical Methods
Exercise 3, Spring 2018

Note: The solution must be handed in no later than **April 27$^{\text{th}}$ 2018, 16:00**.
Hint: In the bootstrap exercises you will need to use the R-function `sample`.

Problem A: Comparing $AR(2)$ parameter estimators using resampling of residuals

The data files and pre-programmed R-code can be downloaded from the course webpage. Look in the `probAhelp.R`-file and read the documentation to see how the code works. Load the code and data into `R` with

```
source("probAhelp.R")
source("probAdata.R")
```

In this exercise you should analyse the data in `data3A$x`, which contains a sequence of length $T = 100$ of a non-Gaussian time-series, and compare two different parameter estimators.
We consider an AR(2) model which is specified by the relation

$$x_t = \beta_1 x_{t-1} + \beta_2 x_{t-2} + e_t,$$

where e_t are iid random variable with zero mean and constant variance.
The least sum of squared residuals (LS) and least sum of absolute residuals (LA) are obtained by minimising the following loss functions with respect to $\boldsymbol{\beta}$:

$$Q_{LS}(\boldsymbol{x}) = \sum_{t=3}^{T} (x_t - \beta_1 x_{t-1} - \beta_2 x_{t-2})^2$$

$$Q_{LA}(\boldsymbol{x}) = \sum_{t=3}^{T} |x_t - \beta_1 x_{t-1} - \beta_2 x_{t-2}|$$

Denote the minimisers by $\widehat{\boldsymbol{\beta}}_{LS}$ and $\widehat{\boldsymbol{\beta}}_{LA}$ (calculated by `ARp.beta.est`), and define the estimated residuals to be $\hat{e}_t = x_t - \hat{\beta}_1 x_{t-1} - \hat{\beta}_2 x_{t-2}$ for $t = 3, \ldots, T$, and let $\bar{e}$ be the mean of these. The $\hat{e}_t$ can be re-centered to have mean zero by defining $\hat{\epsilon}_t = \hat{e}_t - \bar{e}$. (Results for $\hat{\epsilon}_t$ obtained by LS and LA can be calculated with `ARp.resid`).

1. Use the residual resampling bootstrap method to evaluate the relative performance of the two parameter estimators. Specifically, estimate the variance and bias of the two estimators.

   You may use `ARp.filter` as a helper function in your resampling code. Use at least $B = 1500$ bootstrap samples, each as long as the original data sequence ($T = 100$). To do a resampling, initialise values for x_1 and x_2 by picking a random consecutive subsequence from the data.

   The LS estimator is optimal for Gaussian AR(p) processes. Explain if it is also optimal for this problem?

2. Compute a 95% prediction interval for x_{101} based on both estimators. That means using the corresponding parameter estimates obtained in part 1), predict for each bootstrap iteration a value for x_{101}. From these B predictions derive a 95% quantile-based confidence interval.

Problem B: **Permutation test**

Bilirubin (see `http://en.wikipedia.org/wiki/Bilirubin`) is a breakdown product of haemoglobin, which is a principal component of red blood cells. If the liver has suffered degeneration, if the decomposition of haemoglobin is elevated, or if the gall bladder has been destroyed, large amounts of bilirubin can accumulate in the blood, leading to jaundice. The following data (taken from Jørgensen (1993)) contain measurements of the concentration of bilirubin (mg/dL) in blood samples taken from three young men.

Individual	Concentration (mg/dL)										
1	0.14	0.20	0.23	0.27	0.27	0.34	0.41	0.41	0.55	0.61	0.66
2	0.20	0.27	0.32	0.34	0.34	0.38	0.41	0.41	0.48	0.55	
3	0.32	0.41	0.41	0.55	0.55	0.62	0.71	0.91			

We will use the F-statistic to perform a permutation test.

Download the data file `bilirubin.txt` from the course webpage and read it into R using

```
 bilirubin <- read.table("bilirubin.txt",header=T)
```

```
> head(bilirubin)
  meas pers
1 0.14   p1
2 0.20   p1
3 0.23   p1
4 0.27   p1
5 0.27   p1
6 0.34   p1
```

The first column, labelled `meas`, contains the concentrations (mg/dL) as shown in the table. The second column, `pers`, is an indicator for the individual.

1. Use a boxplot to inspect the logarithms of the concentrations for each individual. Be careful to use the same y-axis to make the plots comparable. Use the function `lm` in `R` to fit the regression model

$$\log Y_{ij} = \beta_i + \epsilon_{ij}, \qquad \text{with } i = 1, 2, 3 \text{ and } j = 1, \ldots, n_i \tag{1}$$

   where $n_1 = 11$, $n_2 = 10$ and $n_3 = 8$, and $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Use the F-test to test the hypothesis that $\beta_1 = \beta_2 = \beta_3$ and save the value of the F-statistic as `Fval`. Is the hypothesis accepted?

   (Hint: Define the model as `lm(log(meas)~pers,data=bilirubin)`. Then, the F-statistic of the test of interest is contained in the default output of `summary.lm`)

2. Write a function `permTest()` which generates a permutation of the data between the three individuals, consequently fits the model given in (1) and finally returns the value of the F-statistic for testing $\beta_1 = \beta_2 = \beta_3$.

3. Perform a permutation test using the function `permTest` to generate a sample of size 999 for the F-statistic. Compute the p-value for `Fval` using this sample. What do you observe?

Problem C: The EM-algorithm and bootstrapping

Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be independent random variables, where the x_i's have an exponential distribution with intensity λ_0 and the y_i's have an exponential distribution with intensity λ_1. Assume we do not observe $x_1, \ldots, x_n, y_1, \ldots, y_n$ directly, but that we observe

$$z_i = \max(x_i, y_i) \quad \text{for } i = 1, \ldots, n \tag{2}$$

and

$$u_i = I(x_i \geq y_i) \quad \text{for } i = 1, \ldots, n, \tag{3}$$

where $I(A) = 1$ if A is true and 0 otherwise. Thus, for each $i = 1, \ldots, n$ we observe the largest value of x_i and y_i and we know whether the observed value is x_i or y_i. Based on the observed $(z_i, u_i), i = 1, \ldots, n$ we will use the EM algorithm to find the maximum likelihood estimates for (λ_0, λ_1)

1. Write down the log likelihood function for the complete data $(x_i, y_i), i = 1, \ldots, n$. Use this to show that

$$\begin{aligned} E\left[\ln f(x, y|\lambda_0, \lambda_1)|z, u, \lambda_0^{(t)}, \lambda_1^{(t)}\right] &= n(\ln\lambda_0 + \ln\lambda_1) \\ &- \lambda_0 \sum_{i=1}^{n} \left[u_i z_i + (1 - u_i)\left(\frac{1}{\lambda_0^{(t)}} - \frac{z_i}{\exp\{\lambda_0^{(t)} z_i\} - 1}\right)\right] \\ &- \lambda_1 \sum_{i=1}^{n} \left[(1 - u_i) z_i + u_i\left(\frac{1}{\lambda_1^{(t)}} - \frac{z_i}{\exp\{\lambda_1^{(t)} z_i\} - 1}\right)\right] \end{aligned}$$

2. Using the EM algorithm, use the result in 1 to find a recursion in $(\lambda_0^{(t)}, \lambda_1^{(t)})$ for finding the maximum likelihood estimates for (λ_0, λ_1). Implement the recursion and find the maximum likelihood estimates when the data is as specified in tile files "z.txt" and "u.txt" available from the course home page. Explain how you choose the start values for the recursion and visualise the convergence of the algorithm in a plot.

3. Use bootstrapping to estimate the standard deviations and the biases of each of $\widehat{\lambda}_0$ and $\widehat{\lambda}_1$ and to estimate $\text{Corr}[\widehat{\lambda}_0, \widehat{\lambda}_1]$. Present pseudocode for your bootstrap algorithm. Discuss briefly whether you would prefer the maximum likelihood estimates or the bias corrected estimates for λ_0 and λ_1 in this case.

4. An extra question (the solution of this problem does not need to be handed in): For the situation defined here, can you find an analytical formula for $f_{Z_i,U_i}(z_i, u_i|\lambda_0, \lambda_1)$? If yes, can you also find analytical formulas for the maximum likelihood estimators $\widehat{\lambda}_0$ and $\widehat{\lambda}_1$?

Literature

Jørgensen, B. (1993). The Theory of Linear Models. Chapman and Hall

Oral presentations

Date	Problem	Team
23.04.2018	3: Problems A1 and A2 Ask at least one question:	Hilde Heggstad and Mira Lilleholt Vik Sigrid Leithe and Norunn Wankel
23.04.2018	3: Problems B1, B2 and B3 Ask at least one question:	Thomas Frogner and Rasmus Münter Elisabeth Hetlelid and Dag Kristiansen
23.04.2018	3: Problems C1 and C2 Ask at least one question:	Anne Siri Fardal and Bergitte Viste Emma Sofie Skarstein and Sigurd Stenvik
23.04.2018	3: Problems C3 Ask at least one question:	Johan Sokrates Wind Erik Hide Sæternes and Silius Vandeskog