Assignment 1

ADVANCED ECONOMETRICS

Etienne Wijler (Coordinator and Lecturer)

Noah Stegehuis (Coding instructor)

Gabriele Mingoli (Tutorial instructor)

Notes and instructions:

1. This assignment is mandatory.
2. The assignment is to be made individually.
3. The deadline for delivery of this assignment is Thursday, September 12, at 23:59h

   There will be no tolerance period for late deliveries. Deliveries after the assigned deadline imply that you have a final grade of zero for the assignment (AG1 = 0).
4. To get the full score for this assignment, the following three things must be done:

   (a) upload your PDF file in Canvas Assignments containing the report you are asked to make in the assignment with the name **IAreport_2601842.pdf**, where 2601842 is replaced by your VU student number. Make sure it contains your name, student number, email address and the assignment number. Do not just dump all your plots in this PDF, but make good-looking and self explanatory plots. But don't go overboard: too many pages, unnecessary plots and comments will be penalised.

   (b) upload a zip file of your runnable R or Python code in Canvas Assignments. The name of the file should be **IAcode_2601842_language.zip** , where 2601842 is replaced by your VU student number, and language is replaced by the language used (python or R), e.g., IAcode_2601844_python.zip. The code file(s) should be clear, well commented, and directly runnable, so that it reads the datafile and obtains the results of all questions and prints them. Your initial comments in the file should hold your name and student number.

   (c) upload a pdf of your entire code in Canvas Assignments. The name of the file should be **IAcode_2601842_language.pdf**, where 2601842 is replaced by your VU student number, and language is replaced by the language used (python or R), e.g., IAcode_2601844_python.pdf. The file should be well readable, with proper indentations and should not contain pictures/photos/screenshots of code snippets.

5. As a standard anti-fraud measure, we will conduct a detailed plagiarism check on all your submitted files. In addition, we may at random select a number of you to explain your code and answers. Failure to explain your answers will result in a deduction of credits for this assignment.

6. For the support for the assignments, carefully read the announcement we put out at the start of the course and consult the discussion boards on Canvas related to the assignments, where you can ask your questions.

7. Warning: coding might feel as a frustrating exercise at first, certainly if you have not done it often before and if you are making small mistakes every time. But take heart and persevere: completing a coding task also feels very satisfying after all the bits align and you actually made it work completely by yourself!

We wish you success!!

VU VRIJE UNIVERSITEIT AMSTERDAM

Detecting elevated heart rates

Despite numerous technological advances in healthcare, Cardiovascular disease remains a leading cause of death across the world. This makes the accurate and timely monitoring of bio-metric signs, such as heart rates and blood oxygen levels, an essential task for healthcare facilities. However, heart rates vary naturally over time due to many natural causes, including physical exercise, visual stimuli, stress and sleep. Therefore, while it is certainly possible to continuously measure heart rates at small intervals via portable monitoring devices such as smartwatches, it is simply not efficient to react to every visible fluctuation in heart rates. Instead, one may try to use a so-called "observation driven filter" to smooth out the heart rates of monitored patients. These smoothed heart rates should be less sensitive to noise and persistent deviations over time in the smoothed heart rates may indeed signal underlying health issues. In this assignment, you are going to explore the heart rates of four (fictitious) monitored patients to check whether any underlying issues may be present.

Disclaimer: this case is fictitious and presents a simplified view of the complex dynamics in heart rates and there effect on our health. Nonetheless, the analysis you are going to conduct is sensible and could certainly serve as a basis for a richer follow-up analysis.

Data

We consider a sample of four patients who are known to have genetic predispositions for the development of cardiovascular diseases. In a newly developed health program, these patients were given a wearable heart rate monitoring device, that collects accurate information concerning their heart rhythms. For one month, the patients were in close contact with specialists and required to attend weekly check-ups. All patients were considered to be in healthy condition, but were asked to keep wearing the monitoring devices for an extended period of time. The dataset "heart_rates.csv", available on Canvas, contains 5400 measurements of the average heart rates over 10-minute intervals of these patients. The measurements are collected daily between 11:00-20:00 for 100 days, resulting in a total of $T = 5,400$ measurements, of which $t = 1,...,1,620$ concerns the first month during which the patients were actively monitored. You are going to conduct a basic analysis on these heart rates to detect any potential health issues that require follow-up.

Important: For questions 1 to 6, consider only a training dataset that consists of the first 2,000 observations.

Question 1: Using the training data, make a table with the descriptive statistics (number of observations, means, medians, standard deviations, skewness, *excess* kurtosis, minimum, maximum) with a column for each of the four patients. Provide this table in the report.

- Tip: make sure to obey good academic reporting standards. A good example is given by the figures in the pdf version of this paper, and for tables in Table 3 of the same paper. If you have trouble accessing the paper, follow these steps from the University Library to get automatic access with your VU account (a feature that will prove useful at many occasions during your program). Sometimes, the explanatory notes may seem excessive to you, like in Table 3, but that is not true: you should do everything to make the figures and tables stand-alone, i.e., understandable without reading anything else in the paper. (Note, however, that the explanatory note does not give an interpretation of the data in the figure or table.)

Let $x_{i,t}$ denote the average heart rate of patient i observed at the 10-minute interval at time t. Consider the following statistical model for this heart rate:

$$x_{i,t} = \mu_{i,t} + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim NID(0, \sigma_i^2),$$
$$\mu_{i,t+1} = \omega_i + \alpha_i x_{i,t} + \beta_i \mu_{i,t}$$

Here, $\mu_{i,t}$ is to be interpreted as the *slowly varying* true average heart and $\epsilon_{i,t}$ as a collection of external stimuli that pushed the heart rate away from this average during the specific interval at time t.

Question 2: The parameters $\boldsymbol{\theta}_i = (\omega_i, \alpha_i, \beta_i, \sigma_i)$ are unknown and will need to be estimated from the data. However, to get an idea on how these parameters influence the filtered heart rates, consider the following hypothetical values:

(i) $\boldsymbol{\theta}_i^1 = (0.01 \cdot \bar{x}_i, 0.15, 0.84, 4)$,

(ii) $\boldsymbol{\theta}_i^2 = (0.01 \cdot \bar{x}_i, 0.84, 0.15, 4)$,

for $i = 1, \ldots, 4$, where $\bar{x}_i = \frac{1}{100}\sum_{t=1}^{100} x_{i,t}$. Setting $\mu_{i,1} = \bar{x}_i$, compute using the training data for each hypothetical $\boldsymbol{\theta}_i$, the filtered heart rates $\mu_{i,t}$. Create a 2×2 grid, and plot for each patient the original heart rate data and the filtered heart rates corresponding to $\boldsymbol{\theta}_i^1$ and $\boldsymbol{\theta}_i^2$ in a single subplot. Comment on the differences between these two filters and how these are related to the choice of α_i and β_i. Provide your opinion on which filter seems more useful for the detection of persistent elevated heart rates.

- Tip: make sure to obey good academic reporting standards. That means: create complete, clear and beautiful plots, with strong explanatory notes. Make sure to include a title, axis values, axis labels, a description and if plotting multiple series make sure to have a clear distinction with different line types/colors and provide a legend. A good example is Figure 3 in the pdf version of this paper. If you have trouble accessing the paper, follow the steps given in the Tip for Question 1. Also here, the explanatory notes may seem excessive to you, like in Figures 1 to 3 of this paper, but that is not true: you should do everything to make the figures and tables stand-alone, i.e., understandable without reading anything else in the paper. (Note, however, that the explanatory note does not give an interpretation of the data in the figure or table.)

Question 3: To estimate the parameters $\boldsymbol{\theta}_i$, we are going to implement the method of maximum likelihood. Derive and argue (e.g. based on the lecture slides of week 1) that the log-likelihood is given by

$$L(\boldsymbol{\theta}_i \mid x_{i,1},\dots,x_{i,T}) = \sum_{t=1}^{T} \left\{ -\log\sqrt{2\pi\sigma_i^2} - \frac{(x_t-\mu_t)^2}{2\sigma_i^2} \right\}.$$

Question 4: Again setting $\mu_{i,1} = \bar{x}_i$, code up the log-likelihood function. In the report, provide the value you obtain for the log-likelihood for patient 1 using the training data, evaluated at $\tilde{\boldsymbol{\theta}} = (0.7, 0.15, 0.84, 4)$.

- Tip: using the training data for patient 2, the total log-likelihood evaluated at $\tilde{\boldsymbol{\theta}}$ should produce a value of -6206.323.

Question 5: Compute the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_i$ of $\boldsymbol{\theta}_i$ for $i = 1,\dots,4$ based on the training dataset. Report your estimates in a nice table, were each column corresponds to a patient and each row corresponds to a parameter. Add one extra row in which you report the optimal value of the log-likelihood.

- Tip: you will need to use an optimizer to maximize the likelihood. There are plenty of generic optimizers available in R of Python, all with slightly different implementations. In general, we trust the quality of `optim` in R and `scipy.optimize.minimize` in Python, so those would be relatively safe choices for you with plenty of flexibility. Note, these are both minimizers.
- Tip: in case you require an initial estimate of $\boldsymbol{\theta}_i$, you may consider the values of $\boldsymbol{\theta}_i^1$ given Question 2 (i).
- Tip: before you start optimizing, check whether your likelihood function is working properly. Otherwise it will be unlikely that your optimal estimates are correct. See the tip for Question 4.
- Tip: Of course different optimizers and different programming languages might yield different results with optimisation. We will take this into account when grading and allow for an appropriate bandwidth of error. Note however, that the log-likelihood value

in Question 4 should be exactly reproducible regardless of the programming language, as it is simply a function of the data and given parameters and no optimisation is involved.

- Tip: always check whether the optimiser converged and did indeed find a minimum. In python, if the default `"BFGS"` method gives any problems, try another method (`"SLSQP"` or `"Nelder-Mead"` for example).
- Tip: general advice for convergence problems or unstable results: optimizing the *average* likelihood is easier to handle for the optimizer and can produce more stable results. Furthermore, it is good practice to penalize the log-likelihood function when the optimizer tries undesirable values of the parameter vector (think about stationarity conditions, non-negativity constraints). Furthermore, optimizers can be sensitive to the initial parameter you provide. Always make sure you try a couple of initial values to see how stable the result is, and whether your optimum is not a local minimum.
- Tip: for example scripts of how to optimize a likelihood function in R and Python, see the coding scripts in `Pythonscripts.zip`, `Rscripts.zip` we provide on canvas ("Course Manual" > "Additional learning support" > "Coding support").
- Tip: for directions on how to create a table that adheres to academic standards, consider the tip below question 1.

Question 6: For each patient, use the training data to create a plot with the observed heart rates $x_{i,t}$, the heart rates filtered based on $\boldsymbol{\theta}_i^1$ as in Question 2(i) and the heart rates filtered based on the MLE $\hat{\boldsymbol{\theta}}_i$ from Question 5. Store these plots in a (2×2) grid and copy this to your report.

Question 7: Finally, let us consider now the full sample of data ranging from $t = 1, \ldots, 5,400$. For all four patients, compute the filtered heart rates based on the maximum likelihood estimators from Question 5. Plot the four filtered sequences of heart rates in a single plot. Put a vertical line in your plot at $t = 2,000$, the end of your estimation sample. Copy the plot to your report and comment on whether a certain patient may require a follow-up.