Take Home Assessment 2022

Takoua Jendoubi

2022-03-22

Rules for the take home assessment

The deadline for submission is 16:00 Friday 29 Apr 2022

Please read the following carefully before proceeding to your take home assessment.

General

- Your Take Home Assessment (THA) this year is a group assessment.
- For your group component, each group will submit, via the Submit your take home assessment (group component) Moodle link, one PDF file containing your report and one R program file. It is enough that one group member clicks their "Submit" button in order for the group's submission to be successful and final.

Plagiarism

- **The Turn-It-In® plagiarism detection system may be used to scan both your group submissions for evidence of plagiarism and collusion.**
- For the group project, you will work together within your group and the usual plagiarism and collusion regulations do not apply to this form of interaction. However, they do apply to collusion with other groups or plagiarism of work from other groups or from other sources.
- Any plagiarism will normally result in zero marks for all students involved, and may also mean that your overall examination mark is recorded as non-complete. Guidelines as to what constitutes plagiarism may be found in Departmental Student Handbooks. The relevant excerpt from the Statistical Science handbook is also posted on Moodle.

Grading

- Your THA grade is worth 75% of your overall STAT0004 grade. All members of a group will be awarded the same mark, except in exceptional circumstances (e.g. a member of a group did not contribute to the project).
- Your group project will be marked out of 50, with allocation as follows:
  - 40 marks for the written report. I will be looking for clarity of writing/figures/tables, appropriate selection of materials, soundness of statistical reasoning and ability to explain your findings in a non-technical language.
  - 10 marks for the accompanied program. I will be looking for the correctness of your code, the readability of your program and elegance/efficiency of your implementation.
- Late submission will incur a penalty unless there are extenuating circumstances (e.g. medical) supported by appropriate documentation. Penalties are set out in the latest editions of the Statistical Science Department student handbooks, available from the departmental web pages.
- Failure to submit this THA may mean that your overall examination mark is recorded as non-complete, i.e., you will not obtain a pass for the course.
- I may ask you to come and discuss your output with me.
- You will receive, via Moodle, feedback on your work and a *provisional grade — grades are provisional until confirmed by the Statistics Examiners' Meeting in summer 2022.*

Data description

The data for the Take Home Assessment, available under the Data Files folder in the Take Home Assessment section on Moodle, contain measurements of temperature and wind velocity of 377 local authorities in the UK and covid infection records of 312 local authorities in England only from 2020-06-15 to 2020-07-06. You are given three files `wind_data.csv`, `temp_data.csv` and `covid_cases.csv`.[1]

- In `wind_data.csv`, each row stores relevant information for daily wind velocity in a specific local authority at a given date and has four comma-separated values:
  - `date`: this is the date wind velocity has been recorded formatted as yyyy-mm-dd.
  - `lads`: these are the local authority district names covering all of the UK.
  - `U1` and `V1`: these are two characteristic values of the wind velocity. These are respectively the eastward and northward components of the 10m wind i.e. at a height of ten metres above the surface of the Earth, in metres per second. Wind velocity can be computed using the Euclidean norm of both `U1` and `V1`.
- In `temp_data.csv` contains information for temperature records in the UK:
  - `date`: this is the date temperature has been recorded formatted as yyyy-mm-dd.
  - `lads`: these are the local authority district names covering all of the UK.
  - `value`: this is the recorded temperature for each lad at a given date in Kelvin.
- In `covid_cases.csv`, each row has five comma-separated values. **Note: You do not need to use all of these variables.**
  - `lads`: the local authority districts names. This only covers England local authority districts.
  - `code`: the identifier of the local authority district, they contain a unique code.
  - `date`: date on which covid infections have been recorded, formatted as yyyy-mm-dd.
  - `cases`: the number of covid infections recorded per day, this is an integer.
  - `population`: total recorded population for each local authority.

[1] Data used in this Take Home Assessment are adapted from ECMWF.

Group Assignment

Tasks

In this group project, you will need to use the 3 data files given to you. As a group, you will describe and analyse all the data in these two files by answering the problems below. Your analysis must be based solely on the data given to you using basic R. Do not introduce other data into your work. Do not import any extra R packages. Also, you do not need to investigate the source of the data further.

1. Describe the wind velocity and temperature measurements of all of the UK. Explore these measurements along with covid infections per 100k population in England as well. Your description should include both univariate and multivariate analysis. Specifically:
   - You should write a function named `compute_velocity` that computes wind velocity using wind components `U1` and `V1`. Wind velocity is the Euclidean norm of wind components `U1` and `V1`.
   - You should create 2 datasets comprising the following column variables respectively: i) only temperature and wind velocity for the UK and ii) temperature, wind velocity and covid infections per 100k population in England. Save these datasets into separate csv files `uk_weather.csv` and `england_data.csv`.
   - You should use techniques such as summary statistics and plots both including univariate and multivariate analysis.
   - You should comment on the plots you produce.
2. A weather presenter claims that England was warmer in the period 2020-06-15 to 2020-07-06 compared to the rest of the UK, whereas the rest of the UK was more windy compared to England during the same period.
   (a) Compute the average temperature and wind velocity across all local authorities i) in England and ii) in the rest of the UK excluding England. **This should give a single measurement per day for i and ii.**
   (b) Carry out an appropriate statistical test (at 5% significance level) separately for each of temperature and wind velocity to verify the presenter's claim.
   (c) Write one to two sentences to explain your findings in non-technical terms.
3. An epidemiologist claims that there are monotonic linear relationships between i) the daily number of covid infections per 100k population in England (averaged across all local authorities) and wind velocity and ii) the daily number of covid infections per 100k population (averaged across all local authorities) in England and temperature during the period 2020-06-15 to 2020-07-06.
   (a) Build an appropriate simple linear model to investigate the epidemiologist's claims at 1% significance level.
   (b) Write one to two sentences to explain your findings in non-technical terms.

For the linear model, you may assume that the residual plots raise no issues about model assumptions or fit and you should not attempt to analyse or study them (I know that this is not what normally happens but I am trying to make your life easier).

What to submit

Please submit two files for your group component:

1. A PDF report named `report.pdf`. The report should be consistent with the following:
   - You must use the Microsoft Word template provided on the Moodle page for your report and not change its font, font sizes or margins. If the template has been changed, up to 4% of marks can be lost and I will reformat the document to the template standard, to which the following point will apply.
   - The report must not be longer than 2 pages in A4 paper, including figures. I will not mark any content beyond the page limit. Note that this doesn't mean that you should aim to fill all the space available to you. Writing more text doesn't necessarily get you more marks.
   - In addition, all groups submit an additional cover page (so that the total number of pages submitted is three) where each group member briefly describes their contribution to the project.
     - You will need to agree this in your groups before submitting the report.
     - If all group members agree that everyone contributed equally, then it is sufficient to write a single sentence to that effect on the third page, or alternatively you are very welcome to describe your own personal contribution to the project.
     - Note that I will not mark this page, nor allocate different marks to different group members based on this. The purpose is to encourage you all to be mindful about contributing to this piece of groupwork.
     - If a group reports that one or more of their members is not contributing fairly, please contact me by email in the first instance BEFORE SUBMISSION of the report.
   - The report must be capable of being read on its own: i.e., it should not refer to the R program but just contain data/plots from the program's output.
   - Please save your report as a PDF file from Microsoft Word (FILE, Export, Create PDF/XPS).
   - It must be written in clear comprehensible English with readable and well-labelled figures.
   - Your report should be anonymous — i.e., there should be no mention of group members' names anywhere in your submission.
2. An R program named `analysis.R`. Your R program should satisfy the following:
   - It should be clearly laid out and well commented.
   - You can assume that all data files are in the working directory, i.e., there should be no `setwd()` command or reference to directories.
   - Should not use non-standard packages (Does not include a `library()` command).
   - It should create an output file named `output.txt`, **containing only the statistics you use in your report**. Your program may investigate other things but the output file should contain all the information you use in your report. The output file itself should **not** be included in your submission. Instead, I will run your program using the `source()` function in R to generate your output file.
   - Create a `.pdf` image file for each plot (or set of plots) that you use in your report, and no others. Name the image files `fig1.pdf`, `fig2.pdf`, ..., following the same order in which they appear in your report. Do not submit the figures. They should be created when I use the `source()` function to run your program.
   - Be anonymous — i.e., there should be no mention of group members' names anywhere in your submission.