

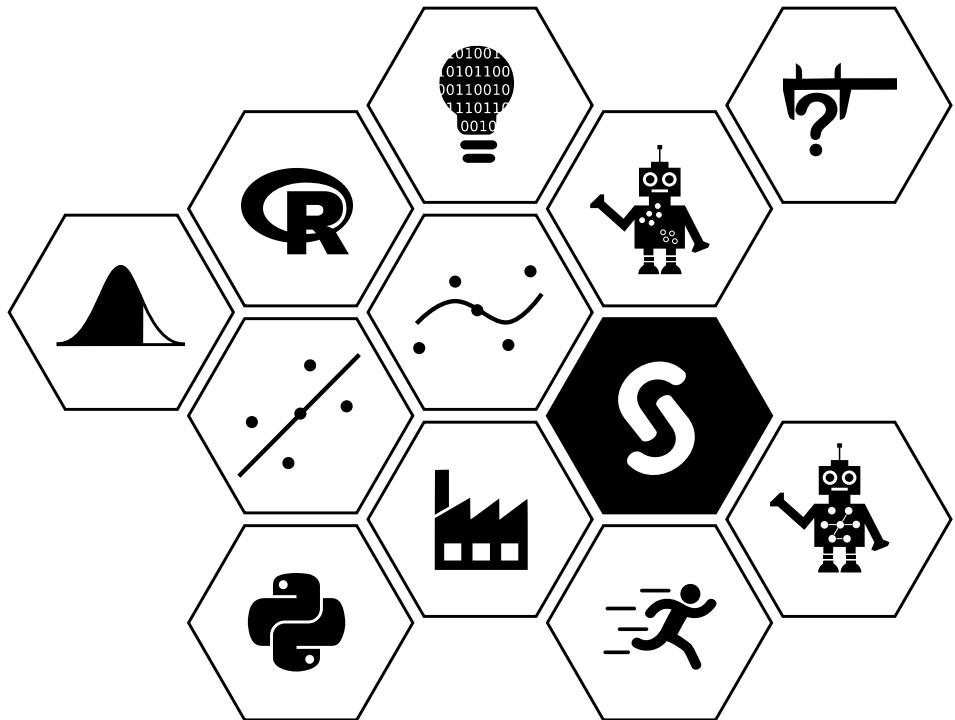
Data Management and Analytics using SAS

Benn Macdonald and Michael Waltenberger

Academic Year 2020-21

DMAS Assessment 3:

Lab Report



Lab Report

Instructions

You are required to submit your lab report as a pdf file. Put your matriculation number at the start of your document and as the name of your pdf file.

For this assessment you must submit your work as a report. There will be questions of interest that you must address and discuss by conducting a statistical analysis in SAS. You are **not** required to print from SAS to pdf using “ods pdf”, copy-pasting from the Results Viewer is completely fine. Please note that any output (i.e. tables and plots) you show in your report **must** be SAS output. You should **not** create your own table (e.g. in Word or LaTeX) and fill in the values. Since you **do not upload/show code** for this assessment, only output directly from SAS can be accepted, to verify you produced it using SAS.

The deadline for the assessment is April 23rd 2021 11.55pm.

The **maximum** page limit for your submission is 20 pages.

Data

It is well known that earthquakes can be devastating: often causing structural damage and loss of life. Being able to better understand the characteristics of earthquakes could lead to better forecasting of frequency and hazard, allowing us to prepare and react more effectively. In the 1930s, the Richter magnitude scale was developed in order to quantify the strength of an earthquake. Although the Richter scale is commonly referred to by news outlets, seismologists refer to various magnitude scales (such as the moment magnitude scale) to deal with an array of circumstances and measurement instruments.

You have been given one dataset: `earthquakes.sas7bdat`:

You can find this on SAS OnDemand using the following path “/courses/dc36fc35ba27fe300/DMASAssessments”. The dataset has also been uploaded to MOODLE.

The dataset contains 23741 independent records about earthquakes over a number of years. The variables you have observations for are given in the following table:

Variable Name	Type	Description
id	Numeric	ID of record
lat	Numeric	Latitude of earthquake (degrees)
long	Numeric	Longitude of earthquake (degrees)
dist	Numeric	Distance travelled by earthquake in a particular direction (km)
depth	Numeric	Depth of earthquake (km)
md	Numeric	Magnitude of earthquake, estimated from the duration of seismic wave-train (Md)
richter	Numeric	Intensity of earthquake (Richter)
mw	Numeric	Moment magnitude scale value of earthquake (Mw)
ms	Numeric	Surface-wave magnitude scale value of earthquake (Ms)
mb	Numeric	Bodywave magnitude value, measured using P-waves and a short-period seismograph in the first few seconds of an earthquake (mb)
country	Character	Country of earthquake
direction	Character	Direction of earthquake

Questions of interest

- a) Sometimes, the largest value of a series of measurements is used to represent the magnitude of an earthquake. Use "xm" to denote the largest magnitude value out of "md", "mw", "ms", "mb" and "richter" for each record. Is there evidence that the average value of "xm" is different to 4.1?
- b) Is there a difference in the moment magnitude scale value of an earthquake (Mw) between countries in which the earthquakes occurred, on average?
- c) Fit a regression with "richter" as the response and consider the other variables in the dataset as **potential explanatory variables**, but do not use the variable "id" or the variable "xm".
- d) A magnitude of 5 and above on the Richter scale is considered to be a moderate or stronger earthquake, causing damage and loss of life. Consider a new variable "serious", where the value of "serious" is 1 if the corresponding Richter scale value is 5 or more and 0 if the corresponding Richter value is below 5. Fit a regression with "serious" as the response and consider the other variables in the dataset as **potential explanatory variables**, but do not use the variables "id", "mw", "richter" or "xm".
- e) Fit a regression with "serious" as the response and "xm" as the only explanatory variable. How does this model compare to your model from part d) in terms of out-of-sample predictive performance (i.e. the model's ability to predict data on which it has not been built)?

Structure/content of your report

Your report should address each of the questions of interest above. Clearly outline what you are doing at each part of your report and discuss anything you present. You will need to include SAS output (tables and plots) containing the **relevant** results of whatever analysis (exploratory or formal) you have performed. **Please note** you must include SAS output corresponding to any results you discuss in your report (i.e. **you will be penalised for discussing results you have not presented output for**). Ensure your output is clearly titled, labelled etc. Captions and titles for plots and tables should be created using SAS.

Ensure your report is self-contained. This essentially means that a reader should not have to hunt around for further information. If you show a plot, for example, explain what the plot is and shows. The same thing goes for tables. All tables and plots should have clear captions.

Start your report with an introduction section. This should contain some background information about the topic of the dataset to give your reader context, summarise the dataset and describe the questions of interest. A good introduction section allows your reader to have a good idea as to the content of the report.

You should then go on to conduct an exploratory analysis and write up an exploratory analysis section. This means you do not perform formal statistical tests or fit formal statistical models at this point. The purpose of an exploratory analysis is to get an initial impression as to the relationships between your variables. It helps to consider the particular questions of interest, so you can keep your exploratory analysis relevant. At this stage you would also describe any data manipulation you may have performed preceding your impending formal analysis.

Next you should conduct a formal analysis and write up a formal analysis section. This will involve performing statistical tests/fitting statistical models. Make sure you conclude and discuss any test/model you include in your report at this point. Ensure you check your modelling assumptions and report them!

Finally, you should include a conclusion section. This summarises your report and main findings. It helps to think of what the specific questions of interest were and how you have answered them. A good conclusion section will allow a reader to know your main findings solely by reading this part of your report.

The type of analyses you perform is up to you, so long as they address the questions of interest. You are not required to perform tests/fit models that we have not seen on the course. If you find that your modelling assumptions are not valid whilst you are conducting your analyses, you should take steps to resolve this. State what resolution you are going to undertake (at the relevant part of your report), justify why and then show the outcome of your resolution. If you are unable to resolve violated modelling assumptions, report what things you tried (include output from an example) and continue on with your lab report. To reiterate, you are not required to perform/fit tests/models that we have not covered on the course, so think about the options that are available to you should you find your modelling assumptions are violated.

A very important point to keep in mind is to make sure you walk your reader clearly through any output or discussion throughout your report. Clarity is key!

-Good luck and have fun!