SOR1232 – SPSS/R Assignment Guidelines

Examiners: Dr Derya Karagoz

Deadline for the Assignment is Strictly: **31st May 2025**

For this assignment you are requested to work in groups.

1. Dataset

- Download a dataset from the internet (below are some sites where you can find datasets).
- Make sure that the dataset you choose has **at least (possibly more) two categorical variables** and **three non-categorical.**
- Once you find a dataset send the link to the dataset by email to Dr. Derya Karagoz on derya.karagoz@um.edu.mt, so that she can confirm that you can use it for the assignment.
- **All groups must have a different dataset.**
- If the dataset you choose has already been assigned to another group you will be asked to choose another dataset. (Datasets are allocated to groups on a first come first served basis).

Some data sets:

- http://lib.stat.cmu.edu/datasets/sleep
- http://lib.stat.cmu.edu/datasets/Plasma_Retinol
- http://lib.stat.cmu.edu/datasets/CPS_85_Wages
- http://archive.ics.uci.edu/ml/datasets/Wine+Quality
- https://dasl.datadescription.com/datafile/cereals/?_sfm_cases=4+59943&sf_paged=6

Websites which contain data sets:

- http://www.statsci.org/datasets.html
- http://www.assda.edu.au/
- http://www.sci.usq.edu.au/staff/dunn/Datasets/
- http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/DataSets
- http://lib.stat.cmu.edu/datasets/
- http://datalib.ed.ac.uk/sources.html
- https://dasl.datadescription.com/datafiles/
- https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html

You can find other sites by searching for data libraries in google. There are also datasets available in R packages.

Note

1. **Group Work**
    - Group assignments are allowed.
    - The group size will not be more than three.
2. **Data Set Allocation**
    - Each group will be assigned a different data set.
    - If your chosen data set has already been assigned to another group, you will need to select a different one.
    - Data sets are allocated on a first-come, first-served basis.
3. **Data Collection**
    - You may assemble your own data set by collecting data from various websites.
    - Inventing or fabricating data is strictly prohibited.
    - All sources used to compile your data set must be properly cited in the reference list.
4. **Feedback Sessions**
    - Two feedback sessions will be provided during the semester:
        - The first feedback will be given at the end of the **Introduction** chapter.
        - The second feedback will be given at the end of the **Statistical Tests** chapter.
5. **Group Meetings**
    - Each group may request up to two organized meetings if needed:

- One meeting to discuss **statistical tests**.
- Another meeting to discuss **statistical modeling**.

Please ensure that you follow these guidelines to complete your assignment successfully.

Example of a Data Set:

Covariate	Covariate	Covariate	Categorical	Categorical	Categorical
Stress level	Age	Hours of Work	Edu Level	Gender	Drive
3.67	6	0	0	1	N
4.58	20	0	2	1	N
10.11	34	39	4	2	Y
12.46	59	45	2	2	Y
7.9	23	10	3	1	Y
13.6	34	41	1	2	Y
...	...	...	...	...	...

2. Formatting

Fonts to be used

Times New Roman 12 for text

Times New Roman 14 bold for headings

Front Page – See Template

Name and Code of Study Unit

Names of authors

Lecturer's Name

3. Write-up Layout

Table of Contents

List of Figures and Tables

It is important to number and label any figures used in the text.

Chapter 1 – Introduction

A brief description of the dataset including a table made up of 4 columns - variable name, description, measurement units and type of variable (categorical or non-categorical).

It is also important to mention from where you got the data. Write down the link to the website and also state the date when the data was last accessed.

Here you will also set a number of scientific questions (the aims of the study) which you will set out to answer by using the appropriate statistical analysis.

Variable	*Description*	*Measurement units*	*Variable Type*
Gender	The gender identification of the participant	1 - Male 2 - female	Categorical Nominal
Exercise	The amount of exercise carried out by the participant	1 – high 2 – moderate 3 - low	Categorical Ordinal
Ran	What the participant did between the 2 pulse rate measurements (ran in place or sat for 1 minute)	1- ran 2- sat	Categorical Nominal
Height	The height of the participant	Centimetres (cm)	Covariate Continuous
Age	How old the participant is	Years	Covariate Discrete
Pulse1	The first measurement of the pulse rate	Beats per minute	Covariate Discrete
Pulse2	The second measurement of pulse rate after standing/sitting	Beats per minute	Covariate Discrete

Table 1.1: *Brief description of the dataset variables used from the data set "Pulse Rates Before and After Exercise".*

Chapters 2- Exploratory Data analysis

Use descriptive statistics and graphical representations to summarize the main characteristics of the different variables present in the dataset. The aim of this section is to gain a better insight on the variables present in the dataset.

Important: If a dataset contains many variables you might want to focus solely on those variables that are the protagonists in the scientific questions listed in the introduction.

Chapter 3 Parametric / Non parametric Tests

Before running any parametric / non-parametric text, it is important to mention why this test is being conducted. It is also important to explain not only your results but also any possible repercussions that these results will have on the analysis of the data set. Don't forget to write the Ho and H1 hypothesis before displaying the results of any statistical test. (You should not attempt to perform all the tests found in the lecture notes. You must only apply those tests which can give you meaningful results.) You should also include the chi-squared test and the table of correlations. The latter contains important information which will be used in the Regression section.

Chapter 4 Statistical Modelling

The contents of this section will vary depending on the choice of your main variable of study. If this variable (sometimes called the dependent variable) is quantitative then you should adopt the procedure outlined in (a). If on the other hand, your dependent variable is qualitative, then you should adopt the procedure outlined in (b).

(a)You should start by constructing a Multiple linear regression (MLR) model. It is important to comment on the outputs contained in the tables. Note you must remove all variables that are not significantly affecting the dependent variable. You must check that all the assumptions are true and check for Outliers. Next you should construct an N-way Anova (using the same dependent variable as the MLR model). It is important to comment on the outputs contained in the tables. Note you must remove all variables that are not significantly affecting the dependent variable. You must check that all the assumptions are true and check for Outliers. Finally, run an ANCOVA model (using the same dependent variable as above) by including the significant quantitative explanatory variables of MLR and the significant qualitative variables of the ANOVA. It is important to comment on the outputs contained in the tables. Note you must remove all variables that are not significantly affecting the dependent variable. You must check that all the assumptions are true and check for Outliers.

(b)If your dependent variable is qualitative and binary, then you should go for a binary logistic regression model. Alternatively, if your dependent variable is qualitative and has more than Page 5 of 6 two levels, then you should go for a multinomial logistic regression model.

Last Chapter-Conclusion

A summary of all the results obtained in the previous sections.

References

A list of all the books and websites which you referred to in your work.

5. Submission

A soft copy of the assignment should be sent by email to Dr Derya Karagoz by the given deadline. In aim to reduce paper use, hard copies are not required. In the correspondence sent, kindly cc all the persons involved in producing the assignment and also cc our departmental secretary Ms Ann Zammit – administrator-stator.sci@um.edu.mt

Important

1. Please note that you will NOT be awarded any marks if you just copy and paste tables from SPSS to your assignment. You will be given marks when the outputs are fully explained. Therefore, no marks will be given to students who submit assignments containing SPSS outputs only. Keep in mind that if a table or diagram is not relevant to the analysis of your data set then it should not be included in the assignment.

2. Copying of assignments is not allowed. Students who are found guilty of a breach of the University Assessment Regulations are liable to disciplinary action which may result in the assignment being cancelled and other consequences.

3. **Copying word for word from any book, website or assignments done by students in previous years is not allowed. Plagiarism is not accepted. Any group caught plagiarising will be heavily penalized. Kindly read through uom's plagiarism guidelines here:** https://www.um.edu.mt/__data/assets/pdf_file/0009/95571/University_Guidelines_on_Plagiarism.pdf

4. Assignments that do not have a signed declaration of authorship will not be accepted.

4. Assignment Deadline

31st May 2025

Marking Scheme

10 marks	Quality of Presentation and flow.
10 marks	Introduction – a clear presentation of the dataset, how the dataset has been obtained and a clear statement of the aims and objectives are expected here.
5 marks	Conclusion – A good summary of the results and implications of the results. Any limitations encountered whilst performing the study should also be specified.
10 marks	EDA
30 marks	Clear motivation and correct use of Statistical Tests. Use of adequate tests to answer the scientific questions posed. Checking of assumptions of the different tests that will be used in the assignment is especially important.
35 marks	Clear motivation and correct use of statistical models. Interpretation of results needs to be given both mathematically (e.g. in hypothesis testing: We do not have enough evidence to reject Ho) as well as in terms of its applicability (e.g. This implies that group A performs better than group B hence the method of teaching applied with group A is more successful). Testing of any assumptions is once again considered to be very important here.