

Problem Solving Assignment 1

Total Marks: 100; Weighting: 20%

Due: 30/3/2023 11.59pm

Instructions:

- Submit only one file in pdf format to the link on the Study Desk.
- Assume that your report will be read by someone familiar with the data set but with limited statistical knowledge. Fully explain plots and when stating statistics or results explain what they mean statistically AND in context of the data.
- Presentation should be neat, consistent, spell-checked and proofread. All questions should be clearly labelled, and all answers should clearly and concisely address the questions.
- If you convert a Word document to pdf for submission check that all symbols, equations etc. have converted correctly, i.e. proof-read your work.
- If you do not use Rmarkdown to compile your submission, where asked to provide R code, paste relevant code within the assignment document and italicise (or otherwise highlight or distinguish from other content). Do not include code in an appendix.
- Do not include an appendix at all. Any work included in an appendix will not be marked.
- Please note that referencing textbooks and other resources is not the goal of this assessment. This work requires students to demonstrate their understanding of the analysis and interpretation, not provide quotes from resources.
- Use only statistical methods covered in this course and do not transform variables to try to normalise them.
- When interpreting output you are expected to do so in context of the data and the method (i.e. ensure that you comment on aspects of the method that affect your interpretation with respect to the variables and sample).
- A maximum of 10 marks will be deducted from your total marks for poor presentation.

Marks:

- Question 1: 32
- Question 2: 29
- Question 3: 39

Data Files

Three data sets will be used for this assignment.

Question 1: film_2023.txt which contains data for the thickness of pieces of plastic film measured in different positions after being cut.

Question 2: iris_2023.txt which contains data for variables of flower characteristics for different species of iris.

Question 3: usair_2023.dat which contains data for air quality variables measured across different United States cities.

Question 1 [32 marks]

Provide R code, output and written interpretation for parts a) to d) of this question. Provide only output that is directly relevant to address each section.

Test for multivariate normality (MVN) by:

a). Provide output from the structure function (0.5 mark) and describe the structure of the 'film_2023.txt' data (2.5 marks). *(3 marks total)*

b). Produce (2 marks) and interpret (4 marks) univariate QQ plots, histograms and univariate Shapiro-Wilks tests of normality for each of the four film thickness variables. What is the default univariate test produced by the mvn function? (1 mark) *(7 marks total)*

c). Produce (2 marks) and interpret (4 marks) perspective and contour plots for the TopRight and TopLeft film thickness variables. What is an inherent problem with using these plots to assess MVN (1 mark)? *(7 marks total)*

d). Do the analysis necessary to provide the results of the Mardia, Henze-Zirkler and Royston tests of MVN based on all four film thickness variables. Include in your interpretation: *(13 marks total)*

- The Chi-Square QQ plot (1 mark) and interpretation (2 marks).
- Describe how the QQ plot is constructed and its relationship to the univariate normal QQ plots (4 marks).
- Output (1 mark) and interpretation (3 marks) for the 3 tests.
- What is a key limitation of these MVN statistical tests (2 marks)?

e). If your data does not meet MVN, why might you need to consider the ratio of cases to variables? (This question does not necessarily relate specifically to this particular data set) *(2 marks total)*

Question 2 [29 marks]

Provide R code, output and written interpretation for parts a), b), c), e) and f) of this question. Assume the data meets the assumption of MVN (do not test for MVN).

a). Use the structure function and describe the structure of the ‘iris_2023.txt’ dataset (1 mark). In the context of MANOVA list the dependent and independent variables and define the relationship that the MANOVA would test (2 marks). What type of variable does SPECIES need to be for MANOVA (1 mark). Make sure you have converted this variable if necessary before attempting the analysis in later parts of Question 2. *(4 marks total)*

b). Produce a draftsman display for four flower characteristic variables (2 marks). Use the function scatterplotMatrix (from week 2) for the draftsman and check the help documentation (?scatterplotMatrix) to help you produce the plot with observations grouped by species using different colours and include the associated legend. Your plot should not include smoothing, regression lines, or distribution curves in the diagonal panels of the plot (1 mark). Interpret these plots (3 marks). What are the y and x axes on plot [3,2] of the scatterplotMatrix (1 mark)? *(7 marks total)*

Hint: to move the legend in scatterplotMatrix try something like: legend=list(coords=“bottomleft”).

c). Using MANOVA in R, test for differences in ‘flower characteristics’ between the three species. Include results using all four test statistics (2 marks) covered in this course and interpret output (2 marks). Include a sentence about what these results mean in terms of within and between group variances in general (2 marks). *(6 marks total)*

d). Which of the four tests used in part c) would be the best to interpret if there are concerns about multivariate normality or covariance equality? *(1 mark total)*

e). Perform analysis that specifically compares each of the species with each other (you should have 3 comparisons) using Hotelling’s T^2 test and a significance level of 0.05. Determine the multiple test corrected significance level (1 mark). Do not provide R output; instead reproduce and complete the following table for all comparisons (3 marks) and interpret (3 marks). *(7 marks total)*

| Comparison | | Hotelling’s p-value | Significant (Y/N) | Significant after correction (Y/N) |
|------------|-----------|---------------------|-------------------|------------------------------------|
| Species A | Species B | | | |
| | | | | |

f). Produce a table of the sample sizes by species (1 mark). Do you think sample sizes could have affected the results from parts c) and e) (2 marks)? In general, do you think deviation from MVN would influence these results (1 mark)? *(4 marks total)*

Question 3 [39 marks]

Provide R code, output and written interpretation for parts a) to f) of this question where relevant. Assume the data meets the assumption of MVN (do not test for MVN).

a). Provide output (1 mark) from the structure function and describe the structure of the 'usair_2023.dat' dataset (1 mark). *(2 marks total)*

b). Produce the correlation (1 mark) and covariance (1 mark) matrices. Explain the difference between these matrices in detail i.e. explain clearly **in words** how the values are adjusted mathematically and the effect of these changes (4 marks). Would using the covariance matrix in PCA on the usair data be appropriate (1 mark)? Why (2 marks)? Describe the output of the correlation matrix (2 marks) *(11 marks total)*

c). Perform PCA analysis on the 5 variables using the prcomp function. Provide the eigenvalues (1 mark), %variation (1 mark), cumulative %variation (1 mark) and scree plot (1 mark). Interpret each of these results (4 marks) and state how many PC's you think should be interpreted (1 mark). Remember to keep in mind the overall purpose of PCA. *(9 marks total)*.

d). Provide the Z equation and interpretation for PC1 (1 mark) and PC2 (1 mark). Describe the relationship between *temp*, *days.precip* and *annual.precip* on PC1 (2 marks) and *annual.precip*, *SO₂* and *wind.speed* on PC2 (2 marks). Comment on which variables most clearly represent PC1 and PC2 (2 marks). Comment on the overall variation explained by this 2 PC solution (1 mark) *(9 marks total)*

e). What is the correlation between the first and second PCs (1 mark) and what does this tell you (2 marks)? *(3 marks total)*

f). Produce a biplot based on the first 2 PCs (1 mark). Explain your understanding of the position of city 1 (2 marks) compared to city 11 (1 mark) and city 9 (1 mark). Relate your interpretation back to the original data and the biplot vectors. *(5 marks total)*

End of Assessment