MATH 4044 – Statistics for Data Sciences

Case Study SP5 2022

Due Sunday 30th Oct 2022

Instructions

- This assignment is worth 35% of your final mark. It is due by **Sunday 30th October**.
- You will need to submit your assignment via learnonline.
- The submitted assignment needs to be a **single file**, in either a Microsoft Word (doc or docx) or pdf file format, **25 pages at most excluding any appendices**.
- The assignment is out of 100 marks. To achieve maximum marks for each question, you should aim to:
  - Complete the requested statistical analysis in SAS using appropriate tasks or procedures (40%).
  - Include only the output most relevant to the question and interpret all key results (40%). Do not include every piece of output produced by SAS!
  - Discuss the results more broadly in the context of the given scenario (20%).
- Assignments submitted late, without an extension being granted, will attract a penalty 5 marks per day or any part thereof beyond the due date and time.

Introduction

HR Analytics helps company to identify trends and pattern in employing, helping company to set strategies and policies for a profitable and healthy working environment. In this case study, we will explore the trend and relationships that affect employee's income. We will analyse a fictional dataset created by IBM scientists to answer the following questions

- What are the income trends for employees in different departments of the company.
- What are the relationship between employees' income and their education level and education field.

Data Description

The data is collected downloaded from `https://www.kaggle.com/datasets/itssuru/hr-employee-attrition`. The data is available on our SAS server as `mydata.HRAttrition`. The data contains many HR information. The variables of interest for our case study are

Variable	Description
`age`	Age of the Employee
`Gender`	employee gender (female/ male)
`Department`	Employee's deparment (*Human Resource, Research & Development and* Sales)
`TotalWorkingYears`	Total working years of employee
`Education`	Employee's highest qualification 1:Below college; 2: College; 3: Bachelor; 4: Master; 5: Doctor
`EducationField`	Fields of education: Human Resource, Life Sciences, Marketing, Medical, Tehcnical Degree, and Other
`MonthlyIncome`	Employee's monthly income

Assignment Tasks

Question 1 (10 marks) Generate the square-root and log transformation for `MonthlyIncome`, namely `sqrtInc` and `logInc` respectively. Among the three variables, select the one that is most suitable for analysis of variance. Denote the variable of your choice as `tIncome`.

Question 2 (55 marks)

(a) **(20 marks)** Carry out a one-way analysis of variance (ANOVA) relating `tIncome` to `EducationField`. Use contrasts to test at least one a-priori hypothesis of your choice. Examine and comment on residuals. Also carry out appropriate post-hoc comparisons and discuss your results.

(b) **(10 marks)** If the assumptions for ANOVA is not satisfied, use a non-parametric method to validate the results in question (a).

(c) **(25 marks)** Use SAS to perform a one-way ANCOVA relating `tIncome` to `EducationField` and `TotalWorkingYears` with `TotalWorkingYears` as a covariate, including appropriate post-hoc comparisons:

- Confirm that there is a linear relationship between the response variable and the covariate (a scatterplot and correlation coefficient plus a comment will suffice);
- Check the two additional ANCOVA assumptions (report and comments only on the parts of the output most directly relevant to condition checking):
  - Independence of the covariate and the treatment effect (perform a one-way ANOVA test);
  - Equality of slopes (add and check significance of the interaction term);
- Report and briefly discuss your results.

Technical note: Make sure you obtain and examine Type III Sum of Squares (ss3). Also obtain estimates of 'least squares means' (lsmeans) which are means by treatment adjusted for the covariate.

Question 3 (25 marks)

Perform and analyse a factorial ANOVA model to determine whether there is statistically significant difference in `tIncome` by `Department` and `Education`. Carry out to test whether there is evidence of interaction between `Department` and `Education`. Examine and comment on residuals. Carry out appropriate follow-up analysis and discuss your results.

Question 4 (10 marks)

Write a summary of your findings from Questions 1–3. Keep the technical details of the analyses that led you to these conclusions to the absolute minimum. Rather,

focus on practical significance and present your findings in non-specialist terms. One to two paragraphs (up to a page) will be sufficient.