ST662 Topics in Data Analytics
Exam
May 2020

Name (PRINT): Nishant Kumar

Name (SIGNATURE): Nishant kumar(19251001)

Student Number: 19251001

- This exam is open book. You may use your notes or online material as you wish.
- Write a single SAS programme to address questions 1 and 2 below. Include comments to clearly indicate which code relates to which question, and what each part of code does. Upload your programme to Moodle before the time is up. Ensure that your name is in the programme file title. E.g. `RAMoral_ST662_SAS_Exam.sas`.

  NB: Getting this done on time is part of the assessment - the time stamp will be visible on Moodle.
- Use the work library only in SAS.
- Where there are boxes below, write out the relevant output. If there is no box, assessment will be on the code you provide. Alternatively, you may submit an accompanying Word or pdf file with typed answers for the boxes.

  You must submit your code/answers by **Friday, the 8th of May, at 4pm**. Moodle submissions will lock at that time, and I will not allow for any extensions. So please, don't leave it until last minute to upload yours.

1. There are three datasets on Moodle: Medical1.csv, Medical2.csv and Medical3.csv.

   - Medical1.csv contains 12,000 rows and three variables that are:
     - Patient: identifier for each patient that takes values 1-1000.
     - Visit: values 1-12 indicating which month the visit took place with 1=January, 2=February etc.
     - Outcome: measurement that has a minimum value of 0 and a maximum value of 27.
   - Medical2.csv contains the ages of each patient on day 1 of the study. Only subjects aged 50-59 on day 1 of the study were to be considered.
   - Medical3.csv contains the date within each month that the visit was intended to take place (the same for each patient).

   Write a SAS programme to address the following questions. You may not alter the csv files prior to reading them into SAS.

   (a) Read the three datasets into the work library in SAS. [4 marks]

   (b) Combine the datasets into one SAS data file called medical. The columns of medical should be Patient, Visit, Outcome, Age, Day. [4 marks]

   (c) Screen the data. Change all error values to either missing, or to another value if appropriate. List any problems encountered and the errors found and state how they were resolved. E.g. Obs XX, out of range, replaced by missing value. [8 marks]

   Identified all the entries based on the given column constrains.
   1. outcome column of the datasets have 5 values which are out of range.
   Resolution :These values are numeric in nature so we replaced with ".".
   2.age column have few values marked as ERROR the count is 36
   and few values are out of range values are 12 in count
   Resolution :these values were marked as missing while cleaning the datasets.

   (d) What is the average age of participants in the study? [2 marks]

   The average age of participants in the study is 54.3965863

   (e) What is the average (across all patients) outcome at visits 1, 6 and 12? [2 marks]

   The average outcome across all the patient is 16.8456304

2. The daily sales for a firm are recorded each month in a file named SalesMonthYear.csv (where month is the text of the month and year is the year in numbers). The files contain the variables day (day of the month) and sales (daily sales in thousands of euro). The sales files for January 2018 (SalesJanuary2018.csv) and February 2019 (SalesFebruary2019.csv) are posted on Moodle.

   (a) Write a SAS macro that can be applied to either dataset to do the following activities: [16 marks]

   - Read the current month of data into SAS and give the dataset a name in SAS specific to the month and year.
   - Generate a line plot of sales versus day, including a title that is specific to the month and year.
   - Print out the highest daily sales from the month, including a title that is specific to the month and year.
   - Print out the highest daily sales from the month for weekdays and weekends separately, including a title that is specific to the month and year. Hint: recall that when the weekday function is applied to a date, it returns values 1 to 7, where 1 = Sunday, 2 = Monday, etc.

   The macro should include the calling variables month (the month in text), monthnum (the number of the month) and year (the number of the year).

   (b) Using the macro, answer the following: [4 marks]

   i. What is the maximum daily sales (over all days) for January 2018?

   The maximum daily sales (over all days) for January 2018 is 84.0000000

   ii. What is the maximum daily sales for weekdays only in January 2018?

   The maximum daily sales for weekdays only in January 2018 is 74.0000000

   iii. What is the maximum daily sales (over all days) for February 2019?

   The maximum daily sales (over all days) for February 2019 is 96.0000000

   iv. What is the maximum daily sales for weekend days only in February 2019?

   The maximum daily sales for weekend days only in February 2019 is 96.0000000