MA317 Group Coursework

Due in: 12pm(noon) Friday 26th March 2021, week 25

Submission of the project report: submit a copy via FASER.
The **same** report has to be submitted by **all** group members.
Oral presentations will take place on Thursday 25th and Friday 26th March (week 25). A schedule of the presentations will be sent via email and it will also become available in Moodle.

All members of each group should participate in the editing and writing of the submitted version of the project report and in the oral presentation (10 minutes each group). The allocation of the marks between the group members will be based on the written statement listing the contribution of each member of the group, which has to be included in the project report and the contribution of the members of the group at the oral presentation. ***Students are encouraged to equal contributions within groups.***

Suppose you work as a data analyst for an insurance company. You are asked to analyse a dataset of the World Development Indicators (WDI), which are derived from a primary World Bank database for development data from officially-recognized international sources. The dataset is available via moodle.

Task: Each group should investigate the response variable life expectancy in the year 2018 and use other indicators (predictor variables) of the dataset to develop a linear model which explains the life expectancies in 2018. The report should propose a model which explains life expectancy in the world for 2018. You should also discuss if and how the model can be used to predict life expectancies for countries which have not provided data on life expectancy. You should use R in order to conduct your statistical analysis. **You should include the R code as part of an Appendix of your report which should run without errors.** When answering the questions you should explain the statistical methods used and justify your answers. In order to analyse life expectancy complete the following tasks:

1. Analyse using descriptive statistics (both graphical and numerical representations) and R the LifeExpectancyData1.csv dataset. Generate an appropriate table as summary and appropriate graphs - for example boxplots. [10 marks]

2. Many predictors in the dataset contain missing values. Is a complete case analysis (that is, deleting predictor variables with many missing values) an appropriate method to deal with missing values? Choose a method to deal with the missing values. Justify your choice. [10 marks]

3. Collinearity increases the variance of the estimators and hence, reduces the adequacy of the model. When collinearity is present, how do you solve this problem? Investigate collinearity between the predictor variables in the LifeExpectancyData1.csv dataset. [10 marks]

4. To understand better life expectancy and the factors that affect it,

   (a) Suggest a model which predicts life expectancy in 2018. Justify you answer. [10 marks]

   (b) Evaluate the model you suggested in (a). [5 marks]

   (c) Can the model you suggested in (a) be used to predict life expectancy of other countries which have not provided data for 2018 life expectancy already? Employ your suggested model to predict the life expectancy of the countries in the LifeExpectancyData2.csv dataset. Provide your predictions in a separate text or Excel spreadsheet (csv) file. [5 marks]

5. Using the LifeExpectancyData1.csv dataset only, employ one-way Analysis of Variance (ANOVA) to study differences of average life expectancies across continents. Is a method like one-way ANOVA an appropriate for this data? What are the benefits of this method? To asnwer this question, use the following groupings for continents: Africa, Asia, Europe, North America, South America, Oceania. [10 marks]

The dataset includes the following worldbank indicator variables:

Code	Indicator Name
SP.DYN.LE00.IN	Life expectancy at birth, total (years)
EG.ELC.ACCS.ZS	Access to electricity (\% of population)
NY.ADJ.NNTY.KD.ZG	Adjusted net national income (annual \% growth)
NY.ADJ.NNTY.KD	Adjusted net national income (constant 2010 US\$)
SE.PRM.UNER.ZS	Children out of school (\% of primary school age)
SE.XPD.PRIM.ZS	Expenditure on primary education (\% of government expenditure on education)
SP.DYN.IMRT.IN	Mortality rate, infant (per 1,000 live births)
SE.ADT.LITR.ZS	Literacy rate, adult total (\% of people ages 15 and above)
SP.POP.GROW	Population growth (annual \%)
SP.POP.TOTL	Population, total
SE.PRM.CMPT.ZS	Primary completion rate, total (\% of relevant age group)
SH.XPD.CHEX.GD.ZS	Current health expenditure (\% of GDP)
SH.XPD.CHEX.PP.CD	Current health expenditure per capita, PPP (current international \$)
SL.UEM.TOTL.NE.ZS	Unemployment, total (\% of total labor force) (national estimate)
SP.DYN.AMRT.FE	Mortality rate, adult, female (per 1,000 female adults)
SP.DYN.AMRT.MA	Mortality rate, adult, male (per 1,000 male adults)
NY.GDP.MKTP.KD.ZG	GDP growth (annual \%)
NY.GDP.PCAP.PP.CD	GDP per capita, PPP (current international \$)
SP.DYN.CBRT.IN	Birth rate, crude (per 1,000 people)
NY.GNP.PCAP.PP.CD	GNI per capita, PPP (current international \$)
SL.EMP.TOTL.SP.ZS	Employment to population ratio, 15+, total (\%) (modeled ILO estimate)

General rules and hints:

- Follow the guideline: 'Writing Reports: a brief guide'.
- Plan and structure your work. Structure your report, for example: Page 1: cover page (title, your name, date, ...). Page 2: abstract, contents and word count. Pages 3-7: introduction; preliminary analysis; analysis; discussion; conclusion; references. Page 8-10: appendix: R-code with explanations, etc..
- Use R. Put all R code, which was necessary for your report in an appendix and explain your R code (add comments within the R code). Do not include R code of an analysis which is not used for your report. Make sure, that YOU wrote the R code (the use of some R code, without citing the source, can be viewed as plagiarism).
- Use an appropriate word processor (MS Word, Open office, ...) or type setter (Lyx, Latex,...).
- UG: The report can have a length of 1600 to 2400 words (without cover page and appendix). Not more than 8 pages without counting the cover page and the appendix. More than 2400 words or more than 8 pages (without counting the cover page and the appendix) will reduce the marking.
- PG: The report can have a length of 2400 to 3600 words (without cover page and appendix). Not more than 12 pages without counting the cover page and the appendix. More than 3600 words or more than 12 pages (without counting the cover page and the appendix) will reduce the marking.
- Use point size 12, Times New Roman; line spacing 1.5.
- UG: Do not use more than 6 figures and 4 tables within the main text. You may include further figures and tables into the appendix, if necessary.
- PG: Do not use more than 7 figures and 5 tables within the main text. You may include further figures and tables into the appendix, if necessary.
- In addition your report should include a clear account of any assumptions made in the analysis of the data.

Marking: Each project report and project presentation will be marked by two markers independently. The markers agree a final mark for each project. Thereafter, marks for individual students will be based on the mark for their group's project, on the written statement listing the contribution of each member of the group, which has to be included in the project report, and the contribution of the members of the group to the oral presentation. The markers reserve the right to make inquiries about the contributions of the members of the group if they feel they need to. If a member of the group contributes less than other members of the group, the markers will reduce the individual mark. If a member of the group contributes not at all, the individual mark will be zero.

Additionally, marks will be awarded as follows:

Report guide lines:

0 of 10: group did not follow the guide lines.

5 of 10: group followed the guide lines; but understanding of specific parts of the guidelines/report structure is weak; e.g. no table legends, citation style inappropriate, etc.

10 of 10: group followed the guide lines.

Tasks 1, 2, 3, 4(a), 4(b)& 4(c) and 5 (each):

0 of 10: is missing or makes no sense.
5 of 10: group describes a main analysis, which was suggested in the lectures and classes; tables and/or figures should support the results. The discussion and/or conclusion summarises the data analysis and result of the study.
10 of 10: group describes a main analysis, which includes justification of assumptions, provides further tables or figures which support the argument of the report. The discussion effectively communicates the results to the reader.

Oral presentation: The key fact for the markers to a presentation is that the student should demonstrate understanding. The presentation is an integral part of the assessment process. Failure to attend the presentation will result to a mark of 0 out of 30 for the presentation. The markers reserve the right to request an interview in addition to the presentation. Failure to attend the interview, if asked to do so is likely to have serious negative consequences.

0 of 30: No presentation.

10 of 30: A poor presentation of the data and lack of understanding of the results.

20 of 30: A clear presentation of the data and a good understanding of the results.

30 of 30: A clear presentation of the data and a comprehensive understanding of the results.