MA317 Group Coursework

Due in: 12pm(noon) Tuesday 22nd March 2022, week 25

Submission of the project report and presentation slides: submit a copy via FASER.
The **same** report and presentation slides have to be submitted by **all** group members.
All members of each group should participate in the editing and writing of the submitted version of the project report and in the presentation slides. 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. ***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 2019 and use other indicators (predictor variables) of the dataset to develop a linear model which explains the life expectancies in 2019. The report should propose a model which explains life expectancy in the world for 2019. 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.** You should submit your report in pdf format. Zipped folders e.g. .zip or 7z will not be accepted. 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 Life_Expectancy_data1.csv dataset. [14 marks]

2. Many predictors in the dataset contain missing values. 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 and then employ this method to the life expectancy data. Justify your choice. Additionally, there are some countries (cases) where the value of Life expectancy is missing. Explain how you will handle this problem. [14 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. [14 marks]

4. To understand better life expectancy and the factors that affect it, suggest the *best* model which predicts life expectancy in 2019. Evaluate the suggested model. [14 marks]

5. Using the same dataset (Life_Expectancy_data1.csv) and using the new additional feature `Continent`, employ an appropriate experimental design to study differences of average life expectancies across the continents : Asia, Europe, North America, South America, Africa, Australia/Oceania. Justify your choice of experimental design and methods. [14 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.PC.KD.ZG	Adjusted net national income per capita (annual \% growth)
SH.HIV.INCD.14	Children (ages 0-14) newly infected with HIV
SE.PRM.UNER	Children out of school, primary
SE.PRM.CUAT.ZS	Educational attainment, at least completed primary, population 25+ years, total (\%) (cumulative)
SE.TER.CUAT.BA.ZS	Educational attainment, at least Bachelor's or equivalent, population 25+, total (\%) (cumulative)
SP.DYN.IMRT.IN	Mortality rate, infant (per 1,000 live births)
SE.PRM.CMPT.ZS	Primary completion rate, total (\% of relevant age group)
SE.ADT.LITR.ZS	Literacy rate, adult total (\% of people ages 15 and above)
FR.INR.RINR	Real interest rate (\%)
SP.POP.GROW	Population growth (annual \%)
EN.POP.DNST	Population density (people per sq. km of land area)
SP.POP.TOTL	Population, total
SH.XPD.CHEX.PP.CD	Current health expenditure per capita, PPP (current international \$)
SH.XPD.CHEX.GD.ZS	Current health expenditure (\% of GDP)
SL.UEM.TOTL.NE.ZS	Unemployment, total (\% of total labor force) (national estimate)
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)
EG.FEC.RNEW.ZS	Renewable energy consumption (\% of total final energy consumption)
SH.HIV.INCD	Adults (ages 15-49) newly infected with HIV
SH.H2O.SMDW.ZS	People using safely managed drinking water services (\% of population)
SI.POV.LMIC	Poverty headcount ratio at \$3.20 a day (2011 PPP) (\% of population)
SE.COM.DURS	Compulsory education, duration (years)

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,...).
- The report can have a length of 2000 to 3000 words (without cover page and appendix). Not more than 12 pages without counting the cover page and the appendix. More than 3000 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.
- PG: Do not use more than 10 figures and 4 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: 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 prepare the presentation slides. 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 guide-lines/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 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.

Presentation slides: The key fact for the markers to the presentation slides is that the group has effectively chosen what to include in the presentation. **The maximum number of presentation slides is 10 (without the title page and 'Thank you' page).** The markers reserve the right to request an interview in addition to the presentation slides. Failure to attend the interview, if asked to do so is likely to have serious negative consequences.

0 of 20: No presentation slides.

10 of 20: A poor presentation of the data and lack of understanding of the results.

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