Final Assignment: World Health, Nutrition, and Population Statistics

FEM11149 - Introduction to Data Science

Instructor: dr. A. Tetereva
TAs: L. de Wit & M. Praum

October 2025

Introduction

This is an instruction document for the final assignment of the Introduction to Data Science course from the Data Science and Marketing Analytics Master program at the Erasmus School of Economics.
In this final assignment you are not going to be given step-by-step instructions. You are expected to know which techniques are needed to clean and subset data (if needed), run models and their respective diagnostics.

What to deliver

- A PDF file, generated using R Markdown containing
  - A business report (max 4 pages) with your analysis
  - The report should follow the guidelines specified on Canvas. You can add your code as Appendix, where you should aim at not using more than two pages for this. Use the RMarkdown pdf template for code.
  - You are allowed to have an Appendix for tables and/or figures, where you should aim at not using more than one page for this.
- A *.rmd file, used to generate the PDF file above.

You are not allowed to change margins or font size of the 'standard' RMarkdown template, nor are you allowed to have a cover page.

The Assignment

Congratulations! Because of your hard work for in the previous two projects, you are now a worldwide known data scientist. A consortium of several countries is looking for a consultant for a data science job.

In this assignment, we will study the relationship between variables in the Health, Nutrition, and Population Statistics database, from the World Bank (available at `https://databank.worldbank.org/home`). In this dataset, key health, nutrition and population variables are gathered from a variety of international sources. The variable that we are mainly interested in is that of the life expectancy at birth.

You receive three datasets for your work. The first is different for each individual and contains observations of 30 variables for 160 countries. The second dataset contains the variable of interest for all countries: **life expectancy at birth, total (years)**. The third dataset contains observations of 30 variables for 3 countries for which you have to predict the life expectancy at birth.[1]

In the analysis, you will focus on constructing prediction models for the life expectancy variable. Note that this is a first exploratory analysis, and causal questions such as 'what are the determinants of the life

[1] All three datasets are pre-processed as follows: for each variable, we used the most recent observation from the past four years. If there are no observations available over this time period, you find a NaN in your data.

expectancy' cannot be answered with simple models like the ones seen in this course. However, this does **not** mean that you cannot comment at all on which variables seem to be correlated to each other, and to the variable of interest. In short, try to formulate your research question accordingly to the data available and the models that are being used.

Minimum Requirements

You need to investigate what characteristics are related to the life expectancy. For that, you will **construct and compare** two regression models:

- A PCA regression model.
- A penalized regression model using elastic net.
- You also have to report predictions for the countries in **predictions.csv**

For both models, use the best practices you saw in the lectures.

Specifically for PCA, in addition to the three simple criteria above, please use the permutation test to select a meaningful number of principal components. Moreover, apply a bootstrap procedure to Kaiser's rule, i.e. test if the variance explained by each component is significantly larger than 1 or/and apply a bootstrap procedure to VAF, i.e. test if the variance explained by selected components explains at least 70% of the total variance. Use the results of your analysis to name and interpret the selected components.

Note that those are partial requirements and are not sufficient for a full grade.

Everything should be explained and interpreted. Single results without interpretation will not be considered. Be aware that different criteria for selecting the principal components might differ in their conclusions, and the final decision is up to you and need to be justified.

This is an individual assignment, and all students have different datasets.