MATH3030/4068: Coursework, Spring 2024

Prof. R Wilkinson

18/03/2024

- This coursework is ASSESSED and is worth 20% of the total module mark for MATH3030 and MATH4068.
- **Deadline**: Coursework should be submitted via the coursework submission area on the Moodle page by **Thursday 2 May, 10am**.
- Do not spend more time on this project than it merits - it is only worth 20% of the module mark.
- **Format**: Please submit a single pdf document. The easiest way to this is to use R Markdown or Quarto in R Studio. **Do not submit raw markdown or R code** - raw code (i.e. with no output, plots, analysis etc) will receive a mark of 0.
- As this work is assessed, your submission must be entirely your own work (see the University's policy on Academic Misconduct).
- Submissions up to five working days late will be subject to a penalty of 5% of the maximum mark per working day. Deadline extensions due to Support Plans and Extenuating Circumstances can be requested according to School and University policies, as applicable to this module. Because of these policies, solutions (where appropriate) and feedback cannot normally be released earlier than 10 working days after the main cohort submission deadline.
- **Report length**: Your report should not be too long. You should aim to convey the important details in a way which is easy to follow, but not excessively long. Think about your reader, and try to help them quickly understand the key points. Avoid repetition and long print-outs of uninteresting numerical output.
- Please post any questions about the coursework on the Piazza discussion boards. This will ensure that all students receive the same level of support. Please be careful not to ask anything on the discussion boards that reveals any part of your solution to other students.
- I will be available to discuss the coursework at our Tuesday or Thursday sessions during the semester (our regular drop-in session). I will not be meeting students 1-1 to discuss the coursework outside of these times.

Plagiarism and Academic Misconduct For all assessed coursework it is important that you submit your own work. Some information about plagiarism is given on the Moodle webpage.

Grading The coursework will be marked out of 20:

- 10 marks for technical content, use of R, and appropriate methods
- 10 marks for presentation and interpretation of results.

Coursework

The file `UN.csv` is available on Moodle, and contains data from the United Nations about 141 different countries from 1952 to 2007. This includes the GDP per capita, the life expectancy, and the population.

Load the data into R, and extract the three different types of measurement using the commands below:

```
UN <- read.csv('UN.csv')
gdp <- UN[,3:14] # The GDP per capita.
years <- seq(1952, 2007,5)
colnames(gdp) <- years
rownames(gdp) <- UN[,2]

lifeExp <- UN[,15:26] # the life expectancy
colnames(lifeExp) <- years
rownames(lifeExp) <- UN[,2]

popn <- UN[,27:38] # the population size
colnames(popn) <- years
rownames(popn) <- UN[,2]
```

In this project, you will analyse this data using the methods we have looked at during the module.

Exploratory data analysis

Begin by creating some basic exploratory data analysis plots, showing how the three variables (GDP, life expectancy, population) have changed over the past 70 years. For example, you could show should how the average life expectancy and GDP per capita for each continent has changed through time. Note that there are many different things you could try - please pick a small number of plots which you think are most informative.

Principal component analysis

Carry out principal component analysis on the three different variables. It makes sense here to look at each variable type on its own (i.e. do PCA on `gdp`, then on life-expectancy etc, rather than doing PCA on the entire UN dataset). Things to consider include whether you use the sample covariance or correlation matrix, how many principal components you would choose to retain in your analysis, and interpretation of the leading principal components.

Use your analysis to produce scatter plots of the PC scores for gdp and life expectancy, labelling the names of the countries and colouring the data points by continent. You can also plot the first PC score for life expectancy against the first PC score for GDP (again colouring and labelling your plot). Briefly discuss these plots, explaining what they illustrate for particular countries.

Canonical correlation analysis

Perform CCA using log(gdp) and life expectancy as the two sets of variables. Provide a scatter plot of the first pair of CC variables, labelling and colouring the points. What do you conclude from your canonical correlation analysis? What has been the effect of using `log(gdp)` rather than `gdp` as used in the PCA?

Multidimensional scaling

Perform multidimensional scaling using the combined dataset of log(GDP), life expectancy, and log(popn), i.e., using

```
UN.transformed <- cbind(log(UN[,3:14]), UN[,15:26], log(UN[,27:38]))
```

Find and plot a 2-dimensional representation of the data. As before, colour each data point by the continent it is on. Discuss the story told by this plot in comparison with what you have found previously.

Linear discriminant analysis

Use linear discriminant analysis to train a classifier to predict the continent of each country using `gdp`, `lifeExp`, and `popn` from 1952-2007. Test the accuracy of your model by randomly splitting the data into test and training sets, and calculate the predictive accuracy on the test set.

Clustering

Apply a selection of clustering methods to the GDP and life expectancy data. Choose an appropriate number of clusters using a suitable method, and discuss your results. For example, do different methods find similar clusters, is there a natural interpretation for the clusters etc? Note that you might want to consider scaling the data before applying any method.

```
UN.scaled <- UN[,1:26]
UN.scaled[,3:26] <- scale(UN[,3:26])
```

Linear regression

Finally, we will look at whether the life expectancy in 2007 for each country can be predicted by a country's GDP over the previous 55 years. Build a model to predict the life expectancy of a country in 2007 from its GDP values (or from `log(gdp)`). Explain your choice of regression method, and assess its accuracy. You may want to compare several different regression methods, and assess whether it is better to use the raw gdp values or `log(gdp)` as the predictors.