Mini-project 1

S681

Upload your initial submission through the Assignments tab on Canvas by 11:59 pm, Thursday 3rd October.

Submit this project in pairs. Register your group using the "Mini-project 1 groups" tab on Canvas (or ask me or a TA to do it.) If you want to work in a group of size $\neq 2$, ask me for permission. (You can of course discuss and consult with people outside your pair.)

A researcher for a thinktank wants to learn about life expectancy and its relationship to income.He notices the Gapminder website (`www.gapminder.org/data`) has data on life expectancy and income per person (measured as GDP per capita, adjusted for inflation), among many other indicators. He has taken an introductory statistics course using R, but that was a long time ago, so he is outsourcing the exploratory data analysis to YOU.

Data

Go to the Gapminder website and download data for the following variables:

- Life expectancy (years)
- Income per person (GDP/capita, PPP$ inflation-adjusted)
- Population, total

For questions about "continents," use the 5-continent definition used by Gapminder. (You might to merge your data with the data in the `gapminder` R package if you don't want to define this manually; it's okay if you lose some countries in the merge.)

Questions

The researcher's major research question is: **Can the increase in life expectancy since World War 2 be largely explained by increases in GDP per capita?** However, he recognizes this question may be difficult to answer, at least straight away. So he has brainstormed a series of questions he would like you to address, which can be divided into three groups:

1. **GDP and life expectancy in 2018:** How does life expectancy vary with GDP per capita in 2018? Can the trends be well-described by a simple model such as a linear model, or is a more complicated model required? Is the pattern the same or different for every continent? If some continents are different, which ones, and how is the relationship different in those continents?

2. **Life expectancy over time by continent:** How has average life expectancy changed since World War 2 in each continent? Have some continents caught up (at least partially) to others? If so, is this just because of some countries in the continent, or is it more general? Have the changes been linear, or has it been faster/slower in some periods for some continents? What might explain periods of faster/slower change?

3. **Changes in the relationship between GDP and life expectancy over time:** How has the relationship between GDP and life expectancy changed in each continent since World War 2? Can changes in life expectancy be entirely explained by changes in GDP per capita? Does it look like there's a time effect on life expectancy in addition to a GDP effect? Has there been "convergence" in the sense that perhaps GDP and/or continent don't matter as much as they used to? Are there exceptions to the general patterns?

Write a report of no more than six pages, including graphs, for the researcher. The third set of questions is the deepest and will probably require the most attention. Note that some of these questions may not have definitive answers; the researcher recognizes this.

Some constraints:

- The researcher is familiar with elementary methods like linear models, but not with non-parametric methods such as loess and gam. That means that if you want to use those more fancy models, you need to briefly describe what those techniques are doing in words that a non-statistician can understand.
- He is comfortable with transformations, but they would have to be interpretable.
- He took his statistics course from a fairly skeptical lecturer, so he knows all models are wrong. However, he is willing to accept some wrongness in exchange for a simple description of the data.
- He doesn't need to see the R code, but wants to be able to reproduce your work if required.
- The researcher has noticed that student reports on complex real-world phenomena occasionally (accidentally one hopes) say offensive things, and would prefer if you didn't do that.

What to submit

Your initial submission should consist of:

- A report (PDF preferred) of **no more than six pages**, excluding appendices.
- A .Rmd or other file containing your code.
- Any other supplementary files required to reproduce your work.

We will give you feedback, then you will make a final submission by a date to be announced. (You may not get a long turnaround, so it would be prudent to make your initial submission fairly polished.) The grade for your final submission will be the one that counts.

Notes

- There is no one objectively right answer to either part of the project (but there are infinitely many subjectively bad answers.)
- Make sure you justify your answers to the questions (don't just state answers.)
- When analyzing average life expectancy by continent, you should do a weighted average (since there are a lot more people in China than in Bahrain.) You can find this using existing R functions or you can write your own code.
- There aren't many countries in Oceania, so it may not be possible to fit complex models for that continent. You may drop that continent from your analyses should you find that necessary (but only where necessary.)
- It may or may not be worth doing an in-depth examination of one particular continent, to get a feel for the variation of trends within a continent.
- You do not necessarily need one overall model that describes all the data.
- Because there's no correct model, you're free to use multiple models for the same data and question, if you feel that's a good use of your time and page count.
- All the data in Gapminder is estimated. It is certainly possible that some countries fudge their official statistics for their own benefit.
- A large fraction of the points are for communication, so maintain a decent level of professionalism.
- Additional technical graphs such as residual plots can be included in an appendix, which will not count toward the six page limit and which we might not bother to read. Submit your code as a separate file. Also upload any additional sources required to reproduce your work.

Grading

- First set of questions: 5 points
- Second set of questions: 5 points
- Third set of questions: 10 points
- Communication: 10 points. Full credit for communication requires a readable, informative, comprehensive, clearly labeled set of graphs, and a comprehensible write-up with few glaring spelling and grammatical errors that makes the main points of the analysis clear.