Mini-project 1

S681

Upload your project through the Assignments tab on Canvas by 11:59 pm, Sunday 10th February.

You may (but are not obligated to) submit this project in pairs. If you want to work in a pair, email me to let me know by Sunday 3rd Februrary.

Part 1: Apples and oranges (20 points.)

A researcher for the Orange Research Council wants to take a survey to find out about how much people are willing to pay for a pound of oranges. He has data from an old study (in the file `oranges.txt`), which contains the answers of 150 respondents to the following questions:

- "Do you prefer apples or oranges?" (Allowable answers: Apples, Oranges.)
- "How much would you be willing to pay for a pound of oranges?" (Answers are in dollars, with a minimum of one cent.)

He hypothesizes that people who prefer oranges will typically be willing to pay more for a pound of oranges than people who prefer apples. He wants to take a new sample that will be able to show this with 80% power, and asks you what test to do and how large a sample to take. (Note that there are no separate lists of apple fans and orange fans: all individuals must be sampled from the general population.)

Write a plan for the researcher. The researcher does not know any nonparametric tests, so if you propose using one, you'll have to explain what it is, and why you think it should be used, to the researcher.

Part 2: Life expectancy (30 points.)

A researcher for a thinktank wants to learn about life expectancy and its relationship to GDP per capita. He notices there is an R package called `gapminder` that contains a data set of the same name giving the GDP per capita (adjusted for inflation) and life expectancy in 142 countries for a selection of years from 1952 to 2007. 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.

His 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 2007:** How does life expectancy vary with GDP per capita in 2007? 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 over time 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? 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 explain 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.

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.)

For Part 2 in particular:

- 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 are only two 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.
- If you want more data, the website `www.gapminder.org/data` has lots of it.
- Points will be given for presentation, 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 for Part 2 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.

What to submit

- A PDF or other file containing your plan for Part 1.
- A PDF or other file containing your report for Part 2.
- A .Rmd or other file containing your code.
- Any other supplementary files required to reproduce your work.