Stat 470/670 Mini Project: Life Expectancy

Due date: Friday, March 3, 5pm.

Assignment: A researcher for a think tank 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:

- 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? Can differences between continents be simply described by an additive or multiplicative shift, or is it more complicated than that?
- 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?
- 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 it used to? Are there exceptions

to the general patterns?

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.

Constraints:

- The researcher is familiar with elementary methods like linear models, but not with nonparametric methods such as loess. 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.

Notes:

- It's EDA, so there is rarely one objectively right answer (but there are infintely many subjectively bad answers).
- Make sure you justify your answers to the questions (don't just state answers).
- The third set of questions involves three quantitative variables, so you might want to look ahead to the trivariate chapter in Cleveland.
- 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.)
- 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, www.gapminder.org/data has lots of it.

Grading: Submit a report as a PDF, the body of which should be no more than six pages, including graphs. 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.

There will be lots of points for presentation. This includes readable graphs and bothering to spell-check, but mostly concerns making your analysis seem coherent and useful and not just bunch of random models and graphs strung together. An introduction and conclusion would probably help.

Point distribution:

- 5 points for the first set of questions.

- 5 points for the second set of questions.

- 10 points for the third set of questions.

- 10 points for presentation.