SOC 4930/5050: PS-08 - Correlation

Christopher Prener, Ph.D.

November 6th, 2017

Directions

Please complete all steps below. Your well-formatted R Notebook source (the `.Rmd` file) and `html` output along with your LaTeX `pdf` output should be uploaded to your GitHub assignment repository by 4:15pm on **Wednesday, November 15th, 2017**. You will need to have the package `gapminder` installed to access the data for this assignment.

Part 1: Data Preparation

1. Using the data table `gapminder` in the `gapminder` package, create a new data frame that has *only* the following data:
   (a) contains only data for the year 2002,
   (b) contains the country variable,
   (c) contains the continent variable,
   (d) contains a binary variable that is `TRUE` for Asian countries,
   (e) contains a binary variable that is `TRUE` for African countries,
   (f) contains a binary variable that is `TRUE` for countries in the Americas,
   (g) contains a binary variable that is `TRUE` for European countries,
   (h) contains the variable `lifeExp`,
   (i) and contains a version of the variable `gdpPercap` renamed to `gdpPerCap`.

Part 2: Assumption Tests

Using the life expectancy data created above in Part 1, answer the following questions.

2. Report the *appropriate* descriptive statistics for each of the binary variables created in Part 1.
3. Report the *appropriate* descriptive statistics for the variable `lifeExp`.

4. Report the *appropriate* descriptive statistics for the variable `gdpPerCap`.
5. Using a scatter plot, compare the relationship between `lifeExp` and `gdpPerCap` - does it appear to be linear?
6. Using a scatter plot, look at the relationship between `lifeExp` and `gdpPerCap` and assess whether Simpson's paradox appears to be a concern based on continental groupings.
7. Summarize your assessment of how these data meet the assumptions of Pearson's *r*.

Part 3: Pearson's r

Using the life expectancy data created above in Part 1, answer the following questions.

8. Create an appropriately structured[1] correlation matrix in *r* using the `corrTable()` function.
9. Write a paragraph or two summarizing the statistically significant relationships in the correlation matrix. Be sure to report all necessary statistical data when discussing individual relationships.
10. Create a LaTeX version of your correlation matrix. You *do not* have to make the detailed changes to the table that we discussed in class if you do not want to.

[1] *Hint:* Think about missing data!