SOC 4015/5050: PS-05 - Correlation

Christopher Prener, Ph.D.

Fall 2018

Directions

Please complete all steps below. All work should be uploaded to your GitHub assignment repository by 4:15pm on Monday, November 12th, 2018.

Analysis Development

Using RStudio and your operating system's file manager, create an R Project in the *existing* directory in your assignments repository named `Lab-10`. Add a `README.md` file, notebook, and all necessary folders before beginning.[1]

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.

[1] This initial section follows the project workflow that is available in the `lecture-03` repo!

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? Export your plot to the `results/` subdirectory.

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. Export your plot to the `results/` subdirectory.

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[2] correlation table in *r* using the `corrTable()` function. Make sure you write a copy of the table to a `.csv` file, and that you use `knitr::kable()` to ensure that the table in your markdown file prints in an organized fashion.

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.

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