"Essay" 2: World Health Data

PS 270: Understanding Political Numbers

Due Wednesday, April 10, 2019

This assignment will give you an opportunity to practice analyzing data and, crucially, writing up the results. Analysis tasks will include (1) a linear regression, (2) a non-linear regression, and (3) a multiple regression. It uses data from the World Health Organization (WHO). A `csv` data file and a `pdf` codebook are both available on Canvas.

Your final product will include an `R` script and a written report of your results (uploaded in `pdf` format). The script and the report are complements to each other; the `R` file will contain content that doesn't make it into the report, and the report contains writing that shouldn't be in the `R` script. The instructions below ask for specifics of what to include in the ***writing***, but your script will naturally contain more work than just what makes it into the final report (including data transformation, model checking, and so on). You could also be thinking about things that could be important to include but that I don't explicitly ask for. Writing about data analysis requires you to make judgment calls. Do not over-include every little detail, and do not write me a narrative play-by-play of your thought process.[1]

Divide your report into three sections corresponding to the section headings below. The report will probably be about 3 pages, double-spaced (not counting graphics and tables).[2] Don't worry about weaving seamless transitions between sections, but your writing within each section should be clear and easy to understand. Your writing is "showing your work" about the data analysis process. It should be clear what you are asking of the data, what you are doing (graphically, statistically) to answer that question, and what criteria you use to form your conclusions about the data.

Getting started Once you have downloaded the data and codebook, open a fresh session of `R`. Load the `tidyverse`, `here`, `broom`, and `stargazer` packages. After you import the data into R, rename the variables in the codebook so they are easier to work with. Trim extraneous variables out of the dataset; keep only your renamed variables, plus `country` and `regionname`.

[1] Tell me what you did, but don't overshare everything that you are thinking. First-person writing is allowed as long as it is not abused. A good example: "Figure 1 suggests that the variables are logarithmically related, so I estimated the regression where y is a function of $\log(x)$." A bad example: "When I made the plot, I thought, *hmm, that doesn't look like a straight line*, and so I thought I would try transforming the x variable using…"

[2] You can create regression summary tables using `stargazer(model_name, type = "text")` and paste the results into your word processing program.

1 Linear relationships

We sometimes hear about periods in history or places around the world where life expectancy is something like 40 years of age. These figures don't come from people not living past 40; they are usually dragged down by high infant mortality (but also wars).[3]

Evaluate the relationship between life expectancy (y) and infant mortality (x). Create a scatterplot, estimate a linear regression, and interpret both in the writing. Report the estimated equation, and interpret what the intercept and slope tell you about $\hat{y}$. Is infant mortality's relationship to life expectancy statistically significant? How do you know?

2 Nonlinear relationship

Plot the relationship between life expectancy (y) and GNI per capita (x). The relationship should appear logarithmic. Why does it make sense that the relationship is logarithmic?

Estimate a regression *using the appropriate transformation of GNI per capita*, interpret the estimated regression equation. Create a plot that contains both the raw data and the model's predicted values (from the `augment()` function). As long as x is on the log scale, it's hard to interpret, so don't plot x on the log scale. This means you probably need to create a new variable after `augment()` that transforms logged GNI back to its original scale for plotting.

3 Multiple regression

Plot life expectancy (y) against health expenditures as a percentage of GDP (x). You would think this relationship would be positive and fairly strong. What might be happening instead?

Estimate a model where life expectancy is a function of health expenditures as a percentage of GDP, adding GNI per capita (using the appropriate transformation) and infant mortality as control variables. Interpret each coefficient and whether these partial relationships are statistically significant. For your final plot, show the model's predicted values of life expectancy as a function of the health care percentage of GDP, holding all other variables at their means.[4]

[3] For example, John Adams died at 90.

[4] If you want to challenge yourself: plot the confidence interval for $\hat{y}$ as well by creating a MOE using the `augment()` output. Lecture and section code both contain examples of this.