DSC530 Final Project

Joshua Ewer

Bellevue University

August 9, 2024

Statistical/Hypothetical Question

I recently experienced an unexpected death in the family, and it got me wondering if there are any specific variables that contribute to life expectancy more than others. I found several World Health Organization (WHO) datasets that were extremely large and full of sketchy data. Then, I found a sample dataset on kaggle that included some data cleaning to help reduce some of the unrealistic values in the raw dataset.

I chose this dataset to answer the question "What factors have the largest correlation with life expectancy?"

Outcome of EDA

The initial dataset was fairly well-tended, so it required minimal cleanup. There were several columns that were not useful in answering the question because they already answered the mortality likelihood. I found a positive linear relationship between the rates of disease vaccination (e.g. polio, measles) and life expectancy. I found a similar relationship between life expectancy and the expenditure on health by the government.

The regression model had an R-squared value of 0.644, which is better than random, but more time spent researching the different variables, or joining to an external dataset might have yielded results with higher accuracy.

What do you feel was missed during the analysis?

There were approximately 30 different bits of data that could have been analyzed, mostly categorized into medical, financial, and academic categories. I would have loved to spend more time focused on the financial aspects (country GDP, level of poverty), but, due to time, I focused mostly on medical information. You could make the assumption that the poverty and literacy

values are correlated with life expectancy, and I would have liked to spend some time analyzing that hypothesis.

I also had grand plans to do all my analysis with industry-standard tools as opposed to the thinkstats helpers. Unfortunately, I ran into difficulties, particularly in the CDF and analytical distribution, so I resorted to using those helpers. In the future, I would like to explore how to make that analysis without thinkstats code.

Were there any variables you felt could have helped in the analysis?

Inclusion of more lifestyle choices, such as diet and physical activity could be helpful. You can make an assumption that BMI is a coarse reflection of such choices, but, lacking the data, I could not prove that assumption. Also, I noticed the data was not broken down by gender. Separating genders would have been interesting, as, on average, women tend to live longer than men. It's possible that life expectancies could be skewed as a result of not separating statistics by gender.

Were there any assumptions made you felt were incorrect?

When looking at the features in the data set, and before doing any analysis, I initially assumed that the vaccination against disease (stats re: hepatitis, measles, polio, and diphtheria) would be strongly correlated, but that did not seem to be the case. I also assumed that there would be some meaningful difference between the different vaccinations, but it turns out they're all administered at the same time, so using multiple variables wasn't as useful as I had hoped.

What challenges did you face, what did you not fully understand?

The largest challenge was finding a dataset that was clean and well-explained so that I could focus on statistical analysis instead of data wrangling. Once I finally found one that was well-tended and

had explanations attached to it, the analysis became much easier. Regarding statistical analysis, I would have liked to spend more time trying to define the type of relationship between independent and dependent variables. I found no noticeably strong correlations with a specific dataset, so I do not know if that is a result of not doing a proper analysis, the data not being a representative sample, or if there are truly no strong correlations between variables.