Timothy Yeh

Professor Brewer

Word Count: 1441

Methodological Investigation: Data Science Methods Behind Assessing Malnutrition in Ethiopia

Introduction:

Making humans the center of human development requires that we focus on inherent, individual rights because it is these underlying principles that serve as avenues to help meet the end of goal of human development- a satisfactory and enjoyable life. Amartya Sen captures these ideas perfectly in his book discussing development as freedom, explaining that development is the process of expanding real freedoms for people to enjoy. He further explains that when an institutional boundary deprives them of these inherent freedoms, people no longer can make basic yet influential choices that both help them live enjoyable lives and ultimately create a society in which everyone is satisfied. To that end, I believe that a major factor that is necessary for people to attain these inherent freedoms is access to necessities. Without food and water to survive, people can't strive for that these freedoms, thereby preventing human development. Thus, throughout the scope of my research, I seek to assess the extent of malnutrition in Ethiopia while also investigating at the data science methods used to gather and analyze the data. To that end, I will discuss two data science methods I find to be particularly interesting, providing the audience with a deep insight into the actual complexity of the problem.

Analysis of Data Science Methods:

One data science method which I found to be quite prevalent in my research is the use a data analysis method called logistic regression. To start, it is important to understand what regression is and to do so we must look the most its most basic form: linear regression (refer to figure 1). At its most basic level, linear regression is a predictive analysis method that explains the relationship between an independent variable (x) and a dependent variable (f(x)) by fitting a line (the line of best fit) to the observed data in question. This allows one to estimate how a dependent variable change when an independent variable changes. Overall, the idea is to examine whether a set of predictor variables does a good job in predicting an outcome (dependent) variable. The simplest form of regression equation, which in this case is linear regression, is expressed by: y= c + bx [2]. Instead of assessing a best of line fit, logistic regression seeks to estimate a logarithmic S-shaped curve for the data. Moreover, this method is most often used to explain the relationship between one dependent binary variable (0 or 1), and multiple independent variables (weight and age, height and age, weight and height etc…), as opposed to a single of each like it is linear regression. Mathematically, logistic regression groups multiple linear regression models together, producing one coherent function (refer to figure 2) [4].

Example of Linear Regression

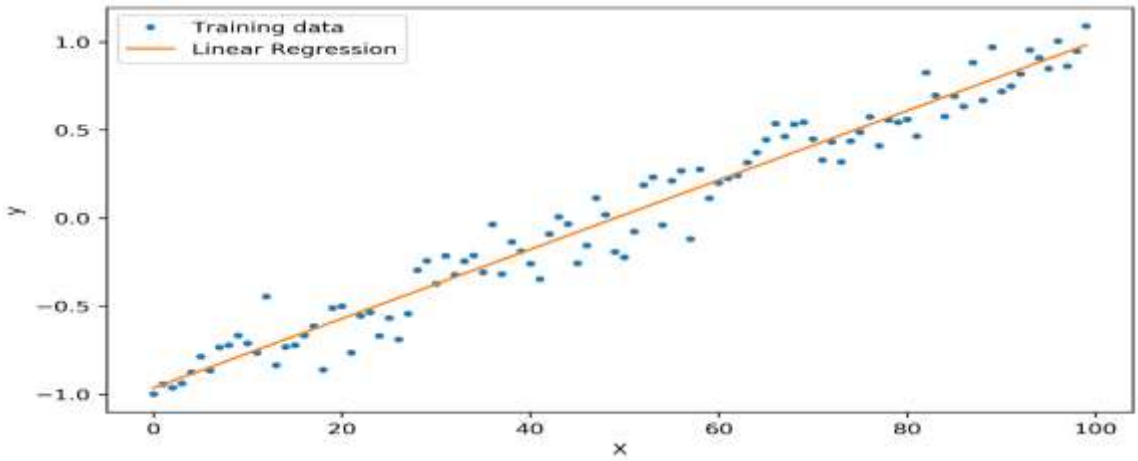

Figure 1: This figure depicts an example of a best of line fit for a given set of data (mlfromscratch.com)

Logistic Regression vs Linear Regression

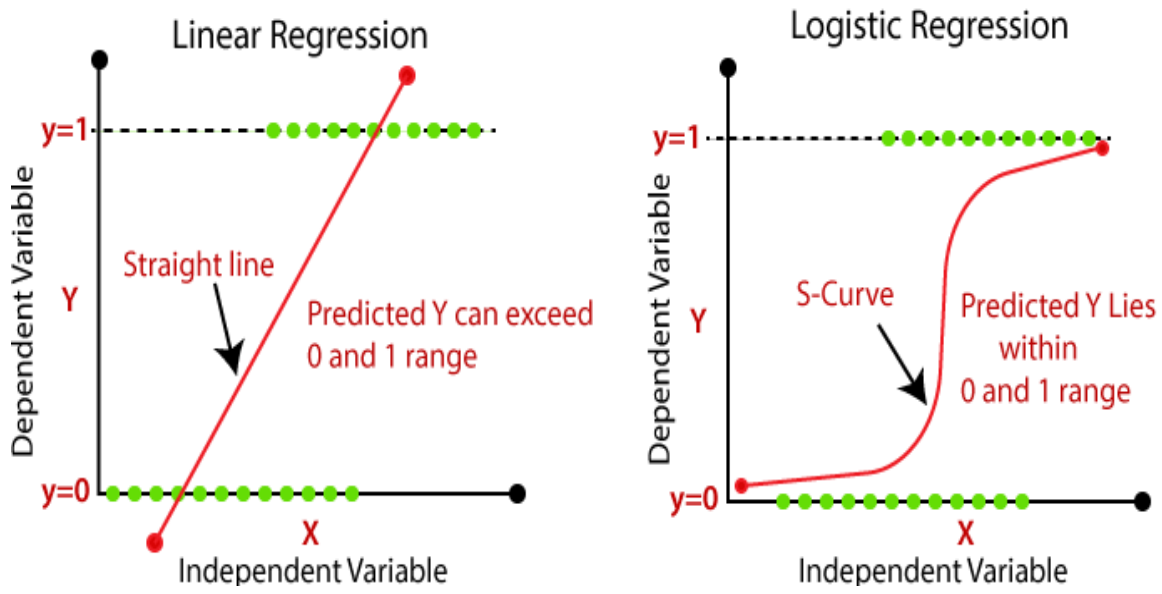

Figure 2: This figure depicts a normal linear and logistic curve. The main difference between the two is that a y value for the logistic curve is limited to a number between 0 and 1, whereas the linear curve is not limited to a certain y value (javapoint.com)

To illustrate the use of this data science method, let me give an example. In 2016, the Central Statistical Agency (CSA) of Ethiopia investigated the magnitude of malnutrition in Ethiopian children under the age of five. Using a 2016 height and weight EDH (Ethiopian Demographic and Health) survey of 9494 child and mother pairings (51.3% male, 48.7% female, 21.4% under 12 months), the agency measures the nutritional status of Ethiopian children based on certain undernutrition indicators that abide by the WHO child growth standard. The measurable indicators in use include stunting, wasting, and underweight, all of which are established based on certain weight and age ratios. The researchers first measure the height and weight a given child and then based on those measurements and its standard deviation ratios, categorize them into one of the three measurable indicators. It is possible for some children to be put in multiple categories others to be put in none. In relation to the logistic regression models, the independent variables are weight and height and the dependent, one of the three indicators.

To quantify the dependent variable (what is being measured, so in this case one of the three indicators), a child is assigned a "0" if he or she doesn't meet the standard for a given indicator and "1" if he or she does, hence the Thus, if a given child's height and weight meet the standard for "stunting" and not the rest, the corresponding y values are 1,0,0. If another child meets both stunting and underweight standards, the corresponding y values are 1,0,1. If a third child meets none of the standards, the y values are 0,0,0. The researchers then determine the best-fitted curves for each of three the regression models, allowing them to draw certain trends from the data. The results ultimately show that the prevalence of stunting was 38.3%, underweight, 23.3%, and wasting 10;1%. Seven out of fifteen (46.5%) suffer from at least on form of malnutrition while 3.1% suffered from all three kinds. Obviously, the explanation above is very bare bones because I truly still don't really have a good understanding of the methodology of this data science method. A deeper understanding requires an exploration of concepts which I believe are beyond the scope of this class. In total, logistic regression serves as a great data scientist to use when dealing dependent variables that are quantified using binary, providing insights into data that will help promote human development [3].

The second data science method I found to be interesting was a electronic search data collection method used in a study that attempted to assess the prevalence of undernutrition among Ethiopian HIV patients. Most data collection methods utilize surveys that seek to draw new data. However, in this study, authors use pre-existing data from previous studies, meaning that data collection involves gathering pre-existing sources for meta-analysis. To gather such data, the authors use electronic search via popular databases such as PubMed, Google Scholar, and Google, all while abiding to the standards set by the Preferred Reporting Items for Systematic Review and Meta-Analyses (PRISMA). Specifically, the authors search for studies

using key words such as “prevalence” OR “burden” AND “malnutrition” OR “undernutrition” OR “malnourishment” OR “underweight” AND “HIV-positive” OR “HIV-infected” AND “Adults” AND “Ethiopia”. By using conjunctions such as “AND” and “OR”, the authors can control how detailed they want the searches to be. The only sources excluded from collection are those which only can be exclusively assessed by the author. After obtaining 424 studies of “data”, they then use a standardized extraction software to extract data from each original study, presenting the data in the form of descriptive summaries and forest plots later to be analyzed [1].

For the second portion of data collection, the researchers’ goal is to now condense the 424 studies into 15 studies that will be used in meta-analysis and serve as the main data points for analysis. To do this, the researchers assess each source and exclude a given source based on the source’s title relevancy, abstract relevancy, and study locations. To aid in narrowing the sources, two researchers assess the quality of each source using Newcastle-Ottwa Scale, a three-step approach for quality assessment. The tool has three main components. The first component grades from five stars and mainly focuses on the methodological quality of each article., the second component grades from two stars and deals with the comparability of the study. The third component grades from three stars, primarily focusing on the outcomes and statistical analysis of each primary research. Articles with stars greater than six are classified as high-quality. In total, this electronic search data collection method and its corresponding subsequent processes provide the audience with a deeper into the lesser known forms of data collection and emphasizes the idea that a data science method doesn’t have to be complicated to be effectual [1].

Gaps in research and Topics for Future Investigation:

Based on my scope research, most research studies use data science to assess the prevalence of undernutrition. However, not enough of it is used to help develop the viable

solutions to malnutrition in. To that end, in the future, I hope to dive deeper into the solutions side of this issue and learn more about how data science is being used to help create solutions. Having synthesized multiple solutions, I hope to eventually determine which solution is best fitted for the given environment in Ethiopia.

Works Cited

[1] Alebel, A., Kibret, G.D., Petrucka, P. *et al.* Undernutrition among Ethiopian adults living with HIV: a meta-analysis. *BMC Nutr* **6,** 10 (2020). https://doi.org/10.1186/s40795-020-00334-x

[2] Bevans, R. (2020, October 26). *An Introduction to Simple Linear Regression*. scribbr.com. https://www.scribbr.com/statistics/simple-linear-regression/

[3] Kasaye HK, Bobo FT, Yilma MT, Woldie M (2019) Poor nutrition for under-five children from poor households in Ethiopia: Evidence from 2016 Demographic and Health Survey. PLoS ONE 14(12): e0225996. https://doi.org/10.1371/journal.pone.0225996

[4] Statistics Solutions. (n.d). *What is Logistic Regression?*. statisticssolution.com. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/