Marcus Chandra, Chris Low, Justin Zhou

Course: Data Science II

Professor: Amy Nussbaum

Date of Submission: May 10, 2023

Analyzing Global Health Factors to Predict Life Expectancy

Investigation:

This dataset is a combination of data from the Global Health Observatory (GHO) repository and the United Nations (UN) Website.[1] The author of this dataset, Kumar Rajarshi, used the Global Health Observatory Repository for all health-related parameters and economic data from the United Nations website for all expenditure and population-related parameters.

The GHO Repository is a reputable public access dataset, constantly updated by the World Health Organization. It reports thousands of different health indicators that are collected through case study reports and broad statistical surveys at health centers around the world.[2] For the UN dataset, data is collected on an annual basis directly from over 230 national statistics offices across member states.[3] Rajarshi selected 22 most relevant economic and health-related parameters from these two datasets in order to predict life expectancy. This dataset is important because it considers a broad scope of factors (economic, health, demographic) affecting life expectancy globally using data from recognized institutions. By analyzing it, researchers can discover which factors contribute to low life expectancy, and thus make recommendations to improve health outcomes in these countries.

---

[1] Kumar, Rajarshi. "Life Expectancy (WHO)." Kaggle, December 17, 2020. Accessed May 10, 2023. https://www.kaggle.com/kumarajarshi/life-expectancy-who.

[2] "Global Health Observatory (GHO) Data Repository." Top Masters in Public Health. Accessed May 10, 2023. https://www.topmastersinpublichealth.com/faq/what-is-the-gho-data-repository/#:~:text=The%20Global%20Health%20Observatory%20.

[3] "Demographic Statistics Database." United Nations Statistics Division. Accessed May 10, 2023. http://data.un.org/datamartinfo.aspx#:~:text=Over%20140%20reporter%20countries%20provide,by%20commodities%20and%20partner%20countries.

Literature Review:

One study that synthesized WHO data and United Nations Data is *2014 Global geographic analysis of mortality from ischaemic heart disease by country, age and income,* authored by Alexandra N. Nowbar et al.[4] In this study, Nowbar used WHO data to study the prevalence of heart disease across countries and used United Nations data to separate countries by GDP per capita.[5] They discovered that age-specific death rates for heart disease have been gradually declining over time, while total death rates have remained constant due to an aging population.[6] This information can help us understand the impact of economic development on health outcomes, which is relevant to our analysis of life expectancy.

Another study that synthesized data from the World Health Organization and the United Nations was *Global healthcare expenditure on diabetes for 2010 and 2030*, authored by Ping Zhang et al.[7] By analyzing data from the World Health Organization and the United Nations, they found that 12% of all health expenditures were expected to be spent on diabetes in 2010.[8] However, this amount varied across countries, with poorer countries spending the least on diabetes.[9] They concluded that more resources should be allocated to these countries for basic diabetes care.[10] In our project, we plan to investigate health spending as a parameter, so this study provides us valuable information on the effect health spending has on disease/mortality rates.

Table 1: Percentage of Missing Values for Each Variable

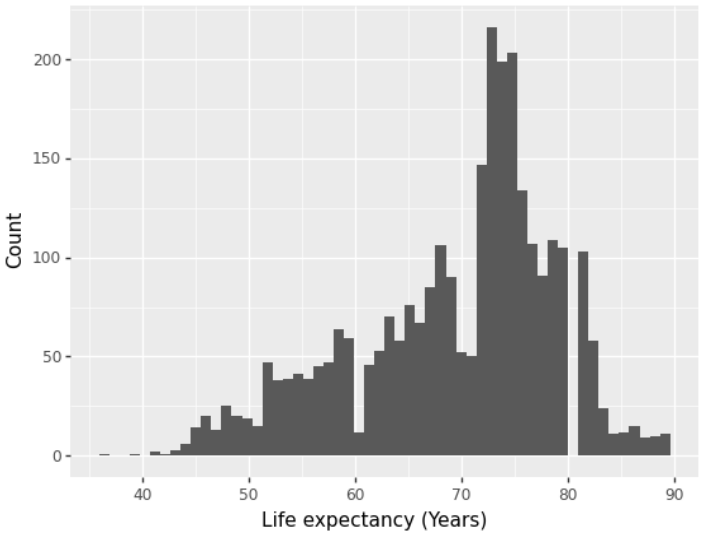
[4] Nowbar, A. N., Howard, J. P., Finegold, J. A., Asaria, P., & Francis, D. P. (2014). Global geographic analysis of mortality from ischaemic heart disease by country, age and income: Statistics from World Health Organisation and United Nations. International Journal of Cardiology, 174(2), 293-298. https://doi.org/10.1016/j.ijcard.2014.04.096
[5] Ibid.
[6] Ibid.
[7] Zhang, Ping, Xinzhi Zhang, Jonathan Brown, Dorte Vistisen, Richard Sicree, Jonathan Shaw, and Gregory Nichols. "Global healthcare expenditure on diabetes for 2010 and 2030." Diabetes Research and Clinical Practice 87, no. 3 (2010): 293-301. Accessed May 10, 2023. https://doi.org/10.1016/j.diabres.2010.01.026.
[8] Ibid.
[9] Ibid.
[10] Ibid.

Data Summary:

This dataset is huge, with numerous NaN cells for multiple variables. From Table 1, which shows the percentage of missing cells for each variable, we decided to examine variables with few missing values. It is important to find a balance between eliminating variables with missing values. We decided to eliminate variables up to "thinness 5-9 years," as the change in missing value percentage from 5% to 1% is significant. Exploratory data analysis was performed on all variables after "thinness 5-9 years" (based on Table 1). From the correlation matrix (not shown on report), "under-five deaths" and "infant deaths" (r = 0.997) are collinear as they both relate to deaths. Thinness 1-19 years and thinness 5-9 years (0.939) are also collinear because both are about thinness. Polio and diphtheria (0.668) has a high correlation relative to the other variables (but not high enough to indicate high collinearity). The rest of the variables have relatively low (< 0.5) correlation.

	% missing values
Population	22.191967
Hepatitis B	18.822328
GDP	15.248468
Total expenditure	7.692308
Alcohol	6.603131
Income composition of resources	5.684139
Schooling	5.547992
thinness 5-9 years	1.157250
thinness 1-19 years	1.157250
BMI	1.157250
Polio	0.646698
Diphtheria	0.646698
Life expectancy	0.340368
Adult Mortality	0.340368
HIV/AIDS	0.000000
Country	0.000000
Year	0.000000
Measles	0.000000
percentage expenditure	0.000000
infant deaths	0.000000
Status	0.000000
under-five deaths	0.000000

Figure 1: Life Expectancy Histogram



If we plot the response variable (Life Expectancy) histogram (Figure 1), we see that it is skewed. We therefore use logarithm to save the response variable. This is also a sign that regularization methods should be performed to improve the dataset. After fitting an OLS model (linear), we examine the scatterplot of residuals against predicted values. From the scatterplot of life expectancy against its residuals (Figure 2), we see that it is not randomly dispersed. At lower life expectancy values, only positive residuals are recorded. And as life expectancy increases, there are significantly more datasets that are dispersed in the center. This shows that a linear regression model without regularization may not be appropriate. Furthermore, upon examining the VIF, there are various high values, such as infant deaths and under five deaths (which is consistent with the correlation matrix). Other high values include Polio and Diphtheria (22), as well as both of the thinness (18).

Figure 2: Life Expectancy Residuals

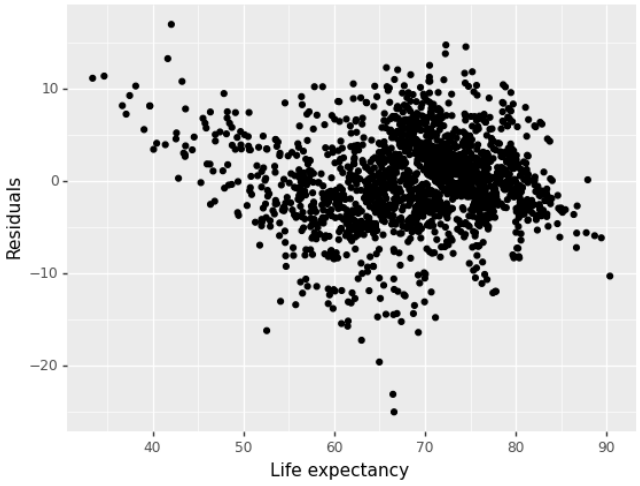
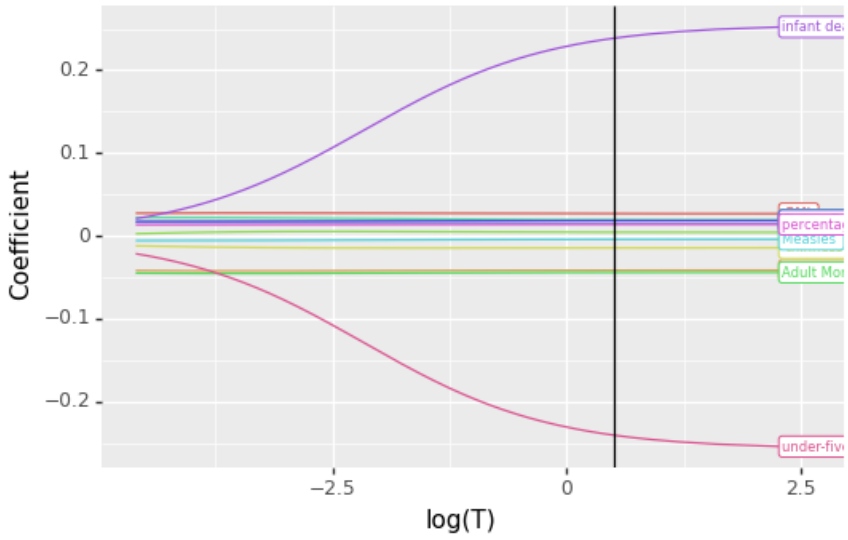


Table 2: VIF Values for All Variables

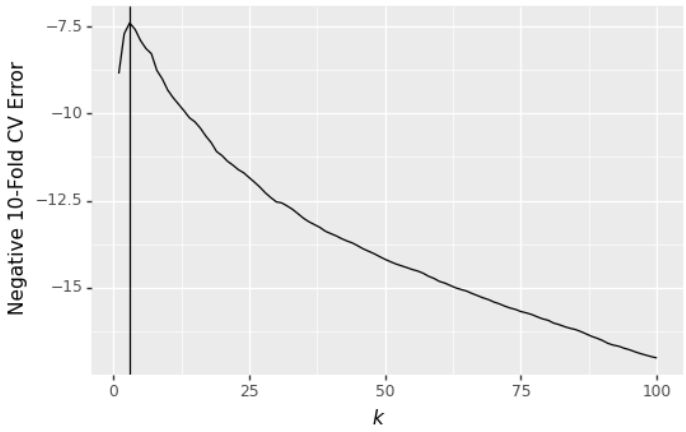
	Feature	VIF
0	Adult Mortality	3.672752
1	infant deaths	170.200175
2	percentage expenditure	1.268766
3	Measles	1.415208
4	BMI	5.823152
5	under-five deaths	172.558327
6	Polio	22.643803
7	Diphtheria	22.355594
8	HIV/AIDS	1.561002
9	thinness 1-19 years	18.729156
10	thinness 5-9 years	18.887266

We now look at three types of regularizations: ridge, kNN, neural network. We determined that ridge gives the model the best fit based on the highest (closest to 0) neg_mean_squared_error of cross_val_score.

Figure 3: Ridge Regression Graph for Tuning Parameters and Corresponding Coefficients



Using Ridge CV from sklearn with 5 fold cross validation, we determine the best alpha to be 0.6 (shown by the black vertical line). The most significant variable is “under-five deaths” (-0.240907), though its effects are negated by “infant death” (0.238711), which, again, is expected due to collinearity. The cross validation score (neg_mean_squared_error) is -0.005085 (nrmse = -0.0713).



For kNN regression, we test k values from 1-100. From the graph, we can conclude that at k =3, the negative 10-fold CV error is the lowest at -7.401981. This, however, is higher than the ridge error.

Figure 4: kNN: k vs Cross Validated Regression Error

Using neural networks, after fitting several different models we found an optimal model with one 10-unit hidden layer, using the linear activation function, and a batch size of 64, and epoch of 50.

It yielded a low negative root mean squared error of (-0.0737). We were reluctant to increase batch size and epoch further because of overfitting, and the improvement in error score was marginal. This approach performed well, but it still had a higher error than our linear ridge model, and the approach seems less appropriate for our relatively small dataset with 2888 observations.

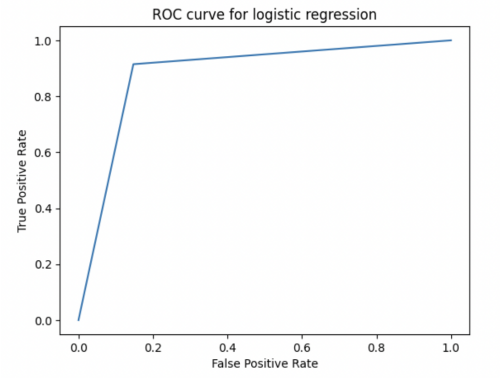


Figure 5: ROC Curve for Logistic Regression

We also experimented with a logistic approach. We took the average world life expectancy for each year, and assigned "low" life expectancy to observations that performed below their respective average. We first fit our logistic model without regularization (Figure 5), which resulted in a AUC of 0.884. We then tried different parameters like LASSO and Ridge penalties, which all yielded high AUC scores (around 0.90). These models performed well, but they are limited in that the output only tells us the likelihood of an observation having below average or above average life expectancy.

Comparing the cross validation scores, we decided ultimately that our linear ridge regression model was most appropriate for creating a Life Expectancy predictor from this data. Our final equation with an alpha of 0.6 is:

Life expectancy = *Adult Mortality* * -0.044287 + *Infant deaths* * 0.238711 + *percentage expenditure* * 0.013742 - *Measles* * 0.004334 + *BMI* * 0.026778 - *Under-five deaths**-0.240907 + *Polio* * 0.015004 + *Diphtheria* * 0.019319 - *HIV/AIDS* * 0.04161 - *Thinness 1-19 years* * 0.014641 + *Thinness 5-9 years* * 0.00426 + *Status_Developed* * 0.018438.

Groupwork statement: Chris performed exploratory data analysis and our KNN and linear ridge regressions, Justin undertook literature review and fitted our linear models, Marcus cleaned the initial dataset and performed neural network and logistic regressions.