COMP562 Final Project

Gaurachandra Das

November 22, 2020

For the final project, I decided to look at data from the Quality of Government (QOG) Basic Dataset from the University of Gothenburg in Sweden. This is a compilation dataset, and the variables were drawn from various academically recognized sources in order to collect a broad set of data with multiple areas of interest. The dataset is structured based on country and year, with various variables for each country, year combination. From this dataset I wanted to look at the Human Development Index variable (HDI), and generate a linear regression model to predict it based on various inputs. HDI is described by the QOG dataset as "the summary of the measure of average achievement in having a decent standard of living, being knowledgeable, and living a long and healthy life". By developing a model to predict this variable, there might be potential benefits in terms of legislature or policy creation to aim to maximize HDI by altering the inputs.

The variables I looked at as inputs were Gross Domestic Product (GDP), Bayesian Corruption Indicator (BCI), Global Peace Index (GPI), and Average Schooling Years (ASY). These are all variables one would not be surprised to find influenced HDI as they all describe something about the standard of living in their respective countries. In addition, a linear model derived from these values might tell us something about their weights in influencing HDI, something which could be of interest in determining what to focus on improving to improve a country's standard of living (which variable has the largest weight).

I first subsetted my data to the year 2010, as years outside of this data had missing values for some or all of the variables. This left 134 observations which I standardized as part of the preprocessing step, where I subtracted each observation by the mean and divided by the standard deviation. I then further split the data randomly into training and test sets with the training set having approximately 80% of observations (104) and the test set having the remaining 20% (30). I collected all the HDI's into an array I labeled Y, and the remaining variables (GDP, BCI, GPI, and ASY) into a 2D array label X with dimensions observations X features (104X4 and 30X4 for train and test respectively).

$$y|x = \beta_0 + \sum_j x_j\beta_j + \epsilon \tag{1}$$

The above equation is representative of the linear regression model, where we aim to solve for the βs. There are two approaches to doing so, one is through gradient ascent and the other is through solving the closed-form equation. For this case, I went with the closed-form equation, which led me to the equation below:

$$\begin{bmatrix}\beta_0^{MLE} \\ \beta^{MLE}\end{bmatrix} = (X_1^T X_1 + \text{diag}(\begin{bmatrix}0 \\ \lambda 1_p\end{bmatrix}))^{-1} X_1^T y \tag{2}$$

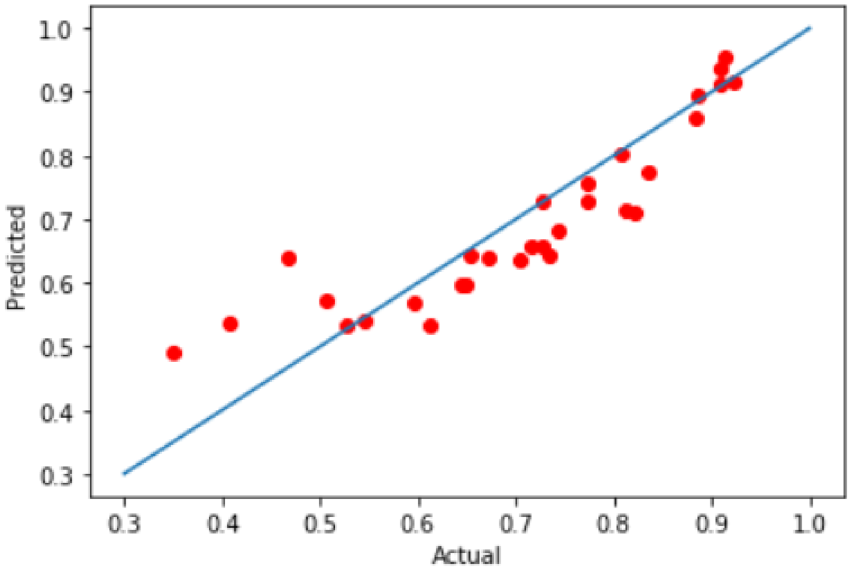
Plugging in the training set into this equation resulted in a set of βs that could be used to generate a prediction ($\hat{y}$) by plugging the βs and test predictors into the equation:

$$\hat{y} = \beta_0 + \sum_j x_j \beta_j \tag{3}$$

Afterwards, I computed the accuracy of my prediction by calculating the root mean-squared error of the predictions and the test results as such:

$$\text{RMSE} = \sqrt{\frac{\sum_n (y_n - \hat{y_n})^2}{N}} \tag{4}$$

I used this error to find the best lambda for the above ridge penalty, by looping over different lambdas from 1 to 25. The best lambda (which had the smallest RMSE) was a lambda of 12, which I used in my final model. I further confirmed my results by plotting my predictions vs the actual test results and observing the difference between a line through y=x as shown below.



The line y=x shows the optimal results that would occur if my predictions were the same as the actual test data. How far off the red points are from the blue line show my error. Looking at the graph, most of the points are quite close to the line, with only a few outliers. This can also be seen by the fact that my RMSE was approximately .4333, In conclusion, using the linear regression model we learned in class, I was able to train a simple model on data obtained from the QOG dataset on various factors of countries. The model I trained was used to predict HDI given 4 predictors; GDP, GPI, BCI, and ASY. I took into account a ridge penalty and tested various lambda values of which I found the value of 12 to provide the smallest error between predicted and actual values. In addtition, the βs I generated were -.163 for BCI, .218 for GDP, -.022 for GPI, .613 for ASY, and -.294e-15 for β_0. This implies to me that, improving ASY (average school years) seems to be the most weighted factor for improving HDI, with GDP being second most weighted. This also implies that BCI (Bayesian Corruption Indicator) is a factor one would want to decrease to improve HDI (which makes sense) and GPI (Global Peace Index) should also be decreased, which doesn't make as much sense but can be explained by the fact that a GPI of 1 is more peaceful than a GPI of 5 (ie GPI is lower for more peaceful countries). The fact that β_0 is quite meaningless is interesting, although that might be explained by that fact that HDI ranges from 0 to 1 so the mean wouldn't be large in the first place.

Overall, the fact that school years has the largest weight out of all the indicators (when standardized) is an interesting conclusion. This might be useful to look deeper into the relationship between education and standard of living and might be meaningful to legislators or lobbyists who want to argue for improving pay for educators and the public education system as the impact is quite visible, even from a simple model such as this.