COMP562 Final Project

Gaurachandra Das

November 21, 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 improve HDI by altering the input variables.

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. GDP relates to a country's economic status in terms of market value of the goods/services they produce. BCI is scaled from 0 to 1 and indicates the level of corruption perceived in the country obtained through surveys with 0 being no corruption and 1 being absolutely corrupt. GPI is scaled from 1 to 5 with 5 being least peaceful and 1 being most peaceful. It is measured based on global conflicts the country is involved in as well as domestic conflicts and militarization. Lastly, ASY is the average years of schooling adults (over 25) in the country posses. 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 for 5 variables of interest which I standardized as part of the preprocessing step: subtracting each observation by the mean and dividing 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 approx 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, we use. To reach this equation, we need 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 project, I went with the closed-form equation method, which led me to the equation below:

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

The above equation is the closed form solution for linear regression with a ridge penalty, which is used to help deal with an ill-posed problem. Plugging in the training set into this equation resulted in a set of βs. We could then generate a prediction ($\hat{y}$) by plugging the βs and test predictors into the following equation:

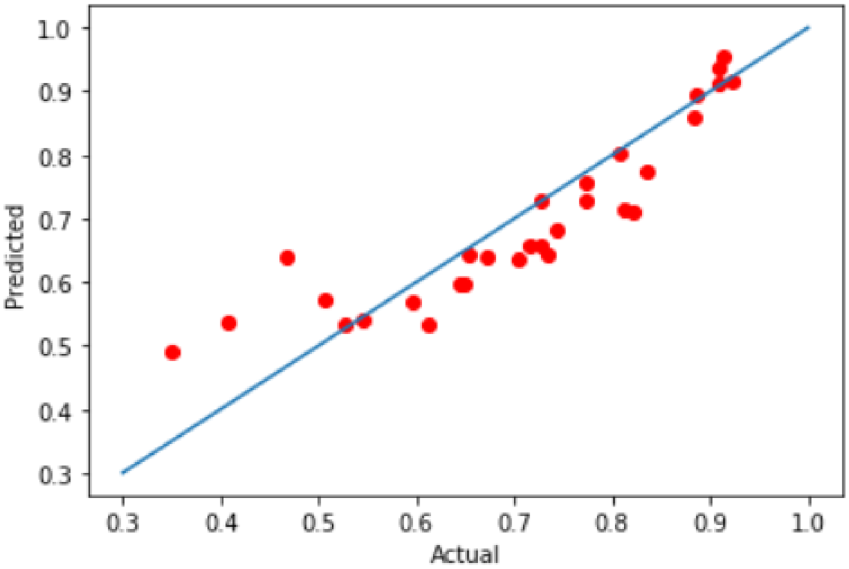
$$\hat{y} = \beta_0 + \sum_j x_j \beta_j \tag{3}$$

which is similar to the one above. After obtaining the prediction($\hat{y}$), I computed the accuracy by calculating the root mean-squared error of the predictions with the test results(y) as such:

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

where N is the number of observations in the testing set. I used this error to tune the hyperparameter λ from the above closed form ridge penalty equation, by looping over different s from 1 to 25 and keeping track of the λ associated with the minimum RMSE. The best λ I found was a λ value 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 differ between a line through y=x.

The line y=x is representative of the optimal case, that would occur if my predictions were the same as the actual test data. We can thus see how well the model performed by looking at how far off the actual, predicted pairs are from the blue line. Looking at the graph, most of the points are quite close to or on the line, with only a few outliers. This suggests that the model performs relatively well in predicting HDI given values for the input parameters. In addition the final RMSE was approximately .4333 which is a small value like we would hope to see,



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 λ values of which I found the value of 12 to provide the smallest error between predicted and actual values. In addition, 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 indicates it might be useful to look deeper into the relationship between education and standard of living. In addition, it might be meaningful to legislators or lobbyists to argue for improving pay for educators and the public education system as the impact on standard of living is quite visible, even from a simple model such as this.

Sources:

Dahlberg, Stefan, Sören Holmberg, Bo Rothstein, Natalia Alvarado Pachon Sofia Axelsson. 2020. The Quality of Government Basic Dataset, version Jan20. University of Gothenburg: The Quality of Government Institute, http://www.qog.pol.gu.se doi:10.18157/qogbasjan20

The dataset I used was from this website: https://www.gu.se/en/quality-government/qog-data/data-downloads/basic-dataset and was collected by members of the Quality of Government Institute at the Unviersity of Gothenburg.