Analyzing Social Expenditure for a Sustainable Society using Machine Learning: Linear Regression

Young Kim

Korea International School

Abstract

Social expenditure, which encompasses cash benefits, direct in-kind provision of goods and services, and tax breaks with social purposes, is an important source of support for disadvantaged or vulnerable groups, such as low-income households, the elderly, the disabled, the sick, the unemployed, and the young. Due to its importance in the socio-economic sphere, it is essential to discover measures for optimizing a country's net social expenditure. As one of numerous data mining techniques, linear regression is an analysis methodology that provides a synthesis of inputs to calculate an output variable. The Corruption Perception Index (CPI), the COVID-19 Case Fatality Rate (CFR), the Gross Domestic Product (GDP), and the Environmental Performance Index (EPI) are the four factors we synthesize for this study. The results indicated that the ideal multilinear regression model for forecasting a country's social expenditure was the combination of all four parameters. In addition, the data suggest that EPI had the highest association with social expenditure, but GDP, which had previously been the primary factor in determining a country's net social expenditure, had the lowest correlation among the four variables examined. In the future, we intend to combine other factors and different prediction techniques, such as feed-forward neural networks, to construct a more accurate prediction model.

Introduction

It is established that the social welfare of citizens, which can be facilitated or hampered by a complex synthesis of factors, is a constant concern of the socio-economic sphere. Therefore, social welfare programs are an essential instrument for redistribution, social cohesion, and solidarity, as they aid underprivileged individuals and families. Stronger welfare regimes — those characterized by

generous social transfers — should improve the health and well-being of their population by implementing extensive redistributive policies based on the principle of social equity. Moreover, studies indicate that education-based health disparities are lower in nations with greater social expenditure. In addition, social expenditure moderates the association between education and health. Lastly, the favorable impact of education on health is minimal in nations with high social expenditures but substantial in nations with low social expenditures [1].

Historically, a country's net social expenditures were evaluated mostly based on its Gross Domestic Product. Despite the fact that GDP statistics measure current economic activity, they overlook wealth variation, international income flows, household production of services, destruction of the natural environment, and numerous determinants of well-being, such as the quality of social relations, economic security, personal safety, health, and longevity. Even worse, the GDP rises when convivial reciprocity is replaced by anonymous market contacts and when increased crime, pollution, or health risks motivate defensive or repair spending, resulting in a weak independent variable for examining social expenditure. Consequently, the practical significance of a measure of social welfare cannot be exaggerated. Policy decisions, cost-benefit calculations, international comparisons, metrics of growth, and studies of inequality continually refer to assessments of individual and communal well-being. The fact that monetary measures continue to predominate in all such contexts is typically regarded as a result of the absence of a superior index rather than a sign of broad consensus. If social and labor market policies are appropriately planned, they contribute to the promotion of social justice as well as economic efficiency and productivity [2].

Given the significance of optimizing social welfare systems, it is crucial to do research on non-economic factors that influence a country's social expenditure. There are other data mining techniques available, however, linear regression was chosen for this study. Linear regression analysis enables us to examine the relation between a number of factors and social expenditures. By definition, linear regression analysis determines the coefficients of the linear equation containing one or more independent variables that predict the value of the dependent variable most accurately. Linear regression involves fitting a straight line or surface that minimizes the discrepancies between the predicted and actual output values. For each model, we can consider regression coefficients, the correlation matrix, part and partial correlations, multiple R, R2, adjusted R2, change in R2, standard error of the estimate, an analysis-of-variance table, the sum of squares of errors (SSE), the sum of squares of the total (SST), predicted values, and residuals [3].

Simple linear regression is a function that enables an analyst or statistician to make predictions about one variable based on the knowledge of another one. Two continuous variables are required for linear regression: an independent variable and a dependent variable. The independent variable is the parameter through which the dependent variable or outcome is calculated. Multiple regression models use several explanatory variables [4]. In this study, we employ a multilinear regression model since, in reality, various factors influence the net social expenditure of a country; metrics such as GDP alone do not dictate social expenditure.

Our research examines four factors from the environmental, economic, social, and healthcare industries: the Corruption Perception Index (CPI), the COVID Case Fatality Rate (CFR), the Gross Domestic Product (GDP), and the Environmental Performance Index (EPI). This study investigates the relationship between social expenditures and each combination of the aforementioned variables. This allows us to determine which (synthetic) factor(s) are ideal for predicting social expenditures. In the past, different types of prediction methods, like feed-forward neural networks, have been used. Employing artificial neural networks, Basaran et al. analyzed projected governmental expenditures in Turkey [5]. Our study, on the other hand, employs a multilinear regression approach, which takes into consideration a more diversified set of factors that can play a role in increasing/decreasing social expenditure. Furthermore, this study examines the association among social expenditures across 45 nations.

First, we shall discuss the terminology and procedures of linear regression analysis. This section will include the development of the linear regression equation as well as an examination of the code used to perform it. Then, we will utilize our linear regression model to do an analysis of a sample data set. After reviewing linear regressions, we will examine the four factors that will be incorporated into our multilinear regression model for predicting social expenditures. We will conclude by analyzing the results and outlining a number of policy suggestions based on our findings.

Linear Regression

Several definitions and key terms will be presented in this report. First, an independent variable (commonly denoted by x) is a variable whose variation is not reliant on that of another variable. A dependent variable is a variable whose value is contingent on that of another variable. Next, a scatter diagram is a form of a graphic or mathematical diagram that uses Cartesian coordinates to represent values for typically two variables inside a data collection. Finally, an error term is a difference between the expected price at a given time and the actual price observed.

Assumption

a. In every linear regression model, there is an independent variable and dependent variable
b. For each x value, there is a probability distribution of the y value that follows.
c. The probability distribution of the y value moves in relation to the x value.
d. There is linearity in the regression model.

Predictions

y=bx+a explanation

The predicted value y ˆ can be represented with coefficients a and b like the following:

y ˆ =b+ax (Equation 1)

First, the object function is set as SSE and the values of a and b are calculated. We temporarily set the value of the object function as SSE so that differentiation can be used to calculate the answer. We can represent SSE as the following:

$$\sum (e_i)^2 = \sum (Y_i - \hat{Y}_i)^2 \quad \textit{(Equation 2)}$$

$$* = \sum (Y_i - b - aX_i)^2$$

$$\because\ y = b + ax \quad \textit{(Equation 3)}$$

We will represent equation 2 as "S" in the following steps

When partial differentiation is applied to the previous equation, the following result is obtained:

$\frac{\partial S}{\partial b}$	$= -2\sum (Y_i - b - aX_i)$
$\frac{\partial S}{\partial b}$	$= -2\{\sum_{i=1}^{n} (Y_i) - nb - a\sum_{i=1}^{n} (X_i)\}$
Since $\frac{\partial S}{\partial b}$ equal zero, organizing the equation yields the following result:	
nb	$= \sum_{i=1}^{n} (Y_i) - a\sum_{i=1}^{n} (X_i)$
The result of dividing each side by n is the following equation:	
b	$= \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n} \times a\sum_{i=1}^{n} X_i$
b	$= \underline{Y} - a\bar{X}$
S	$= \sum_{i=1}^{n} (Y_i - b - aX_i)^2$
$\frac{\partial S}{\partial a}$	$= -2\sum_{i=1}^{n} \{(Y_i - b - aX_i)(X_i)\}$
0	$= -2\{\sum_{i=1}^{n} (Y_iX_i) - b\sum_{i=1}^{n} (X_i) - a\sum_{i=1}^{n} (X_i^2)\}$
0	$= \sum_{i=1}^{n} (Y_iX_i) - \underline{Y}\sum_{i=1}^{n} (X_i) + a\underline{X}\sum_{i=1}^{n} X_i - a\sum_{i=1}^{n} (X_i^2)$
$a\underline{X}\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} (X_i)^2$	$= \underline{Y}\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_iY_i$
a	$= \frac{\sum_{i=1}^{n} X_i \times \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_iY_i}{(\sum_{i=1}^{n} X_i)^2 - \sum_{i=1}^{n} (X_i^2)}$
a	$= \frac{\sum_{i=1}^{n} X_iY_i - n\sum_{i=1}^{n} X_iY_i}{(\sum_{i=1}^{n} X_i)^2 - \sum_{i=1}^{n} X_i^2}$

Example 1. Consider the following table. Once the coefficient and intercept are determined, you can predict a value using your desired factor.

	x	y	x^2	x*y
	27	62	729	1674
	17	62	289	1054
	23	63	529	1449
	37	74	1369	2738
	30	72	900	2160
	30	72	900	2160
	35	83	1225	2905
	47	98	2209	4606
	68	137	4624	9316
	72	152	5184	10944
	99	170	9801	16830
	58	168	3364	9744
	76	177	5776	13452
	91	214	8281	19474
	88	225	7744	19800
	92	237	8464	21804
	100	245	10000	24500
total	990	2311	71388	164610

Table 1. Sample Linear Regression Analysis Data

Using the above data points, we can determine a and b:

$$a = \frac{164610 - 17 \times 58.23529 \times 135.94118}{71338 - 17 \times 58.23529^2}$$

$$= \frac{164610 - 134581.7647}{13735.05882}$$

$$= 2.18625$$

$$b = 135.94118 - 2.18625 \times 58.23529$$

$$= 8.62442$$

With the coefficient values, we can derive the single linear regression equation:

$$\hat{y} = 2.18625x + 8.62442$$

Suppose we wish to estimate the value of y for x = 20. Simply substitute x=20 into the formula above.

$$\hat{y} = 2.18625 * 20 + 8.62442$$

$$\hat{y} = 52.34942$$

A. Goodness-of-Fit test

The Goodness-of-fit test determines if the sample regression equation predicts the value of the dependent variable accurately.

If the sample regression line is accurate, the observed values will cluster around it. The standard error of estimate (SE) quantifies the correlation between the regression line and the observed values. The SE can be solved using the formula below:

$$S_e = \sqrt{\frac{\sum (Y_u - y_i)^2}{n - 2}} = \sqrt{\frac{SSE}{n-2}}$$

$$= \frac{\sum y^2 - a\sum y - b\sum xy}{n-2}$$

n-2 is used as the degree of freedom to clarify the terms, where n is the total number of data points. Y_u is the observed value taken from the sample data, while Y_i is the predicted value based on the regression line. $\sum (Y_u - y_i)^2$ is known as the sum of standard error estimate (SSE). A is the slope of the anticipated linear regression, while b

is the value of the predicted linear regression's y-intercept.

The coefficient of determination is a number between 0 and 1 represented by the variable R. The closer the value of the coefficient of determination is to 1, the more closely the regression model matches the observed data. The following equation represents the coefficient of determination:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

To compute the coefficient of determination, you must first calculate the SSE, SSR, and SST values. Solving the numerator and denominator values would be represented as follows:

$$\frac{a\sum \quad y_i + b\sum \quad x_i y_i - n\underline{y}^2)}{a\sum \quad y_i^2 - n\underline{y}^2}$$

Example 2. Consider the table below. The R squared value can be calculated with SSE and SST.

	y	y^2	y'	y-y'	(y-y')^2	y-E(y)	{y-E(y)}^2
data 1	40	1600	39.48	0.52	0.2704	-15.667	245.4444444
data 2	83	6889	84.24	-1.24	1.5376	27.3333	747.1111111
data 3	62	3844	61.86	0.14	0.0196	6.33333	40.11111111
data 4	48	2304	46.94	1.06	1.1236	-7.6667	58.77777778
data 5	58	3364	54.4	3.6	12.96	2.33333	5.444444444
data 6	43	1849	46.94	-3.94	15.5236	-12.667	160.4444444
Total	334	19850	333.86	0.14	31.4348	1.4E-14	1257.333333

Table 2. Sample Linear Regression Analysis Data

$$SSE = \sum \quad (y_i - \hat{y}_i)^2 = 348.1$$

$$SST = \sum \quad \left(y_i - \underline{y}\right)^2 = 944.2$$

Given the values of SSR, SSE, and SST, we can conclude that the value of the coefficient of determination is 0.631

B. RMSE

Now we must conduct a performance evaluation of the sample regression that we derived. First, we must understand some key terms. First, a residual is a difference between a data point and the regression line. This can be illustrated as follows:

$$e_i = y_i - \hat{y}$$

Second, the MSE is the mean squared sum, which is the average of the square of the residuals. This can be represented as the following:

$$\sum \quad \frac{(y_i - \hat{y})^2}{n}$$

Third, is the MAE, short for the mean absolute error. The MAE is the average of the absolute difference between the data point and regression line (residual). This can be set out as follows:

$$\frac{1}{n}\sum \quad |y_i - \hat{y}|$$

The RMSE, abbreviated for root mean squared sum, is just the square root of the MSE, as

suggested by its name. This can be outlined as follows:

$$\sqrt{\sum \frac{(y_i - \hat{y})^2}{n}}$$

In this research paper, the root-mean-squared error will be employed to evaluate the performance of our regression line. The RMSE is preferred over the MAE because it squares the difference between the data and regression line, highlighting substantial differences. This is useful because in our experiment, big disparities between the data and the regression line are undesired; consequently, the RMSE will capture any errors in our regression more accurately.

Example 3. Regression Analysis using RMSE

Now, we must apply the RMSE calculations to our own regression line. When reviewing the data from Table 2, we find that the sum of the squared residuals is 1257.333333. The outcome of dividing this by n, which in this case is 6, is 209.55555. Taking the square root of this result yields an RMSE of about 14.5.

C. Multiple Linear Regression

Definition: Analyzing the regression of two independent and one dependent variables

For a multilinear regression, there are five key conditions. First, the error term must be independent of the other independent variables. Second, there must be no inaccuracy in the independent variable's measurement. Thirdly, the predicted error value is 0, constituting a normal distribution with a constant variance. Next, the error covariance must be 0. Finally, as a linear function, the independent variables are not totally related to each other.

Multilinear regressions, unlike single linear regressions, involve various x variables with distinct b values. Consequently, the final equation for y can be modeled as follows:

$$y_i = a + B_1X_{1i} + B_2X_{2i} + B_3X_{3i} + \cdots + B_kX_{ki} + \in_i$$

The residual and error equations remain the same.

$$Residual\ e_i = Y_i - \hat{Y}$$

$$\sum E_i^2 = \sum \left(Y_i - \hat{Y}\right)^2$$

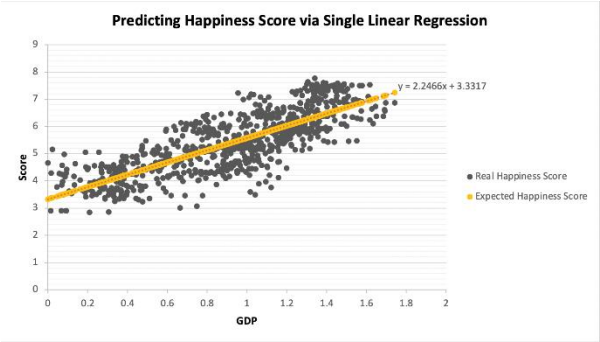
$$= \sum (Y_i - a - B_1X_{1i} - B_2X_{2i} - B_3X_{3i} - \cdots)^2$$

II. Implementation

A. Example Data explanation

Based on this knowledge of data analysis, we may now evaluate actual data samples from the real world. I examined two distinct data samples for this paper. I investigated the connection between GDP and happiness score. My objective was to identify a positive linear regression link between

GDP and happiness. Refer to **Appendix A** for data.



Graph 1. Happiness Score vs GDP Single Linear Regression Model

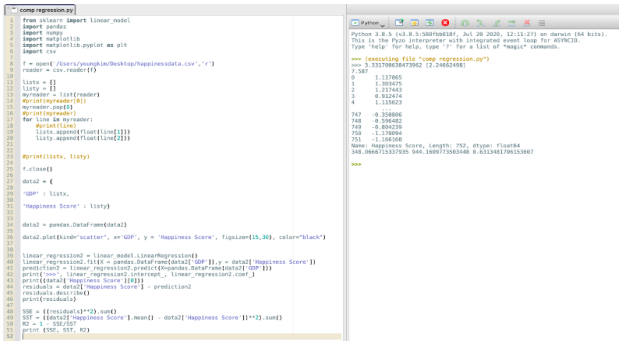


Figure 1. Happiness Score Prediction with GDP, Single Linear Regression Code

B. Analysis

Code explanation

After creating the data from the CSV file, the code constructs an array containing each value for the independent and dependent variables (lines 1 through 35). We split the independent and dependent variables in the data frame, as seen in lines 38–40. Using the imported library, it is possible to calculate the SST, SSE, and coefficient of determination (lines 45–56). The only difference in the code would be the inclusion of the data type. This is evident in lines 20-22, where the third variable is modified based on which column of data is appended.

With a grasp of the code and data analysis, we can evaluate the aforementioned two data sets.

The multilinear regression will compare two independent variables: health and freedom. This code represents the multilinear regression when x2 is set to 'health.'

As indicated by the coefficient of determination (~0.68), the association is relatively robust and has the following SSE and SST values:

$$SSE = 300$$

$$SST = 944.2$$

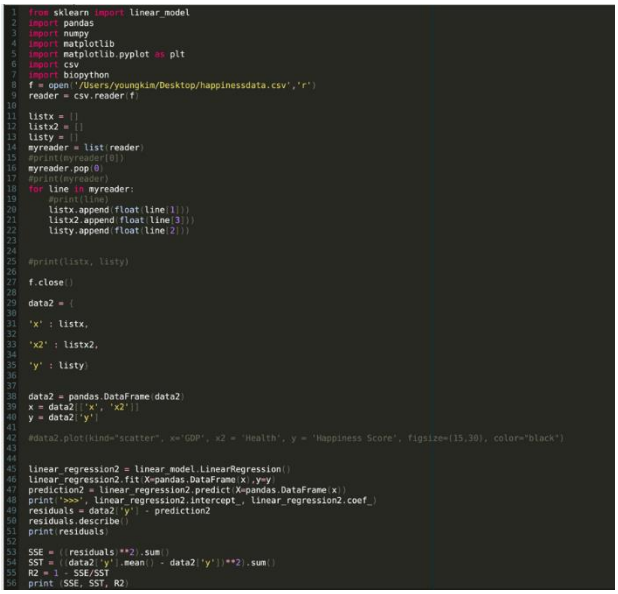
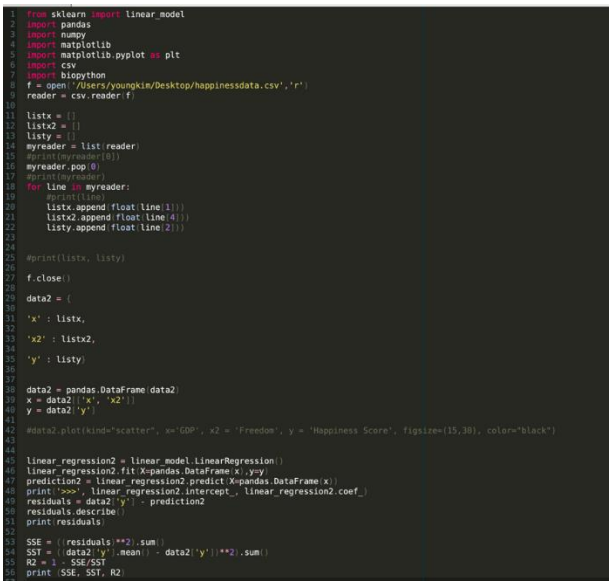


Figure 2. Happiness Score Prediction with GDP and Health, Multi-Linear Regression Code

We can similarly find the multilinear regression for when x2 is freedom.



```
from sklearn import linear_model
import pandas
import numpy
import matplotlib
import matplotlib.pyplot as plt
import csv
import biopython
f = open('/Users/youngkim/Desktop/happinessdata.csv','r')
reader = csv.reader(f)

listx = []
listx2 = []
listy = []
myreader = list(reader)
#print(myreader[0])
myreader.pop(0)
#print(myreader)
for line in myreader:
    #print(line)
    listx.append(float(line[1]))
    listx2.append(float(line[4]))
    listy.append(float(line[2]))


#print(listx, listy)

f.close()

data2 = {

'x' : listx,

'x2' : listx2,

'y' : listy}


data2 = pandas.DataFrame(data2)
x = data2[['x', 'x2']]
y = data2['y']

#data2.plot(kind="scatter", x='GDP', x2 = 'Freedom', y = 'Happiness Score', figsize=(15,30), color="black")


linear_regression2 = linear_model.LinearRegression()
linear_regression2.fit(X=pandas.DataFrame(x),y=y)
prediction2 = linear_regression2.predict(X=pandas.DataFrame(x))
print('>>>', linear_regression2.intercept_, linear_regression2.coef_)
residuals = data2['y'] - prediction2
residuals.describe()
print(residuals)

SSE = ((residuals**2).sum())
SST = ((data2['y'].mean()) - data2['y']**2).sum()
R2 = 1 - SSE/SST
print (SSE, SST, R2)
```

Figure 3. Happiness Score Prediction with GDP and Freedom, Multilinear Regression Code

The coefficient of determination for when x2 is freedom is ~0.67. The SSE and SST values go as the following:

$$SSE = 312$$

$$SST = 944.2$$

When x2 is freedom, we can now interpret the data for the multilinear regression. Given that the coefficient of determination is 0.67 (which is relatively close to 1), we may conclude that there is a strong correlation between freedom and happiness score as well.

III. Factor Analysis

1. Corruption Perception Index (CPI)

CPI is an indicator that rates countries "by their perceived levels of public sector corruption, as determined by expert assessments and opinion surveys." Corruption is commonly defined by the CPI as the "abuse of entrusted power for private gain." Since 1995, the non-governmental organization Transparency International has released the index annually.

The 2012 CPI takes into account sixteen distinct surveys and evaluations from twelve different institutes. The thirteen assessments are either opinion polls of business professionals or performance evaluations from a group of analysts [7]. The institutions are:

- African Development Bank (based in Côte d'Ivoire)
- Bertelsmann Foundation (based in Germany)
- Economist Intelligence Unit (based in the UK)
- Freedom House (based in the US)
- Global Insight (based in the US)
- International Institute for Management Development (based in Switzerland)
- Political and Economic Risk Consultancy (based in Hong Kong)
- The PRS Group, Inc., (based in the US)
- World Economic Forum

- World Bank
- World Justice Project (based in the US)

For this analysis, the 2019 Corruption Perception Index from Transparency International was used. Standardized data sources on a scale from 0 to 100, where 0 represents the highest perceived level of corruption and 100 represents the lowest perceived level of corruption [6, 7].

2. COVID19 Case Fatality Rate (CFR)

The COVID-19 Case Fatality Rate reflects the mortality risk associated with the pandemic. The CFR can be computed as follows [8]:

$$Case\ Fatality\ Rate\ (CFR) = \frac{Number\ of\ deaths\ from\ disease}{Number\ of\ diagnosed\ cases\ of\ disease} \times 100$$

Although other socioeconomic and demographic factors might alter the ultimate fatality rate, the CFR is typical of the characteristics of a country's healthcare system. In most cases, mortality will increase as hospitals become overburdened and deplete resources. COVID-19 CFR data was collected from the Coronavirus Resource Center at Johns Hopkins University [9].

3. Gross Domestic Product (GDP)

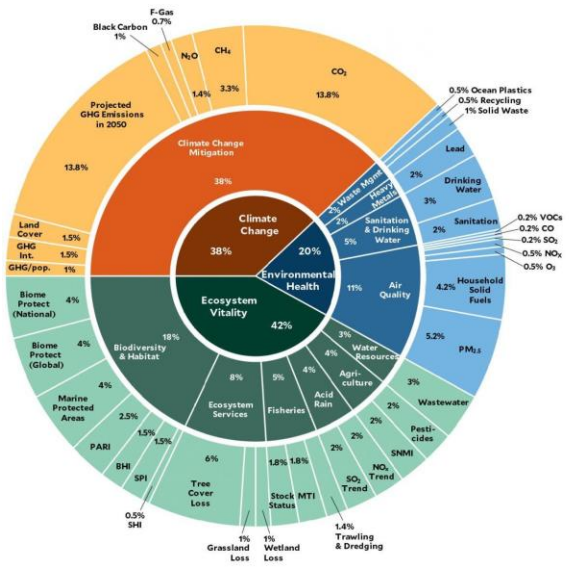
GDP measures the monetary worth of final products and services, i.e., those purchased by the final consumer, produced in a country within a given timeframe (say a quarter or a year). It calculates the overall output generated within a country's borders. In addition to products and services produced for sale on the market, the GDP also includes non-market production, such as government-provided defense or education services [10].

World Bank national accounts data and OECD National Accounts data files were used to compile the GDP for 2019 for 45 nations.

4. EPI

The Environmental Performance Index (EPI) provides an assessment of the global sustainability situation based on statistics. The EPI assesses 180 countries on climate change performance, environmental health, and ecosystem vitality using 40 performance indicators across 11 issue categories. These metrics assess, on a national basis, how closely countries are approaching their environmental policy objectives. The EPI publishes a scorecard that identifies environmental performance leaders and laggards and offers practical suggestions for countries aspiring to move toward a sustainable future [11].

The EPI data for 2019 were obtained from the Socioeconomic Data and Applications Center (SEDAC). Following is a breakdown of the 40 performance metrics that comprise the EPI [12].



Graph 2. 40 Performance Indicators of the Environmental Performance Index

IV. Evaluation

A) Environment

All data collection and coding were run on a MacBook Pro (16-inch, 2019). The Pyzo Python integrated development environment (IDE) was utilized for implementation. The regression analysis code utilized five Python libraries: Scikit-learn, Pandas, NumPy, Matplotlib, and CSV.

B) Results

X1: Corruption Perception Index (CPI)

X2: COVID Case Fatality Rate (CFR)

X3: Gross Domestic Product (GDP)

X4: Environmental Performance Index (EPI)

Factor(s)	R^2	SSE	SST	Intercept
X1	0.073164	922.67870	995.51419	13.88607
X2	0.0094039	986.15249	995.51419	22.23883
X3	0.0001978 7	995.31721	995.51419	21.55860
X4	0.32086	676.09743	995.51419	2.96840
X1, X2	0.073946	921.89951	995.51419	13.25563
X1, X3	0.073937	921.90906	995.51419	13.91753
X1, X4	0.33569	661.33181	995.51419	4.11137
X2, X3	0.0094061	986.15026	995.51419	22.24124
X2, X4	0.32713	669.84919	995.51419	1.52475
X3, X4	0.32437	672.60026	995.51419	2.56363
X1, X2, X3	0.075041	920.80996	995.51419	13.16409
X1, X2, X4	0.33722	659.80957	995.51419	3.21380
X1, X3, X4	0.34239	654.65584	995.51419	3.68018
X2, X3, X4	0.32983	667.16669	995.51419	1.25976
X1, X2, X3, X4	0.34305	654.00766	995.51419	3.10902

Table 3. Social Expenditure Multilinear Regression Results

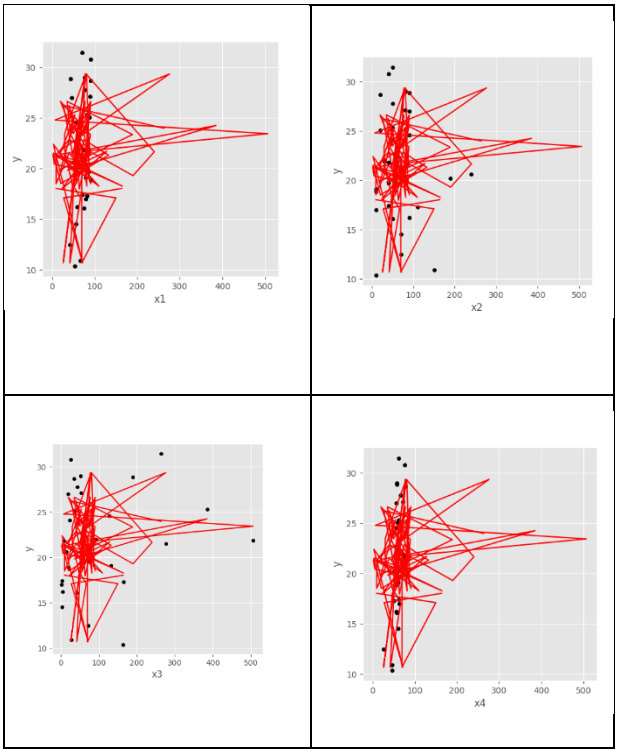
B.1) 4-Factor Multilinear Regression Model

With an R^2 value of 0.34305, the 4-dimensional multilinear regression model best predicts social expenditures. Below are scatter plots for each x value plotted against social expenditure. The

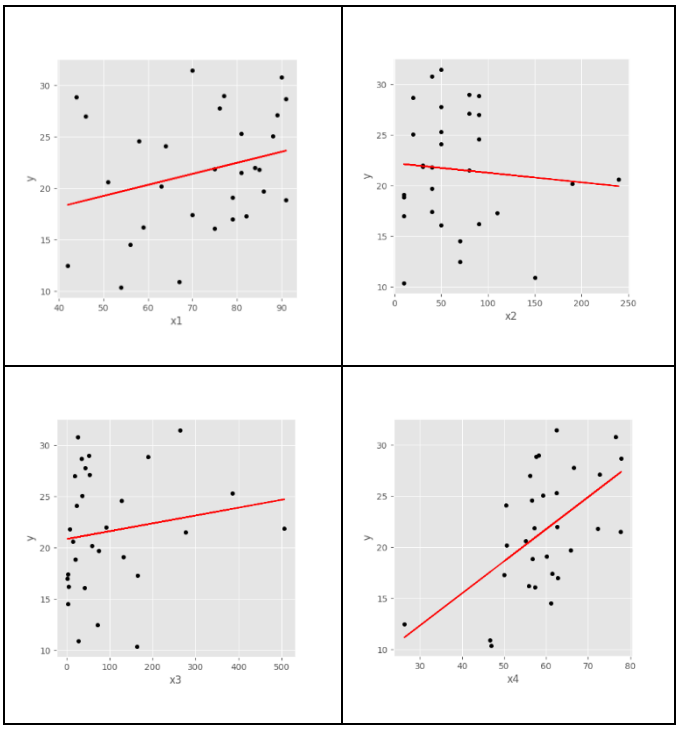
graph for the multilinear regression model is colored red.

B.2) 1-Factor Single Linear Regression Model

We can plot the single linear regression model for social expenditure vs. x_i, $i \in$ to examine the individual correlations between social expenditure and the four corresponding factors [1,4]. The graph below depicts the single linear regression model for each factor, as well as the best-fitting line in red.



Graph 3. 4-Factor Multilinear Regression Model Scatter Plots for Individual Factors



Graph 4. 1-Factor Single Linear Regression Model Scatter Plots

C) Implications (Policymaker's Perspective)

According to the R^2 values in Table 3, EPI has the strongest correlation with social expenditure. With R^2 values less than 0.1, the CPI, COVID-19 CFR, and GDP were all found to have relatively poor associations with social expenditures. Previous research demonstrates that the GDP has the poorest relation of these variables, with an R^2 of 0.00019787. This suggests that Fleurbaey's hypothesis that the GDP is a poor measure of social expenditures is justified. In essence, a nation's overall economic success does not determine its distribution in cash benefits, direct in-kind provision of goods and services, and tax breaks for social purposes.

Similarly, COVID-19 CFR showed minimal association with social expenditure. This makes sense given that external factors, like population size, can impact a country's CFR, making it difficult to use as a measure of social expenditure.

Moreover, it is striking that the link between CPI and social expenditure is quite weak. It is anticipated that a country's willingness to assist low-income households, the elderly, the disabled, the sick, the unemployed, and young people would link directly with the public's perception of corruption. This indicates that corruption in the public sector (as measured by the CPI) is not always connected with social expenditure. Therefore, a public sector's propensity to engage in illegal activity, such as bribery, is not indicative of its social expenditures.

The most intriguing aspect of the results of the single linear regression was that EPI had the highest correlation with social expenditure by a significant margin. Thus, the willingness of a nation to handle environmental issues is highly associated with its social expenditure. This shows that a public sector that is willing to engage in social expenditure is also prepared to invest in environmental protection. This may not be a cause-and-effect relationship, but it demonstrates that a country's efforts in comprehending the factors of environmental progress and improving its environmental policy choices are clearly the result of its commitment to support those from disadvantaged backgrounds.

Observing the R^2 values of all single and multilinear regression models reveals that the 4-factor multilinear regression model has the highest association with social expenditure. The R^2 value for the 4-factor model was 0.34305, which was somewhat higher than the findings of EPI's single linear regression. This indicates that, from a policymaker's perspective, the best way to predict a country's social expenditure is to utilize a mix of all four factors (CPI, COVID-19 CFR, GDP, and EPI).

V. Conclusion and Future Works

In this paper, we first introduced examples of linear regression and then analyzed the many aspects that contribute to a sustainable society. The experiment utilized a total of 5 variables: the Corruption Perception Index (CPI), the COVID-19 Case Fatality Rate (CFR), the Gross Domestic Product (GDP), the Environmental Performance Index (EPI) as factors, and the Social Expenditure as the prediction value. The evaluation was conducted by synthesizing the factors, and the results suggest that the combination of all four factors provides the most accurate linear regression model, with EPI being the strongest individual indicator for social expenditure. Further, we demonstrate that the GDP was a poor measure of social expenditure, supporting Fleurbaey's hypothesis.

In the future, we intend to study an expanded set of variables that may serve as more accurate indicators of social expenditure. The multilinear regression model can also be integrated with feed-forward neural networks, as demonstrated by Basaran et. al.'s study. Although this work presents a multilinear regression model with a relatively high correlation, we hope to identify other parameters that result in a greater causal connection.

REFERENCES

[1] Álvarez-Gálvez J, Jaime-Castillo AM. The impact of social expenditure on health inequalities in Europe. Soc Sci Med. 2018 Mar;200:9-18. doi: 10.1016/j.socscimed.2018.01.006. Epub 2018 Jan 11. PMID: 29355829.

[2] Fleurbaey, Marc. "Beyond GDP: The Quest for a Measure of Social Welfare." Journal of Economic Literature, vol. 47, no. 4, 2009, pp. 1029–75. JSTOR, http://www.jstor.org/stable/40651532. Accessed 8 Jul. 2022.

[3] "About Linear Regression | IBM". Ibm.Com, 2022, https://www.ibm.com/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable. Accessed 8 July 2022.

[4] "Multiple Linear Regression (MLR) Definition". Investopedia, 2022, https://www.investopedia.com/terms/m/mlr.asp#:~:text=Key%20Takeaways-,Multiple%20linear%20regression%20(MLR)%2C%20also%20known%20simply%20as%20multiple,uses%20just%20one%20explanatory%20variable. Accessed 8 July 2022.

[5] Alparslan A. Basaran, Cagdas Hakan Aladag, Necmiddin Bagdadioglu, Suleyman Gunay; Public Expenditure Forecast by Using Feed Forward Neural Networks, Advances in Time Series Forecasting (2012) 1: 40. https://doi.org/10.2174/978160805373511201010040

[6] "2019 Corruption Perceptions Index - Explore The Results". Transparency.Org, 2022, https://www.transparency.org/en/cpi/2019. Accessed 8 July 2022.

[7] (Transparency International (2010). Corruption Perceptions Index 2010: Sources of information (PDF) (Report). Transparency International. Archived from the original (PDF) on 3 December 2010. Retrieved 24 August 2011.)

[8] Ritchie, Hannah et al. "Coronavirus Pandemic (COVID-19)". Our World In Data, 2020, p. ., https://ourworldindata.org/mortality-risk-covid. Accessed 8 July 2022.

[9] "Mortality Analyses - Johns Hopkins Coronavirus Resource Center". Johns Hopkins Coronavirus Resource Center, 2022,

https://coronavirus.jhu.edu/data/mortality. Accessed 8 July 2022.

[10] Callen, Tim. Gross Domestic Product: An Economy's All, International Monetary Fund, 24 Feb. 2020, www.imf.org/external/pubs/ft/fandd/basics/gdp.htm.

[11] "Environmental Performance Index | Environmental Performance Index". Epi.Yale.Edu, 2022, https://epi.yale.edu/epi-results/2022/component/epi. Accessed 8 July 2022.

[12] "Data » Environmental Performance Index (EPI) | SEDAC". Sedac.Ciesin.Columbia.Edu, 2022, https://sedac.ciesin.columbia.edu/data/collection/epi/sets/browse. Accessed 8 July 2022.

Appendix A: Sample Linear Regression on GDP vs. Happiness Score on 753 countries. The GDP vs. Happiness Score of the first 30 countries are shown below:

Country	Economy (GDP per Capita)	Happiness Score
Switzerland	1.39651	7.587
Iceland	1.30232	7.561
Denmark	1.32548	7.527
Norway	1.459	7.522
Canada	1.32629	7.427
Finland	1.29025	7.406
Netherlands	1.32944	7.378
Sweden	1.33171	7.364
New Zealand	1.25018	7.286
Australia	1.33358	7.284
Israel	1.22857	7.278
Costa Rica	0.95578	7.226
Austria	1.33723	7.2
Mexico	1.02054	7.187
United States	1.39451	7.119
Brazil	0.98124	6.983
Luxembourg	1.56391	6.946
Ireland	1.33596	6.94
Belgium	1.30782	6.937
United Arab Emirates	1.42727	6.901
United Kingdom	1.26637	6.867
Oman	1.36011	6.853
Venezuela	1.04424	6.81
Singapore	1.52186	6.798
Panama	1.06353	6.786
Germany	1.32792	6.75
Chile	1.10715	6.67
Qatar	1.69042	6.611
France	1.27778	6.575

Appendix B: Below is a comprehensive dataset of CPI, COVID-19 Case-Facility, GDP (Trillions), EPI (Environmental Performance Index), and Social Welfare Spending (% of GDP) by country. 13 countries (out of a total of 45) did not include the social welfare spending percentage in their forecasts.

A	**CPI**	**COVID CFR** ($\boldsymbol{Rate \times 100}$)	**GDP** **(10,000 Trillions)**	**EPI**	**Social Expenditure**
South Korea	54	10	163.7	46.9	10.4
Chile	67	150	27.7	46.7	10.9
Turkey	42	70	72	26.3	12.5
Latvia	56	70	3.3	61.1	14.5
Ireland	75	50	42.6	57.4	16.1
Lithuania	59	90	5.6	55.9	16.2
Iceland	79	10	2.2	62.8	17
Canada	82	110	164.9	50	17.3
Estonia	70	40	3.1	61.4	17.4
New Zealand	91	10	21	56.7	18.9
Australia	79	10	132.7	60.1	19.1
United States of America	75	120	1954	51.1	19.3
Switzerland	86	40	75.2	65.9	19.7
Poland	63	190	59.6	50.6	20.2
Hungary	51	240	15.6	55.1	20.6
United Kingdom	81	80	276	77.7	21.5
Luxembourg	85	40	7.3	72.3	21.8
Japan	75	30	505.8	57.2	21.9
Netherlands	84	30	91.4	62.6	22
Portugal	64	50	22.8	50.4	24.1
Spain	58	90	128.1	56.6	24.6
Norway	88	20	36.2	59.3	25.1
Germany	81	50	384.6	62.4	25.3

Greece	46	90	18.9	56.2	27
Sweden	89	80	54.1	72.7	27.1
Austria	76	50	43.3	66.5	27.8
Denmark	91	20	35.6	77.9	28.7
Italy	44	90	188.9	57.7	28.9
Belgium	77	80	52.2	58.2	29
Finland	90	40	27	76.5	30.8
France	70	50	263	62.5	31.5
Argentina	39	140	64.3	41.1	NULL
Brazil	37	210	206	43.6	NULL
Colombia	37	230	31.1	42.4	NULL
Costa Rica	59	90	6.0	46.3	NULL
Cuba	47	80	9.7	47.5	NULL
Czech Republic	56	100	24.5	59.9	NULL
Dominican Republic	29	70	8.0	42.2	NULL
Ecuador	32	400	10.4	46.5	NULL
Guatemala	28	210	7.2	28	NULL
Jamaica	44	220	1.4	45.6	NULL
Mexico	29	560	115.9	45.5	NULL
Paraguay	29	290	3.9	40.9	NULL
Peru	37	590	21.1	39.8	NULL
Venezuela	18	110	14.3	46.4	NULL