IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Applying Linear Regression to The World Happiness Report

Ankur A. Dabholkar

Clark High School, 523 W Spring Creek Pkwy, Plano, Texas, 75013, USA; ankur.dabholkar.1@gmail.com
Mentor: Evgeny Goncharov

ABSTRACT: In this paper, we use The World Happiness Report to illustrate the use of linear regression. We perform some linear regression using the dataset values to pinpoint the effectiveness of linear regression in data analysis. The results show that linear regression can be used to precisely define trends in the data between the output variable and the input variables.
KEYWORDS: Mathematics, Statistics; Analysis; Linear Regression; Social Science; World Happiness Report.

Introduction

The World Happiness Report is a measure of the general level of happiness in the world.[1] It measures happiness through a global survey, which includes questions about the demographic, economic, and emotional standpoints of the countries' population. The main question of the survey is rating the survey taker's level of happiness on a scale of 0-10, with 10 being the happiest, and 0 being the least happy. Typically, the countries that have a higher development level have a happiness rating above 7. Countries with a lower development level usually have a happiness rating below 5. For example, Iceland has a happiness rating of 7.554, whereas Kenya has a happiness rating of 4.607. This rating of happiness corresponds with certain statistics recorded by research and census surveys in its respective countries. Some aspects of a country that are included in the dataset are the country's GDP per capita and the country's healthy life expectancy. Intuitively, this means that there is some level of correlation between the aspects of a country that contribute to its development and the overall happiness of that country. With certain mathematical tools, the strength of the correlation between the happiness of a country and the statistics that define a country can be measured and quantified.[2] Calculus is an extremely useful mathematical concept to utilize while calculating the connection between variables and outputs.[3] This type of calculation is called linear regression. It produces an equation correlating columns on a dataset (which represent data points) to the final output column in the dataset. The World Happiness Report is an excellent dataset to run a linear regression on as it contains many general trends between its variables and outputs, which can be found using linear regression. One way to execute linear regression is by writing a program in Python.[4] Python has many tools which make data analysis and linear regression clean and concise. It also offers visual representations of the data in low dimensions.

The rest of the paper is organized as follows. In the Limits and Differentiation, One-parameter Linear Regression, and Multiple-Parameter Linear Regression sections, we introduce limits, differentiation and linear regression, as well as small examples taken from the dataset, so if the reader is familiar with these concepts, they should feel free to skip to the Dataset section.

Methods

Limits and Differentiation:

To introduce linear regression, we need to be able to find the minimums of certain functions which is done via derivatives. Therefore, we remind the reader of some calculus concepts.

Informally, one can think of the limit[3] of a function $f(x)$ at the point x_0 as the value that $f(x)$ approaches as x approaches x_0. The formal definition is as follows.

Definition. The limit of a function $f(x)$ at the point x_0 is equal to L if for every $\varepsilon>0$ there is a $\delta>0$ such that if $0<|x-x_0|<\delta$, then $|f(x)-L|<\varepsilon$. We denote this limit by $\lim_{x\to x_0} f(x)=L$ if it exists. L if it exists.

A derivative at x_0 refers to the slope of the tangent line at that point on the function's graph. The derivative's importance to our research is that derivatives can help with minimizing functions. Minimums are useful for being able to discern which function best fits the data.

≈This determines a function $f'(x)$ that we call the derivative of $f(x)$ if the limit is defined.

$$f'(x_0)=\lim_{h\to 0}\frac{f(x_0+h)-f(x_0)}{h}$$

One-parameter Linear Regression:

Linear Regression with one parameter is a statistical analysis process that predicts and measures the correlation between two variables, the single *input* variable and the output variable that we are trying to predict. It produces a line in the coordinate plane (along with the equation for the line) that correlates best with the given dataset. Linear Regression is performed by minimizing a certain loss function. Intuitively, a loss function[7] minimizes errors of the model based on a training example or a dataset. We shall consider the case of a quadratic loss function which is known as the Least Squares Approximation method.

DOI: 10.36838/v5i6.2

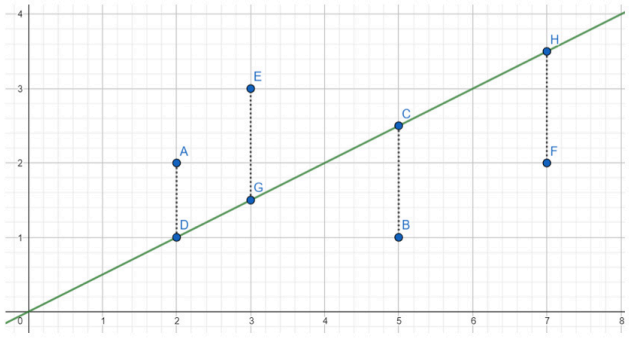

Figure 1: Example of a predicted line for a data set.

Algorithm 1. Suppose that the dataset consists of pairs (x_i, y_i), $i = 1,\ldots,n$. We would like to approximate it by a line of the form $y=kx+b$, that finds the coefficients k and b that best approximate the data. Consider the loss function,

$$L(k,\ b) \coloneqq \sum_{i=1}^{n}(kx_i + b - y_i)^2 = (kx_1 + b - y_1)^2 + \cdots + (kx_n + b - y_n)^2.$$

This function measures how close the actual data points are to the corresponding points on the line. For example, Figure 1 entails a random dataset of 4 points A, B, E, and F. In the figure, the dotted lines are the distances corresponding to the individual terms in the loss function. This function has one local minimum[7] (k_{min}, b_{min}), obtained by solving the system of linear equations

$$\frac{\partial L}{\partial k} = \frac{\partial L}{\partial b} = 0$$

where the ∂ stands for the partial derivative defined as follows.

Definition. We define the partial derivative[5] of function $L(k,b)$ with respect to k as

$$\frac{\partial L}{\partial k} = \lim_{h\to 0}\frac{L(k+h,b) - L(k,b)}{h}$$

This determines a function that is the derivative of the function $L(k,b)$ solely through the change of k , with b acting as a constant. The partial derivative of function $L(k,b)$ with respect to b is defined as

$$\frac{\partial L}{\partial b} = \lim_{h\to 0}\frac{L(k,b+h) - L(k,b)}{h}$$

This equation is like the first partial derivative, however, instead of b acting as a constant, k acts as a constant to make the derivative only depend on the change of b .

The system of equations produces the global minimum of the function corresponding to the best level of correlation between the predicted line and the dataset.

Example 1. In the following example, we are taking a sample of the dataset we are using for this research. We are comparing the GDP column with the Ladder Score column, which essentially compares the economic status of a country to the average happiness rating in that country. The single parameter is the Logged GDP per capita, and the output is the Happiness Rating at the country level. Figure 2 depicts that there is a linear correlation between the two.

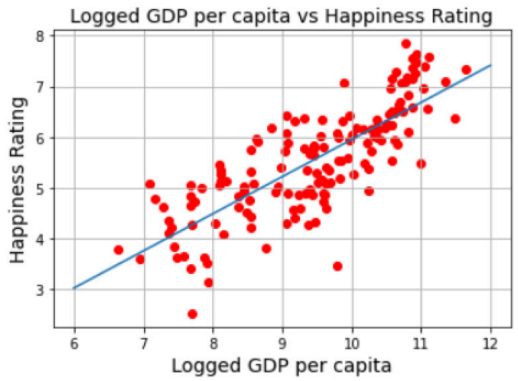

Figure 2: The line-of-best-fit depicts a correlation between GDP and Happiness.

Once linear regression is performed on the data (using Least Squares Approximation) the line-of-best-fit that is produced has

b=-1.3719060741319824, k=0.73203909

As shown in Figure 2, the line does correlate with the data well.

Multiple-parameter Linear Regression:

Linear Regression with multiple parameters $k_0,\ldots,k_n$ can calculate the strength of correlation of multiple variables. There are multiple input variables and only one output variable. It provides a linear function of the form

$$k_0x_0+k_1x_1+\cdots+k_nx_n+b$$

that best approximates the data (using the least squares method to judge how good the approximation is). The method is a direct generalization to Algorithm 1 (the loss function is similar but now depends on $k_0,\ldots,k_n,b$).

Algorithm 2. Suppose that the dataset consists of vectors $(x_{1i},\ldots,x_{ni},y_i)$,$i$=1,…,$m$. We would like to approximate the dataset by a hyperplane[8] of the form $y=k_1x_1+k_2x_2\ldots+k_nx_n+b$, that finds the coefficients $k_1,\ldots,k_n$ and b that best approximate the data. Consider the loss function

$$L(k_1,\ldots,\ k_n,\ b) \coloneqq \sum_{i+1}^{n}(k_{1i}x_{1i} + b - y_i)^2$$

This function has one local[7] minimum ($k_1^{\min},\ldots,k_n^{\min},b^{\min}$), obtained by solving the system of linear equations

$$\frac{\partial L}{\partial k_1} = \cdots = \frac{\partial L}{\partial k_n} = \frac{\partial L}{\partial b} = 0$$

where the ∂ symbol is the partial derivative of the loss function with respect to the variable that follows. The system of equations is the global minimum of the function corresponding to the best level of correlation between the predicted hyperplane and the dataset.

Example 2. In this example, we are once again taking a sample of the dataset we are using for this research. We are comparing the GDP column and the social support column to the Ladder Score column. This compares the correlation between both the economy of a country and the social situation of a country with the average happiness rating in that country. Figure 3 depicts a plane that correlates with the data for these parameters well.

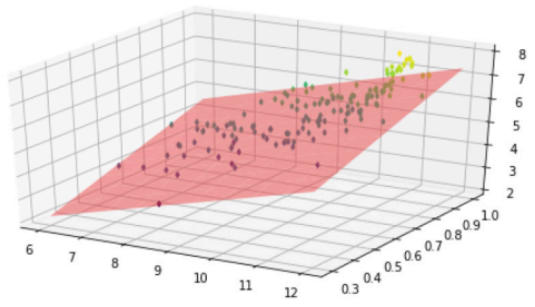

Figure 3: A plane predicted by the data points.

Once linear regression is performed on the data (using Least Squares Approximation) the plane-of-best-fit that is produced (with coefficients approximated to 3 digits) is

$$z=0.472x+3.333y-1.639$$

Dataset:

The original source used for this research contains data from about 150 countries. The dataset consists of columns of data that contain values for each country.[2] The Ladder score specifies the average country-level happiness rating. The dataset has the following columns of importance, which we will use as parameters[1]:

• Logged GPD per capita: The subsequent explanation of the data references how per capita measurements are more accurate than total measurements of GDP, as such the column should be described as a per capita metric.

• Social Support: This column gives a scale on how much support an individual in the country feels they receive from others around them.

• Healthy Life Expectancy: This column provides the average number of years a person born in that country is expected to live based on previous years and interpolation and extrapolation.

• Freedom to Make Life Choices: This column gives a measure of how the government allows the
population to make decisions.

• Generosity: This column is the average level of generosity throughout the population of the country.

• Perceptions of Corruption: This designates a relative level of corruption in a country.

The columns of the 5 happiest countries are shown in Table 1

Implementation:

Table 1: Data about the 5 happiest countries.

Country	Ladder Score	Logged GDP per capita	Social Support	Healthy Life Expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
Finland	7.842	10.775	0.95	72.0	0.949	-0.098	0.186
Denmark	7.620	10.933	0.954	72.7	0.946	0.030	0.179
Switzerland	7.571	11.117	0.942	74.4	0.919	0.025	0.292
Iceland	7.554	10.878	0.983	73.0	0.955	0.160	0.673
Netherlands	7.464	10.932	0.942	72.4	0.913	0.175	0.338

For this research, all coding and linear regression will be done in Python.[4] First, for linear regression to be possible, we will need to import the pandas library (along with other libraries), so the dataset can be read. Lines 1-4 of the code achieve this.

```
import pandas as pd
from sklearn import linear_model
import statsmodels.api as sm
data = pd.read_csv('world-happiness-report-2021.csv')
```

The pandas library is a Python, open-source package focused on data analysis and allows for multiple-parameter linear regression. Once this package is imported, we will then have to import the dataset for analysis. The "hyperplane-of-best-fit" will be found, and the correlation of the data will be measured between certain parameters and the average happiness in that country. The regression coefficients that are produced by the code will show the respective dataset columns' effect on country-level happiness. The full code can be found on request, as the rest of the code is merely formatting and setting up the linear regression.

The main way to measure correlation is Pearson's coefficient.[6] We set the coefficient equal to a variable P (Which it is in the data table in Section 5). If $|P|=1$ then the dataset correlates perfectly with the predicted function. If $1>|P|>.7$, then the dataset correlates strongly with the predicted function. When $.7>|P|>.3$, then the dataset somewhat correlates with the predicted function. When $.3>|P|>0$, then the dataset weakly correlates (and does not correlate at all when $|P|=0$) with the predicted function. The resulting findings will be analyzed by the following code:

```
print(correlation.loc[['Logged GDP per capita','Social support','Healthy life expectancy','Freedom to make life choices','Generosity','Perceptions of corruption'],'Ladder score'])
model = sm.OLS(Y, X).fit()
predictions = model.predict(X)
print_model = model.summary()
print(print_model)
```

Results and Discussion

Upon running the code, we get,

Table 2: Coefficients (coef) and Pearson's correlation coefficient (P) produced by code.

	coef	std err	P
Constant	-2.2372	0.630	
Logged GDP per capita	0.2795	0.087	0.789760
Social support	2.4762	0.668	0.756888
Healthy life expectancy	0.0303	0.013	0.768099
Freedom to make life choices	2.0105	0.495	0.607753
Generosity	0.3644	0.321	-0.017799
Perceptions of corruption	-0.6051	0.291	-0.421140

Approximating three significant digits, the hyperplane of best fit is as follows:

$$f(x_1,x_2,x_3,x_4,x_5,x_6)=0.280x_1+2.476x_2+0.030x_3+2.010x_4+0.364x_5-0.605x_6-2.237$$

where x_1 is the input for Logged GDP per capita, x_2 is the input for social support, x_3 is the input for Healthy life expectancy, x_4 is the input for Freedom to make life choices, x_5 is the input for Generosity, and x_6 is the input for Perceptions of corruption. The regression coefficients indicate the contributions that each input makes toward the Ladder Score. Intuitively, the results mean that all the columns have a positive effect

DOI: 10.36838/v5i6.2

when considering the magnitude of each column's effect on happiness, the regression coefficients' values do not accurately reflect their respective column's impact on happiness. This is because, when considering the scale of the actual data values, social support and Freedom to make life choices are measured on a 0-1 scale, whereas inputs like Logged GDP per capita and Healthy life expectancy are measured on a much larger scale, causing the regression coefficients to be larger with respect to columns with lower data values instead of their actual effect on happiness. When each column is put to the same 0-1 scale (including Ladder score), and linear regression is applied again, the regression coefficients change significantly, and the resulting equation is

$f(x_1,x_2,x_3,x_4,x_5,x_6)=0.415x_1+0.310x_2+0.297x_3+0.249x_4+0.025x_5-0.072x_6-0.285$

It is important to note that following the scale changes, the Pearson correlation coefficients do not change in comparison to the P column in Table 2. Indeed, merely scaling the data does not alter the correlation of the input data with the output data but produces regression coefficients so that they more accurately reflect each column's effect on happiness. The coefficients show that Logged GDP per capita has the largest effect on happiness, followed by social support, and the Perceptions of corruption column is the only data column that has an adverse effect on happiness. This interpretation of the regression coefficients produced by linear regression illustrates the effectiveness of linear regression in showing statistical trends in datasets such as the World Happiness Report.

Conclusion

To use linear regression with multiple parameters effectively, one must first be able to employ derivatives on loss functions to be able to find the minimum of that function. Once this is achieved, linear regression can be used to measure the correlation that parameters have to the overall output. The regression coefficients show that the World Happiness Report contains data values that correlate strongly with country-level happiness, exemplified by the regression coefficients in the equation produced by the linear regression calculation.

Future Research

Some research could involve other aspects of countries that could be quantified. The economy and government system (GDP, Social Support, and Perceptions of Corruption) are emphasized in our research, but some social examples include the Crude Birth Rate (CBR), and Crude Death Rate (CDR), which measure the number of births per 1000 people, and the number of deaths per 1000 people respectively. Along with this, some historical developments, such as economic failures and social unrest, could affect happiness.[9] These parameters could have a relatively strong correlation with the data, as these values tend to change with development. Another way to extend this research is to use an output other than the World Happiness Report. This may include the Human Development Index (HDI)[10] or the Quality-of-Life measure (QoL). This will influence which parameters correlate well, as development and quality of life may depend on parameters that are not equally as important to happiness.

Acknowledgments

I would like to thank Evgeny Goncharov for advising me and giving me invaluable support throughout the process of writing this paper.

References

1. Helliwell, J.; Huang, H.; Wang, S.; Norton, M. Statistical Appendix 1 for Chapter 2 of World Happiness Report 2021. https://happiness-report.s3.amazonaws.com/2021/Appendix1WHR2021C2.pdf (accessed 2023-05-14).
2. Singh, A. World Happiness Report 2021, 2021. https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2021
3. Strang, G. Calculus, 3rd ed.; Wellesley-Cambridge Press: Wellesley, MA, 2017.
4. Real Python. Linear regression in python. https://realpython.com/linear-regression-in-python/ (accessed Sep 10, 2022).
5. Hilton, P. J. Partial Derivatives; Routledge & Kegan Paul PLC: London, England, 1960. (Accessed 2022-09-07).
6. Sedgwick, P. Pearson's correlation coefficient. https://doi.org/10.1136/bmj.e4483 (accessed Sep 10, 2022).
7. Mahendru, K. Understanding Loss Functions to Maximize Machine Learning Model Performance (Updated 2023). https://www.analyticsvidhya.com/blog/2019/08/detailed-guide-7-loss-functions-machine-learning-python-code/#What_Are_Regression_Loss_Functions? (accessed 2023-05-11).
8. DeepAI. Hyperplane. https://deepai.org/machine-learning-glossary-and-terms/hyperplane (accessed 2023-05-16).
9. Ortiz-Ospina, E.; Roser, M. Happiness and Life Satisfaction. https://ourworldindata.org/happiness-and-life-satisfaction (accessed 2023-05-17).
10. Hall, J.; Helliwell, J. F. https://hdr.undp.org/system/files/documents/happinessandhdpdf.pdf (accessed 2023-05-18).

Authors

Ankur Dabholkar is a 10th-grade student at Clark High School in Plano. His areas of interest include data analysis and world systems. He plans to major in computer science or social sciences.