A minor project on

Life Expectancy Prediction

Using Multiple Linear Regression

Neha Gianchandani

The problem

Context

Life expectancy measure is one of the most common ways to find out how long a person if going to live. This usually helps people in making long term decision about this life.

Problem statement

To **predict life expectancy using Multiple Linear Regression** which uses features like but not limited to Alcohol Consumption, various medical issues like Polio, Measles, HIV and also some factors of the country of a subject and its economic status.

Data utilized

Source: https://www.kaggle.com/kumarajarshi/life-expectancy-who/data
This data was provided by WHO.

[21 columns and 2937 rows]

	Country	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria
0	Afghanistan	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	19.1	83	6.0	8.16	65.0
1	Afghanistan	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	18.6	86	58.0	8.18	62.0
2	Afghanistan	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	18.1	89	62.0	8.13	64.0
3	Afghanistan	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	17.6	93	67.0	8.52	67.0
4	Afghanistan	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	17.2	97	68.0	7.87	68.0

Country
Status
Life Expectancy
Polio

Adult Mortality
Alcohol
percentage expenditure
Total expenditure

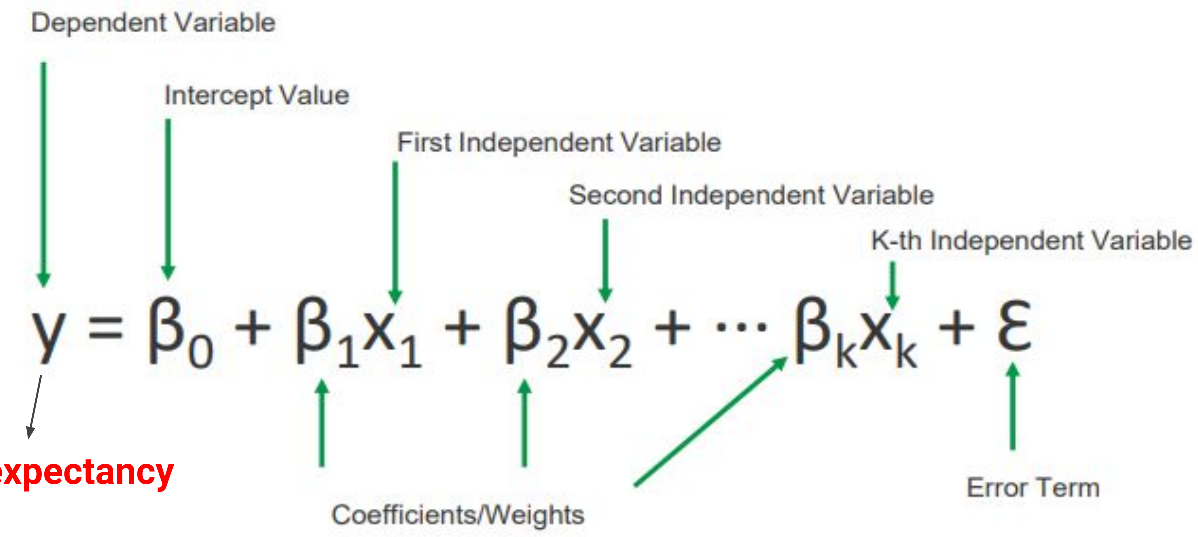
Hepatitis B
Measles
BMI
under-five deaths

Diphtheria
HIV/AIDS
GDP
Population

thinness 1-19 years
thinness 5-9 years
Income composition of resources
Schooling

Approach

Linear Regression → Feature Selection → Multiple Linear Regression



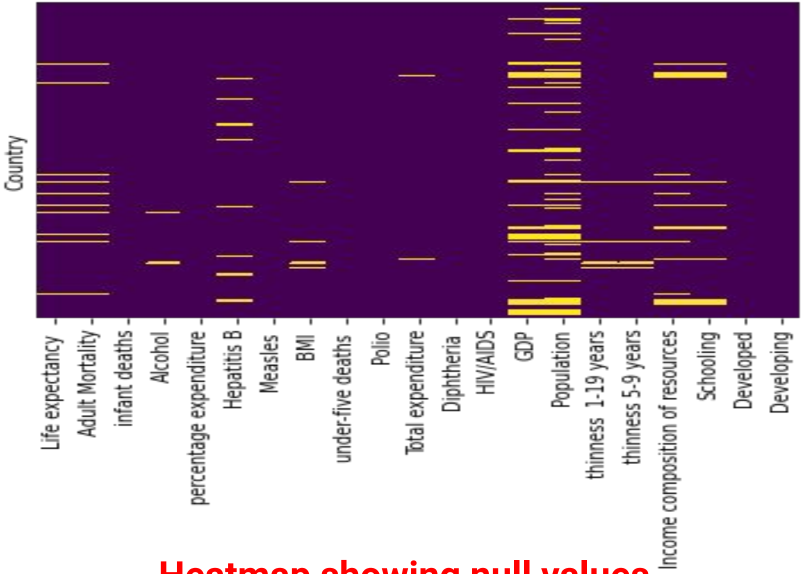
Life expectancy

Linear Regression Model

Steps

1. Data Cleaning/Preprocessing

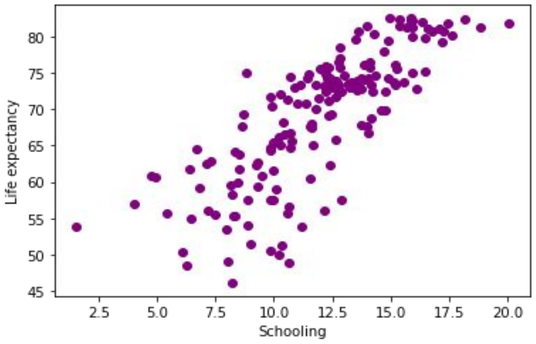
1. Year column is dropped as it has no significance in the prediction process
2. Non-numeric columns:
   "Status" - One hot encoded
   "Country" - Data grouped by country
3. All NaN cells/values are replaced by their respective column mean values

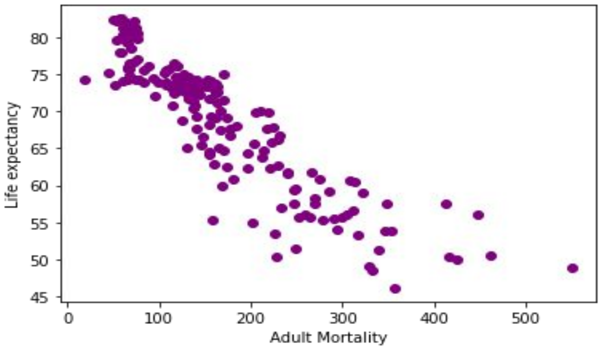


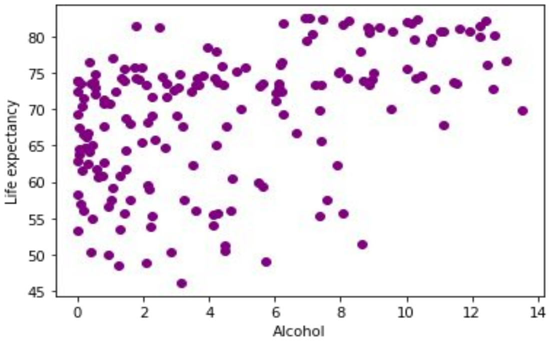
Heatmap showing null values

2. Exploratory Data Analysis

1. Positive correlation between The Percentage of Healthcare Expenditure, Schooling, GDP and BMI and Life Expectancy
2. Negative correlation between Adult Mortality, AIDS and Life Expectancy
3. No correlation between Alcohol, under 5 years – old deaths and Life Expectancy.







3. Linear Regression Assumptions

1. Data must be quantitative
2. No multicollinearity between the features. (Checked with a correlation matrix)

Results of Initial Linear Regression

After processing and cleaning the data, performing exploratory data analysis and taking assumptions we implement Linear Regression on the data set to find out the results. We divide our data in 70:30 ratio to train the model and then test the model.

Following is the result of our first run of Linear regression:

Mean Squared Error	15.52
Mean Absolute Error	2.67
Root Mean Squared Error	3.94
R_Square score	0.79

As you can see, we have very high value of errors along with our R_Sqaure value as 0.79. This means if we use this model to predict the life expectancy only 79% of the variations, it is good but we can increase it by dropping certain features, which have lesser impact on Life Expectancy as compared to other features.

FEATURE SELECTION WITH OLS MODEL

1. We find the p value using the OLS regression model to find the significant parameters.
2. We drop the feature with the greatest p value on each iterative fit.
3. P>|t| is the p-value for this hypothesis test. A low p-value means, that you can reject the null-hypothesis and accept the alternative hypothesis (coef!=0). A p-value of less than 0.05 is considered to be statistically significant. (95% CI)

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-27.6446	1.994	-13.865	0.000	-31.580	-23.709
x2	71.2703	35.033	2.034	0.043	2.123	140.418
x3	2.3082	0.968	2.384	0.018	0.397	4.219
x4	2.9738	3.481	0.854	0.394	-3.897	9.845
x5	-1.8981	1.739	-1.092	0.276	-5.330	1.533
x6	0.9135	2.690	0.340	0.735	-4.396	6.223
x7	3.1599	1.538	2.055	0.041	0.124	6.196
x8	-83.9408	32.262	-2.602	0.010	-147.618	-20.264
x9	-1.5244	3.426	-0.445	0.657	-8.286	5.238
x10	2.4357	1.615	1.509	0.133	-0.751	5.622
x11	8.8239	3.302	2.672	0.008	2.305	15.342
x12	-4.7556	2.058	-2.311	0.022	-8.818	-0.693
x13	0.5736	3.072	0.187	0.852	-5.490	6.637
x14	9.0009	6.743	1.335	0.184	-4.309	22.311
x15	-4.0545	7.614	-0.533	0.595	-19.083	10.974
x16	5.4848	7.695	0.713	0.477	-9.702	20.672
x17	5.7207	1.719	3.328	0.001	2.328	9.113
x18	4.7300	2.235	2.116	0.036	0.319	9.141
x19	62.8089	1.945	32.289	0.000	58.969	66.648
x20	63.0307	1.767	35.677	0.000	59.544	66.518

- Here we observe feature X13 has the maximum p>|t| value, which indicates this feature can be discarded.
- We discard X13 and reiterate OLS regression to find the next insignificant feature, and we repeat this process 4-5 time.

Results of Multiple Linear Regression

After dropping insignificant features from consideration we re run the Linear Regression on the data set to find out the results. Performing the linear regression again after making improvements, is called Multiple Linear Regression.
This time again we divide our data in 70:30 ratio to train the model and then test the model.

Following is the result of our Multiple Linear regression:

Mean Squared Error	8.08
Mean Absolute Error	2.05
Root Mean Squared Error	2.84
R_Square score	0.91

We observe a significant drop in the error values and improvement in R_Sqaure value which is now 0.91. This means if we use this model to predict the life expectancy then 91% of the variations can be explained, which is a great result.

Conclusion

Final accuracy - 91%

Final r^2 - 0.91

Final RMSE - 2.84

After taking 15 features and performing Multiple Linear Regression on the relevant data, our model can now predict the life expectancy of an individual with 91% explanation of variations about the mean.