BANARAS HINDU UNIVERSITY

A Project Report

Submitted for the Partial Fulfillment of the Bachelor’s Degree(Hons) in Statistics

Session: 2022-2025

UNDER SUPERVISION OF:

Dr. Suparna Basu
Head of Department
Statistics Section
Mahila Mahavidyalaya

SUBMITTED BY:

Shiwangi Yadav
Enroll.no. 460885
Roll. No. 22227STA005
Semester sixth

ACKNOWLEDGMENT

It gives me immense pleasure to express my profound gratitude, gratefulness and indebtedness to reverend professor and supervisor Dr. Suparna Basu, Assistant Professor, Department of statistics, MMV, Banaras Hindu University, for her untiring help, constant encouragement, worthy supervision without which it would not have been possible for me to complete this project work.

The guidance and valuable criticism that I received from her during the entire period of this work, has been a great help in the completion of this project.

I would also like to thank him in the validation for this project: "AN ANALYSIS OF RELATIONSHIP BETWEEN LIFE EXPECTANCY AND GDP IN INDIA" Without her passionate participation and input, the project could not have been successfully conducted. Again, I would like to say my sincere thanks to my supervisor and Head of Department, Department of

Statistics, Mahila Maha Vidyalaya Banaras Hindu University (BHU) for providing facilities existing in the Department during my project work.

At last, I am very thankful to my parents to my parents for providing me their valuable time and financial help for completing this project work.

SUBMITTED BY:
Shiwangi Yadav
Enroll.no. 460885
Roll. No. 22227STA005
Semester sixth

CERTIFICATE

This is to certify that the content of this project report entitled **"An Analysis of Relationships between life expectancy and GDP of India"** has been collected, tabulated, analyzed and presented by SHIWANGIAI YADAV, Student of B.A. (Hons.) Statistics, Mahila Mahavidyalaya, Banaras Hindu University, Varanasi221005.

This project has been completed successfully under my supervision and guidance.

UNDER SUPERVISION OF: **Date: 21/04/2025**

Dr. Suparna Basu
Head of Department
Statistics Section
Mahila Mahavidyalaya

An analysis of relationship between life expectancy and GDP in India

Table of content:

Introduction

1.Gross Domestic Product (GDP):

Gross Domestic Product (GDP) is a comprehensive measure of the total monetary value of all goods and services produced within a country's geographical boundaries over a specific period, typically measured annually. It serves as a fundamental indicator of a nation's economic performance and health. Mathematically, GDP is expressed as:

G⍰⍰⍰⍰=⍰⍰+⍰⍰+⍰⍰+(⍰⍰−⍰⍰)

Where:

C = Private consumption

I = Gross investment

G = Government expenditure

X = Exports

M = Imports

GDP can be calculated using three approaches: the production (or output) approach, the income approach, and the expenditure approach. In the context of developmental economics, GDP is frequently used as a proxy for the level of development and a tool to compare economies on a global scale.

2.Definition of Life Expectancy:

Life expectancy at birth is a demographic measure representing the mean number of years a newborn is expected to live under current mortality patterns. It is a continuous variable and serves as a vital indicator of population health, standard of living, and public health infrastructure.

In statistical studies, life expectancy is often used as a dependent or response variable to model the effect of various socio-economic indicators. It is frequently disaggregated by gender to explore differential impacts and assess health disparities between male and female populations.

India's GDP ;

India’s GDP has seen substantial growth over the past two decades. From approximately $476 billion USD in 2000, India’s GDP rose to over $3.1 trillion USD in 2021, making it **one of the world’s fastest-growing major economies**. This period of economic expansion was marked by liberalization reforms, increased foreign investments, the rise of the service and technology sectors, and government initiatives for industrial and rural development.

This increase in GDP has played a critical role in improving healthcare access, nutrition, education, and sanitation—key determinants of life expectancy. Government schemes like the National Health Mission, expansion of immunization programs, and the Swachh Bharat Mission have been possible due to higher public funding, enabled by economic growth.

Why Life Expectancy is Calculated Separately for Males and Females:

Biological Differences: Women usually live longer than men due to natural body and health differences.

Different Life Conditions: Men and women face different risks—like men working in tough jobs or women getting less healthcare in some areas.

Useful for Government Plans: Helps in making better health programs for both men and women.

Different Habits: Men and women have different lifestyles—like smoking, drinking, and food habits—which affect life span.

Better Results: Studying them separately gives more accurate and clear results in correlation and regression analysis.

The Complex Relationship between GDP and Life Expectancy

Statistically, GDP and life expectancy are hypothesized to exhibit a **positive correlation**, where increases in GDP are associated with improved life expectancy due to enhanced access to healthcare, education, sanitation, and living conditions. However, this relationship may vary in magnitude and direction depending on additional latent factors such as inequality, environmental conditions, and policy interventions.

This project models life expectancy (for both males and females) as a function of GDP using **Pearson's correlation coefficient** to assess the strength and direction of linear association, and **simple linear regression** to develop predictive models. This approach allows for an empirical understanding of the economic determinants of health outcomes in India.

Objective of the Study

The primary objectives of this project are as follows:

1.To compute and compare the Pearson correlation coefficients between GDP and life expectancy for males and females in India over the period 2000–2021.

2.To fit separate simple linear regression models for male and female life expectancy using GDP as the predictor variable.

3.To predict life expectancy values for the years 2022 and 2023 based on observed GDP values, thereby assessing the reliability of the model in forecasting demographic outcomes.

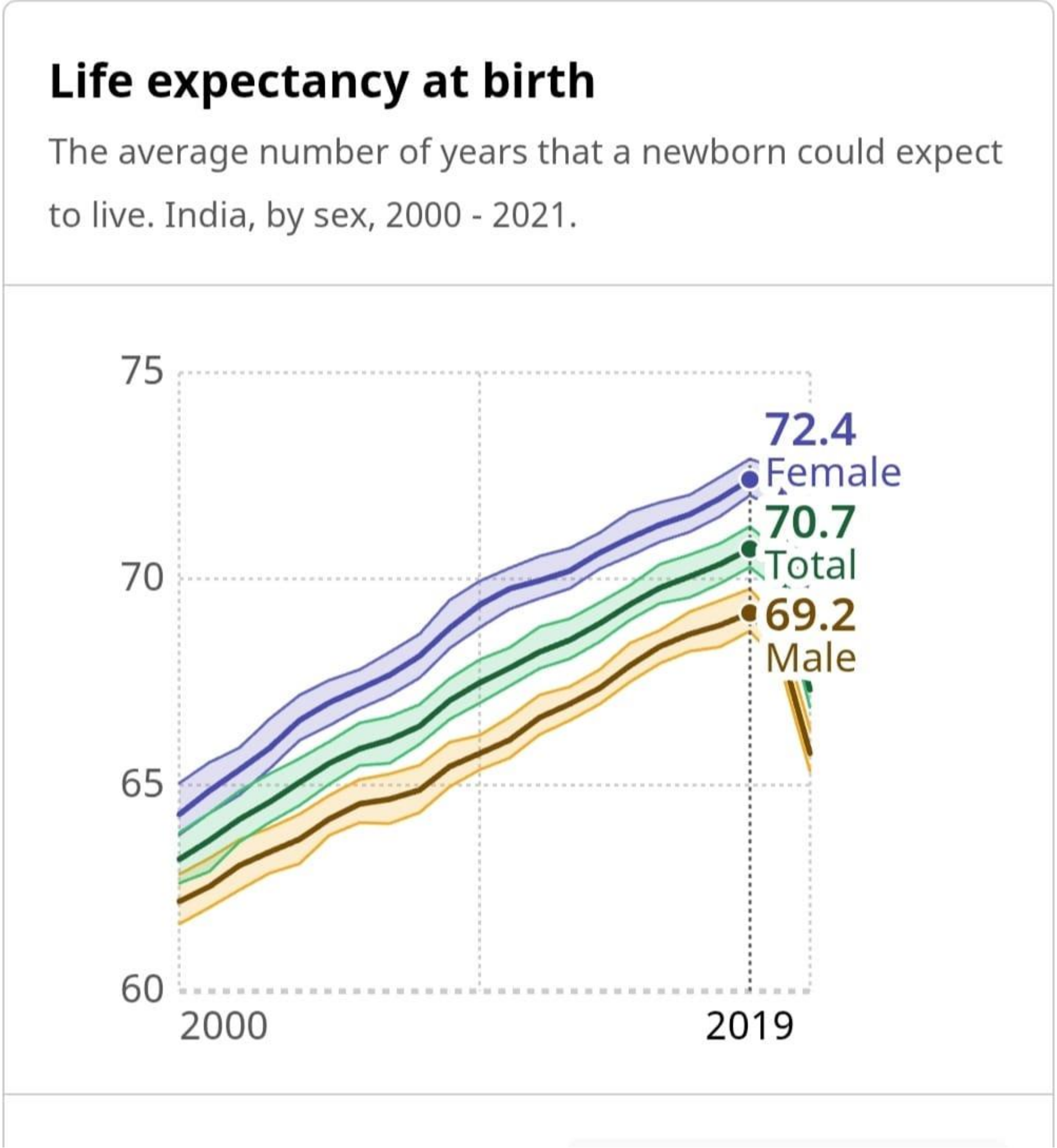
Method of Data Collection

For this project, **secondary data** has been used from **two trusted national and international sources** to analyze the relationship between India’s GDP and life expectancy (male and female separately) from the year 2000 to 2021, and to predict life expectancy for 2022 and 2023 using known GDP values.

1. World Health Organization (WHO)

The World Health Organization (WHO) was established on 7 April 1948 and is a specialized agency of the United Nations focused on international public health. Headquartered in Geneva, Switzerland, WHO provides globally standardized health statistics, including life expectancy at birth for all countries.

In this project, life expectancy data for both male and female populations in India from 2000 to 2021 has been collected from WHO’s official database. These figures represent the average expected lifespan at birth and reflect improvements in healthcare services, nutrition, sanitation, and living conditions over time. By separating the data by gender, we can more accurately analyze how economic progress (GDP growth) has differently impacted the life expectancy of men and women in India.



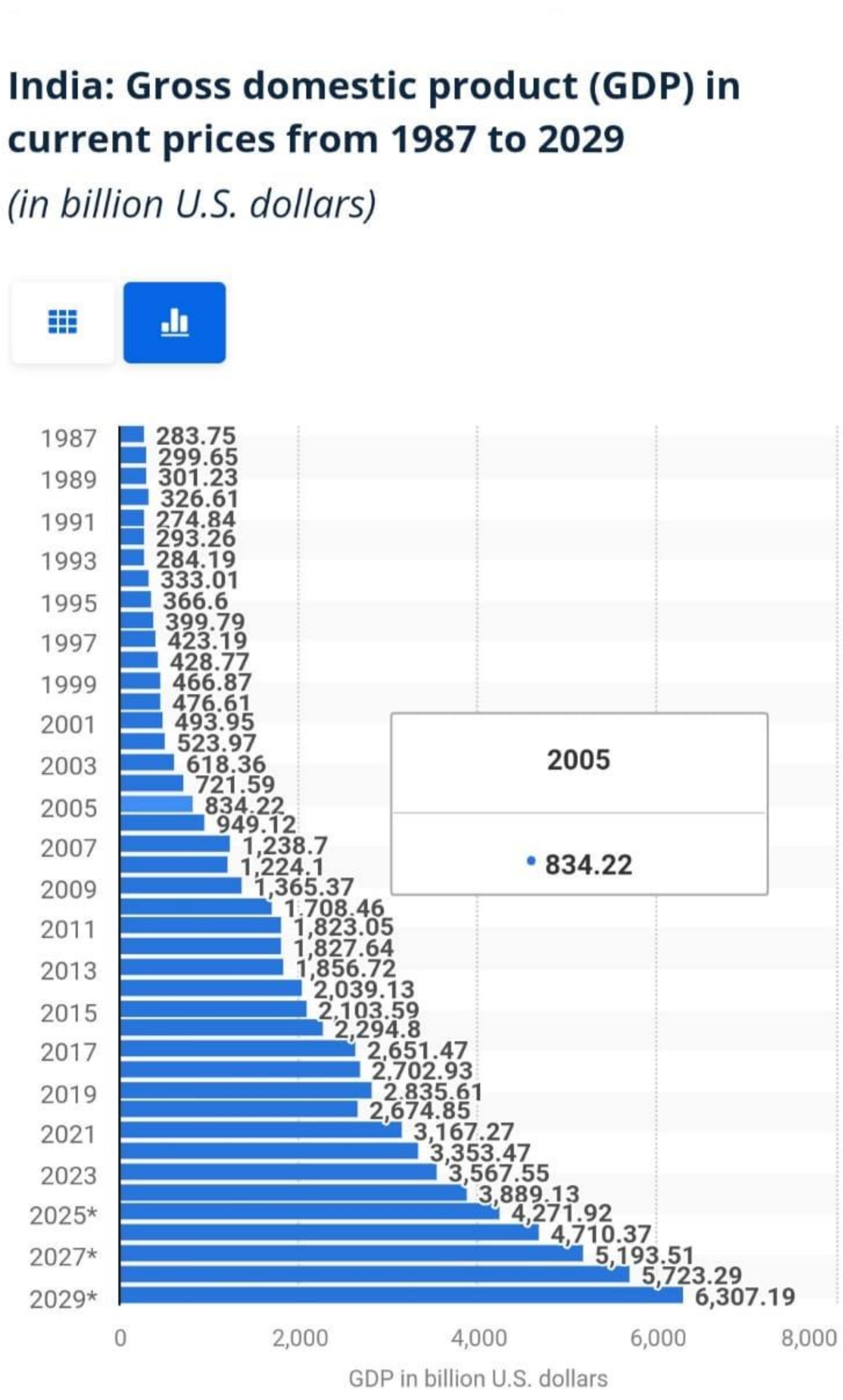
Life expectancy at birth

The average number of years that a newborn could expect to live. India, by sex, 2000 - 2021.

2.**Ministry of Statisticsand programme Implementation(MoSPI)**

The Ministry of Statistics and Programme Implementation (MoSPI), established in 1999, is a key department under the Government of India. It is responsible for collecting, analyzing, and disseminating statistical data related to the economic and social development of the country, including national accounts, surveys, and indicators such as Gross Domestic Product (GDP).

For this project, GDP data for India from 2000 to 2023 has been taken from Statista.com, a global statistics portal that compiles and presents data from official sources like MoSPI, World Bank, and IMF. Statista was chosen for its easy accessibility and consistent formatting of annual GDP figures, reported in USD (constant values).



India: Gross domestic product (GDP) in current prices from 1987 to 2029

(in billion U.S. dollars)

Regression Analysis

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variable. It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them.

Simple linear regression formula

The formula for a simple linear regression is:

$$y = \beta_0 + \beta_1 X + \epsilon$$

- **y** is the predicted value of the dependent variable (**y**) for any given value of the independent variable (**x**).
- **B_0** is the **intercept**, the predicted value of **y** when the **x** is 0.
- **B_1** is the regression coefficient – how much we expect **y** to change as **x** increases.
- **x** is the independent variable (the variable we expect is influencing **y**).
- **e** is the **error** of the estimate, or how much variation there is in our estimate of the regression coefficient.

Linear regression finds the line of best fit line through your data by searching for the regression coefficient (B_1) that minimizes the total error (e) of the model.

While you can perform a linear regression by hand, this is a tedious process, so most people use statistical programs to help them quickly analyse the data.

To find the best-fit line for each independent variable, multiple linear regression calculates three things:

- The regression coefficients that lead to the smallest overall model error.
- The *t* statistic of the overall model.
- The associate p value (how likely it is that the *t* statistic would have occurred by chance if the null hypothesis of no relationship between the independent and dependent variables was true).

It then calculates the *t* statistic and *p* value for each regression coefficient in the model.

Multiple linear regression uses following Null and Alternative hypothesis:

- Null Hypothesis, H_0: $\beta_0 = \beta_1 = \ldots\ldots\ldots\ldots = + \beta_k = 0$
- Alternative Hypothesis, H_1: $\beta_0 = \beta_1 = \ldots\ldots\ldots = + \beta_k \neq 0$
  The null hypothesis says that all the coefficients in the model are equal to zero i.e, none of the

predictor variables have statistically significant relationship with the response variable Y.

We will use a model to show relationship between GDP with Life Expectancy of Male and Female over a span of twenty one year where ,

.Life expectancy = Dependent variable

- GDP= Independent variable

Assumptions of a linear model

Linear regression analysis is based on following four assumptions:

1. The dependent and independent variable show linear relationship between slope and intercept.
2. The residuals are independent.
3. The residuals have a constant value of variance at any level of x.
4. The residuals of models are normally distributed.

The various components of regression analysis

1)Multiple R: This is the correlation coefficient. It tells us the strength of the linear relationship.

- R-squared: It is statistical measure of how close the data are to the fitted regression line. It is also known as coefficient of determination, or the coefficient of multiple determination of multiple regression. It is between 0 and 100%.0 indicates that the model explains none of the variability of the response data around its mean, whereas 1 indicates that the model explains all the variability of the response data around its mean.

The formula for R-squared is:

$R^2=1-RSS/TSS$

Where:

TSS(total sum of square)=$\sum(y^i-y^-)$^2

RSS(residual sum of square)= $\Sigma(yi - \hat{y}i)^2$

Where:

- yi = actual observed value

- ŷi = predicted value

- Σ = summation symbol, indicating the sum of the squared differences

This formula calculates the sum of the squared differences between actual and predicted values, providing a measure of the model’s fit.

2)Adjusted R^2 : The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model actually increases the model fit.

Let’s have a look at the formula for adjusted R squared to better understand its working.

$$Adjusted\ R^2 = \{1 - [\frac{(1-R^2)(n-1)}{(n-k-1)}]\}$$

Here,

- **n** represents the number of data points in our dataset
- **k** represents the number of independent variables, and

- **R** represents the R-squared values determined by the model.

3)Standard error of Regression: The standard error of the estimate is a measure of the accuracy of predictions. Recall that the regression line is the line that minimizes the sum of squared deviations of prediction (also called the sum of squares error). The standard error of the estimate is closely related to this quantity and is defined below:

$$\sigma est = \sqrt{(\sum(Y - Y')^2 / N)}$$

where σest is the standard error of the estimate, Y is an actual life expectancy, Y' is a predicted life expectancy, and N is the number of years.

4)Coefficient: It gives the least square estimate for the regression model.

5)P-value: The p-value for the coefficients indicate whether the dependent variable is statistically significant. When the p-value is less than your

Significance level, you can reject the null hypothesis that the coefficient equals to zero. For 95% confidence level, the p-value is 0.05.

Correlation Coefficient:

The correlation coefficient   measures the direction and strength of a linear relationship. Calculating   is pretty complex, so we usually rely on technology for the computations. We focus on understanding what   says about a scatterplot.

Here are some facts about :

- It always has a value between 0 and 1.
- Strong positive linear relationships have values of closer to 1 .
- Strong negative linear relationships have values of   closer to -1.
- Weaker relationships have values of   closer to 0.

The correlation between two variable X and Y is given by:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}$$

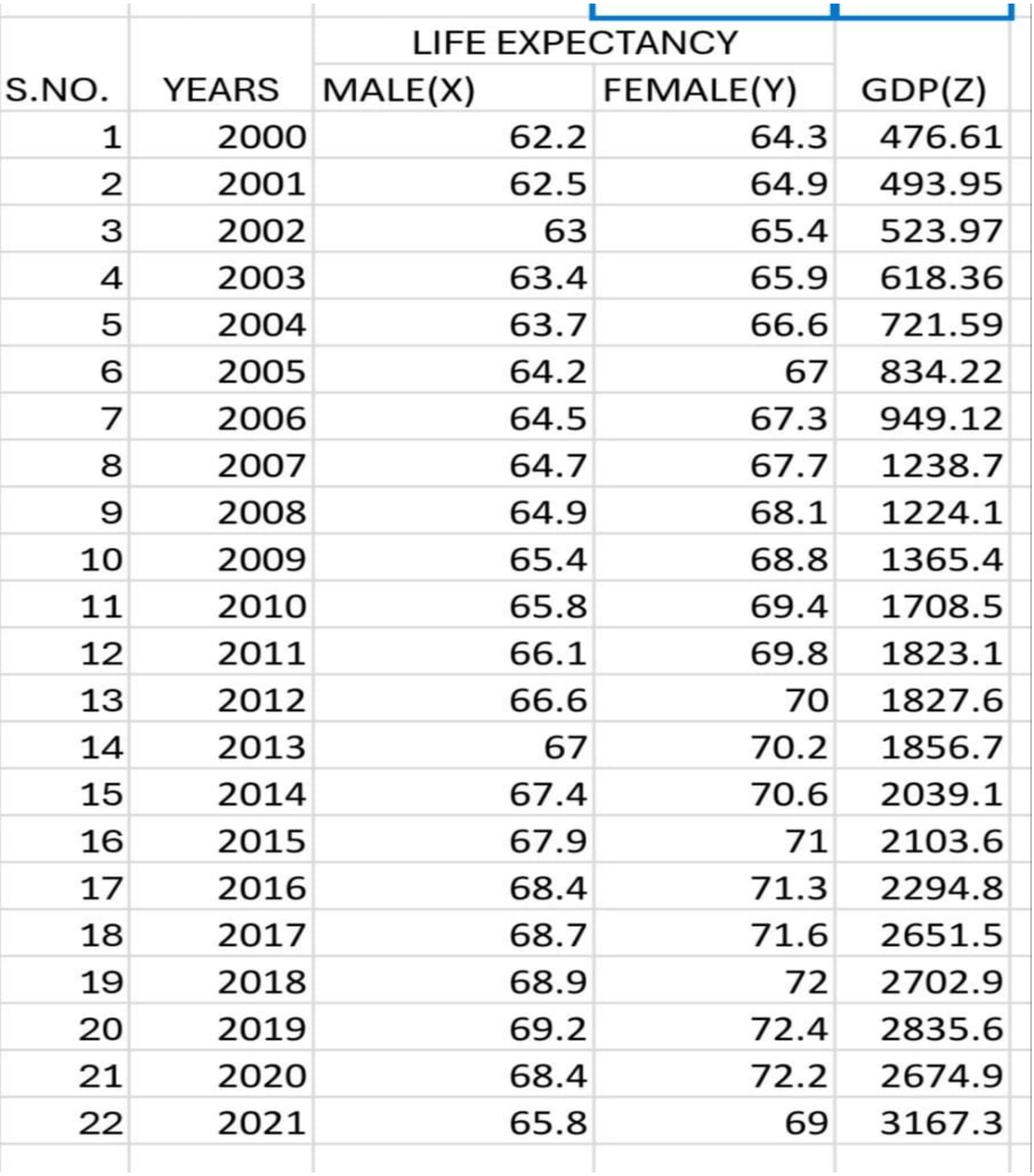
Where,

r_{xz}: correlation coefficient between x & z

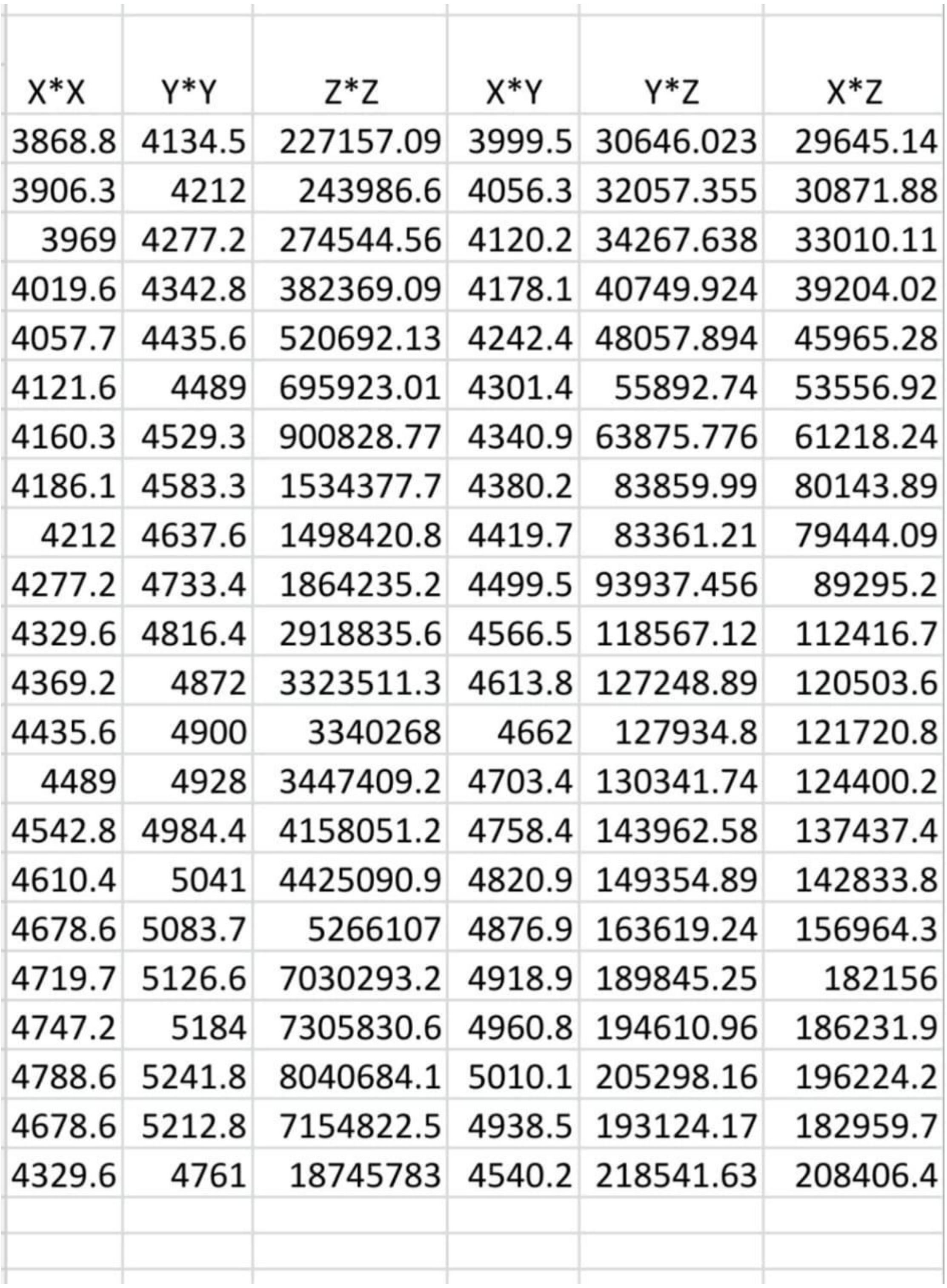
r_{yz}: correlation coefficient between y & z

DATA ANALYTICS

Calculation for correlation coefficient



		LIFE EXPECTANCY		
S.NO.	YEARS	MALE(X)	FEMALE(Y)	GDP(Z)
1	2000	62.2	64.3	476.61
2	2001	62.5	64.9	493.95
3	2002	63	65.4	523.97
4	2003	63.4	65.9	618.36
5	2004	63.7	66.6	721.59
6	2005	64.2	67	834.22
7	2006	64.5	67.3	949.12
8	2007	64.7	67.7	1238.7
9	2008	64.9	68.1	1224.1
10	2009	65.4	68.8	1365.4
11	2010	65.8	69.4	1708.5
12	2011	66.1	69.8	1823.1
13	2012	66.6	70	1827.6
14	2013	67	70.2	1856.7
15	2014	67.4	70.6	2039.1
16	2015	67.9	71	2103.6
17	2016	68.4	71.3	2294.8
18	2017	68.7	71.6	2651.5
19	2018	68.9	72	2702.9
20	2019	69.2	72.4	2835.6
21	2020	68.4	72.2	2674.9
22	2021	65.8	69	3167.3



X*X	Y*Y	Z*Z	X*Y	Y*Z	X*Z
3868.8	4134.5	227157.09	3999.5	30646.023	29645.14
3906.3	4212	243986.6	4056.3	32057.355	30871.88
3969	4277.2	274544.56	4120.2	34267.638	33010.11
4019.6	4342.8	382369.09	4178.1	40749.924	39204.02
4057.7	4435.6	520692.13	4242.4	48057.894	45965.28
4121.6	4489	695923.01	4301.4	55892.74	53556.92
4160.3	4529.3	900828.77	4340.9	63875.776	61218.24
4186.1	4583.3	1534377.7	4380.2	83859.99	80143.89
4212	4637.6	1498420.8	4419.7	83361.21	79444.09
4277.2	4733.4	1864235.2	4499.5	93937.456	89295.2
4329.6	4816.4	2918835.6	4566.5	118567.12	112416.7
4369.2	4872	3323511.3	4613.8	127248.89	120503.6
4435.6	4900	3340268	4662	127934.8	121720.8
4489	4928	3447409.2	4703.4	130341.74	124400.2
4542.8	4984.4	4158051.2	4758.4	143962.58	137437.4
4610.4	5041	4425090.9	4820.9	149354.89	142833.8
4678.6	5083.7	5266107	4876.9	163619.24	156964.3
4719.7	5126.6	7030293.2	4918.9	189845.25	182156
4747.2	5184	7305830.6	4960.8	194610.96	186231.9
4788.6	5241.8	8040684.1	5010.1	205298.16	196224.2
4678.6	5212.8	7154822.5	4938.5	193124.17	182959.7
4329.6	4761	18745783	4540.2	218541.63	208406.4

The correlation coefficient is given by:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}$$

Where r_{xz}, r_{yz}, are correlation coefficient between respective variables, given as:

r_{yz} (correlation b/W Female life expectancy and GDP) =

0.916124703

r_{xz} (correlation b/w Male life expectancy and GDP) =

0.903228592

For Simple linear regression Analysis:

Let X denotes the Male life expectancy (dependent) variable and Y Female life expectancy(dependent) that is linearly related to independent variables GDP denoted by Z.

X=a+bZ   A

Y= c+dZ

Male Life Expectancy and GDP

Simple linear regression model is given as:

$X=a+bZ$

Where:

a is intercept

b is the regression coefficient(slope)

By least square method,

$\sum X= na+b\sum Z$

$\sum XY= a\sum Z+b\sum z^2$

Solving using excel solver:

Regression Statistics	
Multiple R	0.903228592
R Square	0.815821889
Standard Error	0.961906848
Adjusted R Square	0.806612984
Observations	22

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	62.04171222	0.453614759	136.7718114	3.40E-31	61.09548842	62.98793603
GDP(Z)	0.002318816	0.000246361	9.412254631	8.67E-09	0.001804915	0.002832717

Hence, the estimated regression equation will be:

$$X=62.0417122+0.000231881\ Z$$

i.e. Male life expectancy= 62.0417122+0.000231881 GDP

Female Life Expectancy and GDP:

Simple linear regression model is given as:

$Y=c+dZ$

Where:

c is intercept

d is the regression coefficient(slope)

By least square method,

$\sum Y= nc+d\sum Z$

$\sum YZ= C\sum Z+d\sum z^2$

Solving using excel solver:

Regression Statistics	
Multiple R	0.906124703
R Square	0.821061977
Adjusted R Square	0.812115076
Standard Error	1.074421729
Observations	22

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	64.55694573	0.506674378	127.4130854	1.40254E-30	63.50004151	65.61384996
GDP(Z)	0.002636125	0.000275178	9.579694317	6.48078E-09	0.002062113	0.003210138

Hence, the estimated regression equation will be:

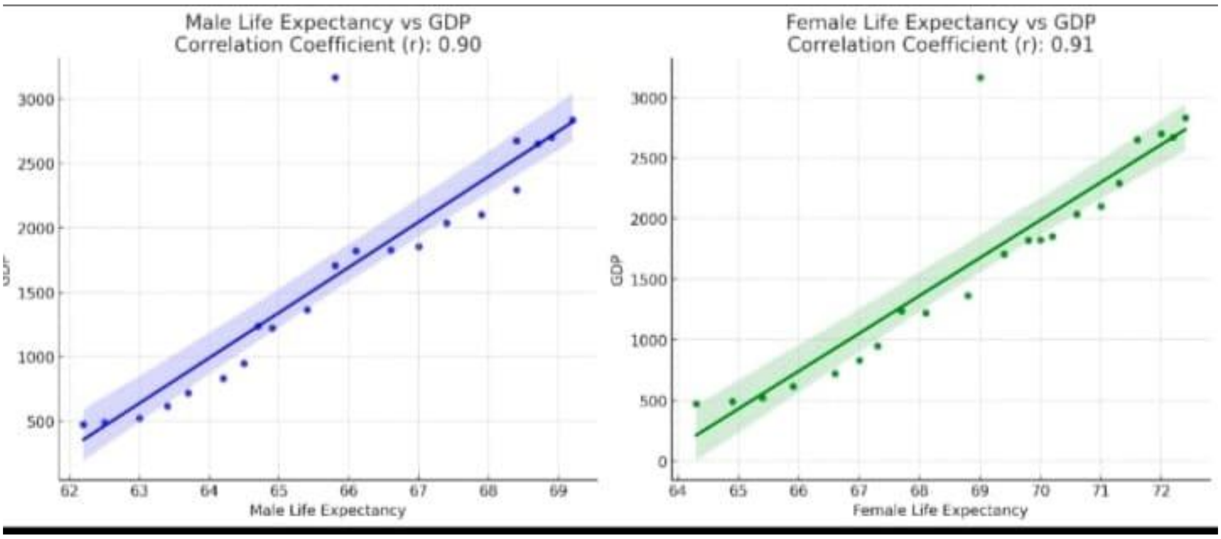
Y= 64.55694573+0.002636125 Z

i.e. Female life expectancy=64.55694573+0.002636125 GDP

Conclusion:

The analysis revealed a **positive** and **strong relationship** statistically between GDP and life expectancy for both genders:

1.The correlation coefficient indicated a strong linear association between economic growth and improved longevity, with a **slightly higher correlation for females.**



2.The **regression models** further validated this trend, with the R-squared value of 0.81 for males and 0.82 for females, indicating that GDP is a strong predictor of life expectancy, particularly for women.

3.The slope coefficients of the regression equations suggest that each unit increase in GDP results in a 0.0023-year increase in male life expectancy and a 0.0026-year increase in female life expectancy.

4.Predictions for 2022 and 2023 based on known GDP values showed continued improvement in life expectancy for both genders, reflecting ongoing benefits of economic progress.

Overall, the findings support the hypothesis that and longer life spans. The slightly **economic development significantly contributes to better health outcomes s**tronger relationship observed for females may reflect targeted public health initiatives, improved maternal care, and **increased awareness of women's health** in recent years.

Limitations:

- **The** main limitations of linear regression analysis is assumption of linearity between the dependent and independent variable. In the real world, the data is rarely linearly separable. It assumes a straight-line relationship between the dependent and the independent variable which is incorrect many times.

Prone to noise and over fitting: If the number of observations are lesser than the numbers of the features, linear it cannot be used, otherwise it may lead to over fit.

Prone to outliers: Linear Regression is very sensitive to outliers(anomalies). So, outliers should be analysed and removed before applying linear regression to the dataset.

Prone to multicollinearity: Before applying Linear regression , multicollinearity must be removed because it assumes that there is no relationship among independent variables.

In summary, Linear regression is great tool to analyse the relationship among the variable but it oversimplifies real world problems by assuming linear relationship among the variables.

References:

- S.C Gupta and V.K Kapoor: Fundamental of mathematical statistics
- Goon AM, Gupta MK. Dasgupta B: Fundamentals or Statistics.
- https://data.who.int/countries/356
- https://www.statista.com/statistics/263771/gross-domestic-product-gdp-in-india/
- https://www.scribd.com/home
- www.wikipedia.org
- www.google.com