I. INTRODUCTION TO MULTIPLE REGRESSION ANALYSIS

Multiple regression is a process to define the relationship between the dependent and multiple independent variables to better predict the dependent variable(Y), unlike linear regression where the relationship is built using just one independent variable.

General Multiple Regression Equation:

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k + \mathcal{E}}$$

Where:

$X_1, X_2 \ldots X_k$ are independent variables,

β_0 is the intercept, The Value of Y if all X's are 0.

$B_1 \ldots, \beta_K$ are the slopes of the regression line.

E is Error Term (Residuals).

II. STATEMENT OF OBJECTIVES AND DESCRIPTION OF VARIABLES

The objective of this project is to analyze the Literacy Rate in 61 different countries in the year 2015 based on multiple independent variables such as Female Literacy Rate in %, Gender Parity Index, Gov. Education Expenditure % of total expenditure, Gender Inequality Index, GDP per capita PPP in $, Human Development Index, Poverty Headcount on 5.5 $ a day, Out of School in Primary education, Rural Illiteracy, Number of Primary Teachers.
Entire operations in this project are performed in the IBM SPSS Statistics application.

i) Data Description

Dataset prepared is sourced from https://data.un.org. 10 different datasets were merged to build the final dataset which is used in this project. The dataset contains 11 columns having 10 independent and one dependent variable.

ii) Variables

The variables used in this regression are as follows:

Variable Name	**Type**	**Class**	**Statistical Measure**
Literacy Rate	Numeric	Continuous Dependent	Scale
Female Literacy Rate in %	Numeric	Continuous Independent	Scale
GPI	Numeric	Continuous Independent	Scale
Government Expenditure	Numeric	Continuous Independent	Scale
GII	Numeric	Continuous Independent	Scale
GDP	Numeric	Continuous Independent	Scale
HDI	Numeric	Continuous Independent	Scale
Poverty	Numeric	Continuous Independent	Scale
OOS	Numeric	Continuous Independent	Scale
Rural Illiteracy	Numeric	Continuous Independent	Scale
NPT	Numeric	Continuous Independent	Scale

GPI - Gender Parity Index
Government Expenditure – Gov. Education Expenditure % of total expenditure
GII - Gender Inequality Index
GDP - GDP per capita PPP in $
HDI - Human Development Index
Poverty - Poverty Headcount on 5.5 $ a day
OOS - Out of School in Primary education

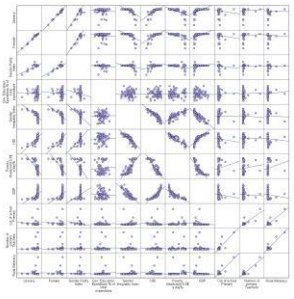
NPT - Number of Primary Teachers

III. ASSUMPTIONS UNDERTAKEN

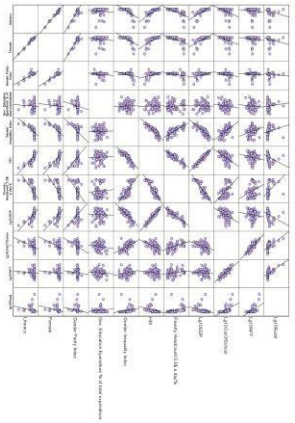
- **Assumption 1: Checking for Linearity:**

There should be a linear relationship between the dependent variable and each of the independent variables. The more is the R^2 (Coefficient of Determination), the more is the linearity. Ver often scatter plots are used to visualize the linearity among the variables.
Graph 1 shows the relation before arithmetically(Lg10) transforming the variables(GDP, Out of School in Primary education, Number of Primary Teachers, Rural Illiteracy), and Graph 2 shows the relation after transforming the variables(Lg10GDP, Lg10OutOfSchool, Lg10NPT, Lg10Rural). After the transformation of the variables, the four variables listed above became more linear visually and mathematically.

Graph 1: **Before performing arithmetic transformation:**



Graph 2: **After arithmetic transformation:**



- **Assumption 2: Independent Observations:**

This means that the residuals do not have a sequence, that the residuals are not strongly correlated. If the residuals are correlated, this is called Autocorrelation., which can be tested using the Durbin-Watson statistic. It ranges between 1 to 3 and is considered appropriate when close to 2. According to the below table, successive residuals are independent.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.998[a]	.997	.996	.008623	2.454

a. Predictors: (Constant), Lg10Rural, Gov. Education Ependiture % of total expenditure, Gender Parity Index, Lg10GDP, Gender Inequality Index, Lg10NPT, Poverty Headcount 5.5$ a day%, Lg10OutOfSchool, Female, HDI

b. Dependent Variable: Literacy

- **Assumption 3: To check Multicollinearity:**

It exists when independent variables are correlated. The standard error of measurement can be skewed by multicollinearity and can also lead to erroneous assumptions as to whether independent variables are statistically important. The standard solution for multicollinearity is to drop one of the highly correlated independent variables and recompute the regression equation. Also, we can use VIF (Variance Inflation Factor) to look for multicollinearity for more precision. A VIF greater than 10 is deemed unsatisfactory, suggesting that the study should exclude the independent variables.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	.230	.036		6.431	.000		
	Female	1.069	.026	1.239	41.648	.000	.068	14.701
	Gender Parity Index	-.343	.032	-.246	-10.775	.000	.116	8.626
	Gov. Education Ependiture % of total expenditure	.063	.026	.021	2.421	.019	.832	1.202
	Gender Inequality Index	-.004	.016	-.004	-.216	.830	.205	4.882
	HDI	-.059	.051	-.055	-1.166	.249	.027	36.468
	Poverty Headcount 5.5$ a day%	.014	.010	.033	1.453	.152	.119	8.404
	Lg10GDP	.019	.012	.057	1.567	.123	.045	22.201
	Lg10OutOfSchool	-.002	.003	-.018	-.712	.480	.091	10.991
	Lg10NPT	.001	.004	.004	.176	.861	.126	7.932
	Lg10Rural	.002	.002	.022	.898	.374	.101	9.908

a. Dependent Variable: Literacy

According to the VIF statistics in the above table, Female Literacy Rate, Human Development Index (HDI), Lg10GDP, and Rural Illiteracy highlighted in yellow are having VIF values greater or equivalent to 10. Therefore, to stick to the assumptions for implementing the multiple regression, exclude these independent variables from the regression analysis. Running the function again in SPSS and there is no collinearity among independent variables according to the result shown below.

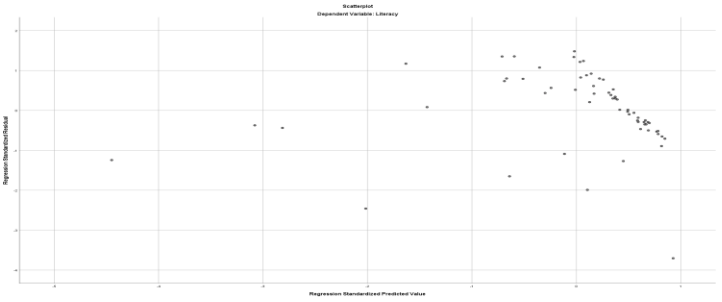
Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	.007	.097		.067	.947		
	Gender Parity Index	.929	.075	.665	12.401	.000	.782	1.279
	Gov. Education Ependiture % of total expenditure	.212	.148	.069	1.430	.158	.964	1.037
	Gender Inequality Index	-.105	.078	-.111	-1.355	.181	.338	2.962
	Poverty Headcount 5.5$ a day%	-.157	.036	-.359	-4.400	.000	.338	2.960
	Lg10NPT	.017	.009	.094	1.897	.063	.923	1.084

a. Dependent Variable: Literacy

- **Assumption 4: To check Homoscedasticity**

The data needs to illustrate homoscedasticity, where the variances remain identical along the best match line as you step along the line. We use the scatter plot by plotting the residuals against the predicted values. Based on the below plot, the variation in residuals are the same for large and small predicted values.



- **Assumption 5: Detection of Outliers:**

No major outliers, high leverage points, or highly powerful points should be present. These points are unusual while multiple regression analysis, hence it needs to be dealt with to prevent the negative effect on the regression equation which is used to predict the dependent variables depending on all independent variables. It can be analyzed using the cook's distance, which should be less than 1.

Initially, the cook's distance in the below table highlighted in yellow is **1.013** for independent variables having no multicollinearity, which is not as per the required assumption, hence analyzed for the outlier in the column Gender Parity Index and replaced it with the average of the column resulting in a cook's distance of **.224**, which is meeting the assumption.

Before analyzing outliers:

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.33114	1.04614	.92291	.132996	62
Std. Predicted Value	-4.450	.927	.000	1.000	62
Standard Error of Predicted Value	.009	.037	.016	.005	62
Adjusted Predicted Value	.39176	1.11106	.92485	.130083	62
Residual	-.195244	.078044	.000000	.050503	62
Std. Residual	-3.704	1.481	.000	.958	62
Stud. Residual	-4.276	1.566	-.016	1.041	62
Deleted Residual	-.260158	.087321	-.001934	.060174	62
Stud. Deleted Residual	-5.163	1.587	-.034	1.115	62
Mahal. Distance	.863	28.325	4.919	4.242	62
Cook's Distance	.000	1.013	.036	.140	62
Centered Leverage Value	.014	.464	.081	.070	62

a. Dependent Variable: Literacy

After analyzing outliers:

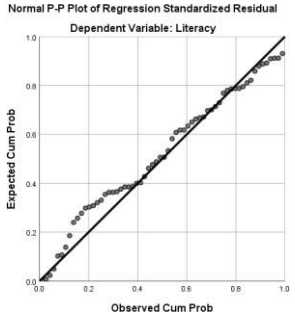
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.29181	1.02009	.92291	.136089	62
Std. Predicted Value	-4.637	.714	.000	1.000	62
Standard Error of Predicted Value	.008	.031	.013	.004	62
Adjusted Predicted Value	.31906	1.02399	.92352	.134401	62
Residual	-.130788	.067118	.000000	.041454	62
Std. Residual	-3.023	1.551	.000	.958	62
Stud. Residual	-3.171	1.631	-.006	1.011	62
Deleted Residual	-.146082	.074200	-.000610	.046303	62
Stud. Deleted Residual	-3.469	1.656	-.019	1.057	62
Mahal. Distance	.864	30.096	4.919	4.284	62
Cook's Distance	.000	.224	.020	.045	62
Centered Leverage Value	.014	.493	.081	.070	62

a. Dependent Variable: Literacy

- **Assumption 6: Normal Distribution of Residuals:**

Normality is the presumption that or roughly, so the underlying residuals are naturally distributed. P-P plots can be used to check for the normal distribution. The data points should be close to the reference line. The P-P plot showing the approval of the statement is below.



IV. Model Building and Diagnostics:

After meeting all the assumptions, based on the dependent and 5 independent variables models will be built. Models can be built by using the linear option from the regression menu under the analyze tab.

- **Descriptive Statistics:**

To have a basic statistical overview of the dataset, we use descriptive statistics in SPSS.

Descriptive Statistics

	Mean	Std. Deviation	N
Literacy	.92291	.142262	62
Gender Parity Index	.97256	.097231	62
Gov. Education Ependiture % of total expenditure	.14963	.046387	62
Gender Inequality Index	.38401	.149660	62
Poverty Headcount 5.5$ a day%	.38935	.325248	62
Lg10NPT	4.7436	.79641	62

- **Model 1:**

First Model is built by all 5 independent variables. Variables entered/removed, Model Summary, ANOVA, and coefficients of the model are represented below:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Lg10NPT, Gender Parity Index, Gov. Education Ependiture % of total expenditure, Poverty Headcount 5.5$ a day%, Gender Inequality Index[b]	.	Enter

a. Dependent Variable: Literacy

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.957[a]	.915	.908	.043265	2.160

a. Predictors: (Constant), Lg10NPT, Gender Parity Index, Gov. Education Ependiture % of total expenditure, Poverty Headcount 5.5$ a day%, Gender Inequality Index

b. Dependent Variable: Literacy

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.130	5	.226	120.705	.000[b]
	Residual	.105	56	.002		
	Total	1.235	61			

a. Dependent Variable: Literacy

b. Predictors: (Constant), Lg10NPT, Gender Parity Index, Gov. Education Ependiture % of total expenditure, Poverty Headcount 5.5$ a day%, Gender Inequality Index

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	-.100	.082		-1.216	.229		
	Gender Parity Index	1.069	.067	.731	15.980	.000	.725	1.378
	Gov. Education Ependiture % of total expenditure	.109	.121	.036	.905	.369	.979	1.021
	Gender Inequality Index	-.060	.064	-.064	-.943	.350	.333	3.000
	Poverty Headcount 5.5$ a day%	-.136	.029	-.311	-4.621	.000	.334	2.995
	Lg10NPT	.009	.007	.051	1.247	.218	.916	1.091

a. Dependent Variable: Literacy

The R^2 value is 91.5 percent, meaning that 91 percent of the variation in the country's literacy rate accounts for the five independent variables. The modified R2 also calculates the frequency of the relationship between the independent variable and the literacy rate, and the number of independent variables in the regression model is also accounted for. Durbin-Watson is also reduced to 2.160 in comparison to the value reported during the assumptions check.

Regression Equation:

Y= -1 + 1.069(GPI) + .109(Government Expenditure)- .060(GII) - .136(Poverty) + .009(Lg10NPT)

Now, a global hypothesis test will be conducted. It will check if any of the regression coefficients are not zero. .05 significance is used.

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
H_1: Not all β's is 0.

The P-value from the above ANOVA table is .000, which is less than the significance level. Hence, rejecting the null hypothesis and accepting the alternate hypothesis of at least one regression coefficient being not zero.

Now, the Individual regression coefficient is evaluated. Referring to the above coefficients table having p-values to test each regression coefficient.
H_0: $\beta_k = 0$
H_1: $\beta_k \neq 0$
K= Different Independent Variables.

p-values in the above coefficients table, written in red are not significant, as they are greater than the significance level, leaving us with two significant variables i.e. Gender Parity Index and Poverty Headcount on 5.5 $ a day. We fail to reject the null hypothesis of the remaining three variables and conclude that regression coefficients of these variables are not related to Literacy Rate and are not effective predictors. Hence, they need to be removed from the regression equation and must be one at a time to analyze the overall effect. The largest p-value will be removed first and so on.

- **Model 2:**

After removing the largest p-value i.e. .369 of Gov. Education Expenditure % of total expenditure, another model is built. R2 changes with a very low percentage without this variable. Repeating the process of removing the p-values in red in the below coefficients table. Below is a summary of the model.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.956[a]	.914	.908	.043196	2.163

a. Predictors: (Constant), Lg10NPT, Gender Parity Index, Poverty Headcount 5.5$ a day%, Gender Inequality Index

b. Dependent Variable: Literacy

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.128	4	.282	151.157	.000[b]
	Residual	.106	57	.002		
	Total	1.235	61			

a. Dependent Variable: Literacy

b. Predictors: (Constant), Lg10NPT, Gender Parity Index, Poverty Headcount 5.5$ a day%, Gender Inequality Index

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	-.075	.077		-.970	.336		
	Gender Parity Index	1.063	.066	.727	15.993	.000	.732	1.366
	Gender Inequality Index	-.063	.064	-.067	-.992	.325	.334	2.992
	Poverty Headcount 5.5$ a day%	-.135	.029	-.308	-4.588	.000	.335	2.987
	Lg10NPT	.009	.007	.048	1.183	.242	.922	1.085

a. Dependent Variable: Literacy

Removing the variable Gender Inequality Index having p-value .325 in red from the above coefficients table does not impact on R2 value.

- **Model 3:**

In this model, removing the last remaining insignificant independent variable (Lg10NPT) from the above model. It makes no difference in the values of R2 and adjusted R2 making these independent variables of no importance in the final regression equation. Below is a summary of the final model.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.954[a]	.911	.908	.043264	2.118

a. Predictors: (Constant), Poverty Headcount 5.5$ a day%, Gender Parity Index

b. Dependent Variable: Literacy

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.124	2	.562	300.288	.000[b]
	Residual	.110	59	.002		
	Total	1.235	61			

a. Dependent Variable: Literacy

b. Predictors: (Constant), Poverty Headcount 5.5$ a day%, Gender Parity Index

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	-.070	.068		-1.043	.301		
	Gender Parity Index	1.082	.065	.739	16.615	.000	.766	1.306
	Poverty Headcount 5.5$ a day%	-.150	.019	-.344	-7.728	.000	.766	1.306

a. Dependent Variable: Literacy

The adjusted R^2 of the model is 91 percent approximately. These two variables collectively describe the variation in Literacy Rate by 91 percent.

Regression Equation:

Y= -.070 + 1.082(Gender Parity Index) - .150(Poverty Headcount 5.5 $ a day%)

Global Hypothesis for ANOVA:
H_0: $\beta_1 = \beta_2 = 0$
H_1: Not all β's is 0.

As per the ANOVA table, the p-value (.000) is less than the level of significance. Therefore, rejecting the null hypothesis and accepting the alternative hypothesis.

Individual Regression Coefficient Evaluation:
H_0: $\beta_k = 0$
H_1: $\beta_k \neq 0$.

As per the coefficients table, the p-values of both the independent variables are significant and are different from zero. Both regression coefficients of these independent variables are related to Literacy Rate.
Durbin-Watson has improved a lot and is very much near to 2. In this model, we do not have any multicollinearity as the VIF value is

much smaller than 10.
Since the model accuracy is very high and all the hypothesis test is passed, making it the final model.
The literacy rate is highly dependent on the Gender Parity Index and less on Poverty Headcount 5.5 $ a day.

V. Final Model Summary

Model 3 is the final model and is used for regression analysis in this project.
Please refer to the Model Summary, ANOVA, and Coefficients screen capture in Model 3 to have a brief description of the model, and below are the details of each of its elements. Based on the model summary, the literacy rate of a country can be predicted with 90 % accuracy having a Gender parity index and prevailing poverty in that country.

R-Value
It is the correlation coefficient between the dependent and independent variables. The R-value of the final model is .954, which denotes a very strong relationship.

R Square

The percentage of the overall difference in the dependent variable Y that the variation in the independent variable X reflects or accounts for. It is also called a coefficient of determination. The R^2 value of the final model is .911.

Adjusted R Square

It is also called an Adjusted Coefficient of Determination. The forecasts are more reliable due to each new independent variable. This makes SSE smaller and SSR bigger, increasing R2 because of the rise in the overall number of independent variables and not because it is a strong indicator. This is the reason Adjusted R square is considered. The value of it in the final model is .908.

Standard Error of the Estimate

A calculation of the dispersion of the values observed for a given value of the predictors along the regression axis. It is .043264 for our final model i.e. Model 3.

ANOVA Table:

This table is majorly used to test the significance of the model. Hypothesis testing is done and explained in the Model 3 creation which is our final model.
Extending the explanation, F-statistics is used to test the significance with k=2 degrees of freedom in the numerator and n-(k+1) = 62-(2+1) =59 degrees of freedom in the denominator. Using the 0.05 significance level, H0 can be rejected if the calculated F (MSR/MSE) is greater than the critical value (3.15).
Since, F-value (300.288) > Critical Value (3.15), the null hypothesis (H0) of having regression coefficients as Zero is rejected and an alternate hypothesis (H1) is accepted for this model. Below are the two hypotheses used in this evaluation.

H0: $\beta_1 = \beta_2 = 0$
H1: Not all β's is 0.

Coefficient

Unstandardized β represents the coefficient by which each independent variable affects the model to calculate the dependent variable if other predictors are constant in the regression equation. The individual hypothesis is done and explained in the model creation topic of model 3 and a detailed explanation of the process is at the start of model building to get the final model which is Model 3.

β_0 (Constant)=-.070
β_1= 1.082, Increase in Gender Parity Index by 1unit accounts for the increase in Literacy Rate by 1.082 %.
β_2 = -.150, Increase in Poverty by 1 % decreases the literacy Rate by .150 %.

Collinearity Diagnostics:

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	Gender Parity Index	Poverty Headcount 5.5$ a day%
1	1	2.673	1.000	.00	.00	.03
	2	.324	2.874	.00	.00	.68
	3	.004	27.593	1.00	.99	.28

a. Dependent Variable: Literacy

EigenValue is used to check the variance among the independent Variables. Condition Index for one variance of a variable is high with a value of 27.593.

Casewise Diagnostics:

Casewise Diagnostics[a]

Case Number	Std. Residual	Literacy	Predicted Value	Residual
1	.812	.990	.95514	.035147
2	.241	.968	.95711	.010417
3	-.173	.993	1.00055	-.007504
4	1.445	.998	.93596	.062498
5	-2.830	.832	.95441	-.122434
6	-.054	.525	.52783	-.002342
7	.079	.920	.91700	.003404
8	-.025	.989	.99041	-.001088
9	-.352	.980	.99523	-.015223
10	-3.117	.525	.65994	-.134849

It diagnoses the actual and predicted values of the dataset by taking it as the parameter to this function.
For example – Actual Literacy - .990, predicted - .95514 for the first country in the dataset.

VI. Conclusion:

This project focused on applying Multiple Regression to predict the Literacy rate of different countries based on multiple factors affecting it. Gender Parity Index and Poverty are the two factors that are affecting the most on Literacy Rate of any country according to the dataset worked upon. The model built is predicting the literacy rate of a country on given factors with 90.8 % accuracy.

VII. References:

[1] T. Jinyu and Z. Xin, "Apply multiple linear regression model to predict the audit opinion," 2009 ISECS International Colloquium on Computing, Communication, Control, and Management, Sanya, 2009, pp. 303-306, doi: 10.1109/CCCM.2009.5267661.