DEVI AHILYA VISHWAVIDYALAYA

SCHOOL OF STATISTICS

NAME : VANSHIKA DHAKAD

ROLL NO : ST4A2108

SUBJECT : ECONOMETRICS ASSIGNMENT

CLASS : B.Sc.(Hons)-ASA

SEMESTER : 4th

SUBMITTED TO: Dr. DEEPSHIKHA AGRAWAL

What is Regression Analysis?

Regression analysis is a statistical tool used to study the relationship between two or more variables. It is used to investigate how a dependent variable depends on one or more independent variables. Regression analysis attempts to determine how the dependent variable is related to a series of other changing variables.

The main objective of regression analysis is to express the response variable as a function variable of the predictor variables.

Let's have an example of linear regression, which is a linear relationship between response variable, Y, and the predictor variable, $X_{i,}$ i=1, 2…., n of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_n x_n + \varepsilon_i$$

where, betas are the regression coefficients (unknown model parameters), and epsilon is the error due to variability in the observed responses.

Source: Eurostat

Year	Y GDP	X1 Education Spend	X2 Unemployment Rate (% of Labor Force)	X3 Employee Compensation
2000	256,376	14,185	7.00	128,564
2001	264,335	15,004	6.60	135,710
2002	273,256	15,821	7.50	141,985
2003	281,200	16,566	8.20	144,669
2004	296,820	16,709	8.40	148,851
2005	310,038	17,646	8.50	153,985
2006	325,152	18,295	8.30	161,393
2007	343,619	18,962	7.50	170,106
2008	351,743	20,133	7.00	179,628
2009	346,473	21,071	7.90	180,906
2010	363,140	21,936	8.30	184,711
2011	375,968	23,356	7.20	193,171
2012	386,175	24,158	7.60	199,806
2013	392,880	25,045	8.40	203,606
2014	403,003	25,436	8.50	206,201
2015	416,701	26,282	8.50	208,128
2016	430,085	26,675	7.80	211,813
2017	444,991	27,853	7.10	219,187
2018	460,419	28,618	6.00	226,300

Performing regression analysis on this data :

X1 – Education Spend in mil.;
X2 – Unemployment Rate as % of the Labor Force;
X3 – Employee compensation in mil.
Y-GDP

APPLYING MULTIPLE REGRESSION MODEL :

Regression Statistics	
Multiple R	0.993733245
R Square	0.987505762
Adjusted R Square	0.985006914
Standard Error	7659.401455
Observations	19

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	69552186643	23184062214	395.1844685	1.71276E-14
Residual	15	879996459.8	58666430.65		
Total	18	70432183103			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29,500.43	31,570.15	0.93	0.364876647	(37,789.76)	96,790.62	(37,789.76)	96,790.62
X1Education Spend	3.98	3.65	1.09	0.29258399	(3.80)	11.77	(3.80)	11.77
X2Unemployment Rate(% of Labor Force)	(2,189.43)	2,466.12	(0.89)	0.388659752	(7,445.83)	3,066.97	(7,445.83)	3,066.97
X3Employee Compensation	1.43	0.56	2.58	0.021035594	0.25	2.62	0.25	2.62

INTERPRETATION :

Multiple R: 0.993: This is the multiple correlation between the dependent variable (GDP) and three independent variables.

R-Square: 0.987: This is the percentage of the variance in the dependent variable explained by the independent variables. The R-Square value of **0.987** indicates that 98.7% of the variation in GDP can be explained by the three variables. This R-Square is also called the coefficient of determination.

Adjusted R Square: 0.985: This R-Square value is adjusted for the number of independent variables in the model. This value is always smaller than the R-Square and will decrease when we use more independent variables.

Analysis of Variance (ANOVA)

Degrees of freedom (df):

Regression df is the number of independent variables in our regression model.

Residual df is the total number of observations (rows) of the dataset subtracted by the number of variables being estimated.

Total df — is the sum of the regression and residual degrees of freedom, which equals the size of the dataset minus 1.

Sum of Squares (SS):

Regression SS is the total variation in the dependent variable that is explained by the regression model. It is the sum of the square of the difference between the predicted value and mean of the value of all the data points.

Residual SS — is the total variation in the dependent variable that is left unexplained by the regression model. It is also called the **Error Sum of Squares** and is the sum of the square of the difference between the actual and predicted values of all the data points.

Total SS — is the sum of both, regression and residual SS

Mean Squared Errors (MS) — are the mean of the sum of squares or the sum of squares divided by the degrees of freedom for both, regression and residuals.

F — is used to test the hypothesis that the slope of the independent variable is zero. Mathematically, it can also be calculated as

F = Regression MS / Residual MS

This is otherwise calculated by comparing the F-statistic to an F distribution with regression df in numerator degrees and residual df in denominator degrees.

Significance F — is nothing but the p-value for the null hypothesis that the coefficient of the independent variable is zero and as with any p-value, a low p-value indicates that a significant relationship exists between dependent and independent variables.

t-Stat — T-Stat – this is the t-statistic for the null hypothesis that the coefficient is equal to zero, versus the alternative hypothesis that it is different from zero;its value is equal to the coefficient divided by the standard error.

p-value — The t-statistic is compared with the t distribution to determine the p-value.
A p-value below 0.05 indicates 95% confidence that the slope of the regression line is not zero and hence there is a significant linear relationship between the dependent and independent variables.

A p-value greater than 0.05 indicates that the slope of the regression line may be zero and that there is not sufficient evidence at the 95% confidence level that a significant linear relationship exists between the dependent and independent variables.

RESULT :

In our model t calculated is less than t tabulated (2.262) , hence we accept null hypothesis that is coefficient is equal to zero.

Also our p value is less than 0.05 that the slope of the regression line is not zero and hence there is a significant linear relationship between the dependent and independent variables.

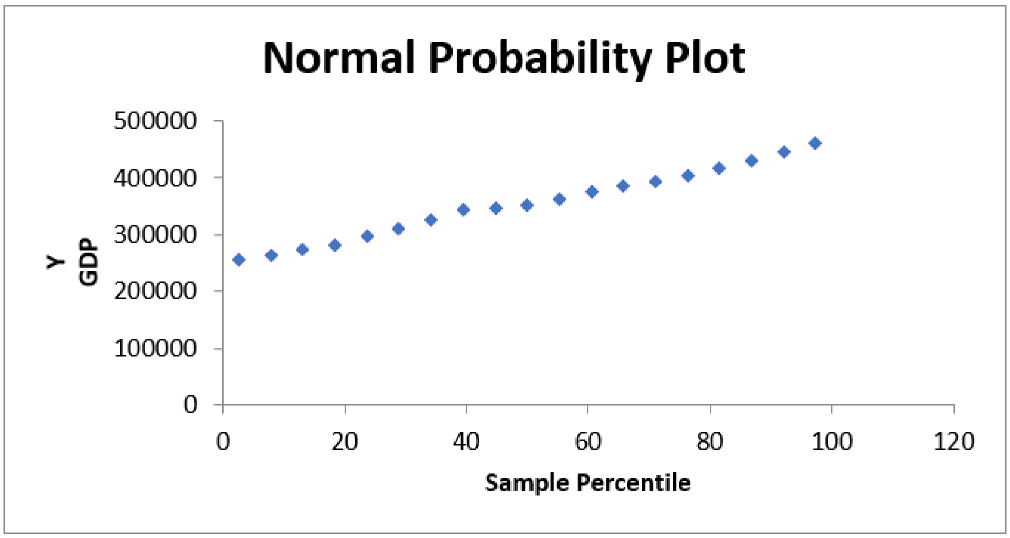
RESIDUAL OUTPUT-

Observation	*Predicted GDP*	*Residuals*	*Standard Residuals*
1	255043.515	1332.484993	0.190704016
2	269430.3434	-5095.343385	-0.729240817
3	279712.5404	-6456.540389	-0.924054069
4	284993.2664	-3793.266445	-0.542888773
5	291125.1215	5694.878484	0.815045724
6	301998.9147	8039.085341	1.150546434
7	315647.863	9504.136964	1.360223262
8	332554.9795	11064.02046	1.583472341
9	351969.9039	-226.9038647	-0.032474271
10	355562.3551	-9089.355051	-1.30086006
11	363585.792	-445.7920078	-0.063801338
12	383780.5969	-7812.596916	-1.11813162
13	395614.6781	-9439.678099	-1.350997969
14	402842.4782	-9962.478215	-1.425820637
15	407901.9372	-4898.937179	-0.701131343
16	414030.7858	2670.214235	0.382158583
17	422414.2341	7588.765901	1.086097133
18	439212.5198	5778.480151	0.827010717

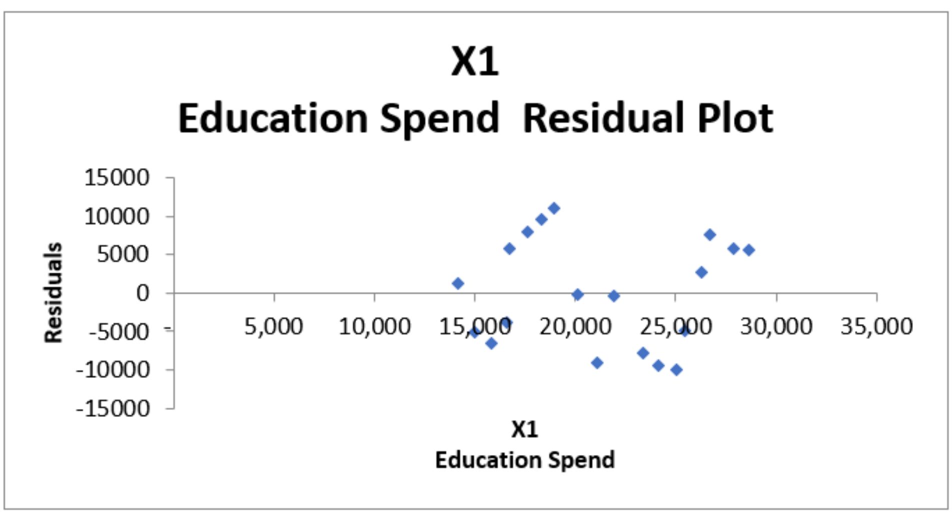
19	454870.175	5548.82502	0.794142687

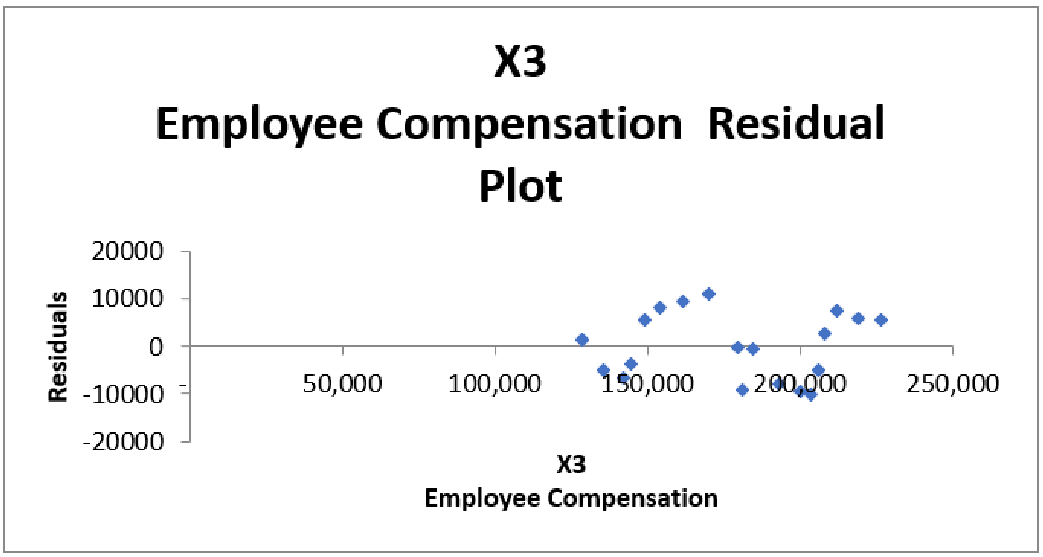
Plots :

Normal Probability Plot: The Normal Probability Plot helps us determine whether the data fit a normal distribution.

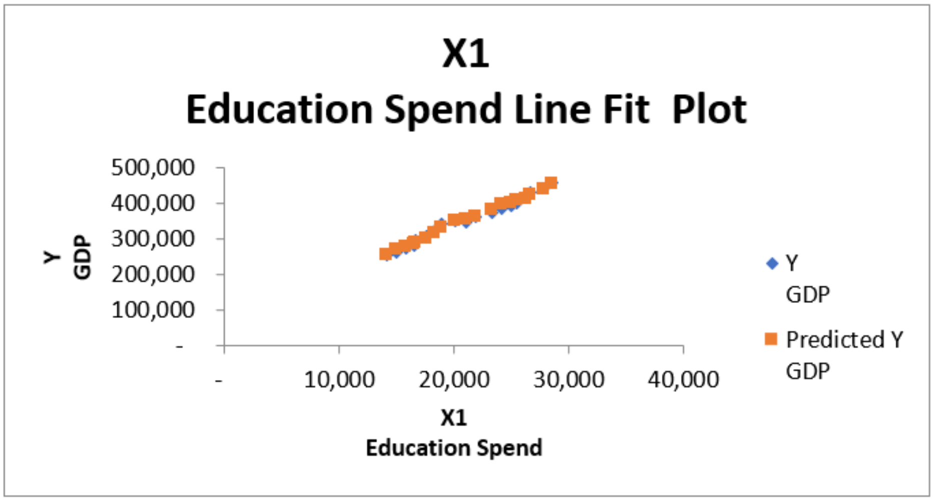


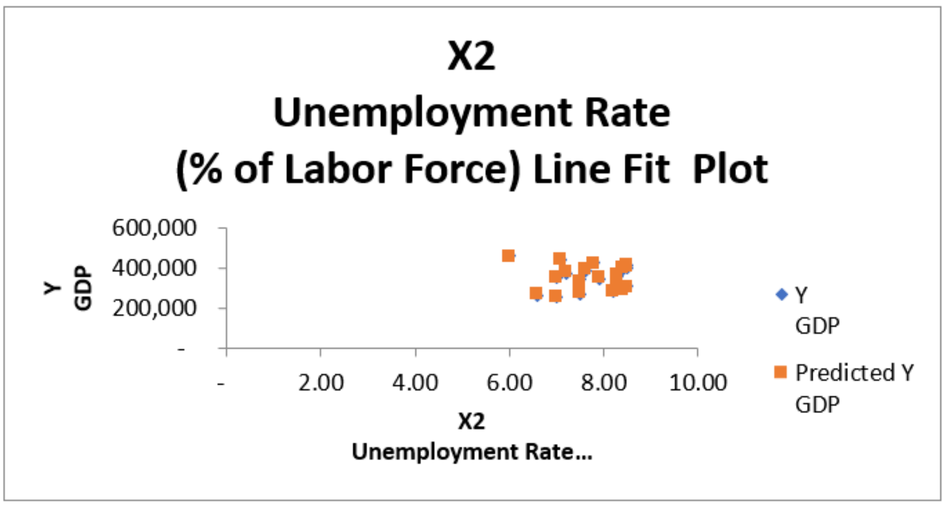
Residual Plots:

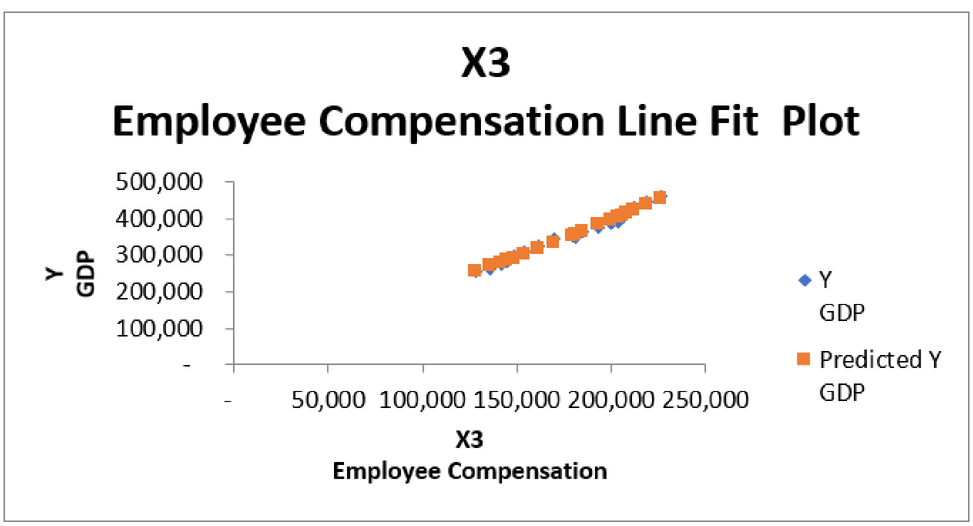




Line Fit Plots: The model provides us with one Line Fit Plot for each independent variable (predictor). This shows the predicted values ($\hat{y}$) versus the observed values (y). The closer these match, the better our model predicts the dependent variable based on the regressors.

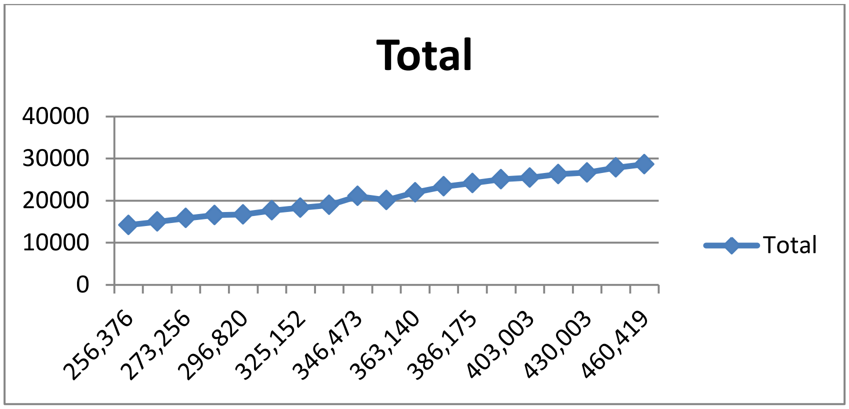

X1
Education Spend Line Fit Plot
Y
GDP
500,000
400,000
300,000
200,000
100,000
-
-
10,000
20,000
30,000
40,000
X1
Education Spend
Y
GDP
Predicted Y
GDP


X2
Unemployment Rate
(% of Labor Force) Line Fit Plot
Y
GDP
600,000
400,000
200,000
-
-
2.00
4.00
6.00
8.00
10.00
X2
Unemployment Rate...
Y
GDP
Predicted Y
GDP


X3
Employee Compensation Line Fit Plot
Y
GDP
500,000
400,000
300,000
200,000
100,000
-
-
50,000
100,000
150,000
200,000
250,000
X3
Employee Compensation
Y
GDP
Predicted Y
GDP

Assumptions:

1. Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.



2. Independence: The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.

3. Homoscedasticity: The residuals have constant variance at every level of x.

Test for detecting **homoscedasticity- THE SPEARMAN'S RANK CORRELATION TEST**

$$\rho = 1 - \frac{6\Sigma\, \mathrm{d}_i^2}{n(n^2-1)}$$

- **Ho : Homoscedasticity is present.**
- **H1 : Homoscedasticity is not present**

Residuals	rank od residual	rank of X1	D	D^2
1332.485	11	1	10	100
-5095.34	6	2	4	*16*
-6456.54	5	3	2	4
-3793.27	9	4	5	25
5694.878	14	5	9	81
8039.085	17	6	11	144
9504.137	18	7	11	144
11064.02	19	8	11	144
-226.904	10	9	1	1
-9089.36	3	10	-7	*49*
-445.792	8	11	-3	9
-7812.6	4	12	-8	64
-9439.68	2	13	-11	144
-9962.48	1	14	-13	169
-4898.94	7	15	-8	*64*
2670.214	12	16	-4	16
7588.766	16	17	-1	1
5778.48	15	18	-3	9
5548.825	13	19	-6	36
				1220

Therefore, spearman’s rank correlation coefficient = -0.068 (putting values in the formula)

and

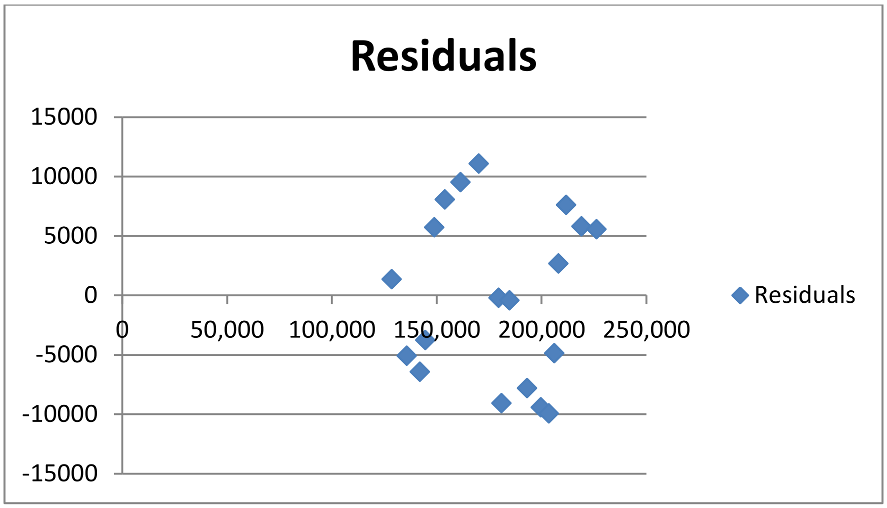
$$t = \frac{r_s \times \sqrt{n-2}}{\sqrt{1-{r_s}^2}}$$

t =-0.2836

also t tabulated=2.262

hence t calculated<t tabulated ; Ho accepted i.e. homoscedasticity is present.

4. No autocorrelation between the disturbances.



5. Normality: The residuals of the model are normally distributed.

6. No exact linear relationship between X2 and X3.