INTRODUCTION:

Our analysis is based on the World Happiness Report. It is a landmark survey of the state of global happiness.

The happiness scores use data from the Gallup World Poll. The scores are based on answers to the main life evaluation question asked in the poll. This question, known as the Cantril ladder, asks respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale. The scores are from nationally representative samples for the years 2013-2016 and use the Gallup weights to make the estimates representative.

The columns used for the estimation of the happiness score reflect the extent to which each of the six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to making life evaluations.

Following is the brief description of the six factors/ variables we have used:

X_1**: Economy**, GDP per capita has been used as a measure for the same.

X_2: **Family**, the closeness between family members and their bond of affection and love has been quantified for the same.

X_3: **Health**, the life expectancy of each country has been used as an indicator.

X_4: **Freedom**, the values of liberty and equality have been quantified for the same.

X_5: **Trust**, data on government corruption has been used as a measure of the same.

X_6: **Generosity**, values of kindness and altruism have been quantified for the same.

All these factors are regressors, which have impact on our regressed variable

Y: **Happiness**, the score for which has been calculated after conduction of surveys by Gallup World Poll.

SIMPLE LINEAR REGRESSION MODEL

1)A Simple linear regression model was fit first relating happiness score against economy (GDP per capita) (X1)

(a)Fitted Model:
y=3.4988097+2.2182271*X1

(b)ANOVA

	Df	*SS*	*MS*	*F*	*Significance F*
Regression	1	125.5399761	125.54	243.9048	1.05E-33
Residual	156	80.29459273	0.514709		
Total	157	205.8345688			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3.4988097	0.133045778	26.29779	3.27E-59	3.236006	3.761613	3.236006	3.761613
X Variable 1	2.2182271	0.142035152	15.61745	1.05E-33	1.937667	2.498787	1.937667	2.498787

(c) By testing for significance of regression coefficient(slope) β_1 , we find there exists a significant linear relationship between x_1 and y.

(d) 95% Confidence interval on slope is (1.937666858, 2.498787). This indicates that 95 times out of 100, we would expect the value of $\beta_1$to lie in this interval.

95% Confidence interval on slope is (3.2360, 3.7616). This indicates that 95 times out of 100, we would expect the value of $\beta_0$to lie in this interval.

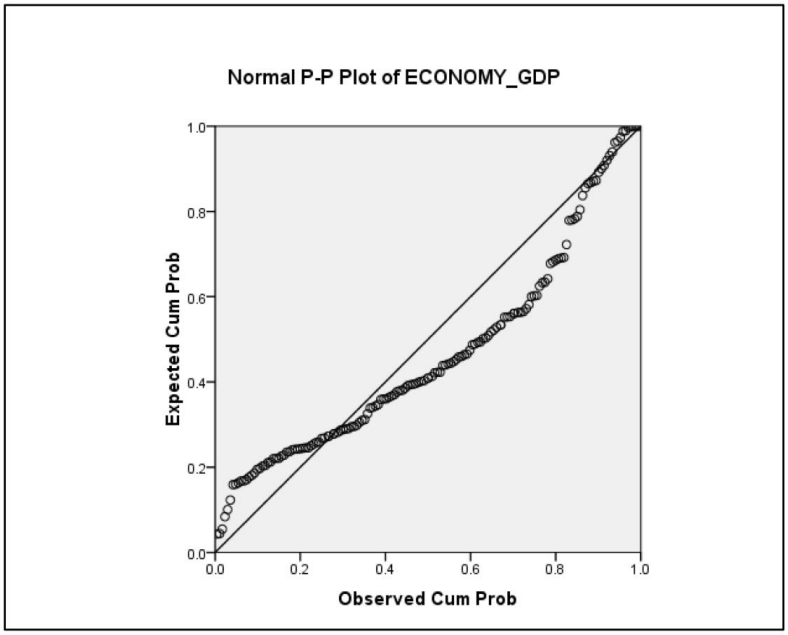
(e) Predicted value of y when x_1=1.5478 (provided, the assumed value of x does not belong to the sample space.)

95% Prediction Interval = (-15.4816, 29.346)

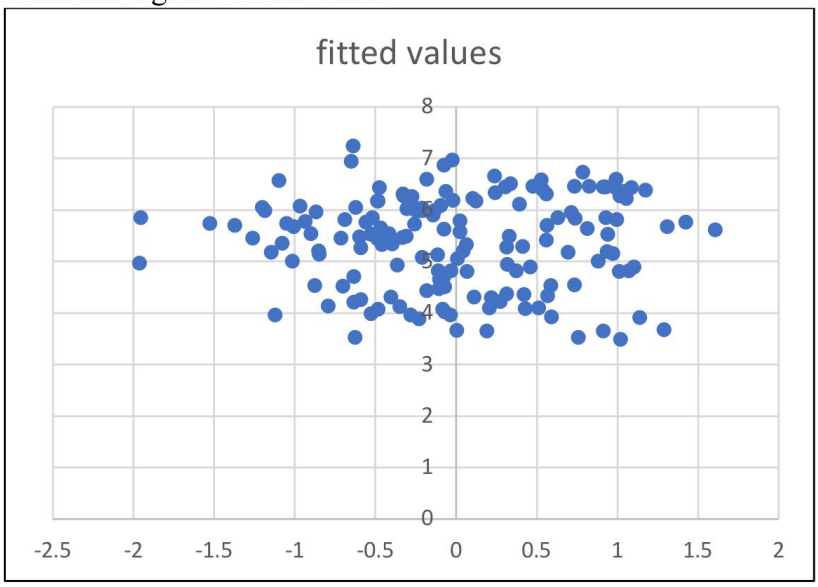
This means that 95 times out of 100, we would expect the value of happiness to fall in this interval when economy(GDP per capita) is limited to 1.5478.

(f) For residual analysis:

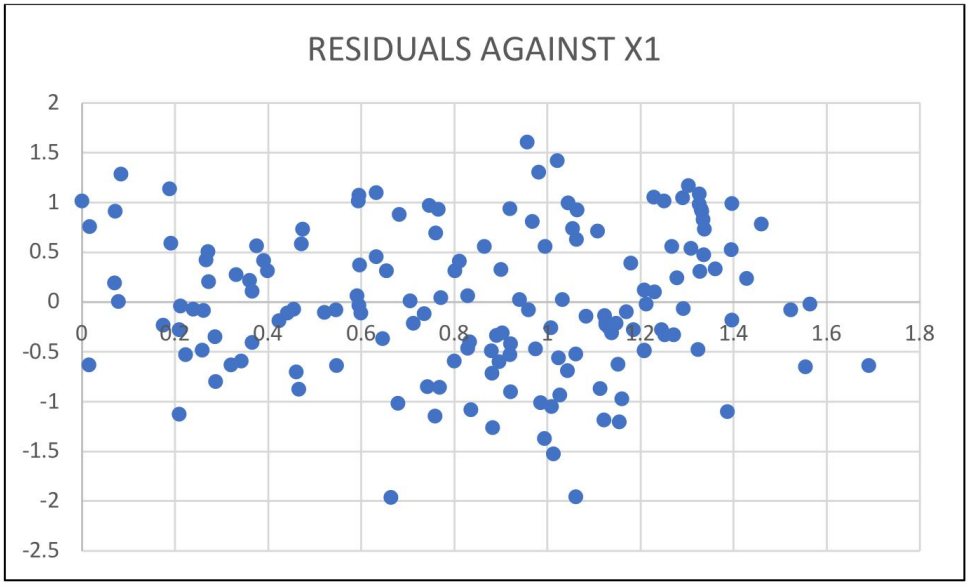
- Normal probability plot



- Residuals against fitted values



- Residuals against regressors



(g) Since $R^2 = 0.6099$, **60.99 %** of total variability in y is explained by this model.

2. A Simple linear regression model was fit first relating happiness score against family (X2)

(a)Fitted Model: y=2.2901=3.1134*X2

(b)ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	112.8994	112.8994	189.5119	9.92E-29
Residual	156	92.93512	0.595738		
Total	157	205.8346			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2.290188	0.232396	9.854675	4.11E-18	1.831139	2.749237	1.831139	2.749237
X Variable 1	3.113424	0.226162	13.76633	9.92E-29	2.666689	3.560159	2.666689	3.560159

(c) By testing for significance of regression coefficient(slope) β_1 , we find there exists a significant linear relationship between x_2 and y.

(d) 95% Confidence interval on slope is (1.8311, 2.7492). This indicates that 95 times out of 100, we would expect the value of $b0$ to lie in this interval.

95% Confidence interval on slope is (2.6666, 3.5601). This indicates that 95 times out of 100, we would expect the value of $b1$ to lie in this interval.

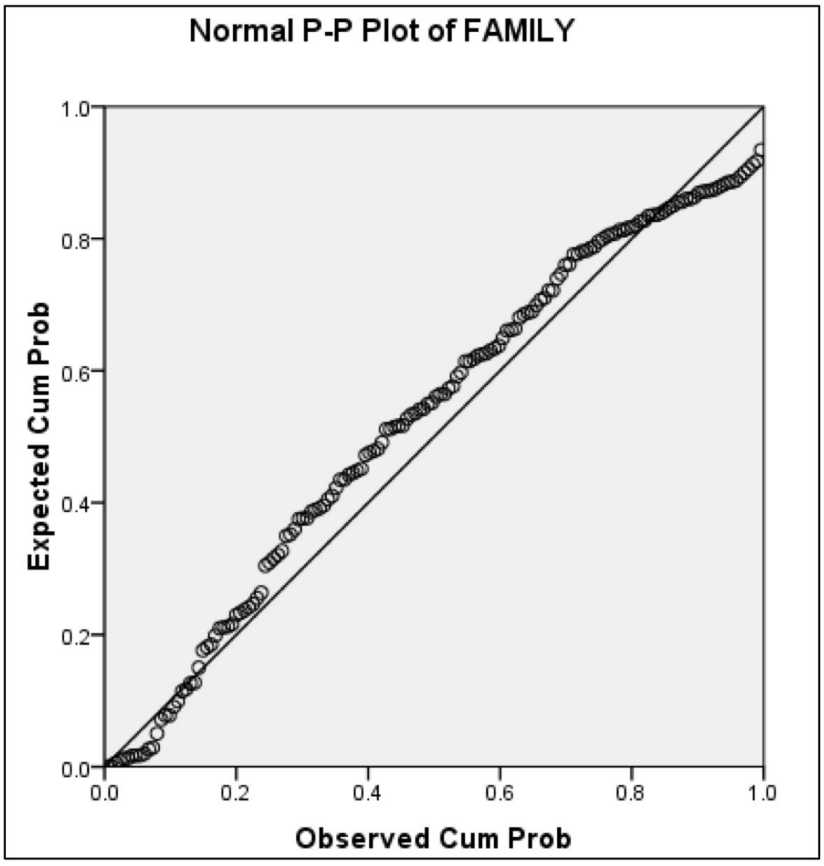
(e) Predicted value of y when x_1=1.3241:

95% Prediction Interval = (4.8760,7.9493)

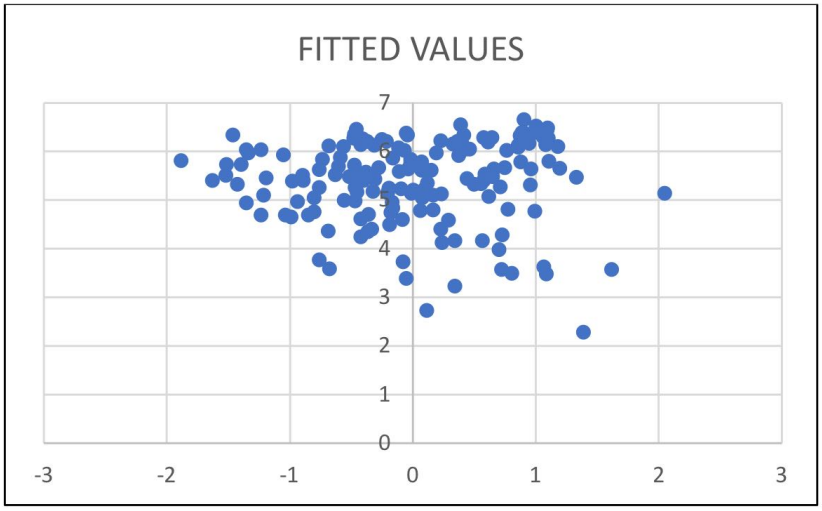
This means that 95 times out of 100, we would expect the value of happiness to fall in this interval when family is limited to 1.3241.

(f) For residual analysis:

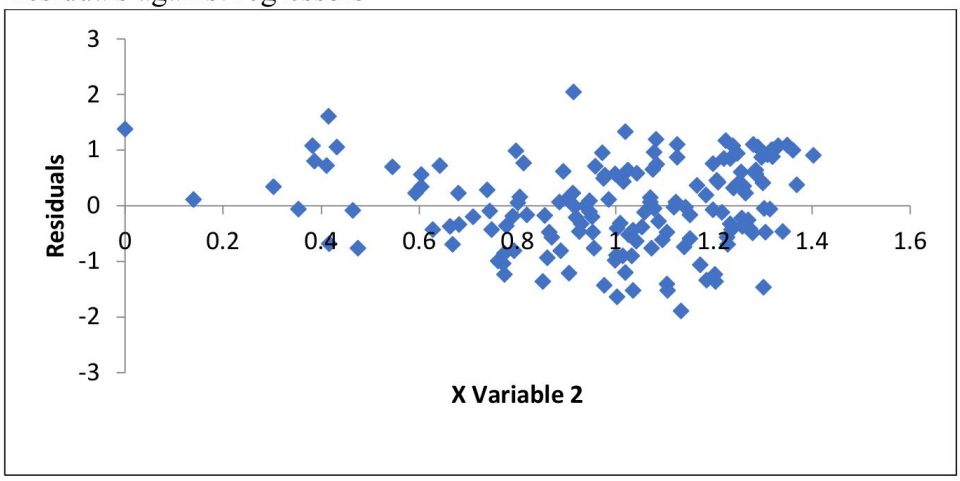
- Normal probability plot



- Residuals against fitted values



- Residuals against regressors



(g) Since $R^2 = 0.5484$, **54.84 %** of total variability in y is explained by this model.

3.A Simple linear regression model was fit first relating happiness score against health (life expectancy) (X3)

(a)Fitted Model:
y=3.2605+3.3560*X3

(b) ANOVA

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	107.953	107.953	172.0516	5.7889E-27

Residual	156	97.88153	0.627446		
Total	157	205.8346			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3.260525	0.173135	18.8323	5.16E-42	2.918534491	3.602516	2.918534	3.602516
X Variable 1	3.356093	0.255861	13.11684	5.79E-27	2.850693044	3.861492	2.850693	3.861492

(c) By testing for significance of regression coefficient(slope) β_1 , we find there exists a significant linear relationship between x_3 and y.

(d) 95% Confidence interval on slope is (2.8506, 3.8614). This indicates that 95 times out of 100, we would expect the value of $\beta_1$to lie in this interval.

95% Confidence interval on slope is (2.9185, 3.6025). This indicates that 95 times out of 100, we would expect the value of $\beta_0$to lie in this interval.

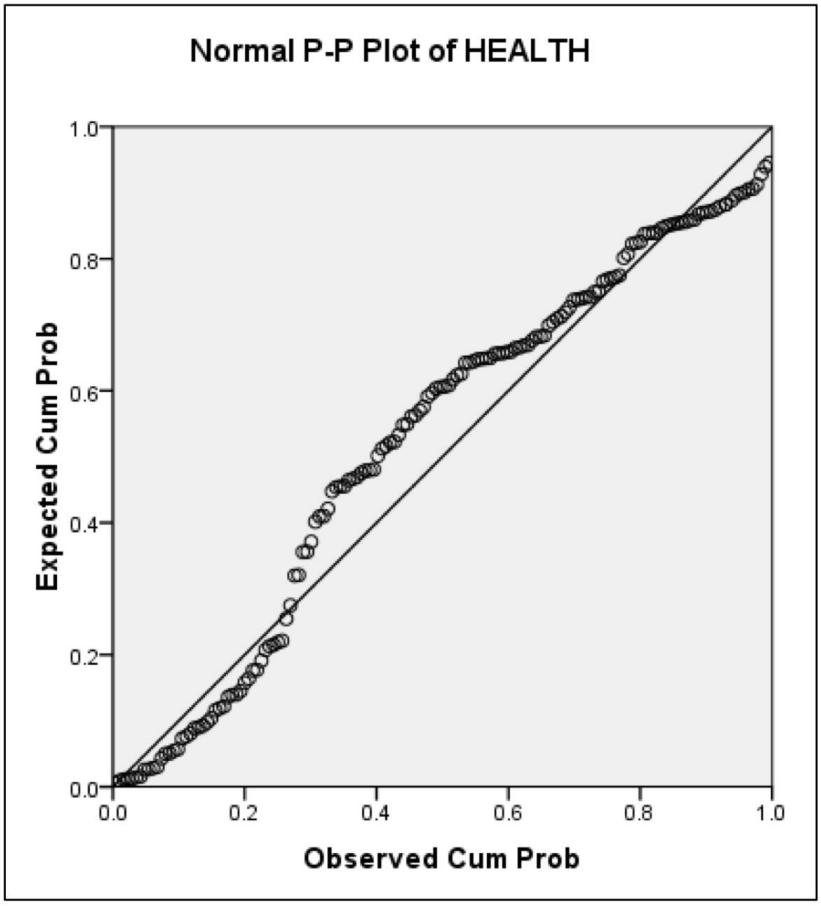
(e) Predicted value of y when x_3=1.0122:

95% Prediction Interval = (5.0761,8.2389)

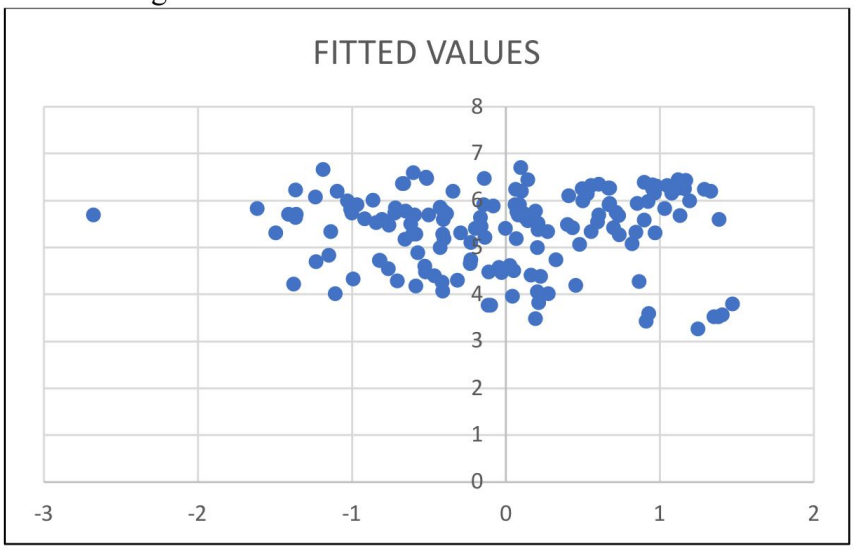
This means that 95 times out of 100, we would expect the value of happiness to fall in this interval when health (life expectancy)is limited to 1.0122.

(f) For residual analysis:

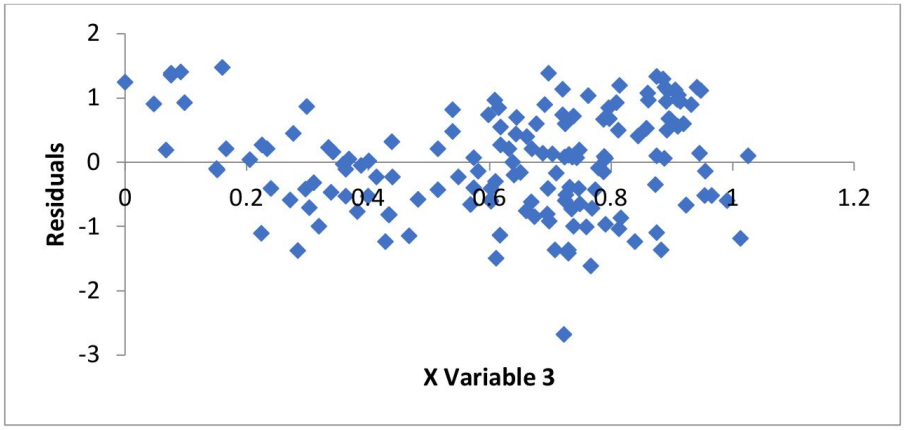
- Normal probability plot



- Residuals against fitted values



- Residuals against regressors



(g) Since $R^2 = 0.5244$, **52.44%** of total variability in y is explained by this model.

4. A Simple linear regression model was fit first relating happiness score against freedom(X4)

(a)Fitted Model:
y=3.5252+4.3174*X4

(b) ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	66.4565	66.4565	74.38195	6.87581E-15
Residual	156	139.3781	0.893449		
Total	157	205.8346			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3.525214362	0.227361	15.50491	2.09E-33	3.07611066	3.974318	3.076111	3.974318
X Variable 1	4.317441268	0.500602	8.624497	6.88E-15	3.32860843	5.306074	3.328608	5.306074

(c) By testing for significance of regression coefficient(slope) β_1 , we find there is a significant linear relationship between x_3 and y.

(d) 95% Confidence interval on slope is (3.3286, 5.3602). This indicates that 95 times out of 100, we would expect the value of $\beta_1$to lie in this interval.

95% Confidence interval on slope is (3.0761, 3.9743). This indicates that 95 times out of 100, we would expect the value of β_0 to lie in this interval.

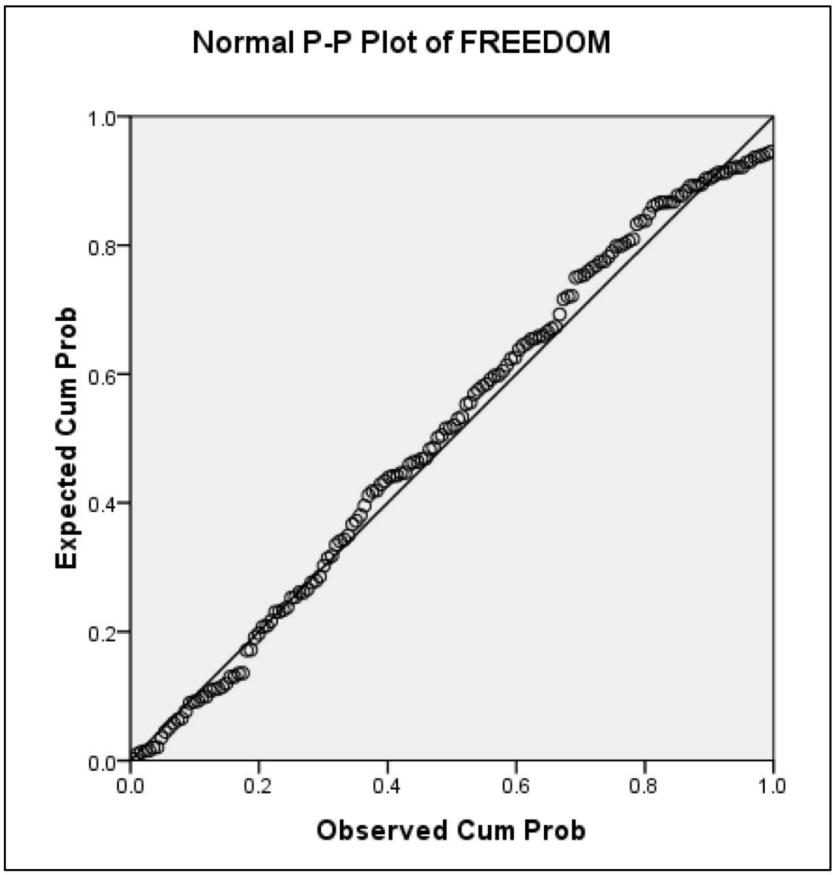
(e) Predicted value of y when x_4=0.5252:

95% Prediction Interval = (3.9173,7.6681)

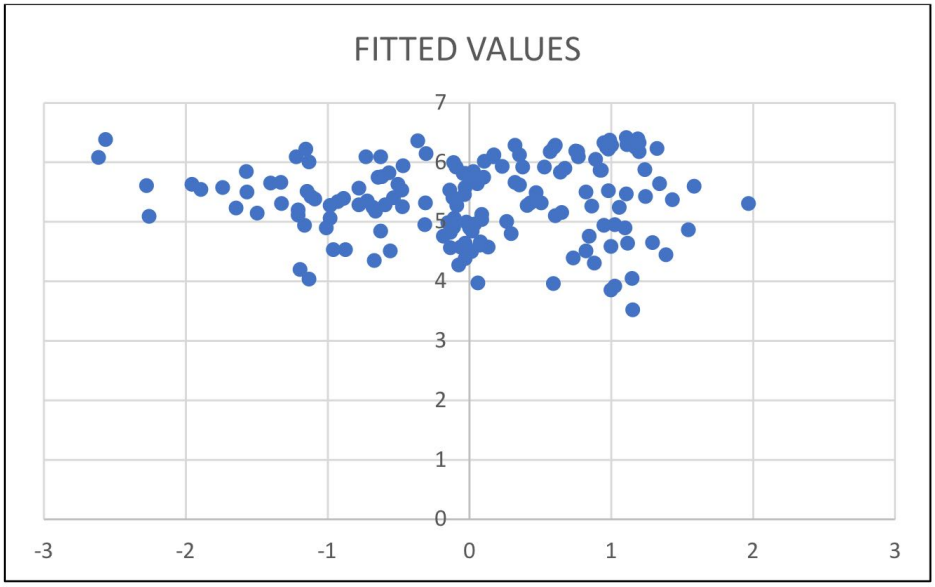
This means that 95 times out of 100, we would expect the value of happiness to fall in this interval when freedom is limited to 0.5252.

(f) For residual analysis:

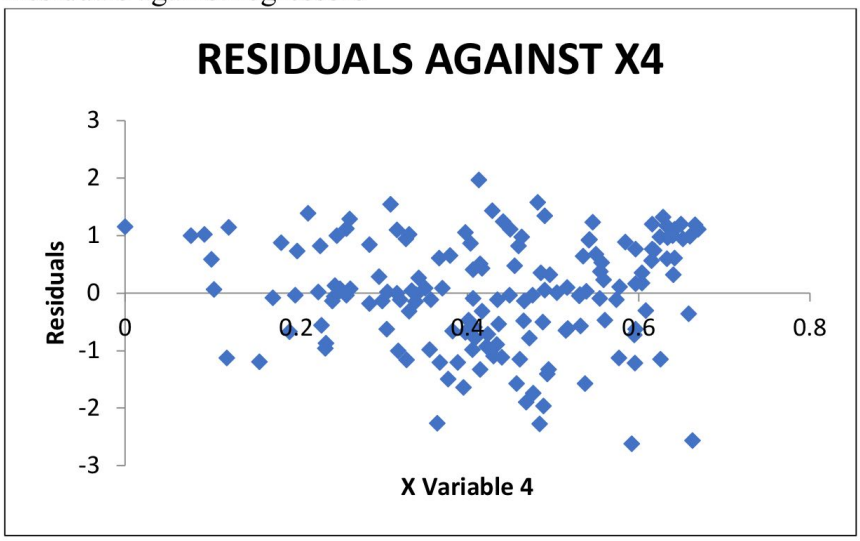
- Normal probability plot



- Residuals against fitted values



- Residuals against regressors



(g) Since $R^2 = 0.3288$, **32.88%** of total variability in y is explained by this model.

5. A Simple linear regression model was fit first relating happiness score against trust (Government corruption) (X5)

(a)Fitted Model:
y=4.8350+3.7698*X5

(b) ANOVA

	Df	SS	MS	F	Significance F
Regression	1	32.1476383	32.1476383	28.87397203	2.76E-07
Residual	156	173.6869305	1.11337776		
Total	157	205.8345688			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	4.835060232	0.131038084	36.89812982	5.81E-79	4.576222342	5.093898122	4.576222342	5.093898122
X Variable 1	3.769816106	0.701563361	5.373450664	2.76E-07	2.384026764	5.155605448	2.384026764	5.155605448

(c) By testing for significance of regression coefficient(slope) β_1 , we find there is a significant linear relationship between x_5 and y.

(d) 95% Confidence interval on slope is (2.3840, 5.1556). This indicates that 95 times out of 100, we would expect the value of β_1 to lie in this interval.

95% Confidence interval on slope is (4.5762, 5.0938). This indicates that 95 times out of 100, we would expect the value of β_0 to lie in this interval.

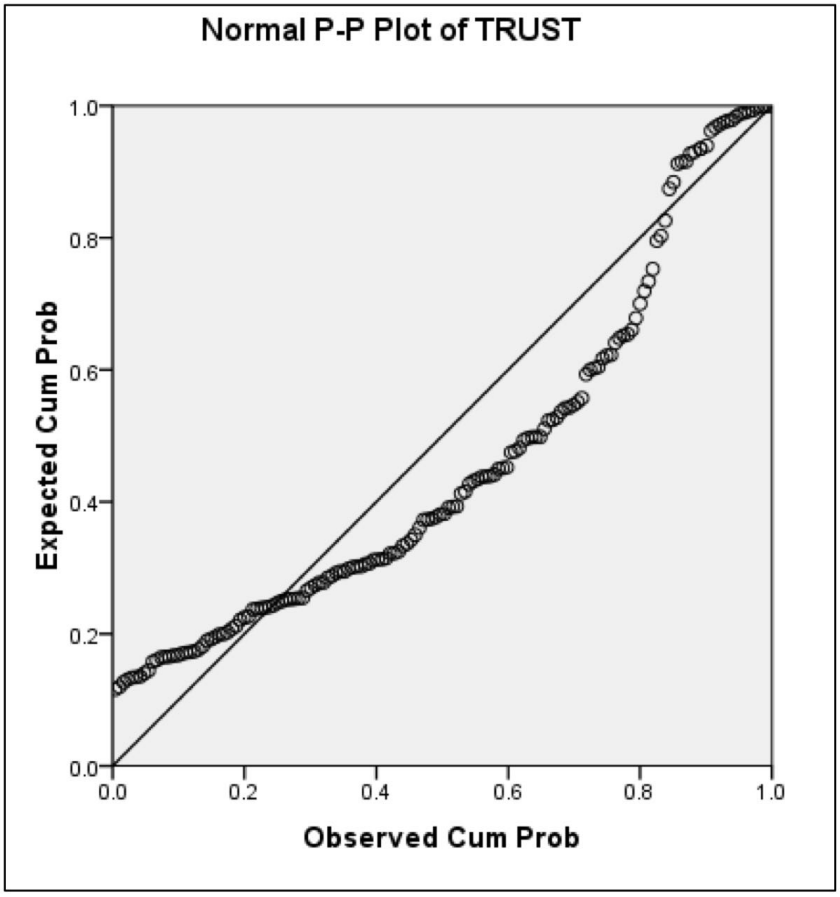
€ Predicted value of y when x_5=:0.3456

95% Prediction Interval = (4.0283,8.2474)

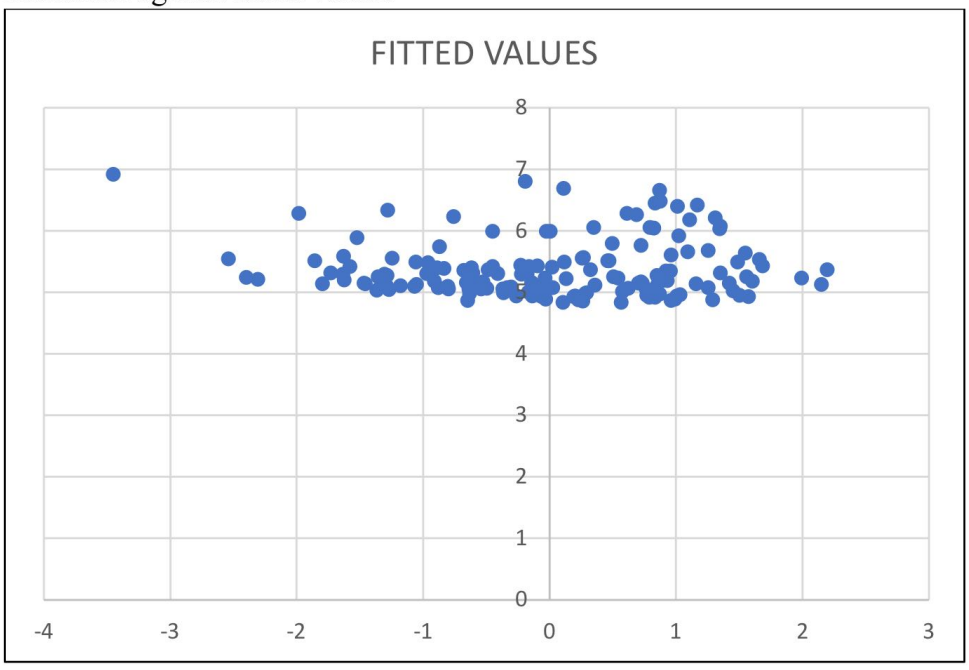
This means that 95 times out of 100, we would expect the value of happiness to fall in this interval when trust is limited to 0.3456.

(f) For residual analysis:

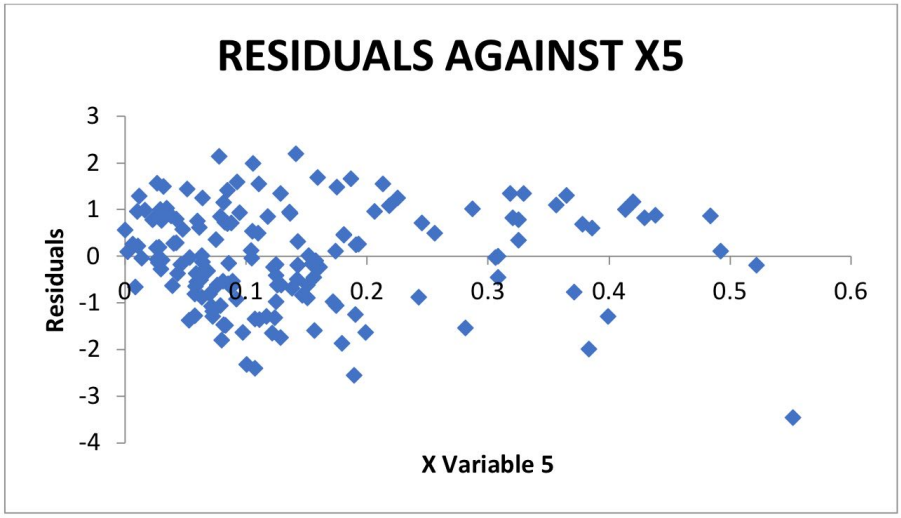
- Normal probability plot



- Residuals against fitted values



- Residuals against regressors



(g) Since $R^2 = 0.1561$, **15.61%** of total variability in y is explained by this model.

6.A Simple linear regression model was fit first relating happiness score against Generosity(X6)

(a)Fitted Model:
y=4.9889+1.6297*X6

(b) ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	6.692664	6.692664	5.242772	0.023379
Residual	156	199.1419	1.276551		
Total	157	205.8346			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	4.988999	0.19133	26.07535	9.63E-59	4.611067	5.36693	4.611067	5.36693
X Variable 1	1.629764	0.711777	2.28971	0.023379	0.223799	3.035729	0.223799	3.035729

(c) By testing for significance of regression coefficient(slope) β_1 , we find there is a significant linear relationship between x_6 and y.

(d) 95% Confidence interval on slope is (0.2237, 3.0357). This indicates that 95 times out of 100, we would expect the value of $\beta_1$to lie in this interval.

95% Confidence interval on slope is (4.6110, 4.3669). This indicates that 95 times out of 100, we would expect the value of $\beta_0$to lie in this interval.

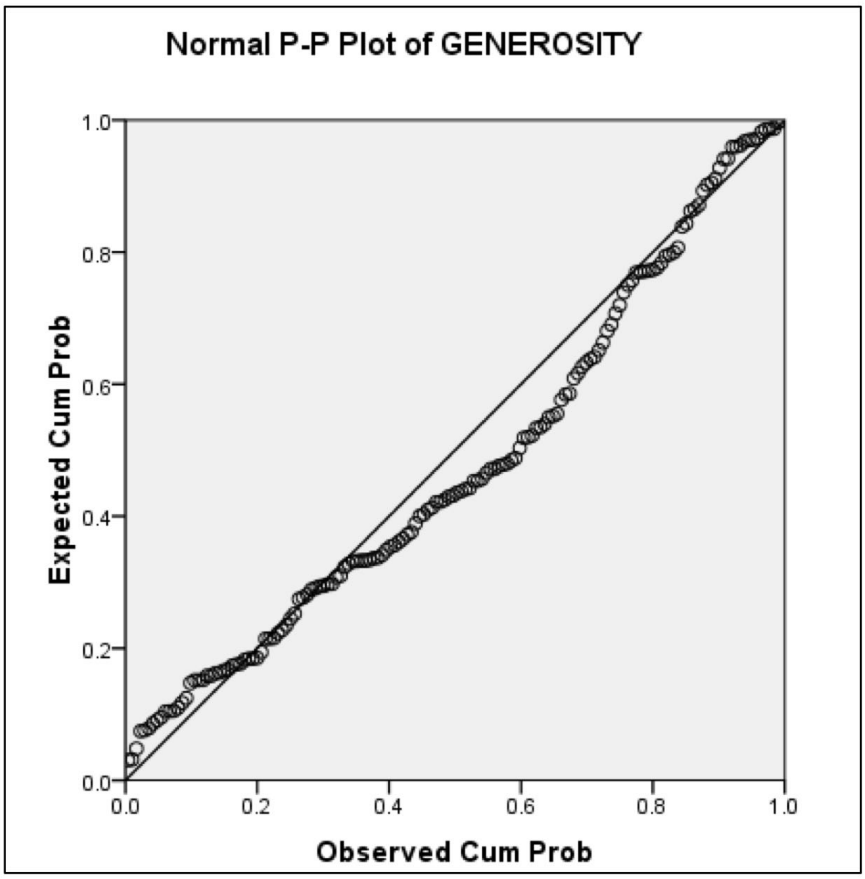
€ Predicted value of y when x_5=:0.2516

95% Prediction Interval = (3.1601,7.6379)

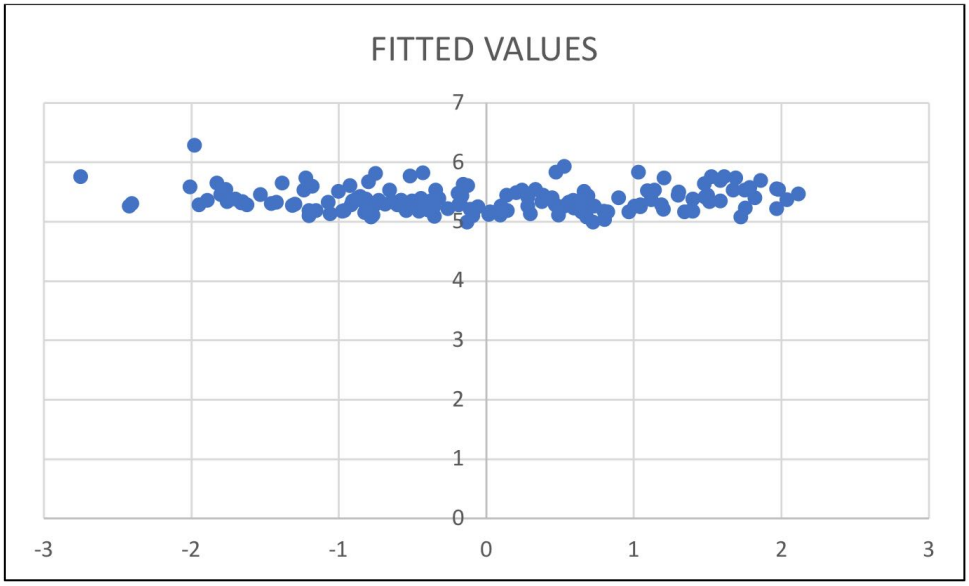
This means that 95 times out of 100, we would expect the value of happiness to fall in this interval when generosity is limited to 0.2516.

(f) For residual analysis:

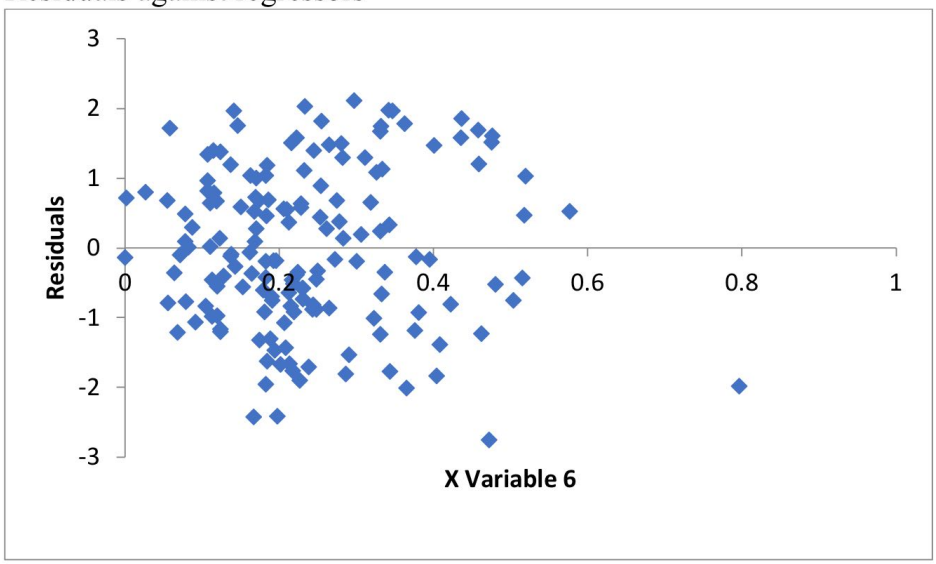
- Normal probability plot
-



- Residuals against fitted values



- Residuals against regressors



(g) Since $R^2 = 0.0325$, **3.25%** of total variability in y is explained by this model.

MULTIPLE LINEAR REGRESSION MODEL

1. FITTING OF A MODEL:

$Y= 1.8601 + 0.8606(X_1) + 1.4088(X_2) + 0.9753(X_3) + 1.3334(X_4) + 0.7845(X_5) + 0.3889(X_6)$

2. HYPOTHESIS TESTING

(i) GLOBAL TEST OF SIGNIFICANCE

Ho: $\beta1=\beta2=\beta3=...=\beta6=0$

H1: At least one of the βj is not equal to 0.

F value=	87.80916386
P Value=	1.04286E-46

Since, F value>P value
We reject H_o at 5% level of significance.

(ii) TEST OF SIGNIFICANCE OF INDIVIDUAL REGRESSORS

1) beta 1

H0: There is no significance due to regression. beta1=0
H1: There is significance due to regression. beta1 not equal 0

beta1=	0.860657204
t tabulated =	3.906550128
t calc =	1.975798924

Since t tab > t calc.
Therefore, we reject H0.

SIMILARLY, IT HAS BEEN DONE FOR ALL OTHER REGRESSORS

(iii) TEST OF REGRESSION FOR SUBSET OF REGRESSORS

1) testing for the significance of x1,x2 given that x3,x4.x5.x6 are already in the model:

H0: beta 1=beta 2=0

H1: at least one of the beta 1 and beta 2 not equals to 0

F0 (CALCULATED)	46.27395151
F0(TABULATED)	3.055161773

SINCE F0 (CALCULATED)> TABULATED, WE REJECT H0 AT 0.05 LEVEL OF SIGNIFICANCE

SIMILARLY, IT HAS BEEN DONE FOR ALL OTHER SUBSETS OF REGRESSORS.

3. 95% CONFIDENCE INTERVALS:

	LL	UL
BO	1.48383	2.23654
B1	0.425366	1.295948
B2	0.96893	1.848853
B3	0.350378	1.60024
B4	0.572719	2.094147
B5	-0.07796	1.647039
B6	-0.38361	1.161476

4. PREDICTION INTERVALS:

LL=5.144207

UL=9.083792

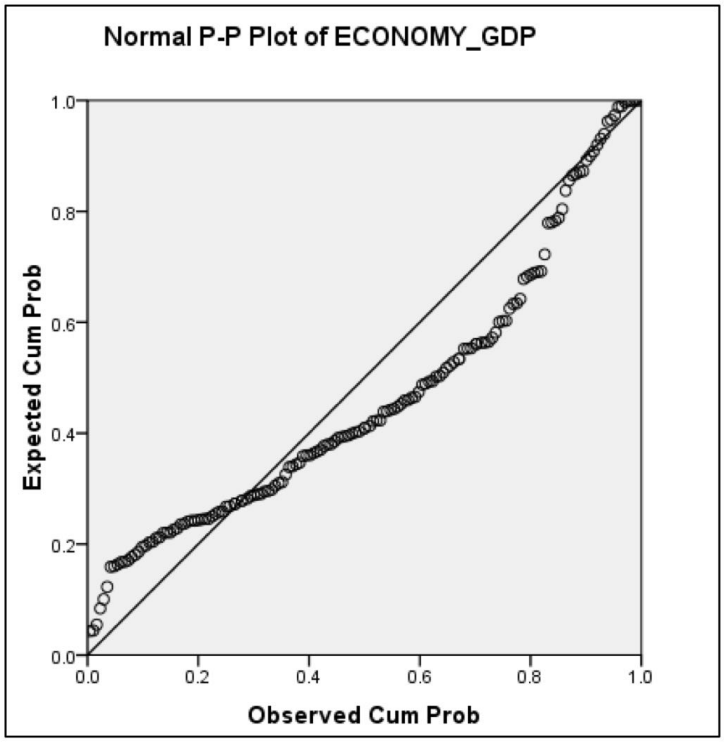
Prediction Interval = (5.144207, 9.083792)

5. RESIDUAL ANALYSIS:

(i) Standardized residuals have been calculated in the excel sheet and no value is greater than 3 or less than -3. So, there is no outlier.

(ii) By residual plots:

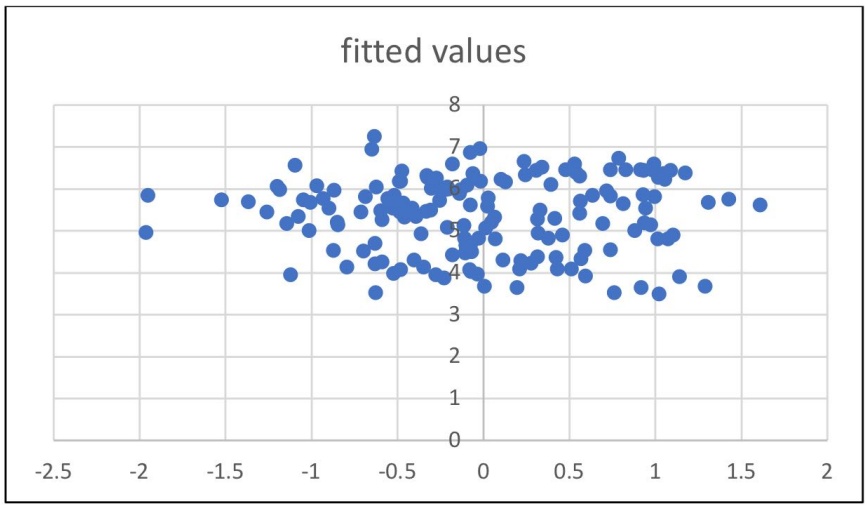
NORMAL PROBABILTY PLOT:



In the P-P plot above, we see that the residuals lie almost along a straight line, so there may not be any violation of the normality assumption for errors.

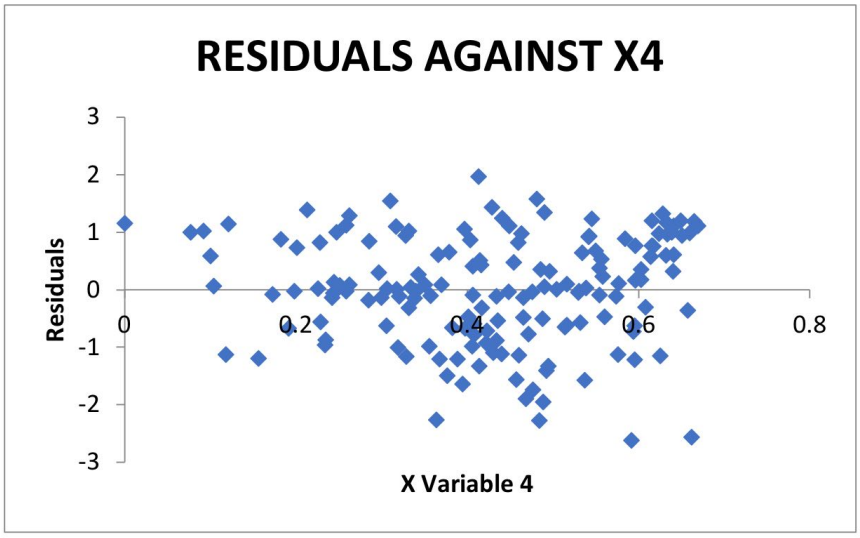
THE NORMAL PP PLOT OF ALL THE VARIABLES WERE ALMOST SIMILAR INDICATING THAT THERE IS NO VIOLATION OF NORMALITY ASSUMPTION, SO ONLY ONE HAS BEEN ATTACHED HERE.

RESIDUALS AGAINST FITTED VALUES:



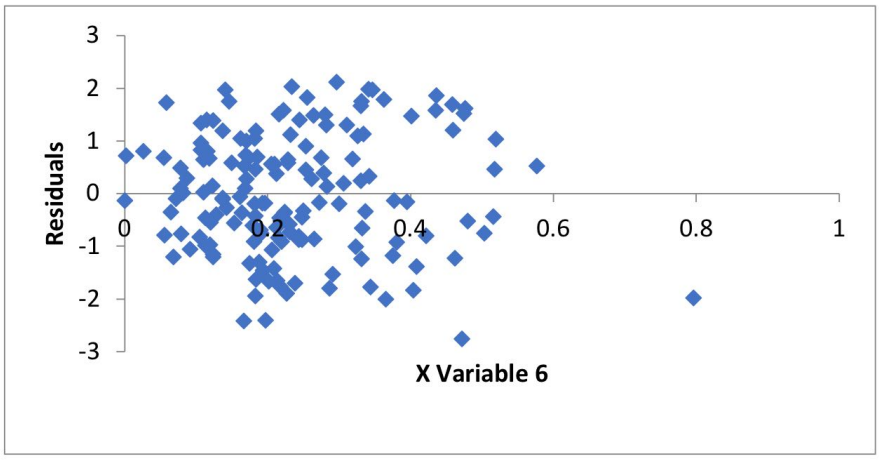
For all the 6 plots, we observe that residuals can be contained in a horizontal band with a random dist. of points. Hence, there are no obvious model deficiencies, particularly concerned with the variance of the errors.

RESIDUALS AGAINST REGRESSORS:



For all the 5 regressors, the residuals can be contained in a horizontal band and hence there is no model inadequacy.

FOR RESIDUALS AGAINST X6:



We suggest that **x6=-0.8** may not belong to the sample due to the fact that it is situated in the different area and has no other point in its proximity. This point should be investigated for a possible outlier.

Except this, the residuals can be contained in a horizontal band and hence there is no model inadequacy.

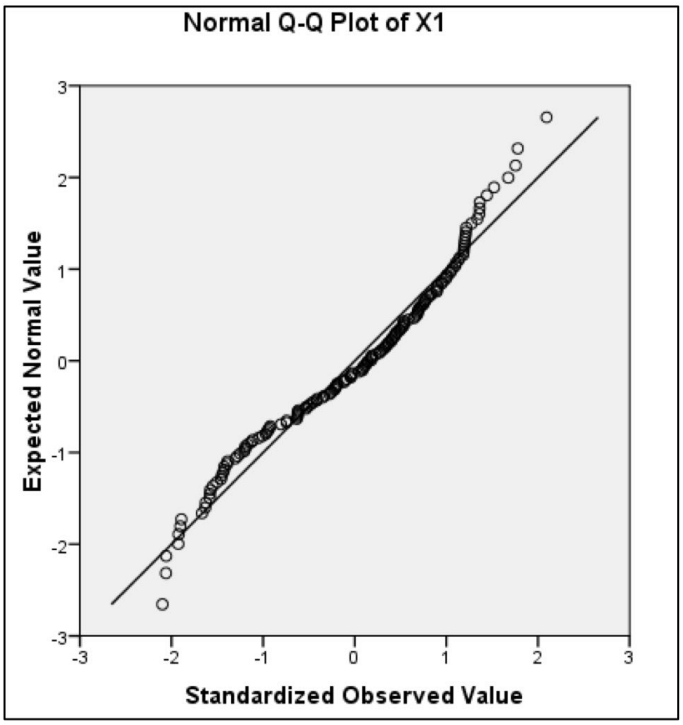
6. LACK OF FIT TEST:

Ho: There is no lack of fit. (Model is adequate)
Mathematically, $E(Y_{ij}) = \beta o + \beta_i(x_i)$

H1: There exists a lack of fit.

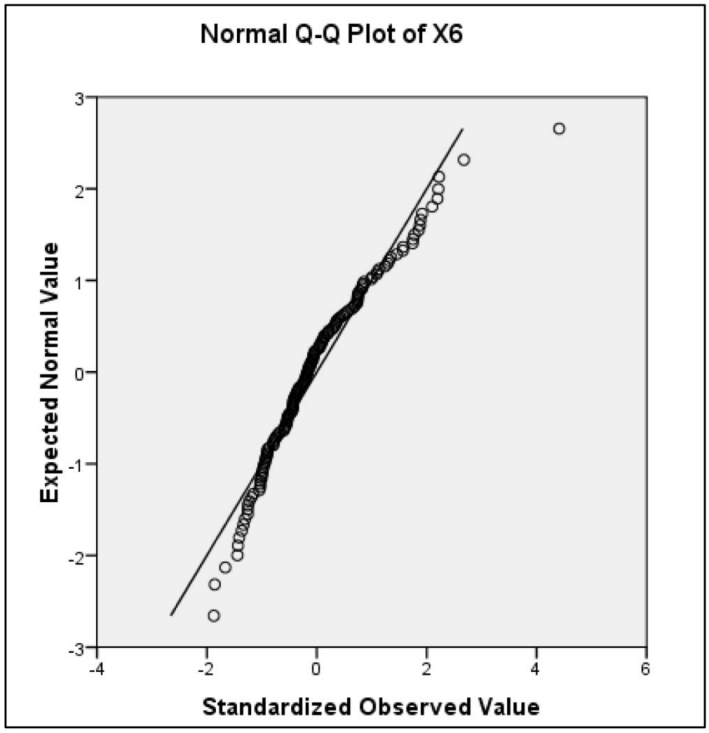
Since, there are no repeating values in our observations, lack of fit test is not applicable and cannot be put to use.

7. Q-Q PLOT:



Since, all the points lie almost close to each other, we see that it is an ordered set of recorded observations.

It is same for all the 5 variables except for x6.



Here, we can see that one point lies away from all the other points, so it must be checked upon.

8. STEPWISE REGRESSION

Model	Variables Entered	Variables Removed
1	X1	.
2	X2	.
3	X4	.
4	X3	.
5	X5	.
a. Dependent Variable: Y		

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.780966	0.609907	0.607406559	0.717432
2	0.84016	0.70587	0.702074318	0.624976
3	0.868778	0.754775	0.74999831	0.572507
4	0.877487	0.769984	0.763970428	0.556279
5	0.880783	0.775779	0.76840313	0.551031
a. Predictors: (Constant), X1				
b. Predictors: (Constant), X1, X2				

c. Predictors: (Constant), X1, X2, X4				
d. Predictors: (Constant), X1, X2, X4, X3				
e. Predictors: (Constant), X1, X2, X4, X3, X5				

We can conclude that the best regression model is:

Y= $1.8601 + 0.8606(X_1) + 1.4088(X_2) + 0.9753(X_3) + 1.3334(X_4) + 0.7845(X_5)$

A model was fit with x1,x3,x4 and x6 and also with x1,x3 and x4 to check what difference does x6 in the model.

But after comparison, it was found that R^2 and adj R^2 are close to each other, also there is no change MSE.

So, model has not become a bad model after removal of X6.

Whereas when the model was fitted with the variables X1, X3 and X4, and X1 and X3. After comparison it was found that the R^2 and adjusted R^2 have decreased and MSE has increased.

If R^2 and adjusted R^2 decrease, and MSE increases, we say that our model is now a bad model.

CONCLUSION

After analysing the data using both SLRM and MLRM, and checking their adequacy by putting various Model adequacy techniques to use, we can infer the following points:

There exists a Strong direct relationship between X1 and Y, i.e., Economy (measured using GDP per capita) and happiness.

There exists a strong direct relationship between X2 and Y, i.e., Family (measured by quantifying the bond of family and their affection) and happiness.

There exists a fairly strong direct relationship between X3 and Y, i.e., Health (measured using GDP per capita) and happiness.

There exists a fairly strong direct relationship between X4 and Y, i.e., freedom (measured by quantifying values of equality and liberty) and happiness.

There exists a weak direct relationship between X5 and Y, i.e., trust (measured using data on government corruption) and happiness.

There exists a weak direct relationship between X6 and Y, i.e., generosity (measured by quantifying the values of kindness and altruism) and happiness.