Course: Data Analysis (task № 3)

Students' Name and Surname ___Rashid Ali_______________________________________

Please do the tasks below.

1. Open top-managers.sav file. Do the test to understand whether the mean ages of top-managers (variable q21) differ significantly in organizations of different industries (variable **qotrasl**). Formulate the hypotheses. Interpret the results of analysis. Do the post-hoc analysis and interpret the results (if applicable). Could we consider the results of analysis as reliable?

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
age	.088	301	.000	.966	301	.000

a. Lilliefors Significance Correction

As the Sig << 0.05, therefore, the age is not distributed normally.

Non Parametric Test
K Independence test (Kruskal-Wallis)

Ranks

	Industry	N	Mean Rank
age	Oil-producing	3	38.50
	Coal	5	172.70
	Ferrous metallurgy	5	201.80
	Non-ferrous metallurgy	6	144.92
	Chemical, Petrochemical	5	137.60
	Engineering, metal	115	174.56
	Wood	22	142.48
	Pulp and Paper	2	140.00
	Textile	16	149.56
	Sewing	27	155.59
	Shoe	6	244.58
	Food	89	116.35
	Total	301	

Test Statistics[a,b]

	age
Chi-Square	37.027
df	11
Asymp. Sig.	.000

a. Kruskal Wallis Test

b. Grouping Variable: Industry

H0 = No difference between means of age and industry

H1 = There is difference between age and industry

Significance is less than 0.05

H0 is reject and H1 is accepted.

So, there is statistically significant difference between means of age and industry

2. Open **world_data.sav** file.
2.1. Do the simple regression analysis. Select the appropriate variables. Write down the regression equation. Assess the goodness-of-fit of the model. Do the diagnostics of the model. Are the residuals normally distributed? Delete the cases than have standardized residuals greater than 3. Will this improve the model?

Population (Population in thousands) and Birth Rate (Birth rate per 1000 people)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.039[a]	.002	-.008	147298.258

a. Predictors: (Constant), Birth rate per 1000 people

R square = 0.002. Only 0.2% population variation is explained by the model.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3536467977.263	1	3536467977.263	.163	.687[b]
	Residual	2321555119950.187	107	21696776821.964		
	Total	2325091587927.450	108			

a. Dependent Variable: Population in thousands

b. Predictors: (Constant), Birth rate per 1000 people

The model is not significantly better than mean in predicting Population.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	59724.597	32903.202		1.815	.072
	Birth rate per 1000 people	-462.938	1146.663	-.039	-.404	.687

a. Dependent Variable: Population in thousands

Equation:

Population= 59724.597 + (-462.938*Birth rate)

Normality of Residuals:

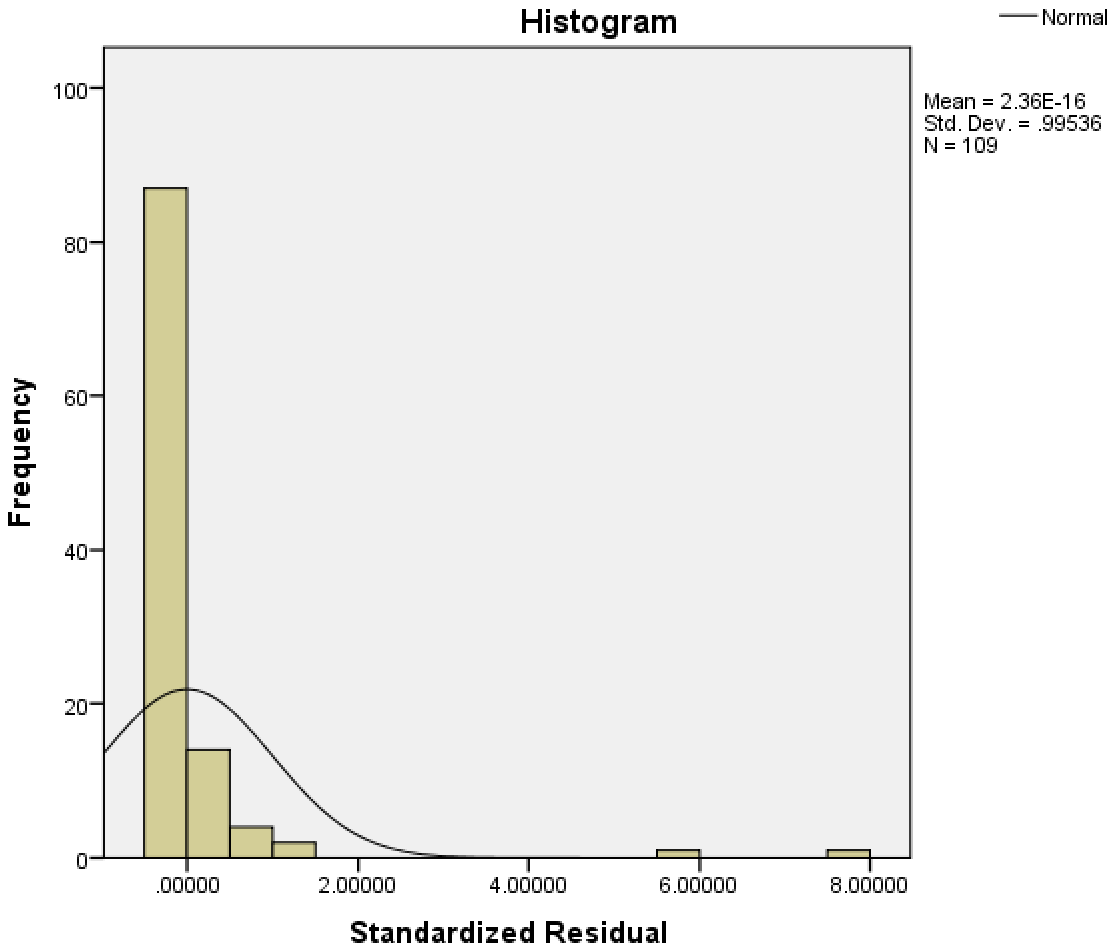
Tests of Normality

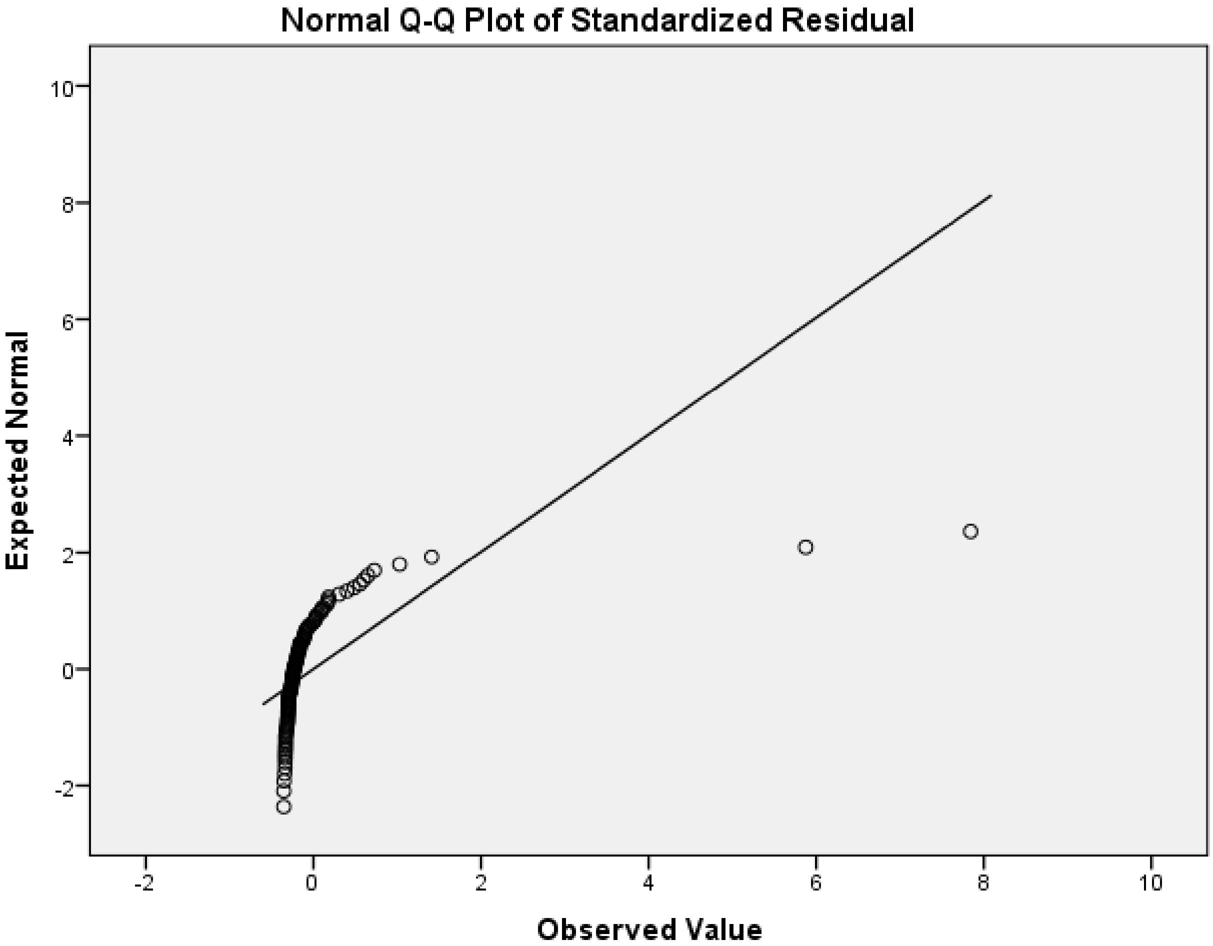
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.361	109	.000	.296	109	.000

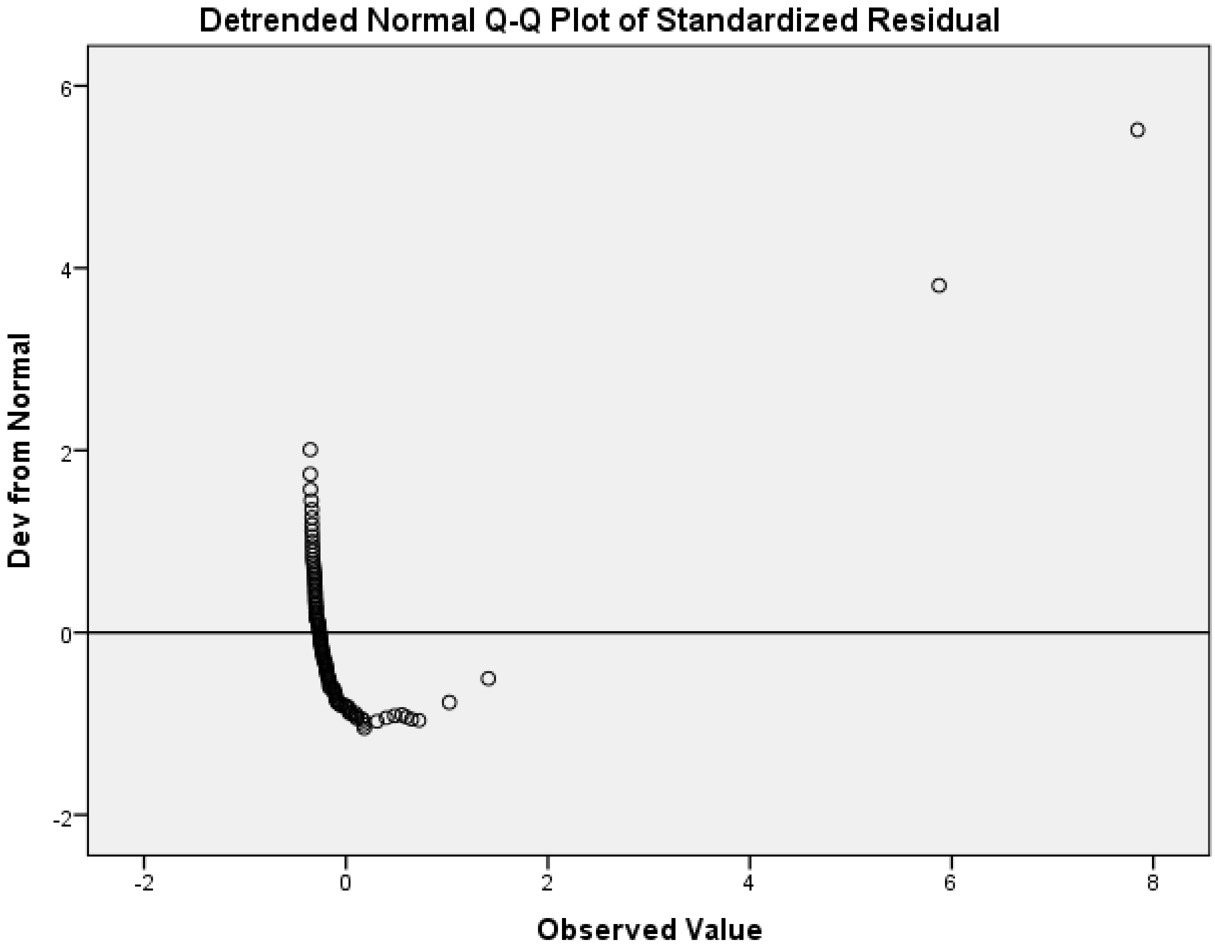
a. Lilliefors Significance Correction

The sig<0.05, so the residuals are not normally distributed.

Standardized Residual




Normal Q-Q Plot of Standardized Residual
Expected Normal
Observed Value
10
8
6
4
2
0
-2
-2
0
2
4
6
8
10



There are two values of standardized residuals which are greater than 3.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.079[a]	.006	-.003	43600.779

a. Predictors: (Constant), Birth rate per 1000 people

b. Dependent Variable: Population in thousands

R Square = 0.006, the value of R Square increased, and only 0.6% population variation is explained by the model.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1265083800.430	1	1265083800.430	.665	.416[b]
	Residual	199607935006.243	105	1901027952.440		
	Total	200873018806.673	106			

a. Dependent Variable: Population in thousands

b. Predictors: (Constant), Birth rate per 1000 people

The model is not significantly better than mean in predicting Population

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	36022.513	9769.592		3.687	.000
	Birth rate per 1000 people	-277.167	339.764	-.079	-.816	.416

a. Dependent Variable: Population in thousands

Equation:

Population= 36022.513 + (-277.167*Birth rate)

So, the model didn’t improve.

2.2. Do the multiple regression analysis using at least two predictors. Select the appropriate variables. Write down the regression equation. Assess the goodness-of-fit of the model. Are all the gradients and intercepts of the model statistically significant? Do the diagnostics of the model. Are the residuals normally distributed? Test the multicollinearity.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.661[a]	.437	.426	4907.830

a. Predictors: (Constant), Population in thousands, Birth rate per 1000 people

b. Dependent Variable: Gross domestic product / capita

R square = 0.437, 43.7% GDP variation is explained by the model.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1981532816.079	2	990766408.040	41.133	.000[b]
	Residual	2553200725.884	106	24086799.301		
	Total	4534733541.963	108			

a. Dependent Variable: Gross domestic product / capita

b. Predictors: (Constant), Population in thousands, Birth rate per 1000 people

As the sig<0.05 so the my model is significantly better than mean in predicting GDP

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	15008.835	1113.053		13.484	.000
	Birth rate per 1000 people	-343.796	38.235	-.656	-8.992	.000
	Population in thousands	-.005	.003	-.112	-1.540	.127

a. Dependent Variable: Gross domestic product / capita

Equation:

GPD= 15008.835+ (-343.796*Birth rate per 1000 people) + (-0.005*population in thousands)

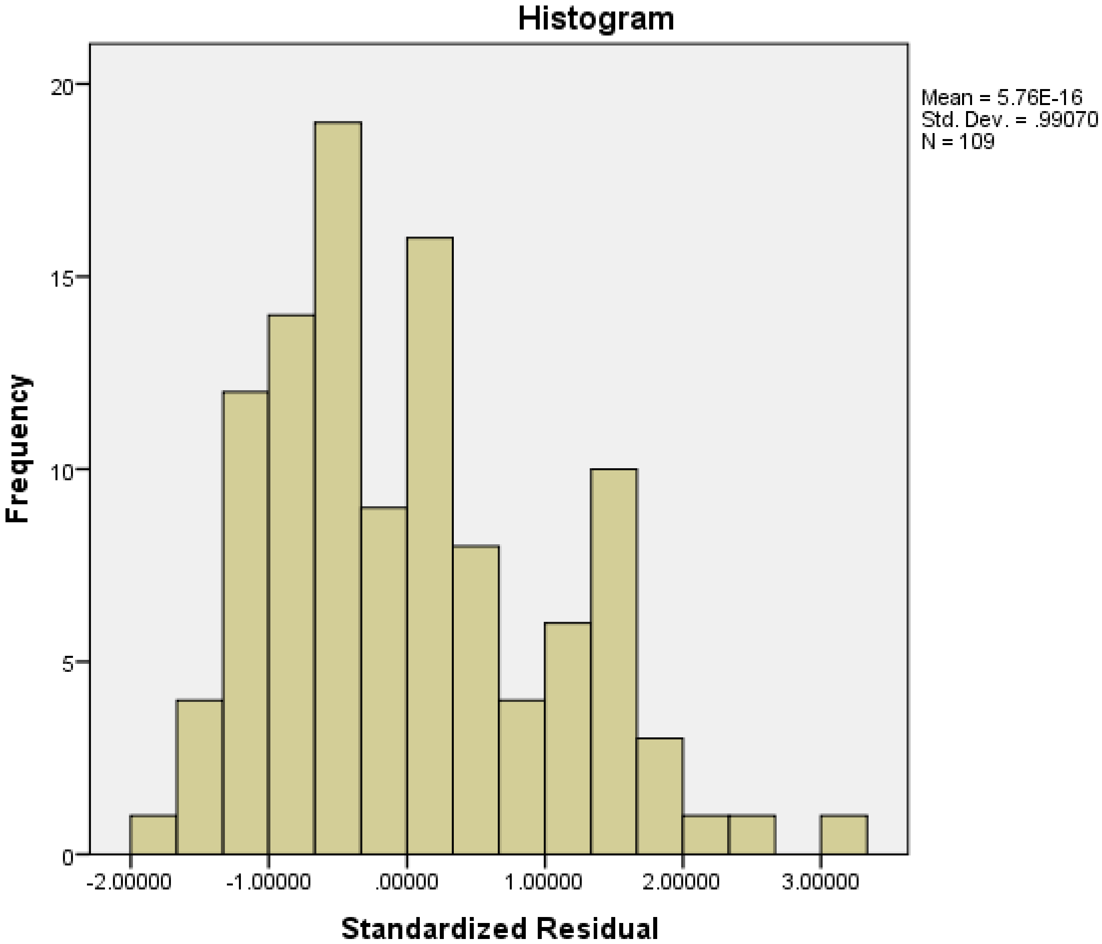
Tests of Normality

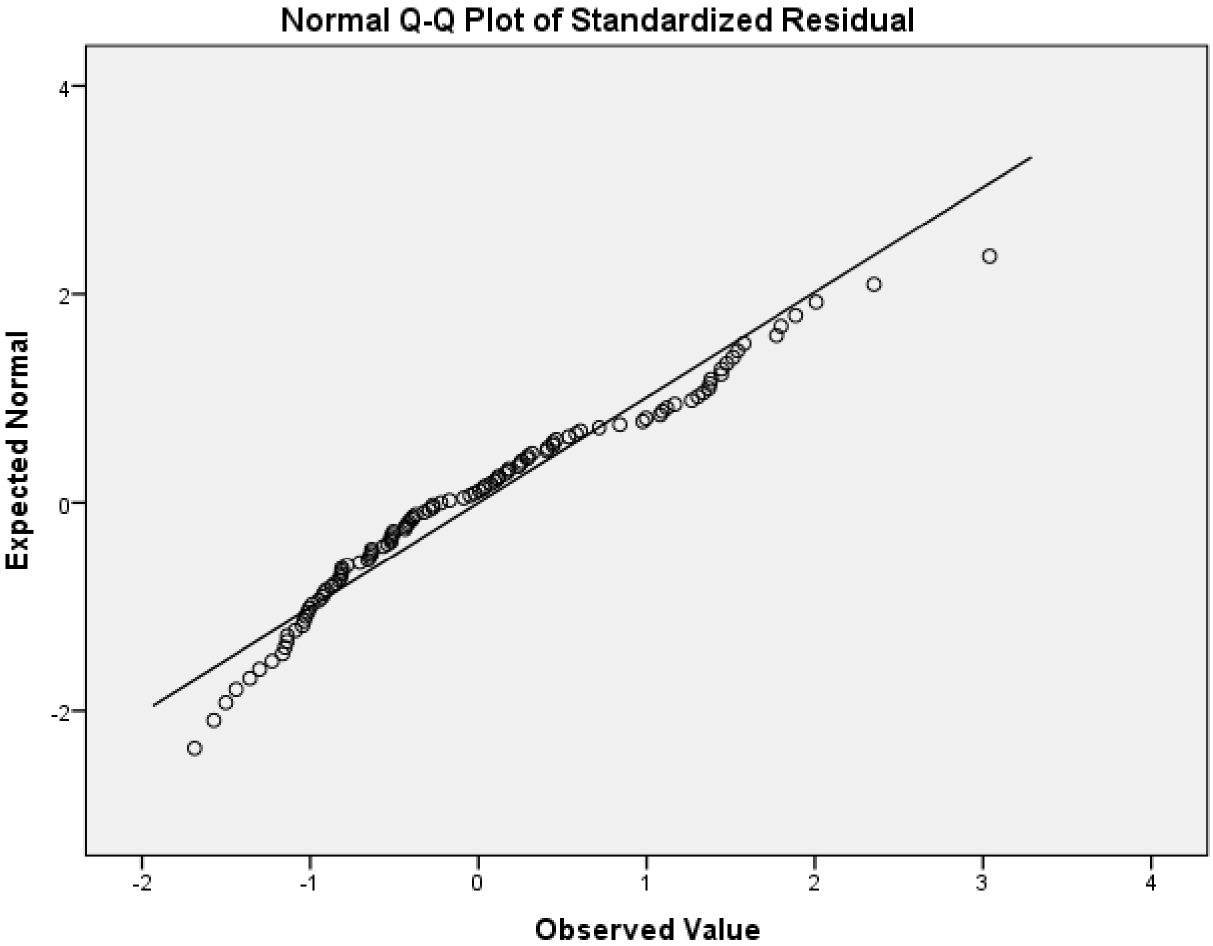
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.105	109	.005	.954	109	.001

a. Lilliefors Significance Correction

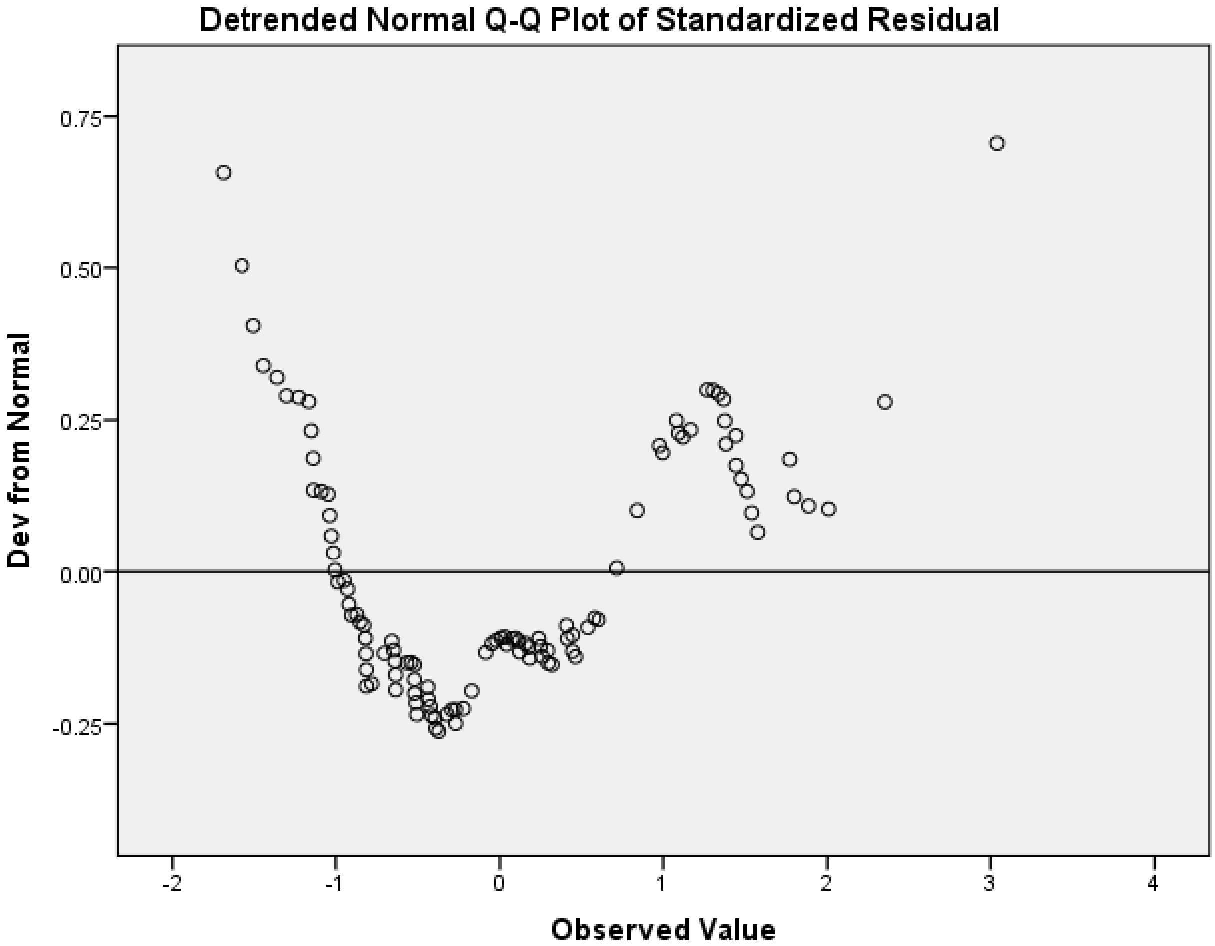
The sig<0.05, so the residuals are not normally distributed.

Standardized Residual





Normal Q-Q Plot of Standardized Residual
Expected Normal
Observed Value
4
2
0
-2
-2
-1
0
1
2
3
4



Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	Birth rate per 1000 people	.998	1.002
	Population in thousands	.998	1.002

a. Dependent Variable: Gross domestic product / capita

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	Birth rate per 1000 people	Population in thousands
1	1	2.060	1.000	.04	.04	.06
	2	.845	1.561	.01	.02	.93
	3	.095	4.649	.95	.94	.02

a. Dependent Variable: Gross domestic product / capita

As the value of tolerance is greater than 0.2, so there is no multicollinearity.