JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3219-3239
https://publishoa.com
ISSN: 1309-3452

The Problem of Linear Multiple in Regression: Concept and Treatment

Sama Saadi Ali Al Hashemi

Middle Technical University, Institute of Administration, al-Rusafa

Samaalhashimi7@gmail.com

Rehab Kadhim Hamza Al-Mafraji

Middle Technical Universit, Institute of Administration al-Rusafa,

Rehabhamza86@gmail.com

Received 2022 March 15; **Revised** 2022 April 20; **Accepted** 2022 May 10.

Abstract:

Multiple linear regressions is consider of one of the most common statistical techniques used by researchers in different fields. Often to multi-linear problem researchers face when building a multi-linear regression model The problem of multi-linear of the big problems facing a researcher at the Regression because they lead him to the wrong conclusions about moral of variables illustrations We found this problem in the data illustration variables for the production of small industrial plants in Iraq (Where we adopted a sample of the production of small industrial plants for the duration of(1997-2009) as come to the statistical group in 2011 [3] [7] We used gradual regression methods to overcome them and found that the two variables(production requirements) and (wages and concessions) can able to explain 91% from the differences in the quantity of production of these plants

Index

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3219-3239
https://publishoa.com
ISSN: 1309-3452

Introduction:

When there is a perfect correlation between two variables or between all the variables involved in the model so that the system becomes determinate matrix (x'x) equals zero. Where it is impossible to find the inverse matrix (x'x) and therefore it is not possible to use the ordinary least squares method OLS or the determinate matrix (x'x) should be close to zero in case of incomplete correlation With which the least squares are able to show the real properties of the model coefficients and have weak predictive capacity [1] To overcome the problem of multi- linear there are several solutions are as follows:

Method of backward, forward stepwise 1-

Method of principle component. (Pc) 2-

Method of Regression-Ridge.3-

Method of Partial least squares method (pls)4-

We have used in this search method number one (1).

Chapter (1)

(Primer strict)

Research aim (1-1):

Focus on the importance of the problem of multi-linear in the multi- linear regression and how to treat them.

The Research hypothesis (1-2):

You can formulate the following hypothesis:

H0: The explanatory variables data for the production of small factories do not

Suffer From multi-linear.

H1: The explanatory variables data for the production of small factories suffer

From Multi-linear, which requires treatment?

Research sample (1-3):

The research data were based on the production of small industrial factories for the period (1997-2009) as reported in the statistical group for 2011 [3] [7]

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3219-3239
https://publishoa.com
ISSN: 1309-3452

Production value of Y	Value of production inputs x_4	Wage and benefits x_3	number of staff x_2	Number of factories x_1
129558484	60478056	7392997	71353	31040
113723687	54067688	7150190	56121	25136
145357017	72347071	9624993	62331	29467
482235777	22646316	44251132	164579	77167
469607969	234176093	33657998	142724	69090
413729835	219855710	31367004	50207	17929
815977845	513071572	67704143	64338	17599
658655361	382254206	55809507	36379	10088
1103756794	617095687	76709079	46494	11620
812441151	467189737	96328627	53679	13406
815953528	389231285	65109035	27780	10289

Research problem (1-4):

When completely absent of a linear relationship between the explanatory Variables called these variables as orthogonal but most regression applications in Which explanatory variables are not orthogonal and strongly correlated so Difficult to estimate the effect of each explanatory variable independently on the Dependent variable, which requires a search for ways to overcome this problem.

Research Methodology (1-5):

There are many traditional approaches such as comparative, descriptive,

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3219-3239
https://publishoa.com
ISSN: 1309-3452

Contemporary, analytical, structural functional and statistical. However, we have

Taken the analytical and statistical approaches because they fit the research method.

Structural Research. (1-6):

The research is divided into four chapters:

1- Chapter (1) (primer structure)

2-Chapter (2) (Theoretical structure) concept of multi-linear and methods of

Detection and treatment

3-Chapter (3) (Applied structure) Contains the application side using the(spas)

Software package.

4- Chapter (4)(Conclusions and suggestion)

Chapter (2)

(Theoretical structure)

Importance of Research (2-1):

The importance of this research in the importance of overcoming the problem of multi-linear in the explanatory variables data for the production of small industrial factories in Iraq This problem hide the vision of significant explanatory variables Because it leads to the magnification of the contrast factor of the regression coefficients, it seems not significant This led the researcher to make the wrong conclusions about the factors affecting the production [2] [8].

The concept of multi-linear (2-2):

The multi-Linear is the existence of strong linear relationships between two or more of explanatory variables in the multiple linear regression models [4]

Methods for detection of multi-linear (2-3):

There are several methods for detecting the multi-linear the simplest is inflation

factor Variance to estimation of regression coefficients (VIF) If (VIF≥5) is a variable with a strong 0f multi-linear with other variables [5] [9].

VIF)=1/1-R^2 -:(variance inflection factor

Ways to treat multi-linear (2-4):

There are several ways to treat multi-linear as follows:

Method of backward, forward stepwise 1-

Method of principle component. (pc)2-

Method of Regression-Ridge.3-

4-Method of Partial least squares method (pls)

Chapter Three (3)

(Side of Applied)

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3219-3239
https://publishoa.com
ISSN: 1309-3452

* The use of spss software package for data analysis [6], where we choose a sample of the production of small industrial factors for the period of 2009-1997 as contained in the statistical group in 2011 [3] [7], the results were obtained by applying the SPSS program as The results are below.

"Results of analysis using spss and statistical analysis"

X_1: Number of factories

Number of staff :x_2

x_3 : Wages and benefits

X_4 : Value of production inputs

y : Production of value

Method/Enter all variables (3-1)

Model	R	R Square	Adjusted R Square	Std. Error Of the Estimate	Change Statistic				
					R Square Change	F Change	df_1	df_2	Sig. F Change
1	.977[a]	.955	.925	90062170.143	.955	32.051	4	6	.000

a. Predictors: (Constant): Production value requirements, number of staff, wages and benefits, the number of factories

b. Dependent Variable : value of Production

* Notes from the Table (1) Model Summary : More than 92% from the variance in the values of the dependent variables explained by input variable

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3219-3239
https://publishoa.com
ISSN: 1309-3452

Table No. (2) analysis of variance (ANOVA)[a]

1	Sum of Square	df	Mean Square	F	Sig.
Regression	1039884585837099900.00	4	259971146459274976.000	32.051	.000[b]
Residual	48667166945727608.000	6	8111194490954601.000		
Total	1088551752782827520.000	10			

a. Dependent Variable: Production value

b. Predictors: (Constant) :Wages and benefits, Number of factories

The value of input production , Number of staff

N* *notes from Table (2) analysis of variance: (F) is very large compared it tabular value This shows the high significant of the model.

a. Dependent Variable: production value.

* Note from Table No. (3) the regression coefficients:

1. The regression coefficients (wages and benefits) and (the value of production inputs) are only significant.

2. The large of inflation variance regression coefficients lead to problem of multi- linear and clear from the column values of inflation variation factor (VIF) this is

3. Can be written the linear regression equation for y on (X1, X2, X3, X4).

$$\acute{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + e$$

$$y = 2635449.527 + 12523.575\ X_1 - 5179.442\ X_2 + 5.411\ X_3 + 1.106\ X_4 + e$$

(88324520.862) (10620.081) (5384.324) (2.176) (0.342)

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3219-3239
https://publishoa.com
ISSN: 1309-3452

Table (3) regression coefficient

Model	Un standardized Coefficients		Standardized Coefficients	T	Sig.	95.0% Confidence Interval for B		Correlations			Co linearity Statistics	
	B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Toler Ance	VIF
(Constan)	2635449.527	88324520.862		.030	.977	-213486867.317-	218757766.37					
Number of factories	12523.575	10620.081	.886	1.179	.283	-13462.826-	38509.976	-.366-	.434	.102	.013	75.764
number of staff	-5179.442-	5384.324	-.678-	-.962-	.373	-18354.407-	7995.523	-.243-	-.366-	-.083-	.015	66.621
Wages and benefits	5.411	2.176	.494	2.486	.047	.086	10.736	.935	.712	.215	.188	5.307
value of input production	1.106	.342	.703	3.232	.018	.269	1.944	.928	.797	.279	.158	6.346

B_0= Represent the regression constant.	**Y= production value (dependent value)**
B_1= Represents a regression coefficient of variable X_1	**X_1=The number of factories (independent**

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3219-3239
https://publishoa.com
ISSN: 1309-3452

	value)
B_2= Represents a regression coefficient of variable X_2	**X_2= number of staff (independent value)**
B_3= Represents a regression coefficient of variable X_3	**X_3= wages and benefits (independent value)**
B_4= Represents a regression coefficient of variable X_4	**X_4= the value of production inputs (independent value)**

	Minimum	Maximum	Mean	Std. Deviation
Predicted Value	125250872.00	1005037952.00	541908858.91	322472415.229
Residual	-118090816.000-	101124208.000	.000	69761857.018
Std. Predicted Value	-1.292-	1.436	.000	1.000
Std. Residual	-1.311-	1.123	.000	.775

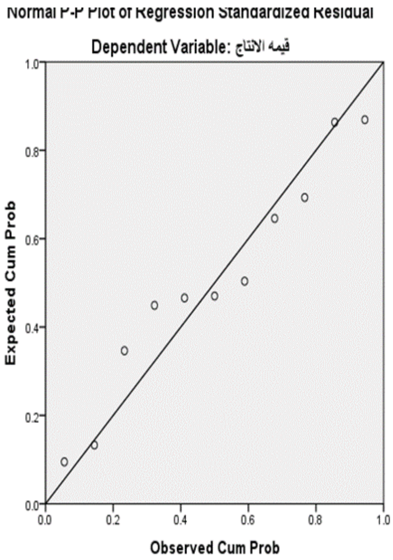
Depen a. Dependent Variable (the value of production)

Table (4) Residuals statistics

*Notes from Table (4) Residuals statistics: There are no outliers' values that are between (-1.311 and 1.123)

"A graph represents the normal distribution of the standard

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3219-3239
https://publishoa.com
ISSN: 1309-3452



Notes from the graph of the residuals standard distributed a normal distribution

Method/backward (3-2)

Table No. (5) (Model Summary)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df_1	df_2	Sig. F Change
1	.977[a]	.955	.925	90062170.143	.955	32.051	4	6	.000
2	.974[b]	.948	.926	89580625.159	-.007-	.925	1	6	.373
3	.965[c]	.932	.915	96351167.259	-.017-	2.255	1	7	.177

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3219-3239
https://publishoa.com
ISSN: 1309-3452

a. Predictors: (Constant) : Value of production input, wages and benefits input,

number of staff, wages and benefits, number of establishments

b. Predictors: (Constant) :Value of production input, wages and benefits, number of

establishments

c. Predictors: (Constant) : Value of input production, wages and benefits

d. Dependent Variable: Value of production.

* It is noted from Table No.(5) the Model Summary : that more than 91% of the variance in the values of the Dependent variables explained by the input variables (wages and benefits, the value of input production)

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3219-3239
https://publishoa.com
ISSN: 1309-3452

Table(6) analysis of variance(ANOVAa)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1039884585837099900.000	4	259971146459274976.000	32.051	.000[b]
	Residual	48667166945727608.000	6	8111194490954601.000		
	Total	1088551752782827520.000	10			
2	Regression	1032378933955921790.000	3	344126311318640580.000	42.883	.000[c]
	Residual	56172818826905752.000	7	8024688403843679.000		
	Total	1088551752782827520.000	10			
3	Regression	1014283373326110340.000	2	507141686663055170.000	54.628	.000[d]
	Residual	74268379456717152.000	8	9283547432089644.000		
	Total	1088551752782827520.000	10			

a. Dependent Variable : (the value of production)

b. Predictors: (Constant) Value of production input, wages and benefits input, number of staff, number of establishments

c. Predictors: (Constant) : Value of production input, wages and benefits, number of establishments

d. Predictors: (Constant) : Value of input production, wages and benefits .

Table No. (7) regression coefficient

Model	Unstandardized Coefficients		Standardized Coefficients	T	Sig.	95.0% Confidence Interval for B		Correlations			Co linearity Statistics	
	B	Sta. Error	Beta			Lower Bound	Upper Bound	Zero order	Partial	Part	Tolerance	VIF
(Constant)	2635449.527	88324520.862		.030	.977	-213486867.317-	218757766.370					
number of factories	12523.575	10620.081	.886	1.17	.283	-13462.826-	38509.976	-.366-	.434	.102	.013	75.764
number of staff	-5179.442-	5384.324	-.678-	-.962-	.373	-18354.407-	7995.523	-.243-	-.366-	-.083-	.015	66.621
wages and benefits	5.411	2.176	.494	2.48	.047	.086	10.736	.935	.712	.215	.188	5.307
Value of production input	1.106	.342	.703	3.23	.018	.269	1.944	.928	.797	.279	.158	6.346
(Constant)	-23344348.850-	83644463.065		-.279-	.788	-221132074.725-	174443377.026					
number of factories	2427.988	1616.868	.172	1.50	.177	-1395.298-	6251.274	-.366-	.494	.129	.563	1.775
wages and benefits	4.637	2.011	.424	2.30	.055	-.119-	9.392	.935	.657	.198	.218	4.581
Value of production input	1.043	.334	.663	3.12	.017	.253	1.833	.928	.763	.268	.164	6.112
(Constant)	77125047.257	53994239.023		1.42	.191	-47385891.207-	201635985.721					
wages and benefits	5.774	2.004	.528	2.88	.020	1.154	10.394	.935	.714	.266	.254	3.931
Value of production input	.743	.288	.472	2.58	.033	.079	1.408	.928	.674	.238	.254	3.931

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3219-3239
https://publishoa.com
ISSN: 1309-3452

a. Dependent Variable: production value ***Note from Table No. (7) Regression coefficients:**

1- The Regression coefficients (wages and benefits) and (value of production inputs) are significant.

2-There is no large inflation in the variations of the regression coefficients, which is indicates to a solution to the multi - linear problem. This is clear from the values of the column of the variance inflation factor (VIF)

$\acute{Y} = b_0 + b_3X_3 + b_4X_4 + e$

$\acute{Y} = 77125047.257 + 5.774\, x_3 + 0.743\, x_4 \quad + \quad e$

$(53994239.023)\ (2.004)\ (0.288)$

B0 = represents the regression constant.	**Y = represents the value of production (dependent variable)**
B3= represents the regression coefficient of the variable X3	**X3 = represents wages and benefits.**
B4= the regression coefficient of the variable X4.	**X4 = represents the value of inputs production.**

*** Notes from Table No. (8) Remaining Statistics: there are no outliers that are**

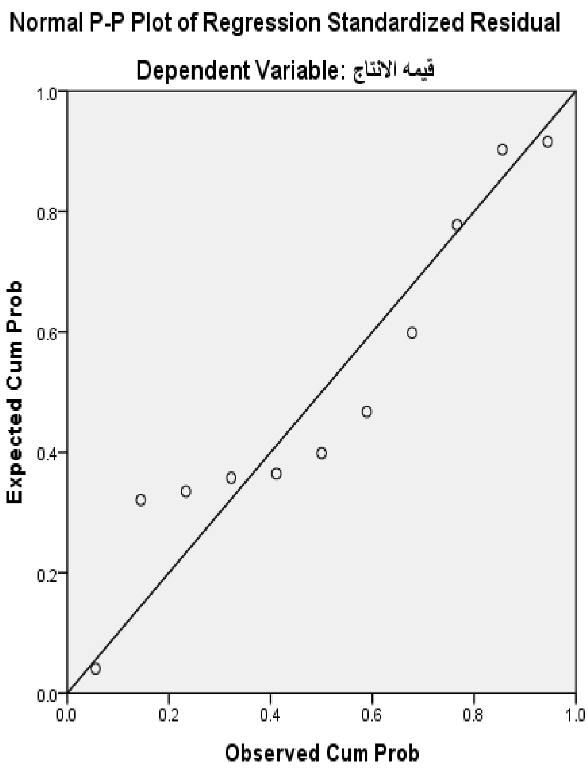
Between (-1.745 and +1.378).

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3219-3239
https://publishoa.com
ISSN: 1309-3452

Residuals Statistic (8) table

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	**158596688.00**	**980574720.0**	**541908858.9**	**318478158.33**	**11**
Residual	**-168133568.00-**	**132767816.00**	**.000**	**86179103.88**	**11**
Std. Predicted Value	**-1.204-**	**1.377**	**.000**	**1.000**	**11**
Std. Residual	**-1.745-**	**1.378**	**.000**	**.894**	**11**

a. Dependent Variable: production value



JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3219-3239
https://publishoa.com
ISSN: 1309-3452

Notes from the graph of the residuals standard distributed a normal distribution

Method/ forward (3-3):

Model Summary table (9)

Model	R	R Square	adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df_1	df	Sig. F Change
1	.935[a]	.875	.861	122942300.551	.875	63.019	1	9	.000
2	.965[b]	.932	.915	96351167.259	.057	6.653	1	8	.033

a. Predictors: (Constant) : wages and benefits

b. Predictors: (Constant) : wages and benefits , value of production inputs

c. Dependent Variable: production value.

(ANOVA) table No. (10) analysis of variance

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	952518469399782530.000	1	952518469399782530.000	63.019	.000[b]
	Residual	136033283383044976.000	9	15114809264782776.000		
	Total	1088551752782827520.000	10			
2	Regression	1014283373326110210.000	2	507141686663055100.000	54.628	.000[c]
	Residual	74268379456717280.000	8	9283547432089660.000		
	Total	1088551752782827520.000	10			

a. Dependent Variable: production value

b. Predictors: (Constant): wages and benefits

c. Predictors: (Constant) : wages and benefits , value of production inputs

*** It is noted from Table No. (10) analysis of variance (F) is very large to its tabular value , indicate the high significance of the model**

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3219-3239
https://publishoa.com
ISSN: 1309-3452

Table No. (11) regression coefficient

Model	Unstandardized Coefficients		Standardized Coefficients	T	Sig.	95.0% Confidence Interval for B		Correlations			Co linearity Statistics	
	B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	partial	Part	Tolerance	VIF
(Constant)	81168893.198	68866598.283		1.18	.269	-74618175.385-	236955961.782					
wages and benefits	10.237	1.289	.935	7.94	.000	7.319	13.154	.935	.935	.935	1.000	1.000
(Constant)	77125047.257	53994239.023		1.43	.191	-47385891.207-	201635985.721					
wages and benefits	5.774	2.004	.528	2.88	.020	1.154	10.394	.935	.714	.266	.254	3.931
Value of production input	.743	.288	.472	2.58	.033	.079	1.408	.928	.674	.238	.254	3.931

a. Dependent Variable : production value

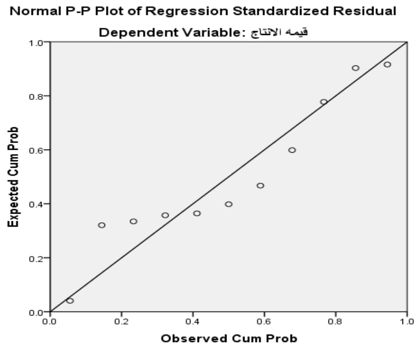
*** Note from Table No. (11) regression coefficients :**

1. The regression coefficients (wages and benefits) and (the value of production input)are significant

2-There is no large inflation in the variations of the regression coefficients, which is indicates to a solution to the multi - linear problem. This is clear from the values of the column of the variance inflation factor (VIF)

$$\acute{Y} = b_0 + b_3X_3 + b_4X_4 + e$$

Graph representing the normal distribution of the standard residuals.



Notes from the graph of the residuals standard distributed a normal distribution

Method/ stepwise(3-4)

Modal	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change sticks				
					R Square Change	F Changed	df_1	df_2	Sig . F Change
1	.935[a]	.875	.861	122942300.55	.875	63.019	1	9	.000
2	.965[b]	.932	.915	96351167.259	.057	6.653	1	8	.033

a. Predictors: (Constant) :(wages and benefits)

b. Predictors: (Constant): (wages and benefits , value of input production)

c. Dependent Variable: production value.

*** It is noted from Table No. (13) the Model Summary : that more than 91% of the variance in the values of the Dependent variables explained by the input variables**

(wages and benefits , value of input production)

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3219-3239
https://publishoa.com
ISSN: 1309-3452

Model		Sum of Squares	dF	Mean Square	F	Sig.
1	Regression	952518469399782530.000	1	952518469399782530.000	63.019	.000[b]
	Residual	136033283383044976.000	9	15114809264782776.000		
	Total	1088551752782827520.000	10			
2	Regression	1014283373326110210.000	2	507141686663055100.000	54.628	.000[c]
	Residual	74268379456717280.000	8	9283547432089660.000		
	Total	1088551752782827520.000	10			

a. Dependent Variable: : production value

b. Predictors: (Constant) : wages and benefits

c. Predictors: (Constant) : wages and benefits .

*** It is noted from Table No. (14) analysis of variance (F) is very large to its tabular value, indicate the high significance of the model**

Coefficients "Table (15) regression coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	T	Sig.	95.0% Confidence Interval for B		Correlations			Co linearity Statistics	
	B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
(Constant)	81168893.198	68866598.283		1.179	.269	-74618175.385-	236955961.782					
wages and benefits	10.237	1.289	.935	7.938	.000	7.319	13.154	.935	.935	.935	1.000	1.000
(Constant)	77125047.257	53994239.023		1.428	.191	-47385891.207-	201635985.721					
wages and benefits	5.774	2.004	.528	2.882	.020	1.154	10.394	.935	.714	.266	.254	3.931
value of input production	.743	.288	.472	2.579	.033	.079	1.408	.928	.674	.238	.254	3.931

1. The regression coefficients (wages and benefits) and (the value of production inputs) are significant

2-There is no large inflation in the variations of the regression coefficients, which is indicates to a solution to the multi -linear problem. This is clear from the values of the column of the variance inflation factor (VIF)

b_0 = Represents the regression constant.	**Y = Represents the value of the output (the dependent variable)**
b_3 =represents the regression coefficient of the variable	**X_3 = Represent wages and benefits.**
B_4=represents the regression coefficient of the variable b_4	**X =Represents the value of input production**

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3219-3239
https://publishoa.com
ISSN: 1309-3452

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	**158596688.00**	**980574720.00**	**541908858.91**	**318478158.329**	**11**
Residual	**-168133568.000-**	**132767816.000**	**.000**	**86179103.881**	**11**
Std. Predicted Value	**-1.204-**	**1.377**	**.000**	**1.000**	**11**
Std. Residual	**-1.745-**	**1.378**	**.000**	**.894**	**11**

$$\acute{Y} = b_0 + b_3X_3 + b_4X_4 + e$$

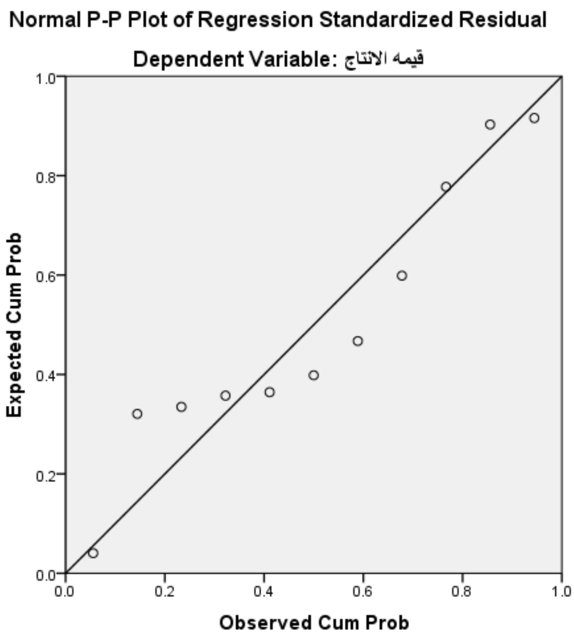
Ý = (77125047.257) + (5.774) + (0.743) **(53994239.023) (2.004) (0.288)**

Table No. (16) Residuals Statistics

a .Dependent Variable: production value.

*** Notes from Table No. (4) Remaining Statistics: there are no outliers that are between (-1.745 and +1.378)**

.



Graph representing the normal distribution of the standard residuals

Chapter Four (4)

(Conclusions and suggestion)

*** Conclusions (4-1):**

1- The gradual regression methods lead to removing the problem of multi- linear by dropping the explanatory variables that cause it.

2. That all the modalities of the gradual regression led to the same selection of variables that explain more than 91% of the differences in production values and small industrial facilities.

3. The variables that explain most of the differences in the production of small industrial facilities are (value of input production) and (wages and benefits) .

Suggestion (4-2):

We recommend to adoption the methods of gradual regression in the removal the problem of multi-linear in the data of industrial facilities in Iraq.

Reference : (4-3)

**** Arabic reference ****

1. Ismail Muhammad Abdul Rahman (Linear Regression Analysis), Saudi Arabia / Second Edition (2000).
2. Al-Baldawi. Dr. Abdul Hammed Abdul Majeed (Statistics Methods) * Wael Publishing House / First Edition - (2009).
3. Habib, Mustafa, teacher of Applied Statistics, Rusafa Institute of Management / Department of Statistics (2016).
4. Samor. Khalid Qasim (Statistics) Dar Al-Fikr / First Edition - (2007).
5. 5- Saleh. Abu Sidra Fathi (Statistics and Econometrics) National Books House / First Edition - (1999).
6. 6-Atiyah. Abdul Qadir Muhammad (Econometrics between theory and practice) * Mecca / First Edition - (2004).
7. 7 – quality mahfoz (advanced statistical analysis using SPSS) Waael Publishing House / First Edition - (2008).
   Foreign reference
8. whit house news "American heart moth".
9. national statistics press release 2006.
10. national vitals statistics reports volume 58 number