AL KUT JOURNAL OF ECONOMIC AND ADMINISTRATIVE SCIENCES
Publisher: College of Economics and Management - Wasit University

A new algorithm for detecting and treating outliers in multiple regression models

Professor Dr. Fayyadh Abdulla Ali

Osol Al Elm university college

Abstract

In multiple regression analysis, we aim to construct a statistical model that describes the relationship between a dependent variable and several independent variables. However, the data may contain observations that differ significantly from the rest of the values. These observations are known as "outliers."

Outliers are data points that fall outside the general pattern of relationships between variables. They may result from measurement or input errors, or they may reflect exceptional cases with real significance. The presence of these outliers can distort the results of the analysis.

Hence, the aim of the research was to build an algorithm to detect the strays present in the data and then delete them from the data to reach the most accurate results. The algorithm was applied to data free of outliers and data containing 10%, 20% and 30% of outliers cases. The algorithm proved its efficiency in all cases.

Keyword: multiple linear regression, outliers

Introduction

In real data, it found that some observations behave in different from the majority. Such data are called "anomalies" in machine learning, and in statistics called "outliers" [7]. Outliers may be coming by errors recorded under unusual circumstances, or may be belong to another distribution. It is very significant to be capable to detect "outliers" cases, If such outliers are not handled properly, these outliers can lead to a mistake model, leading to incorrect conclusions and predictions [2] .

Regression analysis is one of the most usually used statistical techniques. among many regression techniques, the ordinary least squares (OLS) method has been mostly adopted because of ease of computation, the estimation by O.L.S. is best method when assumptions are satisfied, but when data does not meet some of these assumptions, then the estimates and results can be deceptive. Especially, outliers will violate essential assumption of normality distributed of residuals in the least squares regression.

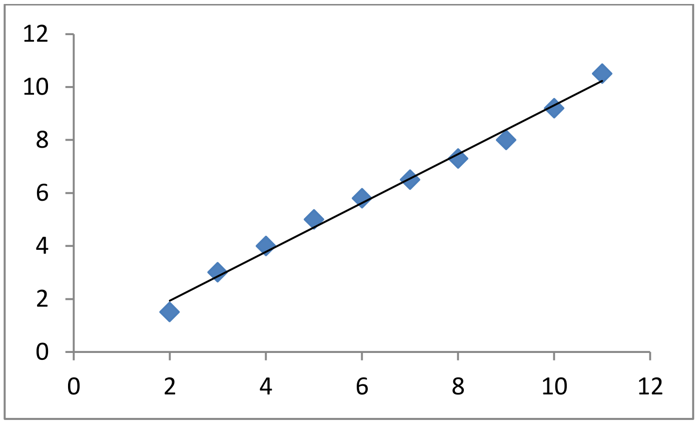
As the definition (Barnett and Lewis 1994), outliers are observations that seem maladjusted with the majority of the data. Often time, such effective points remain unobserved to the researcher, because they do not always detect in plot of the least squares residual.

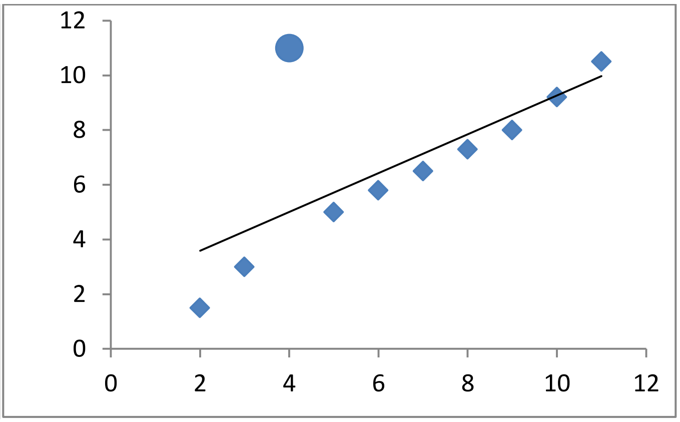
The danger of exist outliers, both in dependent or independent variables, to the least squares regression is that they can have a strong reverse action on the estimation and they may stay unobserved. Therefore, many ways already exist to detect outliers in linear regression model that classified into two groups, namely analytical methods and graphical [4,9]

Effect and rating of outliers.

Let us begin with some simple linear regressions illustrated in Fig. 1. (Fig. 1a) display a simple linear regression [5]. The points are located in a straight line that passes almost during the origin, the slope is 1. The deviation of the points from the line is distributed almost naturally and the deviations are small. Figs. (1b-d) display the itself points, but one of the points changed so that it turns into an outlier. In Figs. 1b-c) the outlier significantly changed the regression line. in both cases, the slope be smaller and the line no longer contacts the origin. Fig. 1d) display an outlier which does not variation the regression, but which still lies away from the else points. The impact here is different: the outlier expands the region where the regression gives valid results. For the first two outliers, forecasting become unreliable due to the existence of the outliers. In the latter, case a forecasting may appear reliable in a definite region in spite of it is not reliable there.

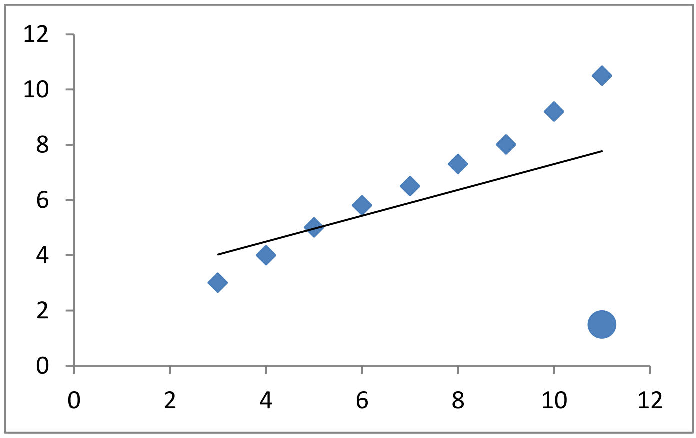
The examples also explain a classification of normal data points and outliers. The reason to the large impact of outliers in linear regression is the minimum least square: the algorithm is to minimize sum squared distances of the points from regression line (classical standard deviation). using of the standard deviation is the direct conclusion of the assumption that the errors are distributed normally. Although many researchers suppose that this assumption is true in most cases, but it is often violated and must be tested in each analysis. In Figs. 1b-c) outliers have a big distance from the original line explain in Fig. 1a). so, the weighting of the outlier is big and the influence on the regression line is large.

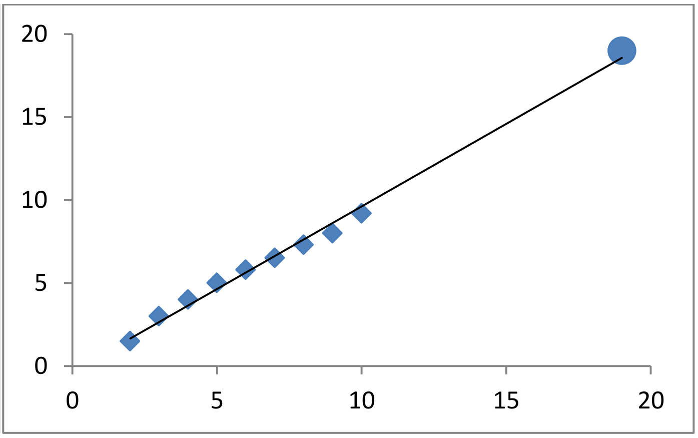




Figure(1.a) regression without outliers

Figure(1.b) One point shifted in y direction





Figure(1.c) One point shifted in x direction

Figure(1.d) One point shifted in both directions

Outliers in Multiple Linear Regression

Rousseeuw and Leroy (1987) introduced in detail a description of detection outlier methods linking to regression methods, highlighting two main approaches. First, the outlier diagnostics way which engages with residuals produced by a standard OLS regression or a standard OLS regression where some data points have been omitted. Residuals calculated by leave-one-out approach (Weisberg 1985) or the application of Cook distance and DFFITS can be efficient to detect effective points, i.e. the observations that are significant affecting on regression plane. The second approach by applying robust regression methods that are less sensitive to the existence of outliers comparing with the OLS [1]. Several methods have been suggested and their efficiency can be assessed use the conception of the 'breakdown point' i.e. the part of distorted points that can have an arbitrarily large influence on estimation of regression parameter [3].

The general linear regression model in expression of the observations can be written in matrices form by $= X\beta + \varepsilon$, where y is 1 a vector$(n \times 1)$ of observed response values, X is the matrix of the predictor variables$(n \times p)$, β is the unknown parameter vector$(p \times 1)$, and ε is a vector of random error $(n \times 1)$. The object of regression analysis is to estimate of unknown parameters. The OLS is used to find the best estimation of β 's that minimizes sum square distances of all of the points from the real observation to the regression surface.

It oftentimes happen in practice that supposed normal distribution model - a linear regression model for normal errors- holds in approximate in that it represent the most of observations, but some observations follow up a different style or no style at all [6].

In the situation when the randomness in the pattern is assigned to observational errors that was the first instance of using of the least-squares method, the fact is that while the action of many sets of data showed rather normal, this held only approximately, with the main different being that a few portion of observations were just atypical by virtue of being far from the majority of the data.

Behavior of this kind is popular across the entire spectrum of statistical modeling applications and data analysis. this atypical data or even one outlier can have a large distorting effect on statistical method (classical) that is optimal under assumption of the linearity or normality.

Outliers detection algorithm in multiple linear regression model

1. Draw subsamples representing 70% of the original sample (integers).

2. Conduct an analysis of variance (ANOVA) and extract an F value for each sample.

3. Plot the F values in descending order.

4. If the F values continue to decline smoothly (despite the presence of a cutoff), this indicates that the data are free of outliers. We then conduct an analysis of variance (ANOVA) for the original data.

5. If there is a cutoff and the F values do not decline smoothly, this indicates the presence of outliers. Their number can be determined based on the number of F values before and after the cutoff.

6. Delete the cases (the corresponding Y values and X values) and conduct an ANOVA for the remaining data.

Application

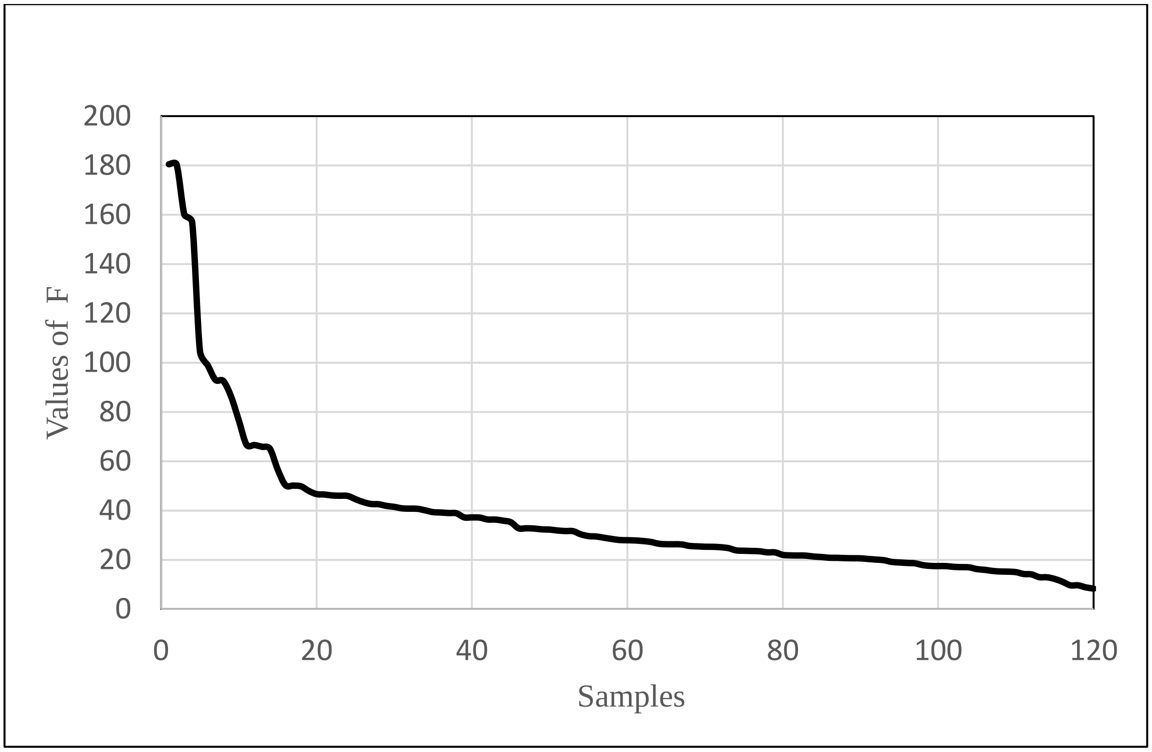
The data of an Albanian economic establishment with foreign capital in the construction and trade sector was approved during 10years [8], which represents the income, the number of employees, and the price of the product, as in the following table

Table (1) data of economic organization of Albania with foreign capital

year	**income**	**Number of employees**	**Price of product**
2006	71	272	15.7
2007	78	269	18.7
2008	81	214	18.6
2009	75	234	18.6
2010	89	232	21.4
2011	82	215	19.9
2012	94	235	22.5
2013	92	233	21.7
2014	79	255	17.5
2015	56	201	15.3

Case (0) No outliers in data

We follow the steps mentioned above (**Outlier detection algorithm in multiple linear regression model)**, where the number of samples drawn is $C_7^{10} = 120$ the graph is as follows:



Figure(2) No outliers

It is noted from the Figure(2) above that the F values continued to decline in a smooth manner, indicating that the data were free of outliers, So we will perform ANOVA on the original data without deleting any case.

Table (2) Analysis of variance for origin data

S.O.V.	**S.S.**	**D.F.**	**M.S.**	**F**	**Sig**
Regression	1025.290	2	512.645	41.338	.000
Residual	86.810	7	12.401		
Total	1112.100				

The regression model is $\boldsymbol{\hat{\imath} = -29.528 + 0.11\,E + 4.375\,P\ and\ R^2 = 0.92}$

Where **I** is income

E: employees number

P:price of product

Case(1) 10% outliers in dependent variable (income)

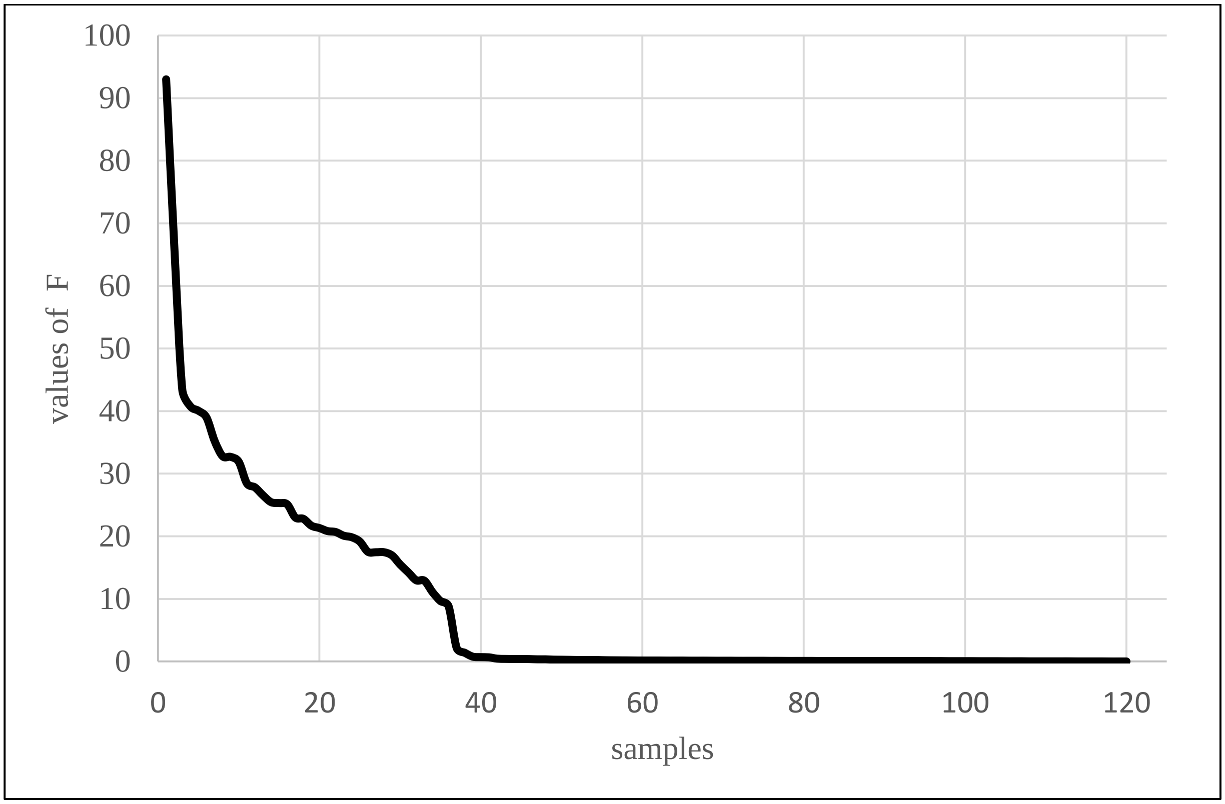
Now we will make the eighth case (2013) an anomaly (changing the income value from 92 to 29) and conduct the analysis of variance as follows:

Table (3) Analysis of variance with 10% outliers(one case)

S.O.V.	S.S.	D.F.	M.S.	F	Sig
Regression	119.558	2	59.779	.139	.873
Residual	3014.842	7	430.692		
Total	3134.400	9			

$\boldsymbol{R^2 = .038}$

Now, following the same steps in the algorithm above but with a 10% anomaly (here is one case in the values of income)



Figure(3) 10% outliers

It is noted from the figure(3) above that the values of F after about 36 values of F stabilized and no smooth decline occurred, which indicates that there is an anomalous case (10%), so we delete the anomalous case and conduct an analysis of variance for the remaining data (9 cases).

Table (4) Analysis of variance after removing only one outlier

S.O.V.	**S.S.**	**D.F.**	**M.S.**	**F**	**Sig**
Regression	**857.979**	**2**	**428.989**	**29.922**	**.001**
Residual	**86.021**	**6**	**14.337**		
Total	**844**	**8**			

The regression model is $\boldsymbol{\hat{\imath} = -28.651 + 0.111\, E + 4.324\ P\ and\ R^2 = 0.91}$

Case(2) 20% outliers in dependent variable (income)

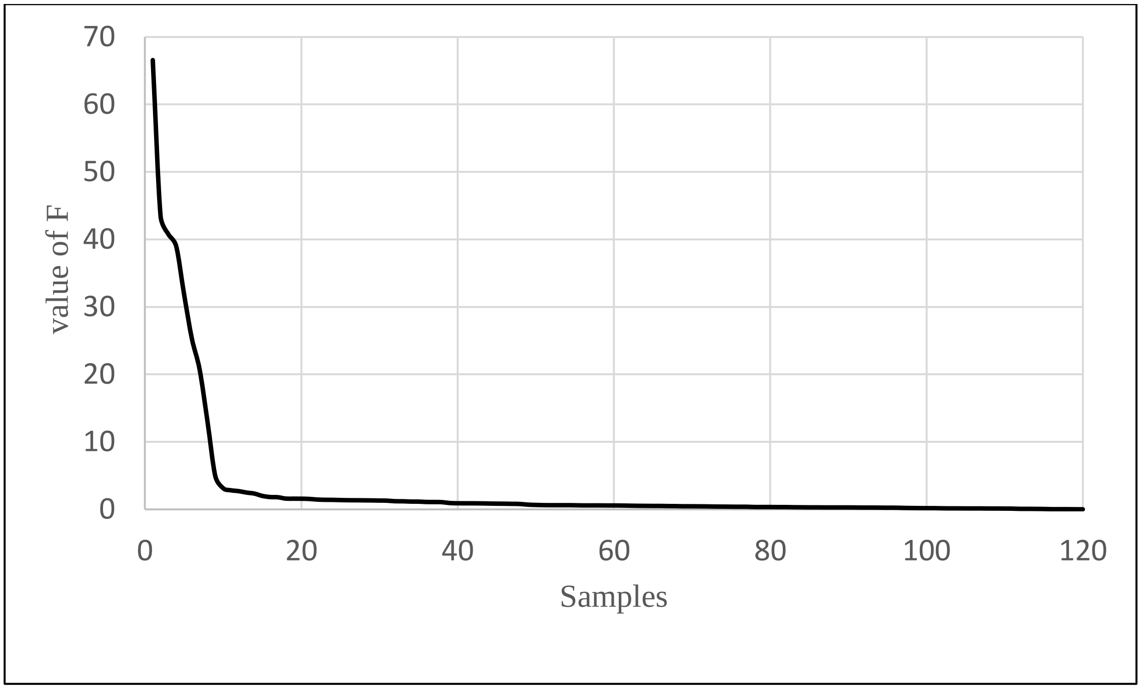
Here we will make the first case (2006)an anomalous case (changing the income value from 71 to 7) in addition to the previous cases and conduct the analysis of variance as follows:

Table (5) Analysis of variance with20 % outliers(two cases)

S.O.V.	**S.S.**	**D.F.**	**M.S.**	**F**	**Sig**
Regression	907.749	2	453.874	.656	.548
Residual	4841.151	7	691.593		
Total	5748.900	9			

$\boldsymbol{R^2 = .158}$

Now, following the same steps in the algorithm above but with a 20% anomaly (here is two cases in the values of income)



Figure(4) 20% outliers

It is noted from Figure (4) above that the F values after about 8 F values stabilized and no smooth regression occurred, which indicates the presence of two anomalous cases (20%). Therefore, we delete the two anomalous cases and conduct an analysis of variance for the remaining data (8 cases).

Table (6) Analysis of variance table after removing two outliers

S.O.V.	**S.S.**	**D.F.**	**M.S.**	**F**	**Sig**
Regression	888.852	2	444.425	47.131	.001
Residual	47.148	5	9.430		
Total	936.000	7			

The regression model is $\boldsymbol{\hat{\imath} = -38.022 + 0.143\, E + 4.373\, P\ and\ R^2 = 0.95}$

Case(3) 30% outliers in dependent variable (income)

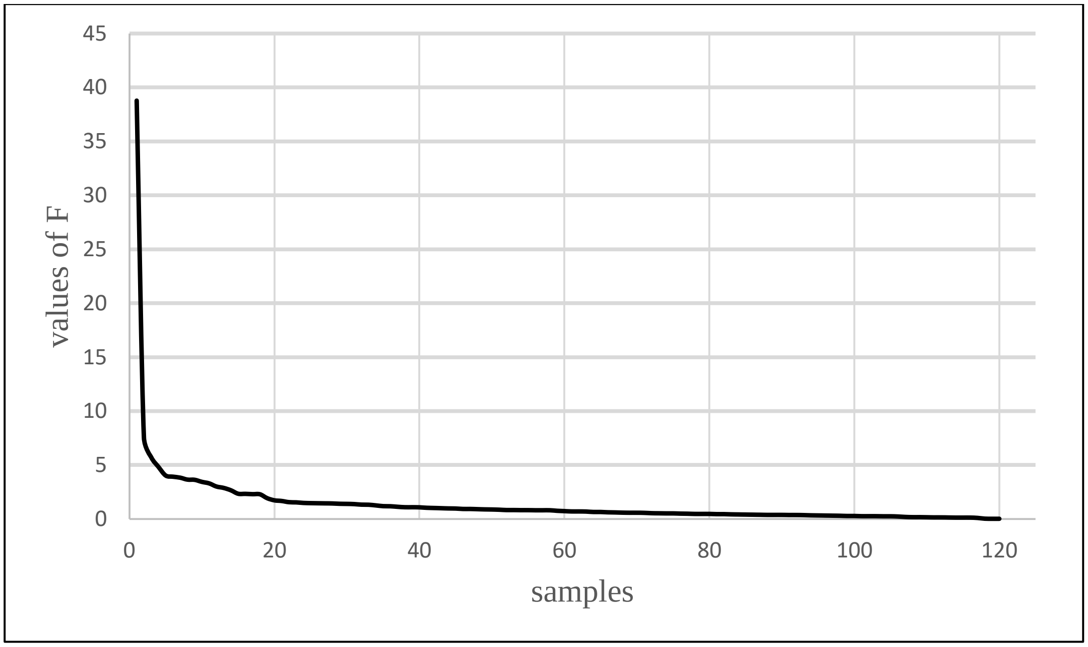
Here we will make the third case (2008) an anomalous case (changing the income value from 81 to 18) in addition to the previous case and conduct the analysis of variance as follows:

Table (7) Analysis of variance with30 % outliers (three cases)

S.O.V.	**S.S.**	**D.F.**	**M.S.**	**F**	**Sig**
Regression	1594.310	2	797.155	.760	.503
Residual	7341.790	7	1048.827		
Total	8936.100	9			

$\boldsymbol{R^2 = .178}$

Now, following the same steps in the algorithm above but with a 30% anomaly (here is three cases in the values of income)



Figure(5) 30% outliers

It is noted from Figure (5) above that the F values after only one F value stabilized and no smooth regression occurred, which indicates the presence of three anomalous cases (30%). Therefore, we delete the three anomalous cases and conduct an analysis of variance for the remaining data (7 cases)

Table (8) Analysis of variance table after removing three outliers

S.O.V.	**S.S.**	**D.F.**	**M.S.**	**F**	**Sig**
Regression	837.046	2	418.523	38.974	.002

Residual	42.954	4	10.738		
Total	880.000	6			

The regression model is $\boldsymbol{\hat{\imath} = -36.775 + 0.123\, E + 4.547\, P\ and\ R^2 = 0.95}$

conclusion

It is clear from the above analysis of variance tables that the presence of one or more outliers in the data leads to a very large deviation in the results, which leads to an incorrect analysis. Removing these outliers corrects this deviation so that the regression model parameters come closer to the regression model of the original data. Therefore, we see it necessary to conduct the outlier test process before conducting the model building process to obtain the best results.

References

1 .Aleng ,Nor Azlida, Nyi Nyi Naing, Norizan Mohamed and Kasypi Mokhtar, Outlier Detection based on Robust Parameter Estimates, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp.13429-13434

2. Brugger, Bernhard, Effects of Outliers in Regression Analysis and How to Handle hem,19 Dec 2024.

3. Farnè, Matteo and Angelos T. Vouldis, A methodology for automatised outlier detection in high-dimensional datasets: an application to euro area banks' supervisory data, No 2171/July 2018 European central bank.

4. Jiao, Xiyu, Felix pretis, Testing the Presence of Outliers in Regression Models,20june 2022, https://doi.org/10.1111/obes.12511

5. Mielke, Andreas, Regression and Outliers, Trufa Science Inside – No. 4, Deloitte Digital GmbH, Mannheim ,2019

6. Özlem Gürünlü Alma, Comparison of Robust Regression Methods in Linear Regression, Int. J. Contemp. Math. Sciences, Vol. 6, 2011, no. 9, 409 – 421

7. Rousseeuw Peter J., and Mia Hubert, Anomaly Detection by Robust Statistics, arXiv:1707.09752v2 [stat.ML] 14, 2017.

8. Shyti,Bederiana and Dhurata Valera),The Regression Model for the Statistical Analysis of Albanian Economy, International Journal of Mathematics Trends and Technology (IJMTT) – Volume 62 Number 2 – October 2018.

9. S, Stephen Raj and Senthamarai Kannan K , Detection of Outliers in Regression Model for Medical Data, International Journal of Medical Research & Health Sciences, 2017, 6(7): 50-56, ISSN No: 2319-5886