International Journal of Research in Computer Science
eISSN 2249-8265 Volume 2 Issue 2 (2012) pp. 23-28

www.ijorcs.org

MULTIPLE LINEAR REGRESSION MODELS IN OUTLIER DETECTION

S.M.A.Khaleelur Rahman[1], M.Mohamed Sathik[2], K.Senthamarai Kannan[3]

[1 2] *Sadakathullah Appa College, Tirunelveli, Tamilnadu, India*
[3]*Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India*

Abstract: *Identifying anomalous values in the real-world database is important both for improving the quality of original data and for reducing the impact of anomalous values in the process of knowledge discovery in databases. Such anomalous values give useful information to the data analyst in discovering useful patterns. Through isolation, these data may be separated and analyzed. The analysis of outliers and influential points is an important step of the regression diagnostics. In this paper, our aim is to detect the points which are very different from the others points. They do not seem to belong to a particular population and behave differently. If these influential points are to be removed it will lead to a different model. Distinction between these points is not always obvious and clear. Hence several indicators are used for identifying and analyzing outliers. Existing methods of outlier detection are based on manual inspection of graphically represented data. In this paper, we present a new approach in automating the process of detecting and isolating outliers. Impact of anomalous values on the dataset has been established by using two indicators DFFITS and Cook'sD. The process is based on modeling the human perception of exceptional values by using multiple linear regression analysis.*

Keywords: *Cut-value, Cook's D, DFFITS, multiple regression analysis, outlier detection.*

I. INTRODUCTION

IN statistics, outlier is an observation that is numerically distant from the rest of the data [1]. Grubbs defined outlier as an observation that appears to deviate markedly from other members of the sample in which it occurs [3]. Occurrence of outliers may be by chance in any distribution, but they are often indicative *either* of measurement error or of one population that has a heavy-tailed distribution. If the occurrence is by chance, robust statistical methods are being used to discard them. In case they indicate that the distribution has high kurtosis then one should be very cautious in using tools or intuitions that assume a normal distribution. A mixture model is frequently used for these two distinct sub-populations. Outlier points can indicate faulty data or erroneous procedures where a certain theory might not be valid. However, in large samples, a small number of outliers is to be expected (and not due to any anomalous condition). Outliers are most extreme observations with maximum or minimum sample. However, the sample maximum and minimum are not always outliers because they may not be unusually far from other observations. Thus, the term "outliers" to values "that lies very far from the middle of the distribution in either direction". This definition is applicable for continuously valued variables having a smooth pdf values. Sometimes, numeric distance is not the only consideration in detecting continuous outliers. "An *outlier* is a single, or very low frequency, occurrence of the value of a variable that is far away from the bulk of the values of the variable". The frequency of occurrence should be an important criterion for detecting outliers in categorical (nominal) data, which is quite common in the real-world databases. A more general definition of an outlier is given in: an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data [1]. The real cause of outlier occurrence is usually unknown to data users and/or analysts. Sometimes, this is a flawed value, resulting from the poor quality of a data set, i.e., a data entry or a data conversion error.

Outlier detection methods have been used to detect and remove anomalous values from data [7]. There are three approaches in the outlier detection process.

Determine the outliers through learning approach. No prior knowledge of the data is similar to unsupervised clustering. This approach processes the data as a static distribution, pinpoints the most remote points and identify them as potential outliers.

Pre-labeled data are marked as both normal and abnormal. This approach is similar to supervised classification .This is a semi-supervised recognition or detection task. This algorithm learns to recognize.

II. RELATED WORK

Mathematical calculations are used to find out whether the outlier came from the same or different population. Many statistical methods have been devised for detecting outliers by measuring how far the outlier is away from the other values. This can be the difference between the outlier and the mean of all points or the difference between the outlier and the

mean of the remaining values, or the difference between the outlier and the next closest value.

Different computer-based approaches have been proposed for detecting outlying data and it cannot be claimed that this is the generic or universally acceptable method. Therefore , these approaches were classified into four major categories based on the techniques used , which are: distribution-based, distance-based, density-based and deviation-based approaches [7]. Distribution-based approaches develop statistical models from the given data and then identify outliers with respect to the model using discordance test. Data sets may follow normal or Poison distribution. Objects that have low probability belong to the statistical model are declared as outliers. However, Distribution-based approaches cannot be applied on multidimensional data because they are univariate in nature. In addition, a prior knowledge of the data distribution is required, making the distribution-based approaches difficult to be used in practical applications. In the distance-based approach, outliers are detected by measuring distance. Main limitations of statistical methods are countered by this approach. Rather than working on statistical tests , objects that do not have enough neighbors are defined based on the distance from the given object. Density-based approaches compute the density of regions in the data and declare the objects in low dense regions as outliers. An outlier score to any given data point, known as Local Outlier Factor (LOF), depending on its distance from its local neighborhood, is assigned. Deviation-based approaches do not use statistical tests or distance-based measures to identify outliers. The objects that deviate from the description are treated as outliers.

Outlier detection methods can be divided as univariate methods, and multivariate methods. Currently many researches are worked upon multivariate methods. Another fundamental taxonomy of outlier detection methods is between parametric (statistical) methods and non-parametric methods that are model-free. Statistical parametric methods work on the assumption that underlying observations are known or, at least, they are based on statistical estimates of unknown distribution parameters [3] [4]. These methods flag those observations that deviate from the model assumptions as outliers. They are often unsuitable for high-dimensional data sets and for arbitrary data sets without prior knowledge of the underlying data distribution .

Regression analysis is a statistical tool for the investigation of relationships between variables [2]. Multiple regression is a technique that allows additional factors to enter the analysis separately so that the effect of each can be estimated. It is valuable for quantifying the impact of various simultaneous influences upon a single dependent variable. Further, because of omitted variables bias with simple regression, multiple regression is necessitated, because we are interested in the effects of one of the independent variables. Outliers are subject to masking and swamping [2][1]. One outlier masks a second outlier. In the presence of the first outlier the second is not considered as an outlier but can be considered an outlier only by itself. Thus, after the deletion of the first outlier the second appears as an outlier. In masking the resulting distance of the outlying point from the mean is short.

III. PROPOSED WORK

In this paper, a new approach for outliers detection to detect the values that are in an abnormal distance from other values is used. A number of suspicious observations are identified according to the indicators. To highlight the suspicious individuals the following indicators are used with a set of cut values. Various indicators and their cut values are

Indicator	Cut value
DFFITS	2 * SQRT (p / n)
Cook's D	4 / (n - p)

Here n is the dataset size, and p is the number of estimated parameters (number of descriptors + 1 for a regression with intercept). This method evaluates the overall influence of each observation.

DFFITS measures how much an observation has affected its fitted value from the regression model. It also measures values larger than in absolute value. These measures are highly influential. In DFFITS **,** the differences in individual fitted values with and without particular cases are calculated.

The following rules[3] are followed to consider the case influential if

- *DFFITS* > 1 for small data sets, or if
- *DFFITS* > 2 (*p*/*n*)1/2 for large data sets

Cook'sD measures aggregate impact of each observation on the group of regression coefficients, as well as the group of fitted values. By this method, differences in an aggregate of all fitted values with and without particular case values are calculated. It measures the influence of cases on all n fitted values . Cook's*D* does not require that *n* different regressions be run. According to Cook's*D*, a particular case can be influential by [5]

- Having a large residual e_i
- Having a high leverage h_{ii}
- Having both

Find the percentile value corresponding to *D* in the *F(p, n-p)* distribution. If the percentile is less than 0.10 or 0.20(not p value), then the particular case is relatively little influential. If the percentile is 0.5 or more, the case has a major influence on the fit.

Most assumptions of multiple regression cannot be tested explicitly but gross deviations and violations can be detected and dealt with appropriately. Outliers bias the results and skew the regression line in a particular direction , thereby leading to biased regression coefficients. Exclusion of a single deviated value can give a completely different set of results and model. The regression line demonstrates the prediction of the dependent variable (*Y*), by giving the independent variables (*X*). Observed points are varied around the fitted regression line. The deviation of a particular point from the regression line (predicted value) is called the *residual* value. Multiple regression analysis is capable of dealing with an arbitrarily large number of explanatory variables. With *n* explanatory variables, multiple regression analysis will estimate the sum of squared errors. Its intercept implies the constant term, and its slope in each dimension implies one of the regression coefficients.

In regression analysis , the term "leverage" is used for an undesirable effect . High leverage values do not make considered an "outlier" has an over proportional effect on the resulting regression curve. This effect may completely corrupt great impact on coefficients. It means that a single data point which is located well outside the bulk of the data and may be considered an "outlier" has an over proportional effect on the resulting regression curve. This effect may completely corrupt a regression model depending on the number of the samples and the distance of the outlier from the rest of the data. This regression model would be good in the absence of an outlier.

IV. RESULTS AND DISCUSSION

Now, we investigate our proposed method by using a synthetic data set. This is an artificial data set with two dimensions and values are entered to demonstrate how regression methods are used for detecting outliers and their influence. This data set has 14 attributes with 47 examples. All 14 attributes are continuous attributes.

This dataset has 47 instances. We want to explain the continuous attribute CrRate(CrimeRate) from four attributes such as Male14-24, Education, Unemp14-24 (Unemployed) and FInc(FamilyIncome). All the four attributes are also continuous in nature.

CrRate	Male14-24	Education	Unemp14-24	FInc
79.1	151	91	108	394
163.5	143	113	96	557
57.8	142	89	94	318
196.9	**136**	**121**	**102**	**673**
123.4	141	121	91	578
68.2	121	110	84	689
96.3	127	111	97	620
155.5	**131**	**109**	**79**	**472**
85.6	157	90	81	421
70.5	140	118	100	526
167.4	**124**	**105**	**77**	**657**
84.9	134	108	83	580
51.1	128	113	77	507
66.4	135	117	77	529
79.8	152	87	92	405
94.6	142	88	116	427
53.9	143	110	114	487
92.9	135	104	89	631
75	130	116	78	627
122.5	125	108	130	626
74.2	126	108	102	557
43.9	157	89	97	288
121.6	132	96	83	513
52.3	130	116	70	486
199.3	**131**	**121**	**102**	**674**
34.2	135	109	80	564
121.6	152	112	103	537
104.3	119	107	92	637
69.6	166	89	72	396
37.3	140	93	135	453
75.4	125	109	105	617
107.2	147	104	76	462
92.3	126	118	102	589
65.3	123	102	124	572
127.2	150	100	87	559
83.1	177	87	76	382
56.6	133	104	99	425
82.6	149	88	86	395
115.1	145	104	88	488
88	148	122	84	590
54.2	141	109	107	489
82.3	162	99	73	496
103	136	121	111	622
45.5	139	88	135	457
50.8	126	104	78	593
84.9	130	121	113	588

Table 1

Table 1 displays the number of outliers detected for two statistical procedures DFFITS and Cook'D. Cook's distance follows an F distribution so that sample size is an important factor. Large residuals tend to have large Cook's distance[5]. Influential observations are measured on the coefficient estimate.

Outliers are very different to the others i.e. they look, do not belong to the analysed population and if we remove these influential values, they lead us to a different model. The distinction between these kinds of points is not always obvious. One hypothesis is formulated about the relationship between the variables of interest, here, Education and CrimeRate, Male14-24 and CrimeRate, Unemp14-24 and Crime Rate, Family Income and CrimeRate. Common experience suggests that less family income is an important factor for inducing unemployed youth to commit more crimes. In our database only four records are discriminated which implies normal assumption is tend to make more money, people commit crime so that crime rate will be high for the group with less family income. But this assumption proved wrong here that male12-24 hailed from high family income commit more crimes. Thus, the hypothesis is that higher level of unemployment causes higher level of Crime but other factors also affect it. To investigate this hypothesis, we analyzed the data on Education and FamilyIncome (earnings) for the individuals at the age group of 14-24. Here CrimeRate is the dependent variable and remaining four variables male-14-24, Education, Unemployed 14-24 and FamilyIncome are independent variables. Let E denote education in years of schooling for each individual, and let I denote that individual's earnings in dollars per year. We can plot this information for all of the individuals in the sample using a two-dimensional diagram, conventionally termed a "scatter" diagram.

Effect of each variable can be estimated separately by using Multiple regression. Impact and influence of different variables upon a single dependent variable can also be quantified. As an investigator, we are interested only in the effects of different independent variables. Four independent variables Male14-24, Unemployed14-24, Family Income, Education and one explanatory variable CrimeRate are used to depict the effect in different scatter diagrams. More influential points in Cook'D have the higher values. In table2 these suspicious values are highlighted in records with number 4, 8,11 and 26.

Statistic	DFFITS	Cook's D
Lower bound	0.6523	-
Upper bound	0.6523	0.0952
1	0.32646009	0.09616140
2	1.78480983	0.39986444
3	0.66405863	0.31023771
4	**2.18146706**	**0.77293313**
5	0.43511620	0.13569903
6	-1.47110713	-0.59439152
7	-0.20111063	-0.04528905
8	**2.95457339**	**1.08224845**
9	0.10356779	0.03245340
10	-0.69109583	-0.20852734
11	**1.89708304**	**0.81378603**
12	-0.44522813	-0.09383719
13	-0.69840181	-0.26410747
14	-0.66027892	-0.21406586
15	0.23852123	0.07178046
16	0.86122960	0.29989931
17	-0.96075284	-0.27731615
18	-0.66088492	-0.21784022
19	-1.01672661	-0.27869084
20	0.62035054	0.24444196
21	-0.30894044	-0.06945060
22	-0.07215249	-0.03140858
23	1.31708193	0.41059065
24	0.31558868	0.15823852
25	-0.58615857	-0.28383282
26	**2.43565369**	**0.75912845**
27	-1.91721714	-0.40625769
28	0.35539761	0.12204570
29	0.18945691	0.06703783
30	-0.50027841	-0.20941073
31	-1.08754754	-0.46675214
32	-0.72746843	-0.18305531
33	0.72281331	0.18214980
34	-0.05855739	-0.01478066
35	-0.60047084	-0.22071476
36	0.45041770	0.14851908
37	-0.37726417	-0.21679878
38	-0.00519032	-0.00182978
39	0.50739807	0.15381441
40	0.80370468	0.14202599
41	-1.01047850	-0.40952945
42	-0.87981868	-0.20346297
43	-0.83315825	-0.34374225
44	-0.36672667	-0.12044833
45	-0.82375938	-0.39888713
46	-1.32948601	-0.44754258
47	-0.42401546	-0.13260822

Table 2

Attribute	Coef.	std	t(42)	p-value
Intercept	-222.752	128.140538	-1.738341	0.089478
Male14-24	1.160999	0.567406	2.046153	0.047035
Education	0.091014	0.674681	0.134899	0.893337
Unemp14-24	0.007371	0.293173	0.025141	0.980062
FamIncome	0.270389	0.089632	3.016646	0.004327

Table 3

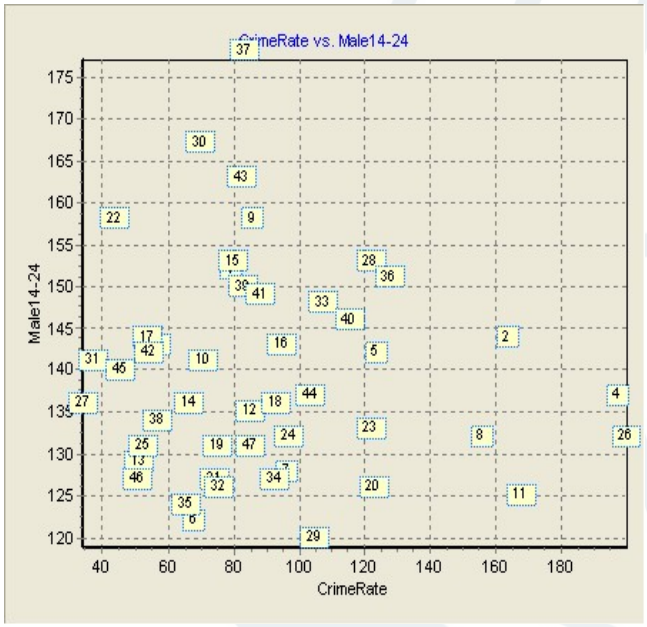

Figure 1

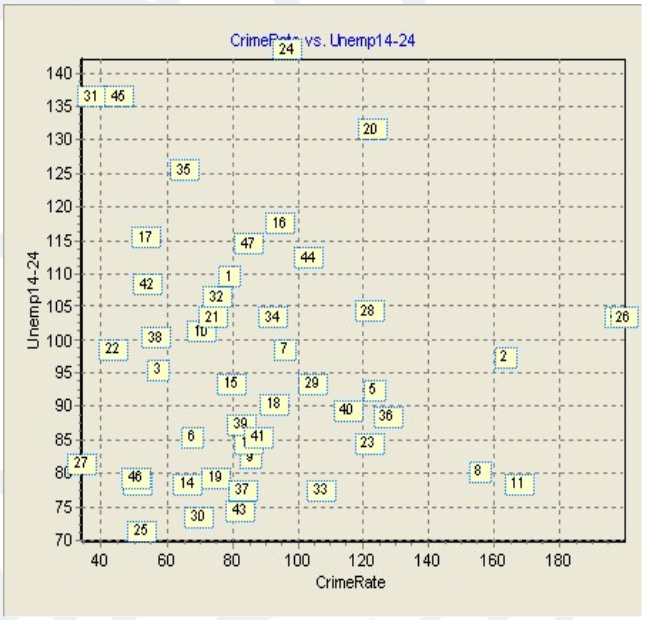

Figure 3

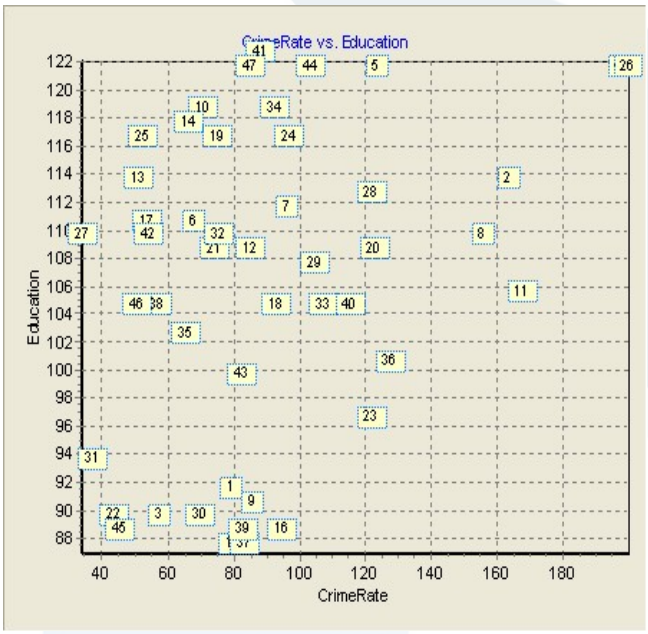

Figure 2

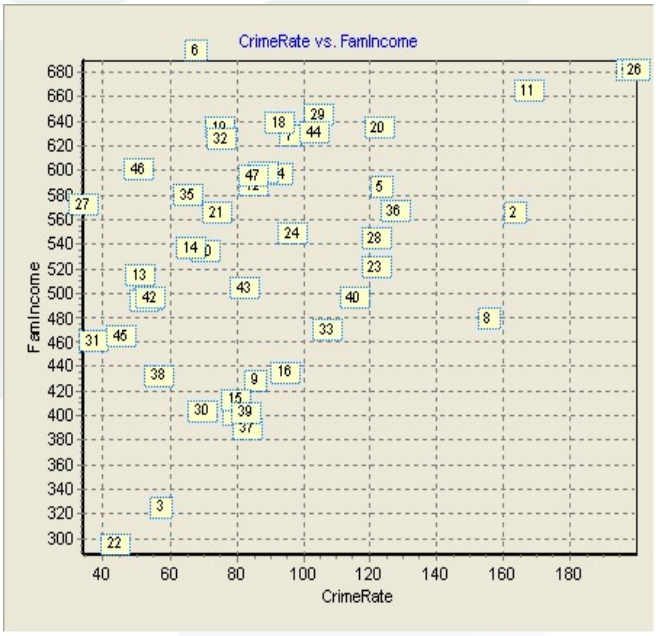

Figure 4

Global results

Endogenous attribute	**CrimeRate**
Examples	47
R^2	0.271953
Adjusted-R^2	0.238860
Sigma error	33.742476
F-Test (2,44)	(8.2178,000928)

Coefficients

Attribute	**Coef.**	**std**	**t(44)**	**p-value**
Intercept	-214.853376	102.132327	-2.103677	0.041157
FamIncome	0.277415	0.069458	3.993974	0.000243
Male14-24	1.151818	0.533284	2.159859	0.036273

V. CONCLUSION

In this paper multiple linear regression with n explanatory variables are used. We proved through examples that Multiple regression allows analysis of additional factors to estimate effect of each variable and quantifying the impact of simultaneous influence upon a single dependant variable. Regression models also explain the variation in the dependent variable well. Here we use this for predictive purposes. Above two statistics Cook'sD and DFFITS after careful consideration, omit influential points and regressions are refitted. Effect of the case can be studied by deleting the particular case from the data and analyzing the rest of the population. Multiple outliers are detected in multiple linear regression model. Distributions are used to get suitable cutoff points. Procedures are proposed and implemented with examples. Regression is not robust if the data set is small and can lead to inaccurate estimates.

VI. REFERENCES

[1] Hawkins. D. Identification of Outliers , Chapman and and Hall , London, 1980
[2] Barnett, V. and Lewis, T.: 1994, Outliers in Statistical Data. John Wiley & Sons., 3rd edition.
[3] Grubbs, F. E.: 1969, Procedures for detecting outlying observations in samples. Technometrics 11, 1–21. doi:10.1080/00401706.1969.10490657
[4] Rousseeuw, P. and Leroy, A.: 1996, Robust Regression and Outlier Detection. John Wiley & Sons., 3rd edition..
[5] Cook R. D and Weisberg S.T. (1982), Residuals and influence in New York Chapman and Hall.
[6] Abraham , B., and A.Chuang. "Outlier Detection and Time Series Modelling." Technometrics(1989). doi:10.1080/00401706.1989.10488517
[7] Jiawei Han and Micheline Kamber "Data Mining concepts and Techniques" Elsever, Second Edition.

How to cite

S.M.A.Khaleelur Rahman, M.Mohamed Sathik, K.Senthamarai Kannan, "Multiple Linear Regression Models in Outlier Detection". *International Journal of Research in Computer Science*, 2 (2): pp. 23-28, February 2012. doi:10.7815/ijorcs.22.2012.018