International Journal of Mathematics and
Computer Science, **8**(2013), no. 2, 55–68

Indices of Distances: Characteristics and Detection of Abnormal Points

Hicham Y. Abdallah

Department of Applied Mathematics
Faculty of Science-1
Lebanese University
Hadath, Lebanon

email: habdalah@ul.edu.lb

(Received September 5, 2012, Accepted November 1 , 2013)

Abstract

Getting a robust regression requires the detection of abnormal points. In this article we give a solution to this problem based on Cook's distance, DFFITS, DFBETAS, among others. We then compare the bounds for those distances which are used to detect abnormal point.

1 Introduction

Statistics is the science whose object is to collect, process and analyze data from the observation of random phenomena, that is to say where the accident occurs.
Data analysis is used to describe the phenomena, make predictions and decisions about them. For this, several statistical methods are available for these studies, but the most used is the regression method. Despite its effectiveness, the problem of influential points and outliers makes it less robust and affects its optimum results. Our objective is to solve this problem by showing the difference between the influential points and outliers.
We begin our study with a definition of influential points and outliers and then we discuss the different methods of detection of abnormal values. In

Key words and phrases: Outliers, influential points, distances.
AMS (MOS) Subject Classifications: 62J07,62J12.
ISSN 1814-0424

addition, we propose a theoretical comparison between the different indices to find the most effective index that helps us detect abnormal points.

2 Influential Points and Outliers

The study of residues $y_i - y_i^*$ can identify outliers observations or comments that play an important role in determining regression, where y^* is the prediction of y.
The two types of abnormal points are outliers and influential points . The first corresponds to observations outside the norm, while the second to points that weigh (unrealistically) on estimates: if removed, the results would be different significantly.
There are several methods for detecting these values according to their types. Once these observations have been identified, it may be better to remove them or use other more robust criteria.
1 - Influential Points:
An influential item weighs heavily in the regression; that is to say, the results are quite different depending on whether or not the point is taken into account in the regression.
The problem of influential values arises in business especially where surveys that collect economic variables have distributions that are highly nonsymmetric. Influential values are problematic because they generally lead to unstable estimators. In other words, including or excluding an influential sample value usually has a significant impact on the volatility of an estimator. It is possible to minimize their impact through an appropriate sampling plan. However, it is generally not possible to completely eliminate the problem of influential values at each step of the plan. As a result, it is important to develop robust estimation methods in the presence of influential values.
2 - Outliers:
Before presenting concepts related to outliers, it is necessary to define them more precisely as many authors have attempted to describe them and the definitions have changed over time. Among these authors, Grubbs has defined these values as follows:
An outlier is an observation that appears to deviate markedly compared to all other members of the sample in which it appears. In 1994, Barnett and Lewis defined an outlier in a data set as an observation (or set of observations) which appears to be inconsistent with other data. That is, outliers correspond to observations outside the norm of the population studied. These points can

distort the results of the regression.

3 The methods for detection of abnormal values

1 - Detection of outliers:
The indices that help us detect outliers are:
a-standardized Residue: e_i^*
This residue compares y with y^* the residue in the presence of the ith observation. An abnormally large value of a standardized residual is considered suspected. More precisely, the model is correct 95% of the time if one has:

$$|e_i^*| \leq t(0.025; n-p)$$

or

$$e_i^* = \frac{e_i}{s\sqrt{1-h_{ii}}}.$$

The observations that qualify as outliers belong to the same region:

$$|e_i^*| \leq t(0.025; n-p).$$

b-Student residue: e_i^{**}
This residue compares y with y^* without the ith observation. In this case, a value of abnormally large Student residue is considered suspected. For such observations we have:

$$|e_i^{**}| > t(0.025; n-p-1).$$

In practice, we use the following formulas:

$$e_i^{**} = \frac{e_{(-i)}}{s(-i)\sqrt{1-h_{ii}}},$$

where

$$e_i = y_i - y_i^*$$

s: standard deviation

$$h_{ii} = x_i(X'X)^{-1}x_i$$

where x_i is the ith row of X, $e_{(-i)}$ is the Residue without the ith observation, and $s_{(-i)}$ is the Standard Deviation without the ith observation.

The previous two criteria contribute to detect potentially influential observations by their distance to the size of the residues. This information is synthesized in the criteria directly assessing the influence of an observation on some parameters.
All these indicators suggest comparing an estimated parameter without the ith observation and this same parameter is estimated with all the observations.

C-treatment of outliers

We have already mentioned the main methods of outlier detection to finally get to address these issues and have a specific regression model.
Several methods of treatment are:
Reject: robustly eliminate extreme values determined following the test mismatch.
Incorporate: change the distribution model in order to incorporate outliers.
Identify: keep outliers, since they may represent particularly important features.
Accommodate: adopt statistical methods that minimize the impact of outliers on the statistical analysis.

2-Detection of influential points:

The indices that help us detect influential points are as follows:

a-Leverage point:

In regression analysis, we call a Leverage point an observation i that significantly affects the estimators because its values over other variables differ much of the rest of the data and the indicated distance between observation i and the center of gravity of the cloud of points. Leverage point h_{ii} observation is read from the main diagonal of the matrix Hat Matrix, and it appears as a measure of the influence of the ith observation on its proper prediction. In practice, an observation i is considered as a point Leverage point if $h_{ii} > \frac{2(p+1)}{n}$. Note also that an observation h_{ii} approaching 1 is an observation with a very important Leverage point.

b-Cook's distance:

Cook's distance from an observation is a measure of the influence of this observation on all the set of predictions of a model.
One calculates a distance between the vector$\widehat{\beta}$ of coefficients of the regression and the vector $\widehat{\beta}(i)$ obtained by repeating the regression without observation i.
When Cook's distance is normalized, a value greater than 1 is likely.
However, a limit of $\frac{4}{n-p-1}$ is often better as the calibration of 1 can permit influential values.

In practice, we have the following two formulas in this case

$$D_i = \frac{\sum_{i=1}^{n}(\widehat{y_i} - \widehat{y_i}(-i))^2}{\widehat{\sigma}_\epsilon^2(p+1)}$$

$$D_i = \frac{(\widehat{\beta} - \widehat{\beta}_{(-i)})'(X'X)^{-1}(\widehat{\beta} - \widehat{\beta}_{(} - i))}{\widehat{\sigma}_\epsilon^2(p+1)}$$

c-DFBETAS:
The purpose of this method is to measure the influence of a point on the estimated coefficient. It is normalized so as to be comparable from one variable to another.
A suspected observation is such that DFBETAS$> \frac{2}{\sqrt{(n).}}$
Note: If there are many variables, we first consider the globally influential observations (Cook) and then for this observation the variable(s) causing that influence (DFBETAS).
In practice,

$$(DFBETAS)_i^j = \frac{\beta_j - \beta_j(-i)}{s(-i)\sqrt{(X'X)_{j,j}^{-1}}},$$

where $(X'X)_{j,j}^{-1}$ is the jth position of the diagonal of $(X'X)^{-1}$

3 - Other criteria:
a-The DFFITS:
The DFFITS of an observation is a measure of the influence of this observation on
prediction of its eigenvalue by the model.
It gives the difference between the adjusted value for observation i and the predicted value of y for i in the estimated model without this observation i.
We consider an observation is influential when:

$$|(DFBETAS)_i| > 2\sqrt{\frac{p+1}{n}}$$

or

$$|(DFBETAS)_i| > \frac{\widehat{y_i} - \widehat{y_i}(-i)}{s(-i)\sqrt{h_{ii}}} = e_i^{**}\sqrt{\frac{h_{ii}}{1-h_{ii}}}$$

b-COVRATIO:
The COVRATIO measures disparities between the precision of the estimators; that is to say, the generalized variance of estimators given by

$$Var(\widehat{\beta}) = s^2 \det(X^{'}X)^{-1}.$$

The presence of the observation i improves the accuracy in the sense that it reduces the variance estimators if COVRATIO> 1
Instead, COVRATIO <1 indicates that the presence of the observation i degrades the variance.
But the most common detection rule is:

$$|COVRATIO_i - 1| > \frac{3(p+1)}{n}.$$

V-Theoretical Comparison of distances
Based on the critical regions of these different distances, we can choose the one that gives the best detection result.
Rule:
The method with the smallest critical region is the most accurate.
So compare these regions:
For Leverage point and COVRATIO:
If the critical region for the Leverage point $\frac{2(p+1)}{2}$ and that of $COVRATIO\frac{3(p+1)}{n} + 1$, then $\frac{2(p+1)}{n} < \frac{3(p+1)}{n} + 1$ and so the Leverage point is more accurate than COVRATIO.
For DFBETAS and DFFITS:
If the critical region for DFBETAS is $\frac{2}{\sqrt{n}}$ and that for DFFITS is $2\sqrt{\frac{p+1}{n}}$, then $\frac{2}{\sqrt{n}} < 2\sqrt{\frac{p+1}{2}}$ and therefore DFBETAS is more accurate than DFFITS.
For Cook's distance and Leverage point:
The critical region for the Leverage point is $\frac{2(p+1)}{n}$ and that of Cook's is $\frac{4}{n-p-1}$ and so $\frac{4}{n-p-1} < \frac{2(p+1)}{n}$.
For Cook's distance and DFBETAS:
The critical region is DFBETAS is $\frac{2}{\sqrt{n}}$ and that of Cook's is $\frac{4}{n-p-1}$ and so $\frac{4}{n-p-1} < \frac{2}{\sqrt{n}}$.
Comparing the critical region of the criteria already mentioned above:
$\frac{4}{n-p-1} < 1 < 2$ and $\frac{2}{\sqrt{n}} < 1 < 2$ and $\frac{2(p+1)}{n} < 2$ for $p+1 < n$ and $\frac{3(p+1)}{n} + 1 > 2$
shows that DCOOK is the best distance in detecting suspected points.

4 Relationships between the distances

The purpose of this section is to explain the relationships among the different distances.

Leverage pointage and Cook
The Leverage point arm measures horizontally the difference between the observed point and the mean $\overline{X}$ of the explanatory variable. They depend only on the values of that variable.
As for Cook distances, they measure somehow the overall importance of horizontal and vertical gaps.
In general, a point can be characterized by a significant residue, without being very influential, if the Leverage point arm is not very high. Similarly, a point may have a large Leverage point arm without being particularly influential, if the residue which is characterized is low.
A point is not therefore influential in the sense of Cook's distance if both, its residue and Leverage point arm are important.
DFFITS and Cook's distance
Cook's distance and DFFITS depend on Leverage point and CookD can be represented as a function of Leverage point and Student residue and even DFFITS. This shows that the observations with high Leverage point are the highest values of DCook and DFFITS then have a great influence on the predictions of the model.
COVRATIO, DFFITS and DFBETAS
Observing the rule of COVRATIO and those of DFFITS and DFBETAS, we note that they do not depend on sample size while COVRATIO depends on n.
DCook and DFBETAS
If there are many variables, we first look at the globally influential observations (DCook) and then for these observations with variable(s) causing that influence (DFBETAS).

5 Table summary

The following table summarizes the corresponding case detection of abnormal points.

	Result	**Purpose**
Standardized residuals	$\|e_i^*\| > 2$ whereas i residue is significantly $\neq 0$	Large residue detection thus atypical observation
Student residuals	$\|e_i^{**}\| > 2$ whereas i observation requires an investigation	Detection of large residue thus atypical observation
Leverage point	$h_{ii} > \frac{2(p+1)}{n}$	Measure the influence of observation i on the estimators
Cook's distance	CookD>1 indicates an abnormal effect	Measure the effect of the removal of the observation i on the prediction of n values
DFBETAS	$\|DFBETAS\| > \frac{2}{\sqrt{n}}$	Measure the influence of a point on the estimated coefficient
COVRATIO	$\|COVRATIO - 1\| > \frac{3(p+1)}{n}$	Measure the effect of the ith observation accuracy
DFFITS	$\|DEFITS\| > 2\sqrt{\frac{p+1}{n}}$	Measure the influence of observation i on the prediction of its eigenvalue

6 Practical Application

To illustrate our study, we propose a real example for an example of 100 students at the Faculty of Science of the Lebanese University taking as variable the average score in the Masters, first, second and third years.

In order to detect outliers for this example regression is performed in the first step to explain the marks in Masters according to the two explanatory variables are the grades in the first year and those in the second and third year together and, as a second step, we determine the critical areas of indices cited in the study:

Student Number	Major	Masters Average(AM)	Average of second and third years (A23)	Average of first year (A1)
1	biology	71.48	64.92	50.17
2	biology	76.35	70.82	60
3	Chemical Physics	61.55	64.13	53
4	biochemistry	66.37	66.86	65.17
5	biology	72.13	64.55	67
6	biochemistry	81.18	79.26	79.92
7	biology	70.85	63.13	61.67
8	biochemistry	76	77.54	76.50
9	biology	80.40	79.20	69.50
10	Fundamental Physics	71.37	68.52	60.17
11	Electronics	69.37	63.52	51.33
12	biology	81.57	71.88	57.67
13	biology	74	71	66
14	biochemistry	58.85	62.48	58.25
15	biology	81.08	72.30	60.75
16	biochemistry	64.17	67.86	51.25
17	mathematical	63.50	64.85	69.42
18	chemistry-molecular	66.07	68.44	55.50
19	Molecular Chemistry	77.13	78.58	84.67
20	biochemistry	77.62	74.92	56.33
21	Fundamental Physics	71.58	68.68	55.75
22	Electronics	61.78	61.81	52.83
23	biochemistry	72.18	69.63	64.75
24	Fundamental Physics	66.92	69.66	63.33
25	biology	75.60	66.67	50
26	Biology: Elective	71.37	68.65	50
27	Chemical Physics	67.92	72.19	62.50
28	Computer	63.42	70.90	66.25
29	biology	66.72	65.43	60.67
30	biochemistry	64.83	67.02	59
31	biology	73.07	70.17	50
32	biochemistry	66.30	67.40	59.75
33	biochemistry	59.95	66.28	56.17

Student Number	Major	Masters Average(AM)	Average of second and third years (A23)	Average of first year (A1)
34	chemistry-molecular	75.20	76.25	74.48
35	Electronics	69	61.94	53
36	biochemistry	66.33	68.93	68.92
37	biochemistry	68.85	66.55	57.48
38	Computer	66.02	63.22	58.50
39	Fundamental Physics	74.42	72.54	63.83
40	Chemical Physics	65.63	68.66	63
41	biochemistry	76.17	72.53	61.17
42	chemistry-molecular	67.65	63.63	50
43	biochemistry	59.30	62.28	59.33
44	biochemistry	58.65	61.06	56
45	biology	77.08	71.85	78.25
46	biology	76.23	68.23	54.17
47	chemistry-molecular	58.60	60.67	50
48	Chemistry option Environmental Sciences	81.05	83.88	82.67
49	biology	81.83	80.37	80.83
50	biology	70.88	71.83	51.42
51	biology	73.48	66.62	62.08
52	Electronics	68.63	64.34	52.92
53	biochemistry	68.08	66.31	63.58
54	biology	67.75	62.68	57
55	biology	62.03	68.90	64.83
56	Computer	69.62	77.71	77.67
57	biology	70.80	64.92	66.17
58	Electronics	76.88	72.21	54.50
59	biochemistry	77.33	76.10	78.42
60	Computer	61.03	63.53	66.08
61	biochemistry	76.60	75.39	69.83
62	biochemistry	69.73	64.29	53.50
63	biology	87.97	83.12	82.08
64	chemistry-molecular	67.35	69.60	50
65	Chemical	59.73	63.75	52.50

Student Number	Major	Masters Average(AM)	Average of second and third years (A23)	Average of first year (A1)
66	biology	77.93	70.55	50.50
67	biochemistry	66.45	67.41	59.58
68	biochemistry	62.30	66.28	50
69	biology	76.25	66.63	56.92
70	chemistry-molecular	72.22	71.83	62.33
71	Fundamental Physics	78.53	82.57	74.50
72	chemistry-molecular	69.80	69.81	52.08
73	Computer	68.40	67.17	67.58
74	Fundamental Physics	69.95	66.58	50
75	biochemistry	70.05	66.38	60.83
76	Chemical Physics	66.13	68.80	64.42
77	biochemistry	70.22	64.78	52.83
78	Computer	64.37	71.80	60.42
79	Electronics	68.68	63.58	57.67
80	biochemistry	65.35	60.78	59.58
81	mathematics	74.50	76.20	84.33
82	biochemistry	66.80	67.93	54.92
83	Electronics	65.90	67.24	61.17
84	mathematics	57.30	61.10	67.75
85	Electronics	67.80	63.17	50.58
86	biochemistry	64.07	64.26	56
87	Electronics	68.42	58.47	50.75
88	biochemistry	69.92	68.70	67.17
89	biochemistry	71.73	72.03	73.42
90	Fundamental Physics	64.67	71.22	53.92
91	Computer	64.17	68.44	59
92	biology	71.52	63.87	60
93	Electronics	67.55	64.36	50
94	biology	70.30	65.55	67.50
95	Computer	59.60	60.85	57.42
96	biochemistry	71.95	70.32	59.58
97	biology	73.35	70.33	53.08
98	biochemistry	71.22	67.13	58.75
99	Electronics	77.57	76.54	56.08
100	Biology: Elec-	65.30	62.98	50

7 Interpretation

The proposed study has two variables (p = 2) and 100 observations,
The model is obtained AM= 10.78164 + 0.92954 A23 - 0.07701 A1.
For a threshold using the SAS software, we had the following results:
Using the Student residual and standardized residual, each observation more than two is an abnormal finding. By examining the different values ??of the residues shows that the observation of which 12 are medium (A1 = 57.67, A23 = 71.88, AM = 81.57) with a student residue = 2.0624 is the first atypical value, observation 78 (A1 = 60.42, A23 = 71.80, AM = 64.37) each have 2035 and as the value of -2.0694 STUDENT RESIDUAL RSTUDENT and is the second atypical value.
Using the Leverage point, each greater than $\frac{2(p+1)}{n} = \frac{2(2+1}{100} = 0.06$ observation is an unusual observation. Then the software uses gives us the values ??of the matrix diagram Hat. We takes a few examples:
6 (A1=79.92, A23=79.26, AM=81.18) and the Leverage point is (h=0,0611) ; 9 (A1=69.50,
A23=79.20, AM=80.40) (h=0,0801) ; 48(A1=82.67, A23=83.88, AM=81.05) (h=0,0966) ....
From Cook's distance, each more than $\frac{4}{n-p-1} = \frac{4}{100-2-1} = 0.041$ observation is an unusual observation. While examining the column COOK found that 56 observations whose average are (A1 = 77.67, A23 = 77.71, AM = 60.62) and Cook (D = 0.056) and 63 (A1 = 82.08, A23 = 83.12, AM = 87.97) (D = 0.080) and 84 (A1 = 67.75, A23 = 61.10, AM = 57.30) (D = 0.045);) are just outliers.
The dfbetas indicates that each variable has a value greater than $\frac{2}{\sqrt{n}} = 0.2$ corresponds to a unique value, then the variable A23 has a unique value 5 (A23 = 64.55) ,7 (63.13) ,12 (71.88) , 63 (83.12) ,66 (70.55) ,84 (61.10) ,87 (58.47) ,90 (71.22) and the A1 variable 5 (A1 = 67.00), 12 (57.67) ,25 (50.00), 45 (78.25), 66 (50.50), 84 (67.75) ,90 (53.92).
Using COVRATIO, each observation has $|covratio - 1| > \frac{3(p+1)}{n} = 0.07$ is an unusual observation. Then follows that the observations 6 (Average A1 = 79.92, A23 = 79.26, AM = 81.18), 8 (A1 = 76.50, A23 = 77.54, AM = 76.00), 9 (A1 = 69.50, A23 = 79.20, AM = 80.40), 12 (A1 = 57.67, A23 = 71.88, AM = 81.57), 19 (A1 = 84.67, A23 = 78.58, AM = 77.13), 20 (A1 = 56.33, A23 = 74.92, AM = 77.62) are outliers.
Finally, we note that the outliers will differ from one remote to another this is due to the existing difference between the distances. In our project, comparing outliers obtained using different distances shows that DCook is

best.
The results are given using DCOOK as follows:
For the individual 56, the average in the first three years (A1 = 77.67, A23 = 77.71) is higher than the average Masters (AM = 60.62)
-for the individual 63, the average in the first three years (A1 = 82.08, A23 = 83.12) is lower than the average Masters (AM = 87.97)
-for the individual 84, the average in the first three years (A1 = 67.75, A23 = 61.10) is higher than the average Masters (AM = 57.30)
-for the individual 87, the average in the first three years (A1 = 50.75, A23 = 58.47) is lower than the average Masters (AM = 68.42)
Thus, these results are consistent with our study, as the fact that these points are atypical of the variation is in the opposite direction between the explanatory variables and to explain that.

8 Conclusion

The observations contained in the databases must absolutely be validated because the appearance of outliers is inevitable because of the quality of data processed and the various sources of errors during acquisition.
To ensure high quality information, a search for stragglers or outliers must be done before the use of databases.
So this article has helped us to differentiate an outlier from influential. We also studied several methods for the detection of outliers by showing that the test suitable for the detection of abnormal points is Cook's distance.
Simulations on large files are the subject of current research.

References

[1] Ricco Rakotomolala, Pratique de la régression linéaire multiple (diagnostic et sélection de variable), université lumière Lyon 2, 2009.

[2] Ricco Rakotomolala, Points atypiques et points influents, régression linéaire multiple, université LYON 2, 2009.

[3] David A Belsey, Edwin Kuhroy, Regression diagnostic identifying influential data and sources of colinearity, 2004.

[4] Pierre André Comilon, Eric Matzner Lober, Regression: théorie et applications, Springer, 2007.

[5] Laurent Carraro, Introduction a la régression, 2005.

[6] Mathieu Resche-Rigon, Outills de régression, résidus, mesure d'influence individuelle, 2010.

[7] Stéphane Canu, Diagnostic de la régression, 2011.

[8] Philippe Besse, Pratique de la modélisation statistique, 2000.