BRAC University Journal, vol. V, no. 2, 2008, pp. 31-39

MULTIPLE OUTLIERS DETECTION: APPLICATION TO RESEARCH & DEVELOPMENT SPENDING AND PRODUCTIVITY GROWTH

A. A. M. Nurunnabi[1]
School of Business
Uttara University, Dhaka-1230, Bangladesh
email: pyal1471@yahoo.com

and

Mohammed Nasser
Department of Statistics
University of Rajshahi, Rajshahi-6205, Bangladesh
email: mnasser.ru@gmail.com

ABSTRACT

Multiple outliers are frequently encountered in applied studies in business and economics. Most of the practitioners depend on ordinary least squares (OLS) method for parameter estimation in regression analysis without identifying outliers properly. It is evident that OLS totally fails even in presence of single outlying observation. Single observation outlier detection methods are failed to identify multiple outliers due to masking and swamping effects. This paper analytically and numerically compares the sensitivity of the most popular diagnostic statistics. Data set from Griliches and Lichtenberg (1984) is used to show that we need to take extra care for model building process in presence of multiple outliers.

Key words: Influential observation, masking, outlier, regression diagnostics, swamping.

1. INTRODUCTION

Applied studies in business and economics routinely encounter data that include unusual observations. Researchers often chose to discard or retain these observations depending upon whether they believe the observations are mistakes or simply atypical comparing with bulk of the data. Williams *et al*., (2002) and Liu *et al*., (2004) show, although unusual observations are often considered as an error or noise, they may carry important information. Detected outliers are candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. It is therefore important to identify them prior to modeling and analysis.

Increasingly investigators have come to rely upon post-estimation diagnostics to identify outliers and influential observations. Many statistical software packages (*e.g*., MINITAB, R, SAS, S-PLUS) now include different measures to identify them. Measures are shown to be sensitive to specific type of unusual observations. Some are sensitive to outliers or leverage points, and some are to influential observations. This paper compares and shows different results are come from a specific data set by using different diagnostic statistics side by side.

We have a short discussion on regression and OLS in section II, in section III we write about different fields of outlier detection, its classification, and a short brief about most common outlier detection methods in linear regression respectively. Section IV illustrates multiple outlier detection in application to a well-referred data set. Conclusion is presented in section V.

II. REGRESSION AND ORDINARY LEAST SQUARES METHOD

Regression analysis is a statistical technique, most widely used in almost every field of research and application in multifactor data, which helps us to

[1]For all correspondence

investigate and to fit an unknown model for, quantifies relations among observed variables. Chatterjee and Hadi (2006) point out; it is appealing because it provides a conceptually simple method for investigating functional relationship among variables. We use the customary multiple regression notation:

$$Y = X\beta + \varepsilon \qquad (1)$$

where X is an $n\times k$ matrix of carriers (regressors), including the constant term, Y is an $n\times 1$ column vector for the response variable, β is a $k\times 1$ ($k = p+1$) coloumn-vector of parameters, and ε is an $n\times 1$ column-vector of error terms.

Rousseeuw and Leroy (1987) mention, the most popular regression estimator dates back to Gauss and Lgendre (see Plackett, 1972; and Stigler, 1981), and corresponds to

$$\underset{\beta}{Minimize} \sum_{i=1}^{n} r_i^2 \quad , \qquad (2)$$

where i-th residual, $r_i = y_i - \hat{y}_i$, and $\hat{y}_i$ is the i-th estimated value of y_i . Till date it (OLS) is most popular for its computational simplicity and mathematical beauty. But now a day it faces a huge criticism, it proves unreliable and falsify itself in presence of outliers. It is well-known that OLS stands on a specific set of assumptions. The implicit assumption, all observations are equally reliable and should have an equal role in determining the least squares results and influencing conclusions (see Chatterjee and Hadi, 1988), claims for the identification of outliers/influential observations. Since OLS has a 0% breakdown value (meaning that an arbitrary small percentage of bad observations can change the OLS coefficients to any value at all from -∞ to +∞; see Hampel 1971, 1974 for the concept of breakdown point), even a small proportion of deviant observations in a large sample can cause systematic distortion in OLS estimates (Rousseeuw and Wagner, 1994). In small sample OLS residuals are of little help in identifying outliers. Rousseeuw and Leroy (1987) present many real data sets in which OLS residuals fail to detect unusual data although big outliers exist.

III. UNUSUAL OBSERVATIONS: MOTIVATION, DEFINITIONS AND IDENTIFICATIONS

Outlying observations are unusual in the sense that they are exceptional, they have extra role on model building process, or they may come from other population(s) and do not follow the pattern of the majority of the data. The presence of unusual observations could make huge interactive problems in inference. Because some times they can unduly influence the results of the analysis, and their presence may be a signal that the regression model fails to capture important characteristics of the data. What are outliers and what is the outlier problem? An interesting answer is found in the following quotation (see Barnett and Lewis, 1995).

In almost every true series of observations, some are found, which differ so much from the others as to indicate some abnormal source of error not contemplated in the theoretical discussions, and the introduction of which into the investigations can only serve … to perplex and mislead the inquirer.

Outlying observations do not inevitably ‘perplex’ or ‘mislead’; they are not necessarily ‘bad’ or ‘erroneous’, and the experimenter may be tempted in some situations not to reject an outlier but to welcome it as an indication of new and important findings.

A. Outliers and Influential Observations

To statisticians, unusual observations are generally either outliers or ‘influential’ data points. In regression analysis, generally they categorize unusual observation (commonly saying as outliers) into three: outliers, high leverage points and influential observations. Johnson and Wichern (2002) defines an outlier, as an observation in a data set which appears to be inconsistent with the remainder of the set of data. In other words, Hawkins (1980) point out, an outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.

In regression, outliers can deviate into three ways (i) the change in the direction of response (Y) variable (ii) the deviation in the space of explanatory variable(s), deviated points in X-direction called leverage points, and (iii) the other is change in both the directions (direction of the explanatory variable(s) and the response variable). According to Belsley *et al.* (1980), influential observation is one which either individual or together with several other observations has a demonstrably larger impact on the calculated values of various estimates than is the case for most of the other observations. Chatterjee and Hadi (1986) point out, 'as with outliers, high leverage points need not be influential, and influential observations are not necessarily high-leverage points'. When an observation is considered to be both an outlier and influential, regression results are usually reported with and without the observation. When observations are not outliers but are influential, it is less clear what should be done.

B. Taxonomy of Outlier Detection Methods

Outliers detection have been suggested for numerous applications, such as credit and fraud detection, clinical trials, voting irregularity analysis, network intrusion, severe weather prediction, geographic information system, and other data mining tasks (Barnett and Lewis, 1995; Fawcett and Provost, 1997; Hawkins, 1980; and Penny and Jolliffe, 2001).

The identification of outliers and influential observations in regression diagnostics is relatively new topic in business and economic studies but is rapidly gaining recognition and acceptance to the analyst as a supplement to the traditional analysis of residuals. Outlier detection methods can be divided into two, univariate and multivariate (usually form most of the current body of research) methods. Another fundamental taxonomy of outlier detection methods is between parametric methods and nonparametric methods that are model-free (*e.g.*, Williams *et al.*, 2002). Statistical parametric methods either assume a known underlying distribution of the observations (*e.g.*, Bernett and Lewis, 1995; Rousseeuw and Leroy, 1987) or, at least, they are based on statistical estimates of unknown distribution parameters (Caussinus and Roiz, 1990; Hadi, 1992). These methods flag as outliers those observations that deviate from the model assumptions. They are often unsuitable for high-dimensional data sets and for arbitrary data sets without prior knowledge of the underlying data distribution (Papadimitriou *et al.*, 2002). Within the class of non-parametric methods one can set apart the data mining methods, also called distance-based methods. These methods are usually based on local distance measures and are capable of handling large databases (e.g., Hawkins *et al.*, 2002; Knorr *et al.*, 2000, 2001; Knorr and Ng. 1997, 1998; Williams *et al.*, 2002; and Williams and Huang, 1997). Another class of outlier detection methods is founded on clustering techniques, where a cluster of small sizes can be considered as clustered outliers (Acuna and Rodriguez, 2004; Kaufman and Rousseeuw, 1990; and Shekhar *et al.*, 2001, 2002).

For the sake of time and study limitations and to show the necessity in economic applications, we briefly describe the most common statistical parametric outlier detection methods as follows.

Identification of Outliers

In the scale parameter context, by an outlier we mean an observation which is so much larger than the bulk of the observations that it stands out. A rule suggests: an observation as outlying if it is more than three times the inter-quartile range from the median (Staudte and Sheather, 1990). A good deal of outliers and influential observations detection methods are suggested in regression literature (Atkinson and Riani, 2004; Belsley *et al.*, 1980; Chatterjee and Hadi, 1986, 1988; and Rousseeuw and Leroy, 1987). Among those, measures based on residual or some functions of residuals (standardized and Studentized residuals) are for outliers, the diagonal elements of hat matrix are for high-leverage values and the Cook's distance and the difference in fitted values (DFFITS) for influential observations are generally used in identification purpose. We can express the above measures as follows.

By the OLS method the vector of estimated parameter is

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad . \qquad (3)$$

and the corresponding residual vector is,

$$r = Y - X\hat{\beta} = (I - H)\varepsilon \quad , \qquad (4)$$

where $H = X(X^T X)^{-1} X^T$ is the leverage or prediction or hat matrix. In scalar form, *i*-th residual is

$$r_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j; \quad i = 1,2,...,n \quad . \qquad (5)$$

Clearly, if the h_{ij} are sufficiently small, r_i will serve as a reasonable alternative of ε_i.

Chatterjee and Hadi (1988) mention that, the ordinary residuals are not appropriate for diagnostic purpose; a transformed version of them is preferable. A logical scaling for the residuals is the standardized residuals

$$r_i^* = \frac{r_i}{\sqrt{MS_{res}}}; i = 1,2,...,n \qquad (6)$$

where MS_{res} is the mean squared residuals, r_i^* have mean zero and approximate variance equal to one, consequently a large standardized residual potentially indicate an outlier. Since the residuals have different variances and they are correlated, the variance of the i-th residual is

$$V(r_i) = \sigma^2(1-h_{ii}) \quad , \qquad (7)$$

where h_{ii} is the i-th diagonal element of H and σ is an estimate of $MS_{res.}$ Daniel and Wood (1971) introduce a type of (i-th internally Studentized) residual as

$$e_i = \frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}}; i = 1,2,...,n \qquad (8)$$

where $\sigma_i = \hat{\sigma}\sqrt{1-h_{ii}}$. But many of the authors feel that internally Studentized residuals are over estimated by the extreme observations and they have suggested the i-th externally Studentized residuals,

$$e_i^* = \frac{r_i}{\hat{\sigma}^{(-i)}\sqrt{1-h_{ii}}}; \quad i = 1,2,...,n \qquad (9)$$

where $\sigma_i = \hat{\sigma}^{(-i)}\sqrt{1-h_{ii}}$ and

$$\hat{\sigma}^{(-i)2} = \frac{Y^{(-i)T}\left(I - H^{(-i)}\right)Y^{(-i)}}{n-k-1} \qquad i = 1,2,...,n$$

is the residual mean squared error estimate when the i-th observation is omitted and

$H^{(-i)} = X^{(-i)}\left(X^{(-i)T}X^{(-i)}\right)^{-1}X^{(-i)T}$, *i=1, 2,..., n*

is the prediction matrix for $X^{(-i)}$. Atkinson (1981) prefers e_i^* over e_i for detecting outliers. The diagonal elements of H denoted by h_{ii} and defined by

$$h_{ii} = x_i^T(X^TX)^{-1}x_i, \quad i = 1,2,...,n \qquad , (10)$$

are termed as leverage values. Observations corresponding to excessively large values of h_{ii} are treated as high-leverage points. Velleman and Welsch (1981) consider leverage values greater than *3p/n* as high-leverage points. Most popular identification techniques of influential observations are Cook's distance (Cook, 1977) and DFFITS (Belsely *et al.* 1980) defined as

Cook's distance,

$$CD_i = \frac{(\hat{\beta}^{(-i)} - \hat{\beta})^T X^T X(\hat{\beta}^{(-i)} - \hat{\beta})}{k\sigma^2}, \quad i = 1,2,...,n \qquad (11)$$

and

$$DFFITS = \frac{\hat{y}_i - \hat{y}_i^{(-i)}}{\hat{\sigma}_{(-i)}\sqrt{h_{ii}}}, \quad i = 1,2,...,n \qquad (12)$$

Observations greater than 1 for Cook's distance and $|DFFITS| \geq 2\sqrt{k/n}$ are treated as influential. Imon and Ali (2005) mention that, residuals together with leverage values may cause masking and swamping for which a good number of unusual observations remain undetected in the presence of multiple outliers and multiple high-leverage points. Imon (2005) proposes GDFFITS for identifying multiple influential observations based on the idea of group deletion techniques (Atkinson, 1994; Hadi and Simonoff, 1993) as

$$GDFFITS = \sqrt{h_{ii}^* t_i^*}, \quad i = 1,2,...,n \qquad (13)$$

where

$$h_{ii}^* = \begin{cases} \dfrac{h_{ii(R)}}{1 - h_{ii(R)}} & for \quad i \in R \\ \dfrac{h_{ii(R)}}{1 + h_{ii(R)}} & for \quad i \notin R \end{cases} \qquad (14)$$

and

$$t_i^* = \begin{cases} \dfrac{y_i - x_i^T\hat{\beta}_{(R)}}{\hat{\sigma}_{R-i}\sqrt{1 - h_{ii(R)}}} & for \quad i \in R \\ \dfrac{y_i - x_i^T\hat{\beta}_R}{\hat{\sigma}_R\sqrt{1 + h_{ii(R)}}} & for \quad i \notin R \end{cases} , \quad (15)$$

where

$$\hat{\beta}_{(R)} = (X_{(R)}^T X_{(R)})^{-1} X_{(R)}^T Y_{(R)} \quad ,$$

$$h_{ii(R)} = x_i^T (X_{(R)}^T X_{(R)})^{-1} x_i \quad ,$$

and $\hat{\sigma}_R$ is the estimated standard error for remaining (R) (without suspected unusual observations) group of regular observations. Imon suggests observations

$$|GDFFITS| \geq 3\sqrt{k/(n-d)}$$

as influential. He shows with a number of examples GDFFITS performs well for identifying multiple influential observations even in presence of masking and/or swamping effects.

IV. APPLICATION

A number of economists (*e.g.*, Franses and Biessen, 1992; and Griliches and Lichtenberg, 1984) have studied the relationship between productivity growth in the U.S. and R & D spending. Here we consider the data set in Griliches and Lichtenberg (1984) (see in Appendix) which is also taken as an illustration of detecting multiple outliers in Reiss (1990). Griliches and Lichtenberg use several regressions to analyze the relationship between private (PRIV) and federal (FED) expenditures on R & D, and total factor productivity growth (TFPG) for 27 industries. One common form is

$$TFPG = \beta_0 + \beta_1 PRIV + \beta_2 FED + \varepsilon(error) \quad , \qquad (16)$$

where β_C is a constant term, and the coefficients β_1 *and* β_2 represent excess social rates of return to private and federal R & D respectively. Plots of the dependent variable against each of the regressors are provided in Figures 1 and 2. The full sample regression estimates are reported in Table 3, that are similar to those in Griliches and Lichtenberg (1984). OLS estimated line in Figure 2, TFPG versus FED, shows how only case 2 (missiles and spacecraft) affect the rest of the cases. Clearly it is outlying in x-space (*i.e.*, high-leverage point) and evident by the respective hat-value. The figure shows how OLS has a destructive consequence by a single unusual observation in model estimation process. Results in Table 3 imply a significant 34.6% social excess rate of return to private R & D and an insignificant, 1% rate of return to federal R & D. Seeing the Figures 1 and 2 we can easily understand the presence of unusual observations.

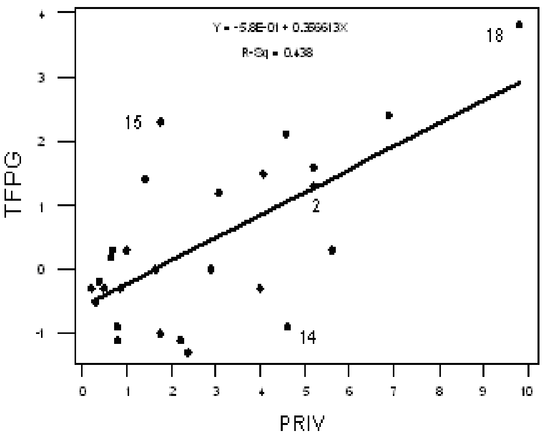


Figure 1: Scatter plot with OLS line; TFPG versus PRIV

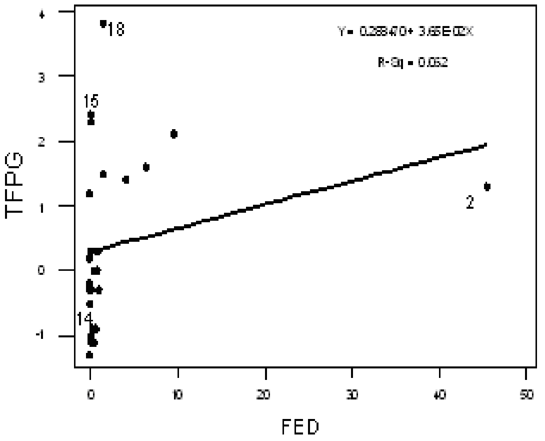


Figure 2: Scatter plot with OLS line; TFPG versus FED

An analysis of single case deletion diagnostics is given in Table 1; residual measures (standardized and Studentized) pick out cases 14 (engines) and 15 (farm machineries) as outliers, observations 2 (missiles and spacecrafts) and 18 (office, accounting and computing machines) as high-leverage points by the hat-values in Table 1 (column 4 and 10). Residuals and hat values give different results and it is also slight confusing. We calculate Cook's distance and DFFITS those who take into account both leverage and residuals. Though Cook's distance identifies only missiles and spacecraft (case 2) as influential but DFFITS identifies two cases missiles and office machines (cases 2 and 18) as influential observations. By single case deletion diagnostic measures we may consider that there is a group of observations are responsible to falsify the OLS estimates.

Table 1: Single case deletion diagnostics for Griliches and Lichtenberg (1984) data

Ind	|Std. Res| (2.00)	|Stu. Res| (2.50)	HAT val (0.22)	Cook Dis (1.00)	|DFFITS| (0.666)	Ind	|Std. Res| (2.00)	|Stu. Res| (2.50)	HAT value (0.22)	Cook Dist (1.00)	|DFFITS| (0.666)
1	1.48	1.52	0.05	0.04	0.358	15	**2.31**	**2.57**	0.04	0.08	0.555
2	-1.64	-1.70	**0.94**	**14.78**	**-6.916**	16	-1.05	-1.05	0.04	0.02	-0.227
3	0.22	0.21	0.08	0.00	0.063	17	0.54	0.53	0.06	0.01	0.131
4	-0.03	-0.03	0.08	0.00	-0.007	18	1.25	1.26	**0.41**	0.37	**1.064**
5	-1.12	-1.13	0.11	0.05	-0.397	19	-0.03	-0.03	0.06	0.00	-0.007
6	0.73	0.72	0.04	0.01	0.153	20	-1.15	-1.15	0.05	0.02	-0.273
7	-1.57	-1.62	0.04	0.03	-0.335	21	-0.44	-0.44	0.04	0.00	-0.089
8	0.65	0.65	0.18	0.03	0.301	22	0.00	0.00	0.05	0.00	0.001
9	-0.82	-0.81	0.06	0.01	-0.209	23	0.32	0.31	0.08	0.00	0.092
10	0.57	0.56	0.07	0.01	0.149	24	-1.31	-1.33	0.04	0.02	-0.276
11	0.25	0.25	0.07	0.00	0.070	25	0.66	0.65	0.06	0.01	0.171
12	0.12	0.11	0.07	0.00	0.031	26	1.03	1.03	0.07	0.03	0.291
13	-0.62	-0.61	0.06	0.01	-0.156	27	0.68	0.67	0.05	0.01	0.157
14	**-1.99**	**-2.13**	0.07	0.10	-0.590						

Now we apply generalized DFFITS (GDFFITS) by deleting suspect group of all four cases (2, 14, 15, and 18). As a multiple outlier diagnostic measure GDFFITS identifies three cases 2, 18 and a new case 8 (drugs and medicines) as influential observations. It shows that drugs-medicines is masked before in presence of the above mentioned suspect group of four. Figure 3 (standardized residual versus fitted value) gives the proper indication in favor of the outcome from the GDFFITS in Table 2. As a result, when the cases 2, 8, and 18 are deleted, the estimation results show a significant parameter ($\beta_2 = 0.249$) for FED and an insignificant one for PRIV ($\beta_1 = 0.058$). When we delete the four (cases 2, 14, 15, and 18) after single case deletion diagnostics, results show both PRIV and FED have significant contribution to TFPG, and that is totally reverse findings when none of them is deleted. For the full sample we get private R & D has a stronger effect on TFPG than federal R & D, after deletion of the cases 2, 8 and 18.

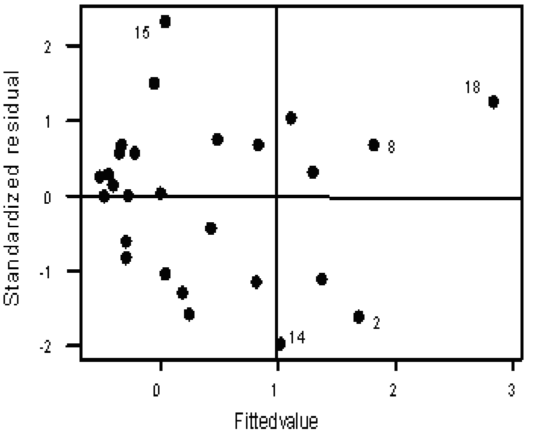
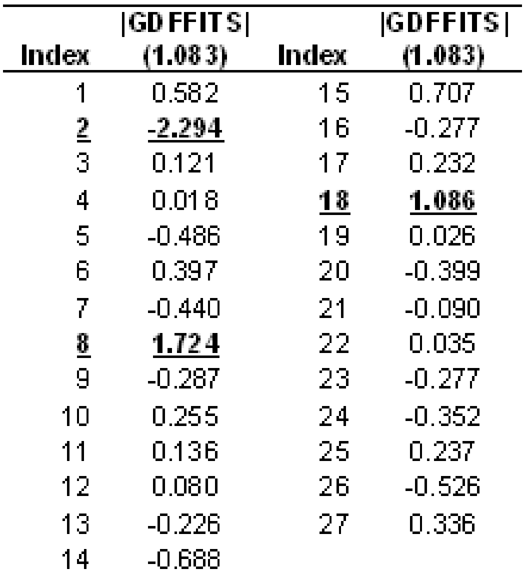


Figure 3: Scatter plot standardized residual versus fitted values

Table 2: Generalized DFFITS diagnostics for Griliches and Lichtenberg (1984) data



Index	|GDFFITS| (1.083)	Index	|GDFFITS| (1.083)
1	0.582	15	0.707
2	**-2.294**	16	-0.277
3	0.121	17	0.232
4	0.018	18	**1.086**
5	-0.486	19	0.026
6	0.397	20	-0.399
7	-0.440	21	-0.090
8	**1.724**	22	0.035
9	-0.287	23	-0.277
10	0.255	24	-0.352
11	0.136	25	0.237
12	0.080	26	-0.526
13	-0.226	27	0.336
14	-0.688		

Table 3: Regression results: with and without unusual observations

Observations deleted	Regression equation	R^2 value
None	*TFPG = -0.579 + 0.346 PRIV +0.0103 FED* *St.Dev* (0.295) (0.086) (0.023) *t-value* (-1.9628) (4.0478) (0.4420) *P-value* (0.0614) (0.0005) (0.662)	0.4426
2, 14, 15, 18	*TFPG = -0.6195 + 0.259 PRIV +0.179 FED* *St.Dev* (0.242) (0.089) (0.073) *t-value* (-2.558) (2.9068) (2.4694) *P-value* (0.0188) (0.0087) (0.0227)	0.5441
2, 8, 18	*TFPG = -0.2794 + 0.058 PRIV +0.249 FED* *St.Dev* (0.2896) (0.1206) (0.088) *t-value* (-0.9649) (0.4766) (2.8275) *P-value* (0.3456) (0.6385) (0.0101)	0.3682

Table 4: Sequential sum of squares

Sum of squares	**Observations deleted**		
Source	None	2,14,15,18	2, 8,18
Regression	19.211	13.366	9.289
(PRIV+FED)	(19.014+0.197)	(9.951+3.415)	(3.222+6.067)
Error	24.196	11.199	15.936
Total	43.407	24.565	25.225
F and P- value	9.527, 0.0009	11.94, 0.0003878	6.12, 0.00805

V. CONCLUSION

Single case deletion diagnostics are failed to identify multiple outliers. Group deletion diagnostic measure GDFFITS is fruitful for identifying multiple outliers (unusual observations) properly. Practitioners have to take extra care about multiple outliers in model building process. Any conclusion drawn from the model in presence of multiple outliers should be treated with a maximum care by using appropriate group deletion diagnostics measure.

REFERENCES

[1] A. C. Atkinson: “Two graphical displays for outlying and influential observations in regression” *Biometrika*, 68, pp.13 – 20. (1981)

[2] A. C. Atkinson: “Fast very robust methods for the detection of multiple outliers” *Journal of the American Statistical Association,* 89, pp. 1329 – 1339. (1994)

[3] A. C. Atkinson, and M. Riani: *Robust Diagnostic Regression Analysis*, New York, Springer. (2004)

[4] E. Acuna, and C. Rodriguez: A meta analysis study of outlier detection methods in classification, *Technical Paper*, Department of Mathematics, University of Puerto Rico at Mayaguez, available at *academic.uprm.edu/~eacuna/paperout.pdf*, *Proceedings IPSI* 2004, (2004)

[5] V. Barnett, and T. B. Lewis: *Outliers in Statistical data*, New York: Wiley. (1995)

[6] D. A. Belsley, E Kuh, and R. E. Welsch: *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York: Wiley. (1980)

[7] H. Caussinus, and A. Roiz: “Interesting projections of multidimensional data by means of generalized component analysis”, in *Computational Statistics,* 90, pp. 121-126. (1990)

[8] S. Chatterjee, and A. S. Hadi: “Influential observations, high leverage points, and outliers in regression”, *Statistical Science*, 1 (3), 379-416. (1986)

[9] S. Chatterjee, and A. S. Hadi: *Sensitivity Analysis in Linear Regression*, New York: Wiley and Sons. (1988)

[10] S. Chatterjee, and A. S. Hadi: *Regression Analysis by Example,* New York: Wiley. (2006)

[11] R. D. Cook: “Detection of influential observations in linear regression”,

Technometrics, 19, pp. 15-18. (1977)

[12] C. Daniel, and F. S. Wood: *Fitting Equations to Data,* New York: Wiley. (1971)

[13] P. H. Franses, and G. Biessen: "Model Adequacy and Influential Observations", *Economics Letters*, 38, pp. 133-137. (1992)

[14] Fawcett, T., and F. Provost: Adaptive fraud detection, *Data Mining and Knowledge Discovery*, 1, 3, pp. 291-316. (1997)

[15] Z. Griliches, and F. Lichtenberg: R & D and productivity at the industry level, in *Z. Griliches ed., R & D, Patents and Productivity*, Chicago: University of Chicago Press. (1984)

[16] A. S. Hadi: "Identifying multiple outliers in multivariate data" *Journal of the Royal Statistical Society*, B, 54, pp. 761-771. (1992)

[17] A.S. Hadi, and J.S. Simonoff: "Procedure for the identification of outliers in linear models", *Journal of the American Statistical Association*, 88, pp. 1264 – 1272. (1993)

[18] F. R. Hampel: "A general qualitative definition of robustness", Annals *of Mathematical Statistics*, 42, pp. 1887-1896. (1971)

[19] F. R. Hampel: "The influence curve and its role in robust estimation", *Journal of the American Statistical Association*, 69, pp. 382-393. (1974)

[20] D. C. Hoaglin, and R. E. Welsch: "The hat matrix in regression and ANOVA", *American Statistician,* 32, pp. 17-22. (1980)

[21] D. Hawkins: *Identification of Outliers*, Chapman and Hall: London. (1980)

[22] S. Hawkins, H. X. He, G. J. Williams, and R. A. Baxter: "Outlier detection using replicator neural networks", Proceedings of the *Fifth International Conference on Data Warehousing and Knowledge Discovery,* Aixen Province, France. (2002)

[23] A. H. M. R. Imon: "Identifying multiple influential observations in linear regression", *Journal of Applied Statistics*, 32, 73-90. (2005)

[24] A.H. M. R. Imon, and M. Ali: "Simultaneous identification of multiple outliers and high-leverage points in linear regression", *Journal of Korean Data and Information Science Society,* 16, 2, pp. 429-444. (2005)

[25] R. A. Johnson, and D. W. Wichern: *Applied Multivariate Statistical Analysis*, India: Pearson Education. (2002)

[26] L. Kaufman, and P. J. Rousseeuw: *Finding groups in Data: An Introduction to Cluster Analysis*, New York: Wiley. (1990)

[27] E. Knorr, and R. Ng: A unified approach for mining outliers, *Proceedings of Knowledge Discovery KDD*, pp. 219-222. (1997)

[28] E. Knorr, and R. Ng: Algorithms for mining distance-based outliers in large dataset, *Proceedings of 24th Intl. Conference Very Large Databases*, pp. 392-403. (1998)

[29] E. Knorr, R. Ng, and V. Tucakov: "Distance based outliers: algorithms and applications", *VLDB Journal: Very Large Data Bases*, 8(3-4): pp. 237-253. (2000)

[30] E. Knorr, R. Ng, and R. H. Zamar: Robust space transformations for distance based operations, *Proceedings of The 7th International Conference on Knowledge Discovery and Data Mining*, pp. 126-135, San Francisco, CA. (2001)

[31] H. Liu, S. Shah, and W. Jiang: "On-line outlier detection and data cleaning", *Computers and Chemical Engineering*, 28, pp. 1635-1647. (2004)

[32] S. Papadimitriou, H. Kitawaga, P. G. Gibbons, and C. Faloutsos: LOCI: Fast outlier detection using the local correlation integral, *Intl. Research Laboratory Technical report*, No. IPR=TR-02-09. (2002)

[33] K. L. Penny, and I. T. Jolliffe: "A comparison of multivariate outlier detection methods for clinical laboratory safety data", *The Statistician*, 50, 3, pp. 295-308. (2001)

[34] R. L. Plackett: "Studies in the history of probability and statistics XXIX: The discovery of the method of least squares", *Biometrika*, 59, pp. 239-251. (1972)

[35] P. J. Rousseeuw, and A. M. Leroy: *Robust Regression and Outlier Detection*, New York: Wiley and Sons. (1987)

[36] P. J. Rousseeuw, and J. Wagner: "Robust regression with a distributed intercept using least median of squares", *Computational Statistics and Data Analysis*, 17, pp. 65-76. (1994)

[37] S. Shekhar, C. T. Lu, and P. Zhang: Detecting graph based spatial outlier: Algorithms and Applications, *Proceedings of the 7th ACM-SIGKDD Int'l Conference on Knowledge Discovery and Data Mining*, SF, CA. (2001)

[38] S. Shekhar, C. T. Lu, and P. Zhang: "Detecting graph based spatial outlier", *Intelligent Data*

Analysis: An International Journal, 6(5), pp.451-468. (2002)

[39] R. G. Staudte, and S. J. Sheather: *Robust Estimation and Testing*, New York: Wiley. (1990)

[40] S. M. Stigler: “Gauss and the invention of least squares”, *Annals of Statistics*, 9, pp. 465-474. (1981)

[41] P. F. Vellman, and R. E. Welsch: “Efficient computing in regression diagnostics”, *American Statistician*, 35, pp. 234-242. (1981)

[42] G. J. Williams, R. A. Baxter, H. X. He, S. Hawkins, and L. Gu: A comparative study of RNN for outlier detection in data mining, *IEEE International Conference on Data Mining (ICDM’02)*, Maebashi City, Japan. (2002)

[43] G. J. Williams, and Z. Huang: Mining the knowledge mine: the hot spots methodology for mining large real world databases, in *Advanced Topics in Artificial Intelligence*, Volume 1342 of lecture notes in *Artificial Intelligence*, pp.340-348, Springer.(1997)

APPENDIX

The Data

Total Factor Productivity Growth and R & D
for 27 U. S. Manufacturing Industries

Index	**Industry**	**1969-73 To 1974-76 TFPG**	**1969-73 FED**	**PRIV**
1	Ordnance	1.4	4.1888	1.4112
2	Missiles and spacecraft	1.3	45.3882	5.2118
3	Food and kindred products	-0.3	0.0000	0.2000
4	Textile mill products	-0.5	0.0000	0.3000
5	Plastics, resins and fibers	0.3	0.0912	5.6088
6	Agricultural chemicals	1.2	0.0341	3.0659
7	Other chemicals	-1.3	0.0264	2.3736
8	Drugs and medicines	2.4	0.1120	6.8880
9	Rubber and misc. plastics	-1.1	0.4116	0.7884
10	Stone, clay and glass	0.2	0.0525	0.6475
11	Ferrous metals and products	-0.2	0.0152	0.3848
12	Nonferrous metals and products	-0.3	0.0190	0.4810
13	Fabricated metal	-0.9	0.6160	0.7840
14	Engines and turbines	-0.9	0.3800	4.6200
15	Farm machinery and equipment	2.3	0.1444	1.7556
16	Construction, mining and materials, machinery and equipment	-1.0	0.1444	1.7556
17	Mattel working machinery and equipment	0.3	0.0836	1.0164
18	Office, computing and accounting machines	3.8	1.5618	9.8382
19	Other machinery	-0.3	0.1230	0.8770
20	Electric transmission and distribution equipment	-0.3	1.0914	4.0086
21	Electrical industrial apparatus	0.0	0.7918	2.9082
22	Other electrical	0.0	0.4494	1.6506
23	Communications and electronics	1.6	6.3800	5.2200
24	Motor vehicles	-1.1	0.0805	2.2195
25	Other transportation	0.3	0.8280	0.6720
26	Aircraft and parts	2.1	9.6276	4.5724
27	Instruments	1.5	1.5456	4.0544