RESEARCH PAPER | Statistics | Volume : 5 | Issue : 5 | May 2015 | ISSN - 2249-555X

Usage of Graphical Displays to Detect Outlying Observations in Linear Regression

KEYWORDS: Influential point, outlier,scalar measures and robust regression.

Mintu Kr.Das
Research Scholar, Department of Statistics, Dibrugarh University Assam, India.

Bipin Gogoi
Professor, Department of Statistics, Dibrugarh University Assam, India

ABSTRACT *Here we try to explain the use of graphical displays for outlying and influential observations in regression analysis. An attempt has been made to classify the observations with the help of some two dimensional plots, which will be comprised of four quadrants.Depending upon the location of the observations, these will be labeled either as good or bad observations. The observations falling in the upper left and lower right quadrants will be subjected for further scrutiny. Also the findings are compared with robust regression. For ensuring proper use of such plots we use some well-referred datasets.*

1. Introduction: The large amount of work on different types of influence measures with respective cutoffs has been enriching the statistical literature through several decades. Being a function of the sample size and number of predictors, there arise some confusing situations to use these cutoffs. Also validity of the cutoff value is subjected to some additional conditions. Now-a-days the major statistical packages avail the values of some outlier diagnostics, which make the use of those diagnostics much wider. But the usage needs some proper guidelines regarding the benchmark value. Regarding the usage of these measures Kuntur et al. (2004) warned that we should try to examine the existence of a gap between the leverage values for most of the cases and the unusually high leverage value(s). This kind of gap can obvious through graphical displays in a lucid manner. As a preliminary diagnosis, the residual analysis isvery useful which reflect those maverick observations that pull the regression line disproportionately. Apart from the unstandardised version, the other types of residuals have their own advantages in detecting anomalous observations.

In this paper we have made an attempt to carve out the best use of thesescalar diagnostic measures and residuals. For this purpose different graphical displays are constructed using three well-referred datasets. Most of the times plotting the measures against the observation label is sufficient. Here we search for suitable pairs of measures whose two-dimensional plot can candidly track down outlying and influential observations.

2. Model Specification:

Let us consider the linear regression model with intercept in matrix form

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad (1)$$

where $\boldsymbol{y_{n\times 1}}$ is the response vector $\boldsymbol{X} = (x_{ij})_{n\times(p+1)}$ with $\boldsymbol{x_{i0}} = 1$ is the design matrix; $\boldsymbol{\beta}_{(p+1)\times 1}$ is the vector of parameters and $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. The ordinary least squares (OLS) estimate of β is given by $\boldsymbol{\beta} = (\boldsymbol{X^TX})^{-1}\boldsymbol{X^Ty}$ and the vector of fitted values as $\hat{\boldsymbol{y}} = \boldsymbol{X\hat{\beta}} = \boldsymbol{Hy}$ where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X^TX})^{-1}\boldsymbol{X^T}$ where is the ***hatmatrix***. The vector of OLSresidualsis $\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$. The different types of measures with different versions of scaled residuals are as follows.

3. Diagnostics based on residual analysis: Since residual analysis deals with studying departures from assumption so it is useful to work with the scaled residuals. In ideal conditions residuals have zero mean and their average variance is approximated by-

$$\sum_i^n (e_i - \bar{e}) \Big/ (n-p-1) = \sum_i^n e_i^2 \Big/ (n-p-1) = SS_{Res}/(n-p-1) = MS_{Res} = \sigma^2$$

The different types of residuals are outlined below:

3.1.Studentized Residuals: Instead of the constant error variance if the exact variance is used then we have Var where is the diagonal element of the projection matrix. Consequently the studentised residuals are dedined by-

$$r_i = e_i / \left[s(1-h_{ii})^{1/2}\right] \text{where} s^2 = MS_{Res} \qquad (2)$$

It is also known as Internally Studentized Residuals.Belsey et al. (1980) recommended a cutoff value of for .

3.2. Externally Studentized Residuals: The Externally Studentized Residuals (also known as R-Student and Jackknife Residuals) are defined by-

$$t_i = e_i / \left[S^2_{(-i)}(1-h_{ii})^{1/2}\right] \qquad (3)$$

where $S^2_{(-i)}$ is defined by-

$$S^2_{(-i)} = \frac{(n-p-1)MS_{Res} - e_i^2/(1-h_{ii})}{(n-p-2)} \qquad (4)$$

The jackknife residuals respond more strongly to the presence of a single outlier than the standardized residuals. It provides an indication of the presence of a bad -value (Atkinson, 1981).

3.3. Adjusted Residuals: Adjusted residuals are nothing but a simpler transformation of the ordinary residuals that possess the same ordering as that of the studentized residuals (Marasinghe, 1985). These are defined by with usual cutoff

$$t_i^a = e_i / \left[(1-h_{ii})^{1/2}\right]; \text{i=1,2,2…n} \qquad (6)$$

4.*ScalarMeasures of Influence Statistics:*

4.1 Cook's Distance: Cook (1997) proposed a measure using the information from the studentized residuals and the variances of residuals and predicted values. Denoting the

Volume : 5 | Issue : 5 | May 2015 | ISSN - 2249-555X

LSE of without the i[th] observation as the statistic is given by,

$$D_i = \frac{\left(\hat{\beta} - \hat{\beta}_{(-i)}\right)' X^T X \left(\hat{\beta} - \hat{\beta}_{(-i)}\right)}{(p+1)s^2} = \frac{r_i^2}{p} \frac{h_{ii}}{(1-h_{ii})} \quad (7)$$

where and. Here combines residual magnitude and the location of the i[th] point in X-space to access influence. examines the changes occurred in estimates for when some cases are deleted. This is the basic idea in influence analysis as introduced by Cook (Su et al., 2012). A rule of thumb of has been suggestedby Kenneth and Robert (1990), Cook and Weisberg (1982). Others have suggested . Also one noticeable thing is that, if all values of are similar, then there is likely no influential point (Vinoth andRajarathinam, 2014).

4.2.COVRATIO$_i$: This measure uses the concept based on the role of the i[th] observation on the precision of estimation and it is defined by

$$COVRATIO_i = \frac{\left|\left(X_{(-i)}^T X_{(-i)}\right)^{-1} S_{(-i)}^2\right|}{\left|(X^T X)^{-1} MS_{Res}\right|} = \frac{\left(S_{(-i)}^2\right)^{(p+1)}}{MS_{Res}^{(p+1)}} \left(\frac{1}{1-h_{ii}}\right) \quad (8)$$

Here *COVRATIO$_i$*>1 indicates that i[th] observation improves the precision of estimation and *COVRATIO$_i$*<1 indicates degradation. Belsley et al. (1980) recommended that the i[th] observation is influential if {(Montgomery et al., 2001).

4.3.DFFITS$_i$: It is a measure of influence introduced by Belsley, Kuh and Welsch (1980), which measures how the deletion of the i[th] observation influence the predicted or fitted values. It is given as

$$DFFITS_i = \left(\hat{y} - \hat{y}_{(-i)}\right) \Big/ \sqrt{S_{(-i)}^2 h_{ii}} = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} t_i \quad (9)$$

The rule of thumb is that an observation for which $|DFFITS_i| > 2\sqrt{(p+1)/n}$ warrants attention. Vellman and Wesch(1981) suggested that values greater than 1 to 2 warrant special attention (Montgomery et al., 2001).

4.4. Hat matrix diagonals: The hat matrix plays an important role for detecting influential observations as it determines the variance and covariance of residuals $\{Var(e_i) = \sigma^2(1-h_{ii})\}$ and that of fitted responses. $\{Var(\hat{y}) = \sigma^2 H\}$ The diagonal is the standardized measure of the distance of the i[th] observation from the centre (or centroid) of the X-space (Seber and Lee, 2003).Generally with $h_{ii} > 2(p+1)/n$an average value $\overline{h_{ii}} = (p+1)/n$ and the data points with may be regarded as outliers in X-space (Su et al., 2012). Larger the value of, smaller is the. So large leverage will lead to the closer fit of $\hat{y}_i$ to y_i. Leverage reflects the position of an observation in the multidimensional space of the carriers or predictors (Velleman and Welsch, 1981). In extreme case $h_{ii} = 1$, the error variance becomes zero. That's why a point which is located extremely in the -space may not always influential unless it has an unusual value in -space. Another guideline is that indicate very high leverage, whereas $0.2 \le h_{ii} \le 0.5$ indicate moderate leverage. Additional evidence of an outlying case is the existence of a gap between the leverage values for most of the cases and the unusually high leverage value(s) (Kunturet al., 2004).

An outlier in predictor space may create a point of high influence $(h_{ii} \cong 1)$. Such a point may have a large effect on the fitted model, but since the standardized residuals all have same variance; a residual plot will not reveal such a point (Atkinson, 1983).

4.5. Potentials: Hadi (1992) found that if there is a high leverage point then the information matrix might have broken down and consequently the observations may not have the appropriate leverage. He introduced a single case deleted measure of leverage known as potentials defined as

$$p_{ii} = x_i^T \left(X_{(-i)}^T X_{(-i)}\right)^{-1} x_i \quad (10)$$

where$X_{(-i)}$ is the data matrix with i[th] row deleted. Its relationship with hat diagonals is given by.$p_{ii} = h_{ii}/(1-h_{ii})$ Those observations with very large potentials are considered as high leverage points (Imon, 2005). Hadi suggested using the cut off as.$Mean(p_{ii}) + c \times st.dev(p_{ii})$ Here C is a constant appropriately selected such as 2 and 3. Also realizing the fact that mean and s.d. are non-robust even for one extreme observation, Hadi suggests using median and median absolute deviation (MAD) respectively.

4.6. Robust Residual analysis: Another approach is the robust regression which is said to be insensitive to such wild points. But interestingly Hurber (1977) illustrated that in case of outliers in the predictor space robust approach may be inefficient.The routine application of robust regression automatically identifies the suspicious points. So whenever a LS analysis is performed, it is advisable to perform a robust fit also. If the residuals of the two procedures are in substantial agreement, then LS should be used, otherwise robust one. And reason for these differences should be identified. Observations in the robust fit should be carefully examined (Montgomery et al. 2001). Again regarding performance two properties of robust regression are to be examined, viz. breakdown and efficiency. And there are different estimators. Here we'll examine the plot of robust residuals versus robust fitted values obtained through M-estimator.

5. Graphical Displays:A common practice is to plot the LS residuals or the studentized or jackknifed residuals against variables such as the fitted responses or one or more explanatory variables to detect outliers. These plots suffer from the fact that the impact of an outlier is not confined to inflating only its own residuals; it may inflate or deflate the residuals of the other observations too, perhaps making itself more or less conspicuous in the detection process (Kianifard and Swallow (1989)).Meloun and Militky (2001) used some types of plots, like, the graph of predicted residuals, the William graph and the Rankit Q-Q plot.

The various diagnostic measures discussed above are differently capable of showing the outlyingness either in the predictor space or in the response. Here we are trying to classify the observations by some 2-D plots, which will be comprised of four quadrants. Two different diagnostic measures will be chosen first, and their ordered values will be plotted along the two-axes. Corresponding to the cut-offs two lines parallel to the horizontal and vertical axes will be drawn at the cutoff points. The observations falling on the lower left quadrant will be good observations. Those lying in the upper right quadrant will be either influential or outlier depending upon the plotted measure. The observations falling in the upper left and lower right quadrants will be subjected for scrutiny. For ensuring proper use of such plots we use some well-referred datasets. The selected datasets are examined in § 7. Although there may be plotted many combination along the two axes, here we

Volume : 5 | Issue : 5 | May 2015 | ISSN - 2249-555X

display only those which are able to met our objective.

6. Application of the Plots to the Well-Referred Data-sets:

6.1.Stack loss Data: This classic dataset originally given by Brownlee (1965) consist of 21 observations with 3 predictors. These observations are from 21 days' operation of a plant for the oxidation of ammonia as a stage in the production of nitric acid. The carrier variablesare:

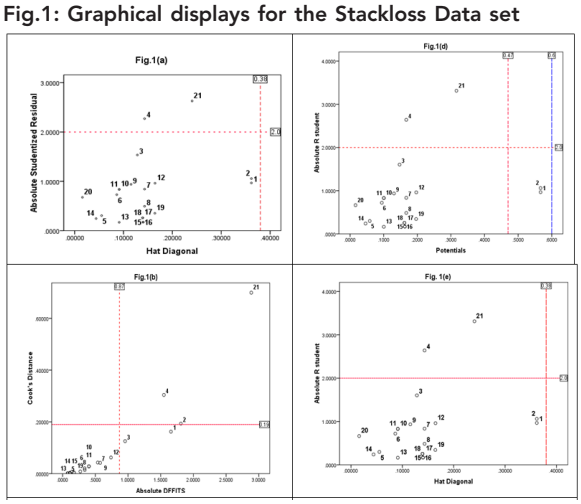
=air flow, =cooling water inlet temperature(°C), =acid concentration(%), and the response variable is Y=stack loss. Here Stack loss is the percentage of the ingoing ammonia that escapes unabsorbed (David et al. 1993). The data were minutely analysed by Daniel and Wood (1980), who used LS method and concluded that observations 1st , 3rd, 4th& 21st were outlier. Andrews (1974) came to a similar conclusion using robust technique. The deletion of 21st observation drops the from 178.83 to 105.6, i.e., approximately a 41 % decrease.As recommended by Daniel and Wood (1971), the data with 21 observations is to be fitted to a model consisting two linear terms and a quadratic term. Marasisinghe(1985) using the multistage procedure and Paul and Fung (1991) using GESR procedure showed that 4th and 21st observations are outliers. So according to our suggested plot the outlying 4 observations should be in the upper right quadrant.

6.2. Longley Data: Longley presented a dataset consisting six economic variables regressed on the total derived employment for the years 1947 to 1962 (Cook, 1977). Here the observations corresponding to the years 1951 and 1962 have the greatest impact on the parameter estimates. After proper investigation it was found that 1951 was the first full year of the Korean conflict. **6.3. Hawkins, Bradu and Kass (HBK) Data.** Hawkins et al. (1984) presented a data set containing 75 observations, with a dependent variable y and three predictor variables x_1,x_2, and x_3. The data set was created so that it has extensive masking. Since the position of the bad points are known, such artificial data help us to get rid of some controversies that are inherent in the real data (Rousseeuw and Leroy 1987). The data set consists of 75 observations in four dimensions. The first 10 observations are bad leverage point, and the next 4 points are good leverage points.The first ten observations were spurious, with low values of h_{ii}. The next four were genuine observations, but with high leverage ("good leverage points"). Thus this data set contains 18.6% outlying observations, 5.3% outliers in X-space, no outliers in y-space, and 13.3% Xy-space outliers.

7. Results and discussion: The three data sets are different from each other in respect of dimension, degree of fitted model and nature of contamination, so same plot is expected to behave differently. From the fig.1(a) we see that the 4th and 21st observations are doubtful with high , fig 1(b) clearly indicate that the points 2nd ,4th and 21st are influential and 1st observation is doubtful with high DFFITS value. Fig.1(c) labeled all the four observations as influential, fig1(d) is same as fig1(a) but it shows 1st and 2nd observation as good leverages. Fig1(e) indicates 4th and 21st as doubtful. The COVRATIO<1 for the observations 21st,4th,3rd also shows that their inclusion degrade the model, but with highest CORATIO for the 1st observation does not improve the model. Fig 1(f) is the residual-fitted values plot from robust regression. It shows 4th, 3rd ,21st as outliers. Thus comparing with the standard results we see that the COVRATIO-Hat diagonal plot is sufficient. Also COVRATIO is partially powerful to detect the true outlier. The plot in fig.1(e), also known as Williams graph (Vinoth andRajarathinam, 2014) is able to label only two observations (21st , 4th) as outlier.Examining our plotted graphs we see that fig2(a) do not flag out outliers, it only shows 10th observation as doubtful with high studentized residual and 5th ,2nd ,16th as high leverage. Fig.2(b) shows 5th and 16th as influential but at the same time unnecessarily high values of DFFITS carries meaningless information. Fig 2(c) confirms 2ndobservation as influential and 5th ,6th as good leverages, while COVRATIO is misleading. Fig2 (d) shows 10th as influential and 16th as a high potential point. Fig2 (e) track 10th point as doubtful while 5th,10th ,16th as good leverages. Fig 2(f) is the residual-fitted values plot fromrobust regression. This plot shows some suspicsion on the points 4th and 10th .Here we observe that the observation corresponding to the year of Korean conflict has been traced with high Cook's distance, high DFFITS but the robust residuals are not in agreement with these. From fig.3(a) we notice that the observations 12th,13th,14th are influential 11th, 7th are doubtful. Fig3(b) confirms 14th as outlier even if we use the cutoff of . Using cutoff we found 11th, 12th, 13th, 14th as influential. The Williams graph i.e., fig3 (e) conveys the same information as 3(a). Fig 3(f) is the residual-fitted values plot from robust regression. For the large data set of HBK the observations 11th ,12th ,13th ,14th are clearly detected but the first ten observations of the data are located in a separate corner.

8. Conclusion: As warned by Hurber (1977) that in case of outliers in the predictor space robust approach may be inefficient, the 21st observation in the stackloss data were detected as outliers but not the 1st and 2nd . COVRATIO can be applied in such situation where the model contains a quadratic term. For the Longley data COVRATIO is not appropriate and the outlier is masked by the others due to the explanation of Huber (1977). Thus it can be inferred that without having the

Fig.1: Graphical displays for the Stackloss Data set



Volume : 5 | Issue : 5 | May 2015 | ISSN - 2249-555X

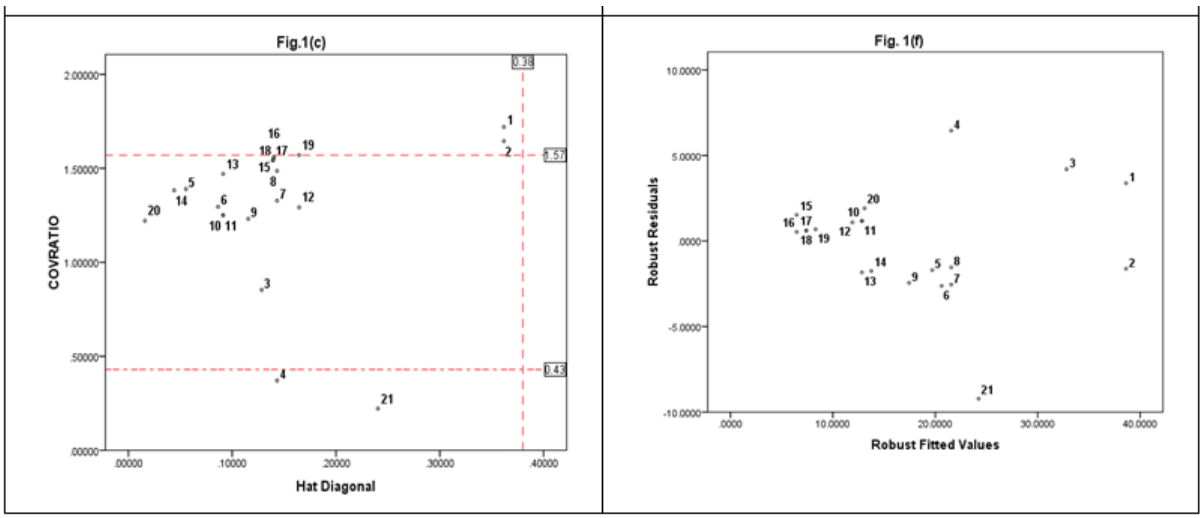


Fig.2: Graphical displays for the Longley Dataset

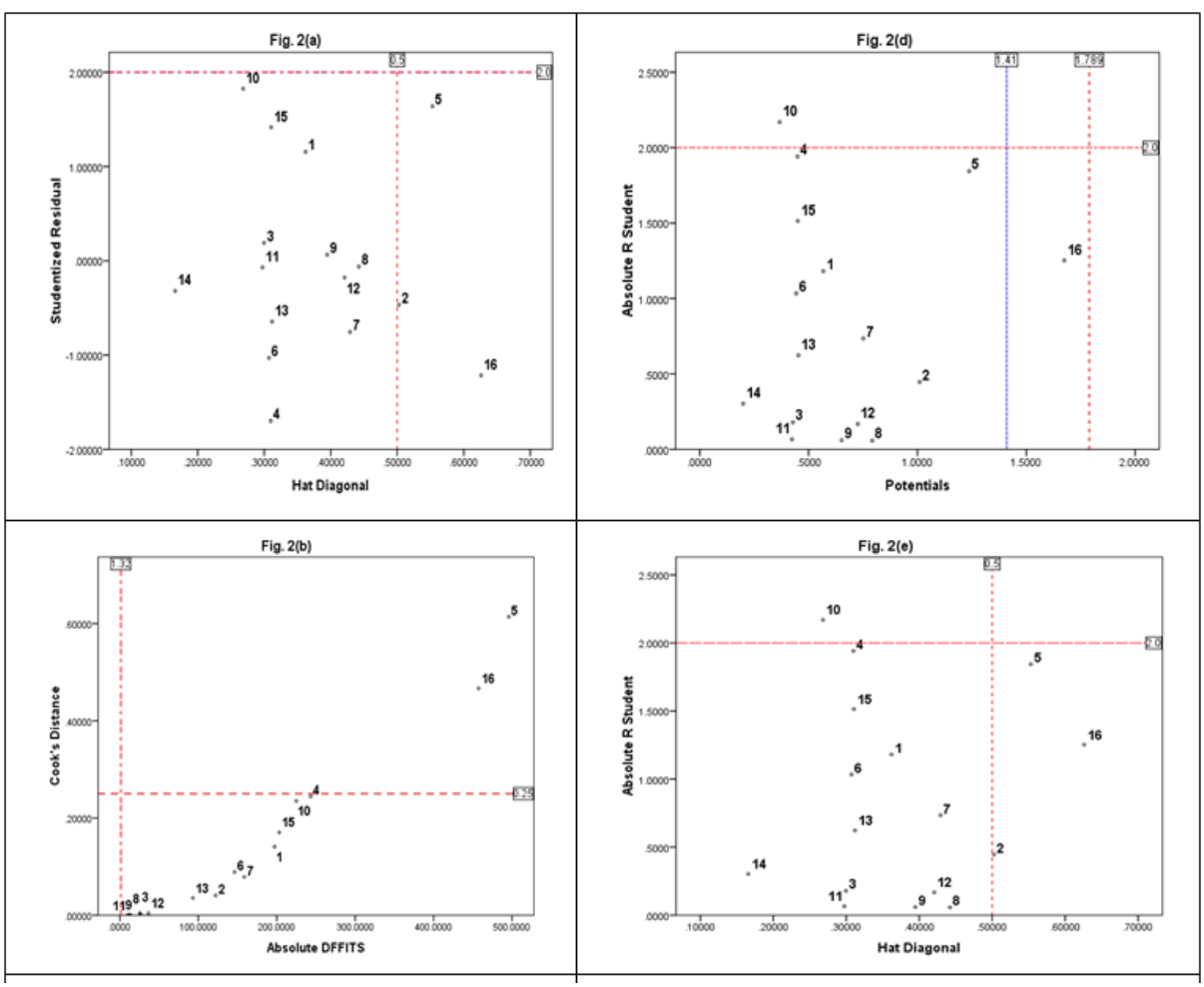


Fig.3. Graphical displays for the Hawkins, Bradu and Kass (HBK) Dataset

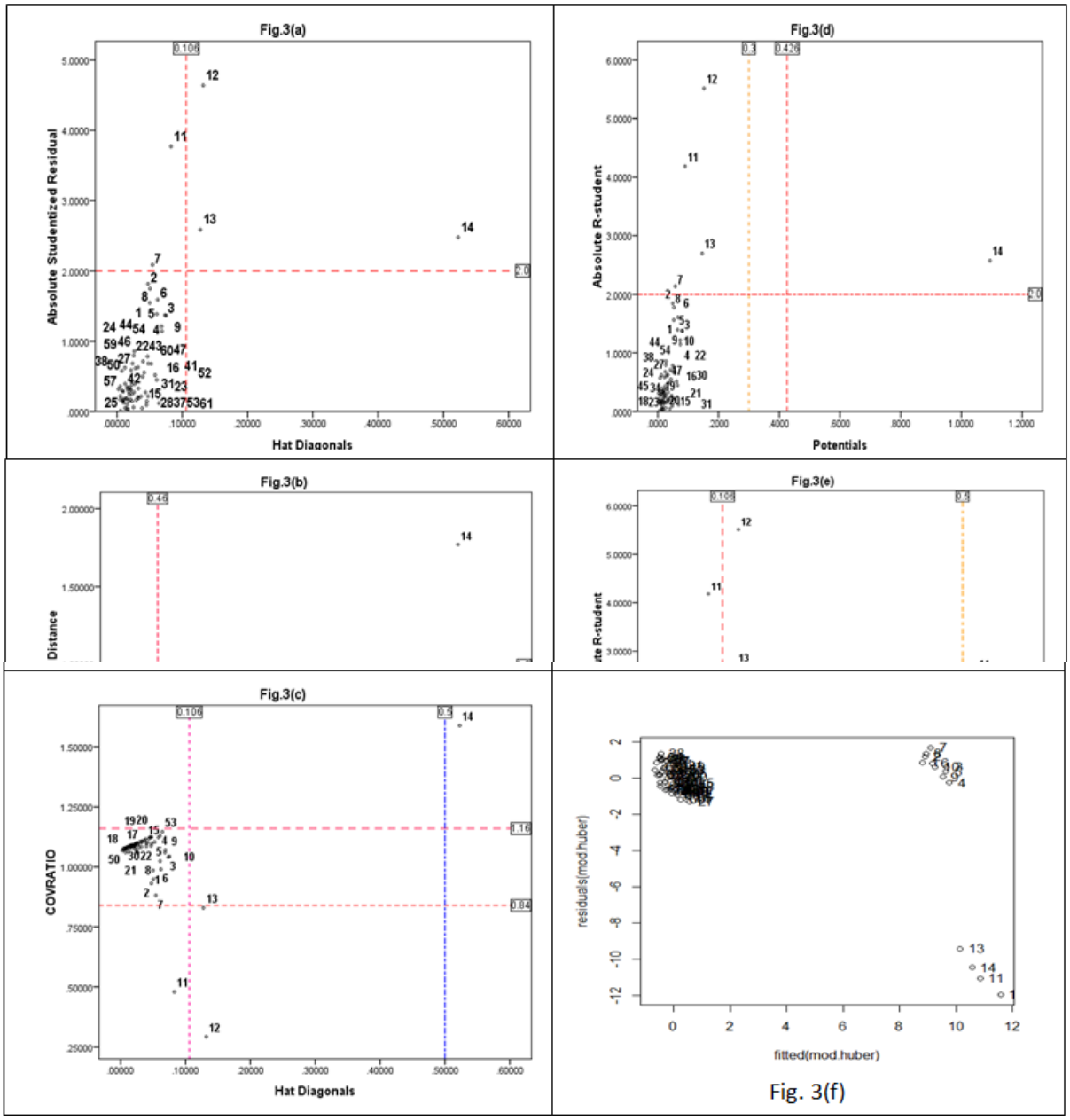


Fig. 3(f)

practical experience one analytics cannot separate out doubtful data as outliers, they can only label them as suspicious. Here though simulation is not performed, still the conclusion is valid. Because the three data sets are different from each other in respect of dimension, degree of fitted model and nature of contamination.The robust fit to the Longley data was not appropriate.For the large artificial HBK dataset only 1 out of 14 outliers are detected using cutoff of and using cutoff of only 6 outliers are detected. But plot of robust residual against fitted values may be useful to separate out the outliers. Being an artificial one, the HBK dataset implies that robust fit should always be performed with least squares method and the plot of residual against the fitted response should be examined. As mentioned at the outset that we aim is to get plot(s) or equivalently measure(s) which is (are) effective in these particular situations. Further validity of the plots will be subject to the pattern of new real life data. Our further work will be based on simulated study for analogous real world problem.

Volume : 5 | Issue : 5 | May 2015 | ISSN - 2249-555X

REFERENCE [1] Atkinson, A.C., (Feb., 1983); "Diagnostic Regression Analysis and Shifted Power Transformations", Technometrics, Vol. 25, No. 1 , pp. 23-33. | [2] Behnken, D. W. and Draper, N. R., (1972). Residuals and their variance patterns. Technometrics, 14, 102-111. | [3] Billor, N. and Kiral, G.(2008). "A Comparison of Multiple Outlier Detection Methods for Regression Data", Communications in Statistics -Simulation and Computation,37:3,521-545. | [4] Cook, R. D. and Weisberg, S. (1994) "An Introduction to Regression Graphics". Wiley series in probability and mathematical statistics. | [5] Draper, N.R., Smith, H.,(2011). "Applied Regression analysis", 3rdedn, Wiley series in probability and statistics. | [6] Gentleman, J.F. and Wilk, M.B. (1975). "Detecting Outliers in a Two-Way Table: I. Statistical Behavior of Residuals". Technometrics, Vol. 17, No. 1, pp. 1-14 | [7] Hand, D. J. , Daly, F., McConway, K., Lunn, D., Ostrowski, E., (Nov-1993), "A Handbook of Small Data Sets". Volume 1, CRC Press, - Mathematics . | [8] Imon, A. H. M. R., (2005). "Identifying multiple influential observations in linear regression" ,Journal of Applied Statistics,32:9, pp.929- 946. | [9] Kianifard F. and Swallow W. H., (1989). "Using Recursive Residuals, Calculated on Adaptively-Ordered Observations, to Identify Outliers in Linear Regression". Biometrics, 45(2), pp. 571-585. | [10] Kuntur, M.H., Nachtsheim, C.J., Neter, J. (2004); Applied linear regression models, 4thedn, McGrow Hill. | [11] Meloun, M., and Militky, J. (2001). Detection of single influential points in OLS regression model building. AnalyticaChimicaActa 439, 161-191. | [12] Montgomery, D.C., Peck, E.A., Vining, G.G. (2001); Introduction to Linear Regression Analysis,3rdedn, Wiley series in probability and statistics. | [13] Paul, S.R., and Fung, K.Y.(1991). "Generalized Extreme Studentized Residual Multiple-Outlier-Detection Procedure in Linear Regression", Technometrics, Vol. 33, No. 3, pp. 339-348. | [14] Peña, D. A., (2005). "New Statistic for Influence in Linear Regression", Technometrics, Vol. 47, No. 1, pp. 1-12. | [15] Prescott, P., "An Approximate Test for Outliers in Linear Models". Technometrics, 17(1), Feb 1975, pp. 129-132. | [16] Rousseeuw, P. J., and Leroy, A. (1987), Robust Regression and Outlier Detection, New York: Wiley. | [17] Seber, G.A.F., Lee, A. J., (2003). "Linear regression analysis", 2ndedn, Wiley series in probability and statistics. | [18] Su, X., Yan, X., Tsai, C.L., "Linear regression", WIREs Comp Stat, Vol.4, pp.275-294, (2012). | [19] Velleman, P.F., and Welsch, R.E., "Efficient Computing of Regression Diagnostics", The American Statistician,35:4, pp.234-242., (1981). | [20] Vinoth.B, andRajarathinam. A,.(February 2014) "Outlier Detection in Simple Linear Regression Models and Robust Regression– A Case Study on Wheat Production Data", International Journal of Applied Research, Vol. ,Issue 2, pp. 531-536.