Science and Technology RMUTT Journal

Vol.6 No.2 (2016) : 115-138
www.sci.rmutt.ac.th/stj
Online ISSN 2229-1547

AN ALTERNATIVE DISTRIBUTION OF INTERNAL STUDENTIZED RESIDUAL AND IDENTIFICATION OF OUTLIERS THROUGH EVIDENCE PLOTS

G.S. David Sam Jayakumar[1*] and A.Sulthan[1]

[1]Jamal Institute of Management, Tiruchirappalli, India
*E-mail : samjaya77@gmail.com

Abstract

This paper proposed the exact distribution of internal studentized residual which used to evaluate the outliers in X and Y space in linear multiple regression analysis. The authors explored the relationship between the internal studentized residual in terms of two independent t-ratio, F-ratio's and they show the derived density function of the residual in terms of Gauss hyper-geometric function. Moreover, the new form of the distribution is symmetric, first two moments of the distribution are derived and the authors computed the critical points of internal studentized residual at 5% and 1% significance level for different sample sizes and varying number of predictors. Evidence plots were also proposed to evaluate the exact position and location of the outliers. Finally, the numerical example shows the results extracted from the proposed approaches are more scientific, systematic in identifying the outliers in both spaces(X and Y) and its exactness gives more insights than the traditional Weisberg test.

Keywords: Internal studentized residual, X,Y space, outliers, t-ratio, F-ratio, Gauss hyper-geometric function moments, critical points, evidence plots, Weisberg test

Received: April 17, 2016
Revised: September 01, 2016
Accepted: September 02, 2016

1. Introduction and Related work

A studentized residual is the quotient resulting from the division of a residual by an estimate of its standard deviation. Typically the standard deviations of residuals in a sample vary greatly from one data point to another even when the errors all have the same standard deviation, particularly in regression analysis; thus it does not make sense to compare residuals at different data points without first studentizing. It is a form of a Student's t-statistic, with the estimate of error varying between points. This is an important technique in the detection of outliers. It is named in honor of William Sealey Gosset, who wrote under the pseudonym Student, and dividing by an estimate of scale is called studentizing, in analogy with standardizing and normalizing. Studentization, is the adjustment consisting of division of a first-degree statistic derived from a sample, by a sample-based estimate of a population standard deviation. The term is also used for the standardisation of a higher-degree statistic by another statistic of the same degree (Kendall and Stuart [15]). José A. Díaz-García, et al., [14] find the distributions of normalized, standardized and studentized (internally and externally studentized) residuals, assuming normal and elliptical distributions. In addition, they propose an alternative approach to the results published by Ellenberg [10] and Beckman and Trusell [13], The distribution of an arbitrary studentized residual and effects of updating in multiple regression. In least-squares fitting it is important to understand the influence which a data y value will have on each fitted y value. A projection matrix known as the hat matrix contains this information and, together with the studentized residuals, provides a means of identifying exceptional data points (Hoaglin and Welsch, [12]). The studentized residuals, t_i , (i.e. the residual divided by its standard deviation) have been recommended (Behnken and Draper [2], Davies and Hutton [8], Huber [13]) as more appropriate than the standardized residuals (i.e., the residual divided by the square root of the mean square for error) for detecting outliers. Also, approximate critical values for the maximum absolute studentized residual are available (Lund, [19]). Cook [5] has been the first to establish a simple measure, D_i that incorporates information from the X-space and Y-space used for assessing the influential observations in regression models. The problem of outliers or influential data in the multiple or multivariate linear regression setting has been thoroughly discussed with reference to parametric regression models by the pioneers namely Cook [5], Cook and Weisberg [7], Belsley et al. [4] and Chatterjee and Hadi [6] respectively. In non-parametric regression models, diagnostic results are quite rare. Among them, Eubank [11], Silverman [20], Thomas [21], and Kim [16] studied residuals, leverages, and several types of Cook's distance in smoothing splines, and Kim and Kim [17], Kim et.al [18] proposed a type of Cook's distance in kernel density estimation and in local polynomial regression. The phrase 'influence measures' has glimpsed a great surge of research interests. The developments of different measures are investigated to identify the influential observation from the early criteria of Cook's to the present and a definition about influence, which appears most suitable, is given by Belsley et al. [4]. Cook's statistical diagnostic measure is a simple, unifying and general approach for judging the local influence in statistical models. As far as the influence measures concern in the literature, the procedures were designed to detect the influence of observations on a specific regression result. However, Hadi [1], proposed a diagnostic measure called Hadi's influence function to identify the overall potential influence which possesses several desirable properties that many of the frequently used diagnostics do not generally possess such as invariance to location and scale in the response variable, invariance to non-singular transformations of the explanatory variables, it is an additive function of measures of leverage and of residual error, and it is monotonically increasing in the leverage values and in the squared residuals. Recently, Díaz-García and González-Farías [9] modified the classical cook's distance with generalized Mahalanobis distance in the context of multivariate elliptical linear regression models and they also establish the exact distribution for identification of outlier data points. Considering the above reviews, the authors proposed the exact distribution of internal studentized residual which need to exactly identify the Outlying data points in both spaces with the help of the evidence plots and it is discussed in the subsequent sections.

2. Relationship among Internal studentized residual, student's-t and F-ratio

The multiple linear regression model with 'p' regressors and random error is given by

$$Y = X\beta + e \qquad -(1)$$

Where $\underset{(nx1)}{Y}$ is the vector of values of the dependent variable, $\underset{(nx(p+1))}{X}$ is full column rank matrix of predictors, $\underset{((p+1)x1)}{\beta}$ is the vector of beta co-efficients or partial regression co-efficients and $\underset{(nx1)}{e}$ is the vector of error followed normal distribution N $(0, \sigma_e^2 I_n)$ From (1), statisticians concentrate and give importance to the error diagnostics such as outlier detection, identification of leverage points and evaluation of influential observations. Several error diagnostics techniques exist in the literature proposed by statisticians, but studentized residual attracts the statisticians to scrutinize the outliers in the Y-space. Studentization can be done in two ways namely internal studentization and external studentization of the regression residuals. Many authors believe internal studentization of the residual which followed approximate student's t-distribution with $n-p-2$ degrees of freedom and Weisberg [22] provided a monotonic transformation of the internally studentized residual which followed the exact t-distribution. All the works in the literature show that the transformation of residuals to any forms which always helps to evaluate the outliers in Y-space.

Some authors like Cook [5] and Hadi [1] proposed measures to find the influential observations and potential outliers in the X-space as well as in the Y-space. The authors of this paper argue that the internal studentized residual comprised of information about the position of each observation in the X-space and the Y-space and traditionally the general form of the internal studentized residual (r_i) of the i^{th} observation is given as

$$r_i = \frac{\hat{e}_i}{S_e\sqrt{1-h_{ii}}} \quad - (2)$$

Where $\hat{e}_i$ is the estimated i^{th} regression residual, S_e is the unbiased standard deviation of the estimated residuals and (h_{ii}) is the hat values or the diagonal elements of the hat matrix $(H = X(X'X)^{-1}X')$ which involves the set of predictors respectively. Usually, the (r_i) was compared with critical values of student's t-ratio for $n-p-2$ degrees of freedom and if the computed (r_i) exceeds, then the observation is said to be an outlier in the Y-space. The authors' redefined internal studentized residual is the ratio of the two terms namely $\left(\hat{e}_i / S_e\right)$ and $(1-h_{ii})^{1/2}$,where the first term conveys information about the individual observation in the Y-space and latter term visualizes the position of the observations in the X-space. If the $\left(\hat{e}_i / S_e\right)$ is close to 0, then (r_i) approaches zero, then the observation is an inlier in Y-space and if $\left(\hat{e}_i / S_e\right)$ is very large, then (r_i) will also become very large, then the observation is far away and it is said to be an outlier in Y-space. Similarly, we know the hat values lie between $1/n$ and 1 and it is used to evaluate the influential observations in the X-space. If the hat values close to $1/n$, then the residual and error variance of each observation will be larger and the observation will be remote or formally it will be an influential in X-space. Similarly, if it is close to 1, then the residual of the particular observation will be smaller and it is said to be a leverage point. This hat values always played a vital role in reflecting the information about the position of the observation in the X-space. Rewrite (2) in terms of the true standard deviation (σ_e) of the residual as

$$r_i = \frac{\hat{e}_i / \sigma_e}{\left(S_e / \sigma_e\right)\sqrt{1-h_{ii}}} \quad - (3)$$

From (3) if $(n-p-1)S_e^2/\sigma_e^2$ follows chi-square distribution with $n-p-1$ degrees of freedom, $x^2_{(n-p-1)}$ then $\hat{e}_i/\sigma_e$ follows normal distribution with mean 0 and variance 1 and the quantum S_e/σ_e is equal to $\sqrt{\chi^2_{(n-p-1)}/(n-p-1)}$. Therefore (3) will be further modified as

$$r_i = \frac{z_i / \sqrt{\chi^2_{(n-p-1)}/(n-p-1)}}{\sqrt{1-h_{ii}}} \quad - (4)$$

Where $z_i = \hat{e}_i / \sigma_e$

From (4), we know ratio $\left(z_i / \sqrt{\chi^2_{(n-p-1)}/(n-p-1)}\right)$ followed student's t-distribution with $n-p-1$ degrees of freedom, $z_i, (n-p-1)$ and by definition it is equal to the student's-t ratio, then (4) can be written as

$$r_i = \frac{t_{i(n-p-1)}}{\sqrt{1-h_{ii}}} \quad - (5)$$

From (5), it is the most important form identified by the authors. The t-ratio explained the position of the observation in the Y-space and the hat values represents the status of the observations in the X-space. The explanation of the above discussion is given detailed in section 4. Traditionally, statisticians always concentrating on the distribution of the dependent variable and the residual and they completely ignore the distribution of the predictor set. Over the past decades, studentized residual are frequently used to identify the outliers in Y-space and it is the first attempt made by the authors and they proved that the internal studentized residual comprised of the information about the both spaces (X-space and Y-space). If we incorporate and consider the distribution of dependent, predictor set simultaneously, we can use the internal studentized residual as a measure to evaluate and identify the outliers in both the spaces. In order to prove this scientifically, we proposed the exact distribution of the internal studentized residual by utilizing the relationship among the internally studentized residual (r_i), t-ratio and hat elements (h_{ii}). From (5), the terms t_i and h_{ii} are independent, because the computation of (t_i) involves the error term $e_i \sim N(0,\sigma_e^2)$ and h_{ii} values involves the set of predictors $(H = X(X'X)^{-1}X')$. Therefore, from the property of least squares $E(eX)=0$, so t_i and h_{ii} are also uncorrelated and independent. Using this assumption, we identify the distribution of h_{ii} based on the relationship proposed by Belsley et al [4] and they showed when the set of predictors in a linear regression model followed the multivariate normal distribution with (μ_X, Σ_X), then

$$\frac{(n-p)\left(h_{ii}-(1/n)\right)}{(p-1)(1-h_{ii})} \sim F_{(p-1,n-p)} \qquad - (6)$$

From (6) it follows the F-distribution with $(p-1,n-p)$ degrees of freedom and it can be written in an alternative form as

$$h_{ii} = \frac{\left(((p-1)/(n-p))F_{i(p-1,n-p)}\right)+1/n}{1+((p-1)/(n-p))F_{i(p-1,n-p)}} \qquad - (7)$$

In order to derive the exact distribution of (r_i), substitute (7) in (5), we get the r_i in terms of the independent t-ratio and F-ratio which followed $n-p-1$ and $(p-1,n-p)$ degrees of freedom respectively and the relationship is given as

$$r_i = \frac{t_{i(n-p-1)}}{\sqrt{\frac{n-1}{n}\left(1/\left(1+\frac{p-1}{n-p}F_{i(p-1,n-p)}\right)\right)}} \qquad - (8)$$

From (8), it can be further simplified and (r_i) is expressed in terms of independent t-ratio and beta variable θ_i of the first kind by using the following facts

$$\frac{1}{1+\frac{p-1}{n-p}F_{i(p-1,n-p)}} = \theta_i \sim \beta_1\left(\frac{n-p}{2},\frac{p-1}{2}\right) \qquad - (9)$$

Then, without loss of generality (9) can be written as

$$r_i = \frac{t_{i(n-p-1)}}{\sqrt{((n-1)/n)\theta_i}} \qquad - (10)$$

Based on the identified relationship from (10), the authors derived the exact distribution of the internal studentized residual and it is discussed in the next section.

3. Exact Distribution of Internal Studentized Residual

Using the technique of two-dimensional Jacobian of transformation, J, the joint probability density function of the t-ratio and the beta variable of Kind-1 namely t_i, θ_i were transformed into density function of r_i and it is given as

$$f\left(r_i, u_i\right) = f\left(t_i, \theta_i\right)|J| \qquad -(11)$$

From (11), we know t_i and θ_i are independent then rewrite (11) as

$$f\left(r_i, u_i\right) = f\left(t_i\right) f\left(\theta_i\right)|J| \qquad -(12)$$

Using the change of variable technique, substitute $\theta_i = u_i$ in (12) we get

$$r_i = \frac{t_{i(n-p-1)}}{\sqrt{((n-1)/n)u_i}} \qquad -(13)$$

Then partially differentiate (13) and compute the Jacobian determinant in (12) as

$$f\left(r_i, u_i\right) = f\left(t_i\right) f\left(\theta_i\right)\left|\frac{\partial\left(t_i, \theta_i\right)}{\partial\left(r_i, u_i\right)}\right| \qquad -(14)$$

$$f\left(r_i, u_i\right) = f\left(t_i\right) f\left(\theta_i\right)\begin{vmatrix} \frac{\partial t_i}{\partial r_i} & \frac{\partial t_i}{\partial u_i} \\ \frac{\partial \theta_i}{\partial r_i} & \frac{\partial \theta_i}{\partial u_i} \end{vmatrix} \qquad -(15)$$

From (15), we know t_i and θ_i are independent, then the density function of the joint distribution of t_i and θ_i is given as

$$f\left(t_i, \theta_i\right) = \frac{1}{\sqrt{n-p-1}\,B\left(\frac{1}{2}, \frac{n-p-1}{2}\right)}\left(1+\frac{t_i^2}{n-p-1}\right)^{-\left(\frac{n-p-1}{2}+\frac{1}{2}\right)}$$
$$\times \frac{1}{B\left(\frac{n-p}{2}, \frac{p-1}{2}\right)}\theta_i^{\frac{n-p}{2}-1}\left(1-\theta_i\right)^{\frac{p-1}{2}-1} \qquad -(16)$$

where $-\infty < t_i < +\infty$, $0 \le \theta_i \le 1$, $n, p > 0$, $B(a,b) =$

and

$$\left|\frac{\partial\left(t_i, \theta_i\right)}{\partial\left(r_i, u_i\right)}\right| = \begin{vmatrix} \sqrt{\frac{n-1}{n}u_i} & \frac{r_i}{2}\sqrt{\frac{n-1}{nu_i}} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{n-1}{n}u_i} \qquad -(17)$$

Then substitute (16) and (17) in (15) in terms of the substitution of u_i, we get the joint distribution of r_i and u_i as

$$f\left(r_i, u_i\right) = \frac{1}{\sqrt{n-p-1}\,B\left(\frac{1}{2}, \frac{n-p-1}{2}\right)}$$
$$\times\left(1+\frac{1}{n-p-1}\left(r_i\sqrt{\frac{n-1}{n}u_i}\right)^2\right)^{-\left(\frac{n-p-1}{2}+\frac{1}{2}\right)}$$
$$\times\frac{1}{B\left(\frac{n-p}{2}, \frac{p-1}{2}\right)}u_i^{\frac{n-p}{2}-1}\left(1-u_i\right)^{\frac{p-1}{2}-1}\times\sqrt{\frac{n-1}{n}u_i}\times|J| \qquad -(18)$$

where $-\infty < r_i < +\infty$, $0 \le u_i \le 1$, $n, p > 0$ and $|J| = \sqrt{\frac{n-1}{n}u_i}$. Rearrange (18) and integrate with respect to u_i, we get the marginal distribution of r_i *as* $f\left(r_i\right) = \lambda\left(p, n\right)$

$$\int_0^1 u_i^{\frac{n-p+1}{2}-1}\left(1-u_i\right)^{\frac{p-1}{2}-1}\left(1+\frac{r_i^2}{n-p-1}\left(\frac{n-1}{n}\right)u_i\right)^{-\left(\frac{n-p-1}{2}+\frac{1}{2}\right)} du_i \qquad -(19)$$

Where $-\infty < r_i < +\infty$, $n, p > 0$, $n > p$ and

$$\lambda(p,n)=\sqrt{\frac{n-1}{n(n-p-1)}}\times\left(B\left(\frac{1}{2},\frac{n-p-1}{2}\right)B\left(\frac{n-p}{2},\frac{p-1}{2}\right)\right)^{-1}$$

We know, from (17) and (18)

$$\int_0^1 u_i^{\frac{n-p+1}{2}-1}(1-u_i)^{\frac{p-1}{2}-1}\left(1+\frac{r_i^2}{n-p-1}\left(\frac{n-1}{n}\right)u_i\right)^{-\left(\frac{n-p-1}{2}+\frac{1}{2}\right)}du_i$$
$$=\Omega_1(r_i;p,n)+\Omega_2(r_i;p,n)$$

- (20)

Then substitute (20) in (19) and arrange the terms, we get the density function of r_i as

$$f\left(r_i;p,n\right)=\lambda(p,n)\left(\Omega_1\left(r_i;p,n\right)+\Omega_2\left(r_i;p,n\right)\right)$$

- (21)

where $-\infty<r_i<+\infty$, $n,p>0$, $n>p$ and

$$\lambda(p,n)=\sqrt{\frac{n-1}{n(n-p-1)}}\left(B\left(\frac{1}{2},\frac{n-p-1}{2}\right)B\left(\frac{n-p}{2},\frac{p-1}{2}\right)\right)^{-1}$$

$$\Omega_1(r_i;p,n)=\frac{-2\pi}{B\left(\frac{p}{2},-\frac{1}{2}\right)}\left(\left(\frac{n-1}{n}\right)\frac{r_i^2}{n-p-1}\right)^{-\left(\frac{n-p}{2}\right)}$$
$$\times{}_2F_1\left(\frac{n-p}{2},\frac{-p+2}{2};-\frac{1}{2};-\frac{n(n-p-1)}{(n-1)}\left(\frac{1}{r_i^2}\right)\right)$$

$$\Omega_2(r_i;p,n)=\frac{-2\pi}{B\left(\frac{n-p}{2},\frac{1}{2}\right)}\left(\left(\frac{n-1}{n}\right)\frac{r_i^2}{n-p-1}\right)^{-\left(\frac{n-p+1}{2}\right)}$$
$$\times{}_2F_1\left(\frac{n-p+1}{2},\frac{-p+3}{2};\frac{3}{2};-\frac{n(n-p-1)}{(n-1)}\left(\frac{1}{r_i^2}\right)\right)$$

From (21), it is the density function of internal studentized residual (r_i) which is symmetric in nature and it involves two auxiliary functions namely $\Omega_1\left(r_i;p,n\right).\Omega_2\left(r_i;p,n\right)$ which exhibits in terms of Gauss hyper-geometric function $\left({}_2F_1\right)$ and the normalizing constant $\lambda(p,n)$ in terms of beta functions namely $B(\frac{1}{2},\frac{n-p-1}{2})$, $B(\frac{n-p}{2},\frac{p-1}{2})$ with two shape parameters (p,n), where n is the sample size and p is the number of predictors used in a multiple linear regression model. In order to know the location and dispersion of internal studentized residual, the authors derived the first two moments in terms of mean, variance from (8) and it is shown as follows. Using (8), take expectation and substitute the moments of independent t-ratio and F-ratio, we get the first moment of r_i as

$$E(r_i)=E(t_i)E\left(\sqrt{\frac{n}{n-1}\left(1+\frac{p-1}{n-p}F_i\right)}\right) \quad - (22)$$

$$E(r_i)=0 \quad - (23)$$

From (23), if the moment of the internal studentized residual is zero ($E(r_i)=0$), then the second moment is equal to its variance. Hence, the square (8) on both sides, then take expectation and substitute the appropriate second order moments of independent t-ratio and F-ratio, we get the variance of the r_i is given as

$$V(r_i)=E\left(r_i^2\right)=\frac{n}{n-1}E\left(t_i^2\right)\left(1+\frac{p-1}{n-p}E(F_i)\right)$$

$$V(r_i)=E\left(r_i^2\right)=\frac{n(n-3)(n-p-1)}{(n-1)(n-p-2)(n-p-3)}$$

- (24)

As a proposed approach, the authors adopted the test of significance approach of evaluating and identifying the outliers in a sample. The approach is to derive the critical points of the internal studentized residual by using the following relationship from (8) and it is given as

$$r_{i(p,n)}(\alpha)=t_{i(n-p-1)}(\alpha)\sqrt{\frac{n}{n-1}\left(1+\frac{p-1}{n-p}F_{i(p-1,n-p)}(\alpha)\right)}$$

- (25)

From (25), for different combination of values of (p,n) and the significance probability $p_r\left(|r_i|>r_{i(p,n)}(\alpha)\right)=\alpha$, we computed the critical points of internal studentized residual. If the sample size is very large $(n \to \infty)$, then the limiting distribution of r_i followed standard normal distribution with mean 0 and variance 1. By using the critical points, we can test the significance of the outliers in a multiple linear regression model. The following tables 1,2 exhibits the significant percentage points of the distribution of internal studentized residual for varying sample size*(n)* and the number of predictors *(p)* at 5% and 1% significance (α) calculated by using the software IBM SPSS 22 based on the relationship equation from (25) .

4. Heuristic evidences and Evidence plots

The proposed exact distribution helps to evaluate the outliers in a multiple regression model, but it fails to reveals the exact position of the observations in the X-space or Y-space or in both. For this, the author visualizes heuristic evidences and by using the evidence plots along with test of significance approach, we can identify groups in the observations. We recommend to use a two-dimensional scatter plot to find the evidences of the observation being exact outliers in both spaces. Consider the absolute value of the quantum $\left|\hat{e}_i\right| / S_e$ is the representation of the Y-space and $(1-h_{ii})^{1/2}$ is the good proxy of the X-space. Plot the values $\left(\left|\hat{e}_i\right| / s_e\right)$ and $(1-h_{ii})^{1/2}$ in the two-dimensional space and classify the plots of observations into groups in the spaces by using a predetermined cut-off. From (10) if $U_i = \left(\hat{e}_i / s_e\right)$ follows the t-distribution with n-p-1 degrees of freedom and $V_i = (1-h_{ii})^{1/2}$ follows the beta distribution with shape parameters *p* and *n*, then the following steps need to classify the observations by using the evidence plots and it is given as follows.

Step1: Run a multiple linear regression analysis and compute the internal studentized residuals from (2).

Step2: Use the proposed distribution of internal studentized residual and evaluate the outliers at 5% and 1% level of significance.

Step3: Compute the estimated unstandardized residuals $(\hat{e}_i)$, unbiased standard deviation of the residuals (S_e) and the hat values (h_{ii}).

Step4: Calculate the new variables $U_i = \left(\hat{e}_i / s_e\right)$ and $V_i = (1-h_{ii})^{1/2}$, then plot $|U_i|$ and V_i in a two-dimensional scatter plot by these variables are proxies of the Y-space and X-space respectively.

Step5: We know U_i and V_i are independent, then determine the cut-off separately for both the variables at 5% and 1% level of significance ($U_i \sim t_{(n-p-1)}(\alpha)$ and $V_{i(p,n)}(\alpha)$ follows beta distribution-refer Table 3 and 4) where α is the upper alpha point of the distributions respectively.

Step6: Plot the Cut-offs $(U(\alpha), V(\alpha))$ in the evidence plots, then the two dimensional scatterplot was segregated into 4 quadrants. The following evidence plot shows the classification of quadrants and it's inference.

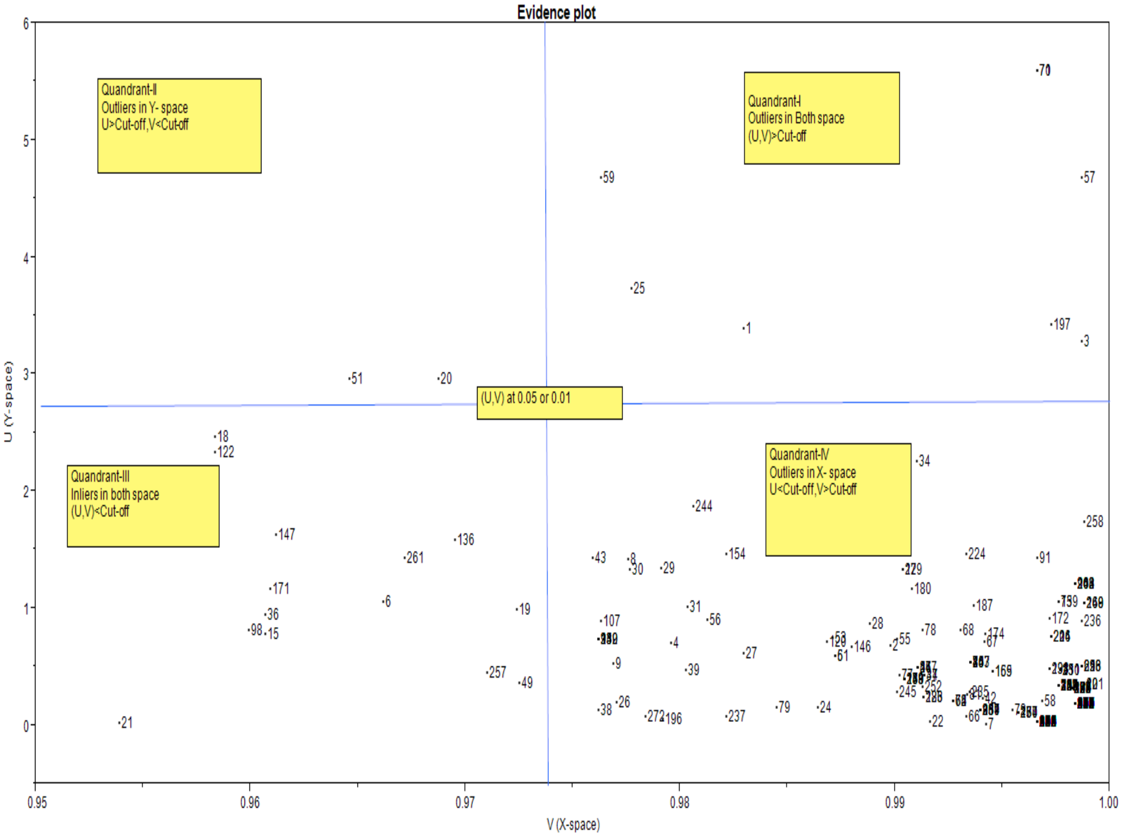


Figure 1 – Evidence plot

Quadrant-I

Any observation plotted in the 1st quadrant will be the outlier in both the spaces,because those observations are far away from the both the axes and remote in the spaces. Moreover,the result of the test of significance shows,the co-ordinates in this quandrant is independently more than the predetermined cut-off at 5% or 1% significance level.

Quadrant-II

The observations plotted in this quandrant will be the outlier in Y-space and the result of the test of significance shows, the U-co-ordinate in this quadrant are independently more than the predetermined cut-off at 5% or 1% significance level and the V-co-ordinate are statistically insignificant.

Quadrant-III

The observations plotted in this quandrant are inliers in Y-space and at the same time it is an leverage point in the X-space. Moreover,the result of the test of significance shows,the co-ordinates in this quadrant is independently less than the predetermined cut-off at 5% or 1% significance level and the authors believe the observations in this quadrant may have a chance to be influential.

Quadrant-IV

The observations plotted in this quandrant will be the influential (leverage or outlier) in X-space and the result of the test of significance shows,the U-co-ordinates in this quadrant are statistically insignificant and the V-co-ordinates are statistically significant at 5% or 1% level.

Step7: Now consider the plotted observations in the classified quadrants are groups and verify the significance of co-ordinates in each group by using $(U_i \sim t_{(n-p-1)}(\alpha)$ and $V_{i(p,n)}(\alpha)$ follows the beta distribution.

Step8: If the U and V co-ordinates are significant at 5% or 1% level in 1st Quadrant, then the observation is said to be outliers in both spaces. If U-coordinates are significant in the 2nd Quadrant, then the observation is treated as an outlier in Y-space. Similarly, if both U-V coordinates are insignificant in the 3rd Quadrant, then the observation is considered as inliers in both the spaces and it may have a chance to be an influential or leverage point. As far as 4th Quadrant is a concern, the V-Coordinates are statistically significant, then the observations are treated as influential in the X-space.

Table 1 Significant two-tail percentage points of Internal studentized residual at $p_r\left(|r_i| > r_{i(p,n)}(0.05)\right)=0.05$

n	*p*									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
3	15.56186	-	-	-	-	-	-	-	-	-
4	4.96828	46.98756	-	-	-	-	-	-	-	-
5	3.55808	10.06304	63.53102	-	-	-	-	-	-	-
6	3.04144	5.96451	12.79392	75.91436	-	-	-	-	-	-
7	2.77655	4.56935	7.26929	14.89824	86.24870	-	-	-	-	-
8	2.61586	3.88431	5.40370	8.29430	16.68392	95.31810	-	-	-	-
9	2.50806	3.48083	4.49206	6.06790	9.17503	18.26740	103.50854	-	-	-
10	2.43074	3.21595	3.95693	4.98051	6.64399	9.96267	19.70747	111.04207	-	-
11	2.37257	3.02909	3.60651	4.34230	5.40721	7.16267	10.68334	21.03874	118.05997	-
12	2.32722	2.89034	3.35980	3.92438	4.68084	5.79347	7.63966	11.35260	22.28363	124.65867
13	2.29086	2.78333	3.17692	3.63014	4.20489	4.98866	6.15019	8.08437	11.98066	23.45767
14	2.26106	2.69830	3.03606	3.41202	3.86958	4.46087	5.27394	6.48389	8.50302	12.57466
15	2.23619	2.62914	2.92427	3.24399	3.62088	4.08875	4.69881	5.54158	6.79890	8.90000
16	2.21513	2.57179	2.83344	3.11065	3.42920	3.81255	4.29300	4.92260	5.79484	7.09829
17	2.19705	2.52348	2.75818	3.00229	3.27702	3.59955	3.99157	4.48551	5.13482	6.03604
18	2.18136	2.48222	2.69483	2.91251	3.15330	3.43033	3.75895	4.16062	4.66841	5.33729
19	2.16763	2.44658	2.64077	2.83693	3.05077	3.29270	3.57404	3.90973	4.32149	4.84319
20	2.15550	2.41549	2.59409	2.77243	2.96441	3.17857	3.42355	3.71017	4.05342	4.47544
21	2.14471	2.38812	2.55340	2.71674	2.89070	3.08242	3.29870	3.54767	3.84007	4.19110
22	2.13505	2.36386	2.51760	2.66819	2.82705	3.00031	3.19347	3.41279	3.66625	3.96469
23	2.12635	2.34219	2.48587	2.62547	2.77153	2.92938	3.10356	3.29904	3.52190	3.78012
24	2.11848	2.32273	2.45755	2.58761	2.72268	2.86749	3.02587	3.20182	3.40011	3.62678
25	2.11131	2.30515	2.43212	2.55382	2.67937	2.81302	2.95806	3.11777	3.29597	3.49734
26	2.10477	2.28920	2.40916	2.52347	2.64071	2.76472	2.89836	3.04438	3.20590	3.38661
27	2.09877	2.27465	2.38832	2.49608	2.60599	2.72159	2.84539	2.97975	3.12722	3.29081
28	2.09325	2.26134	2.36934	2.47121	2.57463	2.68284	2.79809	2.92240	3.05791	3.20710
29	2.08815	2.24910	2.35196	2.44856	2.54618	2.64784	2.75558	2.87115	2.99639	3.13333
30	2.08343	2.23782	2.33599	2.42782	2.52024	2.61608	2.71718	2.82509	2.94140	3.06782
40	2.05018	2.15982	2.22715	2.28837	2.34833	2.40878	2.47067	2.53469	2.60137	2.67124
60	2.01861	2.08795	2.12916	2.16569	2.20061	2.23497	2.26929	2.30389	2.33900	2.37477
80	2.00341	2.05410	2.08375	2.10973	2.13429	2.15820	2.18182	2.20537	2.22900	2.25280
100	1.99446	2.03441	2.05756	2.07770	2.09663	2.11493	2.13291	2.15073	2.16849	2.18628
120	1.98858	2.02153	2.04051	2.05696	2.07235	2.08717	2.10167	2.11599	2.13021	2.14439
∞	1.9599	1.9599	1.9599	1.9599	1.9599	1.9599	1.9599	1.9599	1.9599	1.9599

p-no.of predictors n-Sample Size

Table 2 Significant two-tail percentage points of Internal studentized residual at $p_r\left(\left|r_i\right| > r_{i(p,n)}(0.01)\right) = 0.01$

n	*p*									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
3	77.96327	-	-	-	-	-	-	-	-	-
4	11.46022	235.40271	-	-	-	-	-	-	-	-
5	6.53034	23.21222	318.28371	-	-	-	-	-	-	-
6	5.04353	10.94698	29.51148	380.32293	-	-	-	-	-	-
7	4.35521	7.57721	13.34171	34.36548	432.09685	-	-	-	-	-
8	3.96341	6.09282	8.96079	15.22295	38.48448	477.53358	-	-	-	-
9	3.71176	5.27397	7.04613	10.06222	16.83942	42.13704	518.56681	-	-	-
10	3.53689	4.75939	5.99533	7.81230	11.01752	18.28501	45.45883	556.30902	-	-
11	3.40846	4.40752	5.33739	6.57922	8.48160	11.87764	19.60769	48.52965	591.46797	-
12	3.31020	4.15230	4.88873	5.80782	7.09216	9.08748	12.66861	20.83602	51.40121	624.52675
13	3.23263	3.95896	4.56400	5.28209	6.22295	7.55855	9.64701	13.40607	21.98873	54.10934
14	3.16985	3.80757	4.31845	4.90174	5.63049	6.60179	7.99078	10.17045	14.10031	23.07893
15	3.11800	3.68587	4.12644	4.61421	5.20179	5.94940	6.95392	8.39630	10.66457	14.75860
16	3.07447	3.58594	3.97228	4.38943	4.87765	5.47715	6.24659	7.28512	8.78003	11.13419
17	3.03740	3.50244	3.84583	4.20899	4.62420	5.11995	5.73432	6.52671	7.59918	9.14548
18	3.00546	3.43165	3.74027	4.06101	4.42071	4.84054	5.34667	5.97719	6.79284	7.89883
19	2.97765	3.37087	3.65084	3.93750	4.25379	4.61613	5.04332	5.56114	6.20829	7.04716
20	2.95322	3.31813	3.57411	3.83286	4.11444	4.43199	4.79957	5.23541	5.76552	6.42946
21	2.93159	3.27193	3.50758	3.74310	3.99637	4.27822	4.59950	4.97358	5.41872	5.96137
22	2.91230	3.23114	3.44933	3.66526	3.89507	4.14790	4.43235	4.75857	5.13982	5.59457
23	2.89499	3.19485	3.39791	3.59712	3.80722	4.03606	4.29065	4.57888	4.91070	5.29946
24	2.87938	3.16236	3.35220	3.53698	3.73030	3.93904	4.16901	4.42649	4.71915	5.05694
25	2.86522	3.13310	3.31129	3.48351	3.66241	3.85408	4.06345	4.29563	4.55665	4.85411
26	2.85233	3.10663	3.27447	3.43567	3.60204	3.77907	3.97099	4.18204	4.41705	4.68197
27	2.84053	3.08255	3.24115	3.39260	3.54801	3.71236	3.88934	4.08251	4.29583	4.53404
28	2.82970	3.06055	3.21086	3.35364	3.49937	3.65264	3.81670	3.99460	4.18960	4.40556
29	2.81973	3.04039	3.18320	3.31822	3.45537	3.59888	3.75168	3.91637	4.09573	4.29293
30	2.81050	3.02183	3.15785	3.28587	3.41536	3.55023	3.69312	3.84632	4.01219	4.19337
40	2.74610	2.89449	2.98640	3.07032	3.15276	3.23608	3.32160	3.41027	3.50284	3.60007
60	2.68576	2.77864	2.83412	2.88343	2.93065	2.97717	3.02369	3.07065	3.11833	3.16697
80	2.65700	2.72455	2.76423	2.79905	2.83200	2.86410	2.89585	2.92752	2.95932	2.99137
100	2.64017	2.69324	2.72410	2.75099	2.77627	2.80074	2.82478	2.84862	2.87240	2.89622
120	2.62911	2.67282	2.69806	2.71996	2.74045	2.76021	2.77954	2.79864	2.81761	2.83653
∞	2.5758	2.5758	2.5758	2.5758	2.5758	2.5758	2.5758	2.5758	2.5758	2.5758

p-no.of predictors n-Sample Size

Table 3 Significant two-tail percentage points of V at $p_r\left(V_i > V_{i(p,n)}(0.05)\right) = 0.05$

n	*p*									
	1	2	3	4	5	6	7	8	9	10
3	.81650	.06406	-	-	-	-	-	-	-	-
4	.86603	.27042	.04330	-	-	-	-	-	-	-
5	.89443	.42757	.20000	.03513	-	-	-	-	-	-
6	.91287	.53356	.33630	.16738	.03044	-	-	-	-	-
7	.92582	.60762	.43779	.28880	.14732	.02728	-	-	-	-
8	.93541	.66179	.51380	.38369	.25789	.13330	.02496	-	-	-
9	.94281	.70297	.57225	.45756	.34686	.23554	.12276	.02315	-	-
10	.94868	.73528	.61839	.51613	.41789	.31944	.21833	.11443	.02170	
11	.95346	.76129	.65565	.56351	.47540	.38763	.29789	.20451	.10763	.02049
12	.95743	.78266	.68635	.60255	.52275	.44370	.36343	.28033	.19309	.10193
13	.96077	.80053	.71206	.63524	.56235	.49049	.41797	.34343	.26563	.18342
14	.96362	.81569	.73389	.66300	.59593	.53008	.46396	.39646	.32652	.25308
15	.96609	.82872	.75266	.68685	.62475	.56399	.50324	.44155	.37809	.31196
16	.96825	.84002	.76896	.70756	.64975	.59334	.53715	.48036	.42226	.36214
17	.97014	.84993	.78326	.72572	.67164	.61901	.56673	.51410	.46051	.40538
18	.97183	.85869	.79589	.74175	.69096	.64162	.59276	.54371	.49396	.44304
19	.97333	.86648	.80713	.75602	.70814	.66171	.61583	.56990	.52347	.47613
20	.97468	.87345	.81720	.76880	.72351	.67967	.63642	.59323	.54969	.50546
21	.97590	.87974	.82628	.78031	.73735	.69581	.65491	.61415	.57316	.53164
22	.97701	.88543	.83449	.79073	.74987	.71041	.67162	.63302	.59429	.55515
23	.97802	.89060	.84197	.80021	.76125	.72367	.68677	.65012	.61341	.57639
24	.97895	.89533	.84880	.80886	.77164	.73577	.70059	.66570	.63080	.59567
25	.97980	.89967	.85506	.81680	.78116	.74686	.71324	.67994	.64668	.61326
26	.98058	.90366	.86083	.82411	.78993	.75705	.72487	.69302	.66125	.62937
27	.98131	.90735	.86615	.83085	.79801	.76645	.73558	.70507	.67466	.64419
28	.98198	.91076	.87109	.83710	.80550	.77515	.74550	.71620	.68704	.65786
29	.98261	.91393	.87567	.84290	.81246	.78323	.75469	.72652	.69852	.67051
30	.98319	.91689	.87994	.84831	.81893	.79075	.76325	.73612	.70917	.68225
40	.98742	.93813	.91062	.88714	.86540	.84463	.82445	.80464	.78508	.76565
60	.99163	.95906	.94086	.92536	.91106	.89744	.88426	.87139	.85873	.84622
80	.99373	.96941	.95581	.94424	.93359	.92346	.91367	.90413	.89477	.88554
100	.99499	.97558	.96473	.95550	.94701	.93894	.93116	.92358	.91615	.90884
120	.99582	.97968	.97065	.96298	.95592	.94922	.94276	.93647	.93031	.92426
∞	1	1	1	1	1	1	1	1	1	1

p-no.of predictors n-Sample Size

Table 4 Significant two-tail percentage points of V at $p_r\left(V_i > V_{i(p,n)}(0.01)\right)=0.01$

n	***p***									
	1	2	3	4	5	6	7	8	9	10
3	.81650	.01282	-	-	-	-	-	-	-	-
4	.86603	.12217	.00866	-	-	-	-	-	-	-
5	.89443	.25429	.08944	.00702	-	-	-	-	-	-
6	.91287	.36371	.19667	.07460	.00609	-	-	-	-	-
7	.92582	.44900	.29277	.16776	.06555	.00545	-	-	-	-
8	.93541	.51565	.37240	.25439	.14925	.05925	.00499	-	-	-
9	.94281	.56859	.43761	.32852	.22882	.13600	.05452	.00463	-	-
10	.94868	.61144	.49137	.39091	.29831	.21003	.12585	.05080	.00434	-
11	.95346	.64673	.53617	.44350	.35786	.27562	.19540	.11775	.04776	.00410
12	.95743	.67625	.57396	.48816	.40885	.33256	.25767	.18355	.11107	.04522
13	.96077	.70129	.60620	.52645	.45276	.38191	.31227	.24296	.17369	.10544
14	.96362	.72277	.63400	.55956	.49084	.42483	.36002	.29546	.23059	.16530
15	.96609	.74140	.65819	.58844	.52411	.46241	.40191	.34172	.28121	.22000
16	.96825	.75770	.67942	.61383	.55340	.49552	.43885	.38257	.32608	.26891
17	.97014	.77208	.69820	.63632	.57936	.52487	.47162	.41882	.36593	.31249
18	.97183	.78486	.71492	.65636	.60251	.55106	.50085	.45115	.40147	.35138
19	.97333	.79629	.72989	.67433	.62328	.57455	.52706	.48015	.43332	.38622
20	.97468	.80657	.74339	.69053	.64201	.59574	.55070	.50628	.46202	.41758
21	.97590	.81586	.75560	.70521	.65898	.61494	.57212	.52993	.48798	.44593
22	.97701	.82431	.76672	.71857	.67443	.63241	.59160	.55145	.51157	.47167
23	.97802	.83202	.77687	.73078	.68855	.64838	.60941	.57110	.53310	.49514
24	.97895	.83908	.78618	.74198	.70150	.66303	.62573	.58911	.55283	.51663
25	.97980	.84558	.79474	.75229	.71343	.67651	.64076	.60568	.57096	.53637
26	.98058	.85157	.80265	.76181	.72444	.68897	.65463	.62097	.58769	.55457
27	.98131	.85711	.80997	.77063	.73464	.70050	.66747	.63513	.60317	.57139
28	.98198	.86226	.81678	.77882	.74412	.71122	.67940	.64827	.61753	.58700
29	.98261	.86705	.82311	.78645	.75294	.72119	.69051	.66050	.63089	.60151
30	.98319	.87152	.82902	.79357	.76118	.73050	.70087	.67191	.64335	.61504
40	.98742	.90384	.87186	.84523	.82096	.79803	.77597	.75449	.73342	.71262
60	.99163	.93605	.91467	.89689	.88073	.86551	.85091	.83675	.82291	.80932
80	.99373	.95209	.93604	.92270	.91059	.89920	.88830	.87773	.86743	.85733
100	.99499	.96170	.94885	.93818	.92850	.91940	.91069	.90227	.89406	.88602
120	.99582	.96810	.95739	.94850	.94043	.93285	.92561	.91861	.91179	.90511
∞	1	1	1	1	1	1	1	1	1	1

p-no.of predictors n-Sample Size

5. Numerical Results and Discussion

In this section, the authors shown a numerical study of evaluating the outliers based on the internal studentized residual of the i^{th} observation in a regression model. For this, the authors fitted stepwise linear regression models with different set of predictors in a Brand equity study. The data in the study comprised of 18 different attributes about a car brand and the data was collected from 275 car users.

A well-structured questionnaire was prepared and distributed to 300 customers and the questions were anchored at five point likert scale from 1 to 5.After the data collection is over, only 275 completed questionnaires were used for analysis. The stepwise regression results reveals 4 nested models were extracted from the regression procedure by using IBM SPSS version 22. For each model, the internal studentized residual were computed, comparison of the proposed approach with the traditional Weisberg test of identifying the outliers are discussed through the following tables.

Table 5 Identification of Outliers based on Traditional Weisberg test

Model	*p*	*df* $(n-p-2)$	Traditional Weisberg test			
			Critical $\mathbf{t}(0.05)$	**(n)** $\lvert {}^{*}t_i \rvert > t(0.05)$	**Critical** $\mathbf{t}(0.01)$	**(n)** $\lvert {}^{*}t_i \rvert > t(0.01)$
1	1	272	1.96872	13	2.90292	11
2	2	271	1.96876	13	2.90301	10
3	3	270	1.96879	14	2.90310	10
4	4	269	1.96882	13	2.90319	10

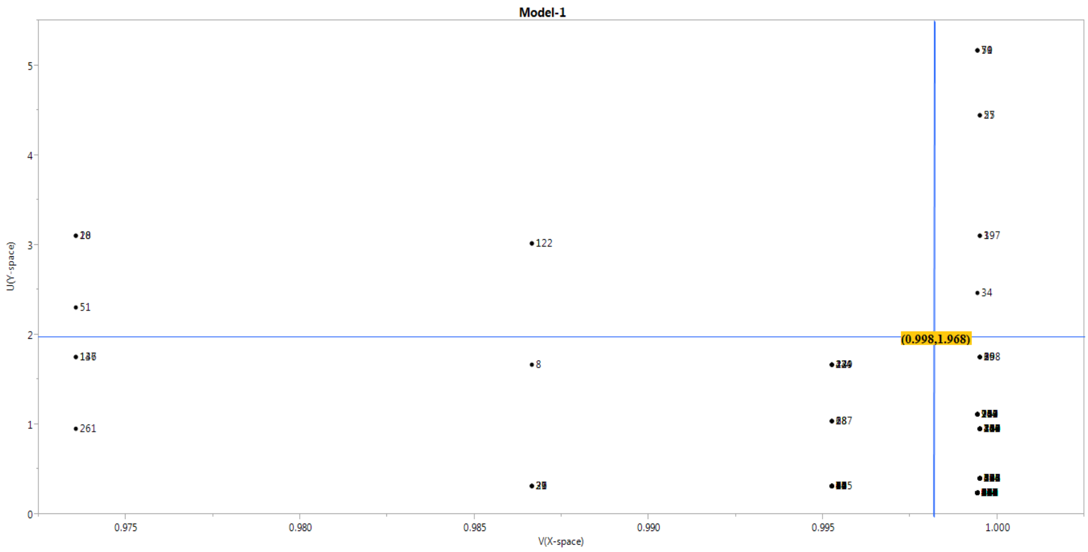
p-no.of predictors n=275 $\quad {}^{*}t_i = r_i\sqrt{(n-p-2)\left(n-p-1-r_i^2\right)^{-1}}$ *df*-degrees of freedom

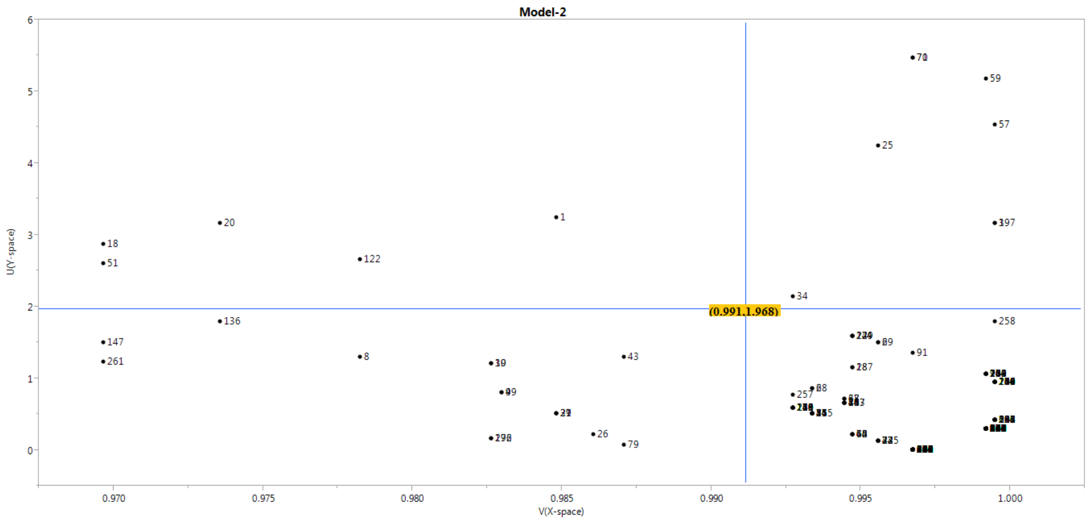
Table 6 Identification of Outliers based on Proposed approach at 5% significance level

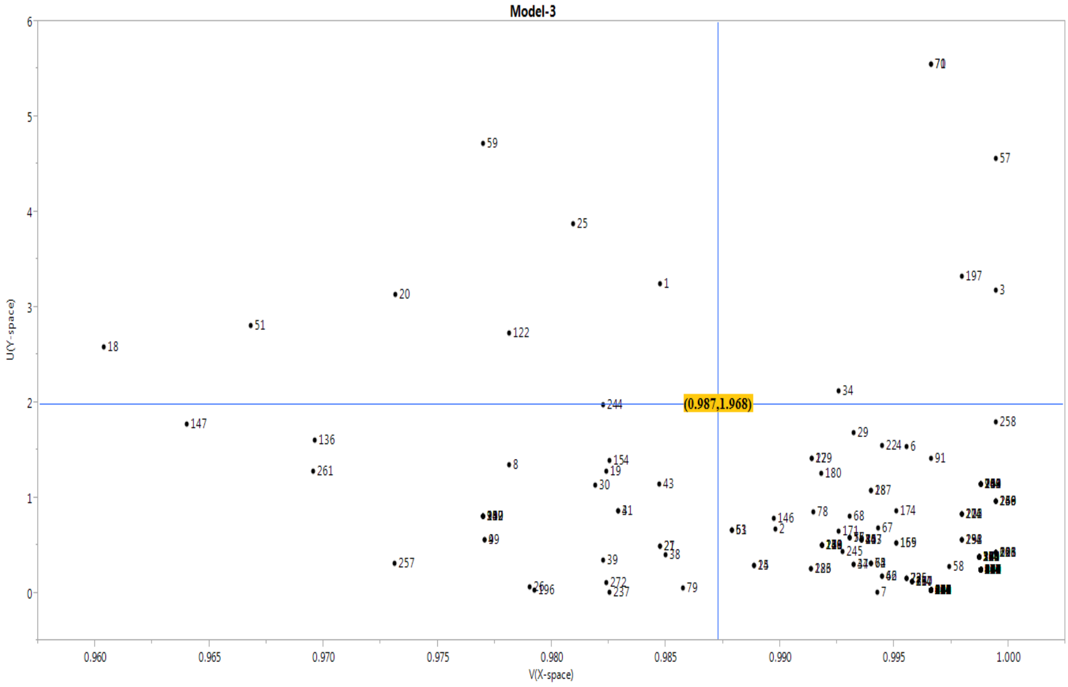
Model	p	Joint test		X-Y space			
		Critical $r(0.05)$	(n) $\|r_i\| > r(0.05)$	(n) [a]Group-1	(n) [b]Group-2	(n) Group-3	(n) Group-4
1		1.97228	13	6	7	-	
2	1	1.98626	13	8	5	-	-
3	2	1.99419	14	6	8	-	-
4	3	2.00097	13	6	7	-	-

p-no.of predictors *n*=275 [a] (t-ratio) p-value<0.05 & (V-test) p-value<0.05 [b] (t-ratio) p-value<0.05

Evidence plots of the Fitted Models at level5% Significance






Model-3
U(Y-space)
V(X-space)
(0.987,1.968)
0.960
0.965
0.970
0.975
0.980
0.985
0.990
0.995
1.000

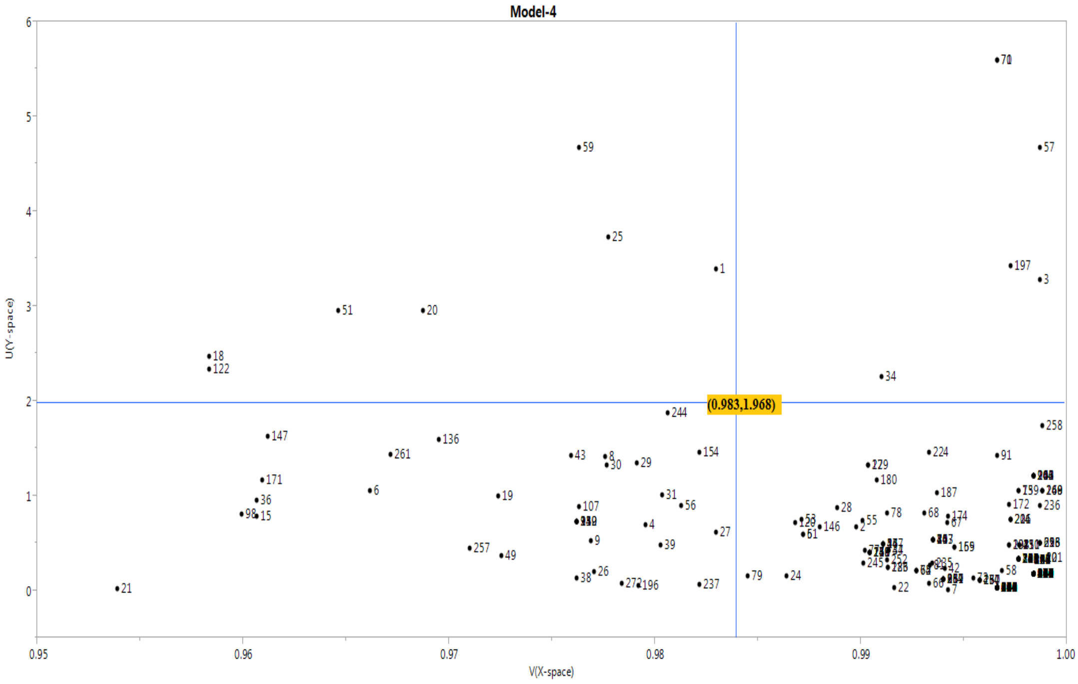
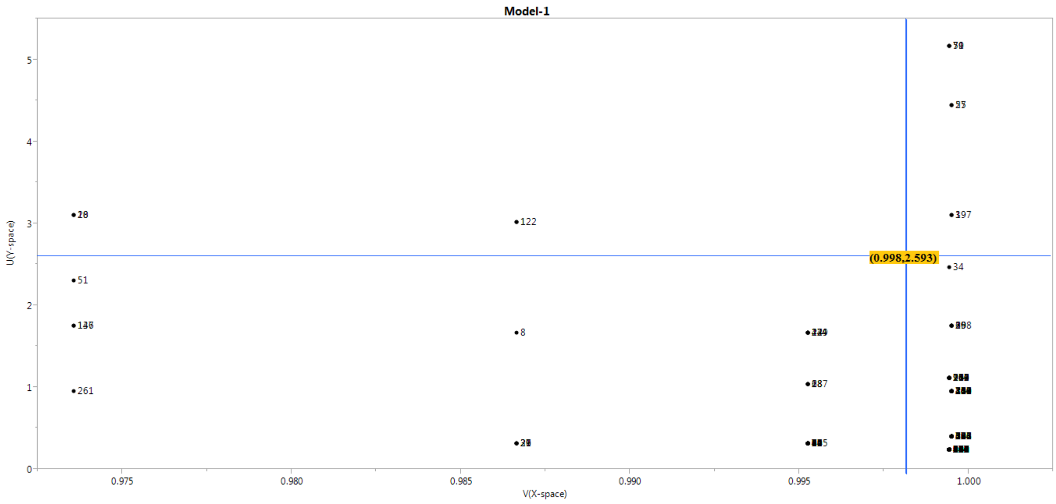

Model-4
U(Y-space)
V(X-space)
(0.983,1.968)
0.95
0.96
0.97
0.98
0.99
1.00

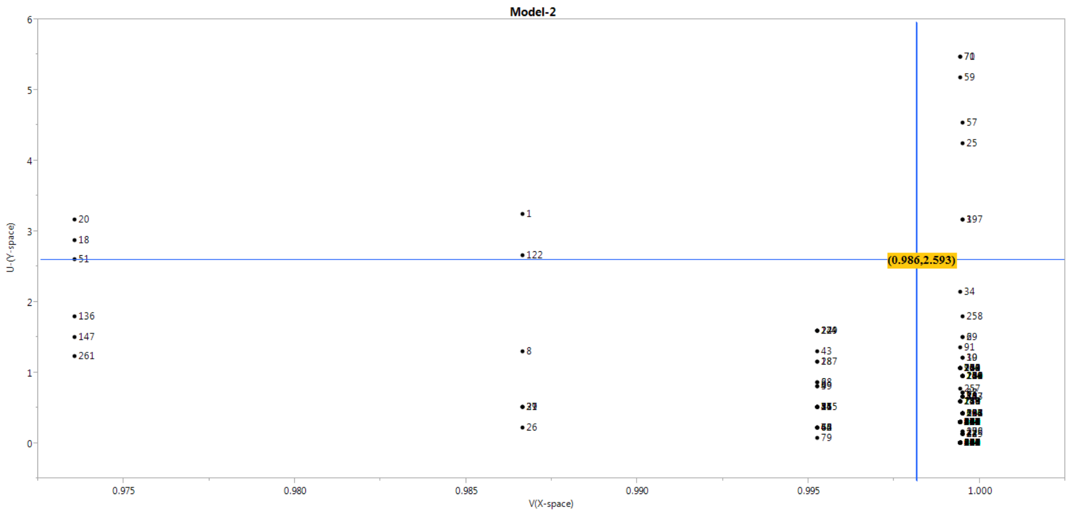
Table 7 Identification of Outliers based on Proposed approach at 1% level

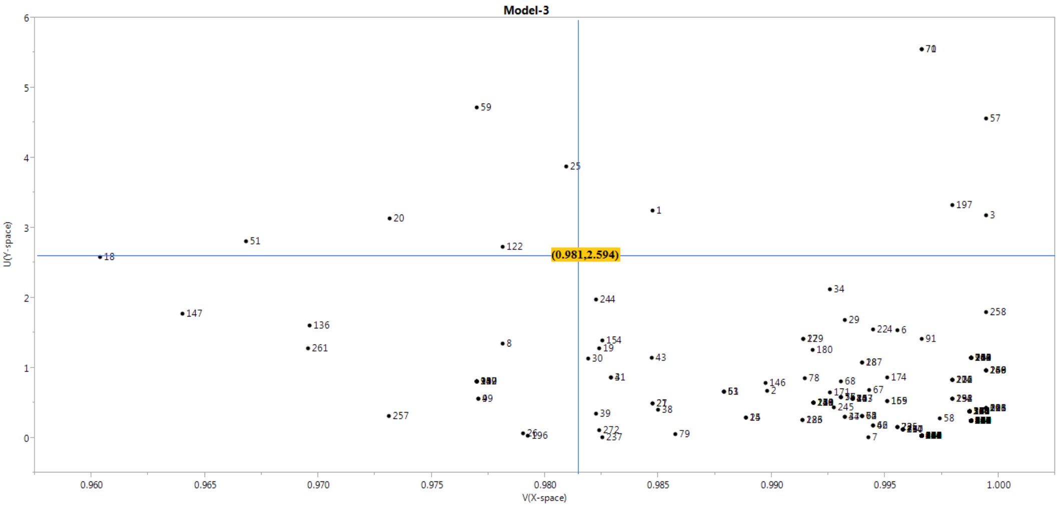
Model	p	Joint test		X-Y space			
		Critical $r(0.01)$	(n) $\|r_i\| > r(0.01)$	(n) [a]Group-1	(n) [b]Group-2	(n) [c]Group-3	(n) [d]Group-4
1	*1*	2.59869	11	7	4	-	-
2	2	2.61714	12	9	2	-	1
3	3	2.62760	12	6	5	1	-
4	4	2.63656	10	7	3	-	-

p-no.of predictors *n*=275 [a] (t-ratio) p-value<0.01 & (V-test) p-value<0.01 [b] (t-ratio) p-value<0.01
[c] (t-ratio) p-value>0.01 & (V-test) p-value>0.01 [d](V-test) p-value<0.01

Evidence plots of the Fitted Models at 1% Significance level






Model-3
(0.981,2.594)
U(Y-space)
V(X-space)

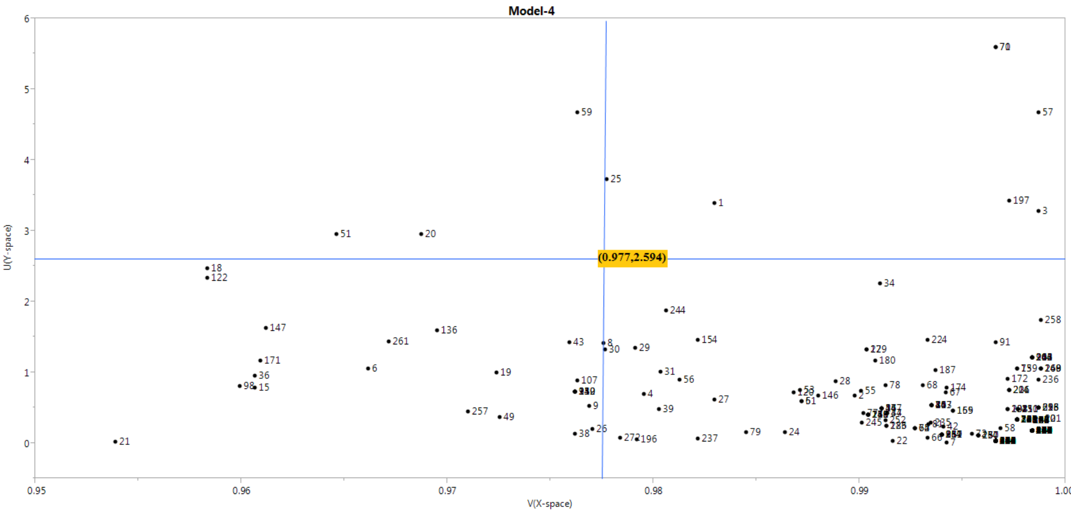

Model-4
(0.977,2.594)
U(Y-space)
V(X-space)

Table 5 clearly visualizes the result of the identification of outliers based on traditional Weisberg test. From the 4 fitted multiple regression models, the Weisberg test helps to identify 13 outliers in model-1, model-2 and model-4 at the 5% level of significance respectively. Similarly it helps to identifies 11 outliers in model-1, 10 outliers in model-2, model3 and model 4 at 1% level of significance. The test results emphasis these identified outliers only exists in Y-space and not in X-Space. On the other hand, table 6 and table 7, exhibits the result of the identification of outliers based on the proposed approach. From table 6, based on the proposed distribution of internal studentized residual, the authors identified 13 outliers in model-1, model-2 and model-4 at the 5% level of significance. Similarly, in model-4 the authors identified 14 outliers at 5% level of significance. These identified outliers may be the outliers in both the spaces. Because the distribution of the proposed test statistic incorporates the information about the Y-space as well as the X-space. But the authors visualize heuristic evidences of each observations and its position thought the evidence plot. Out of 13 outliers in model-1, 6 observations are outliers in both the spaces and it is significant at the 5% level based on the result of both tests (t and V test). Similarly, 7 observations are outliers in Y-space and the result of the t-test confirms the significance at the 5% level. Similarly, out of 13 outliers in model-2, 8 were significantly plotted in the first quadrant and remaining 5 were significantly plotted in the second quadrant. These show 8 observations are outliers in both the space and 5 observations are outliers in Y-space. From, Model-3 out of 14 outliers, 6 were significantly outlying in both the spaces and 8 observations are statistically significant and consider to be an outlier in Y-space only. Finally, in model-4, out of 13 outliers, 6 observations are statistically significant and consider to be outliers in both the spaces and 7 observations are significant outliers in Y-space. From table 7, based on the proposed approach, the authors identified 11 outliers in model-1, 12 outliers in model-2, 3 and 10 outliers in model-4 respectively. Based on the evidence plot, Out of the 11 outliers in model-1, 7 observations are statistically significant at the 1% level and consider to be the outliers in both the spaces. In model-2 there are 12 outliers, where 9 observations are outliers in both the spaces, 2 observations are outliers in Y-space and 1 observation is influential in X-space. As far as model-3 is concerned, among the 12 outliers, 6 observations are outlier in both the spaces, 5 observations are outlier in Y-space and 1 observation is an influential leverage point significantly placed in the third quadrant. In model-4 there are 10 outliers, where 7 observations are outliers in both the spaces and 3 observations are outliers in Y-space. From the above discussion the authors explored and identified some advantages of the proposed approach over the use of traditional approach of evaluating the outliers using internal studentized residuals. At first the proposed distribution of the internal studentized residual comprised of the distribution assumption of both the Y-space and X-space. The evidence plot clearly visualizes the position and outlying nature of the observations in the spaces. Hence, the simultaneous use of internal studentized residual with the evidence plot gave mores insights of the outliers existing in a multiple regression model.

6. Conclusion

From the previous sections, the authors proposed the exact distribution of the Internal studentized residuals which comprised of distribution of X and Y space along with the evidence plot to evaluate the outliers in a multiple linear regression model. At first, the exact distribution of the internal studentized residual was derived and the authors visualized the density function in terms of the Gauss hypergeometric function with two shape parameters namely *p* and *n*. Moreover,the critical percentage points of internal studentized residual at 5 %, 1% level of significance and it is utilized to evaluate the outliers. The evidence plots along with the test of significance (t-test and V-test) helps us to exactly visualize the position of the identified outliers in the spaces. Finally, the proposed approach is more systematic and scientific because it helps to identify the outliers in both spaces and the results were superior when compared it with the traditional approach. So, the authors conclude the proposed approaches over rides the use of traditional approach and we believe that the proposed approaches took the process of identifying the outliers to the next level which helps the statisticians to exactly identify the remote observations in the functional data.

7. References

[1] Ali S. Hadi (1992), A new measure of overall potential influence in linear regression, Computational Statistics & Data Analysis, 14(1), 1-27

[2] Behnken, D. W., and Draper, N. R. (1972), "Residuals and Their Variance Patterns," Technometrics, 14, 101-111.

[3] Beckman, R.J., Trussell, H.J., 1974. “The distribution of an arbitrary studentized residual and effects of updating in multiple regression”. J. Amer. Statist. Assoc. 69, 199–201.

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

[5] Cook, R.D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19, 15-18.

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

[7] Cook, R. D., & Weisberg, S. (1982) *Residuals and influence in regression* (Vol. 5). New York: Chapman and Hall.

[8] Davies, R. B. and Hutton, B., (1975). The effects of errors in the independent variables in linear regression. Biometrika, 62, 383-391.

[9] Dıaz-Garcıa, J. A., & González-Farıas (2004), G. A note on the Cook's distance. *Journal of statistical planning and inference*, *120*(1), 119-136.

[10] Ellenberg, J.H., 1973. "The joint distribution of the standardized least squares residual from general linear regression". J. Amer. Statist. Assoc. 68, 941–943.

[11] Eubank, R.L(1985), Diagnostics for smoothing splines. *J. Roy. Statist. Soc. Ser. B* ***47****, 332–341.*

[12] Hoaglin, D.C,and Welsch, R.E. (1978).The Hat matrix in regression and ANOVA. *The Amer. Statist., 32, 17-22.*

[13] Huber, P. J., (1975). Robustness and Designs. A Survey of Statistical Design and Linear Models. North-Holland, Amsterdam.

[14] José A. Díaz-García, Ramón Gutiérrez-Jáimez , 2007. "The distribution of residuals from a general elliptical linear model". *Journal of Statistical Planning and Inference*, *Volume 137, Issue 7*, *Pages 2347-2354*

[15] Kendall, M.G., Stuart, A. (1973) The Advanced Theory of Statistics. Volume 2: Inference and Relationship, Griffin

[16] Kim, C.(1996), Cook's distance in spline smoothing. *Statist. Probab. Lett. 31, 139–144*.

[17] Kim, C., Kim, W.(1998), Some diagnostics results in nonparametric density estimation. *Comm. Statist. Theory Methods* ***27****, 291–303.*

[18] Kim, C., Lee, Y., Park, B.U.(2001), Cook's distance in local polynomial regression. *Statist. Probab. Lett.* ***54****, 33–40*.

[19] Lund, R. E., (1975). Tables for an approximate test for outliers in linear models. Technometrics, 17, 473-476.

[20] Silverman, B.W.(1985), Some aspects of the spline smoothing approach to non-parametric regression curve 6tting (with discussion). *J. Roy. Statist. Soc. Ser. B 47, 1–52.*

[21] Thomas, W (1991), Influence diagnostics for the cross-validated smoothing parameter in spline smoothing. *J. Amer. Statist. Assoc.* ***86****, 693–698.*

[22] Weisberg, s. (1980). *Applied linear regression*. New York: Wiley.