Journal of Reliability and Statistical Studies; ISSN (Print): 0974-8024, (Online):2229-5666
Vol. 8, Issue 2 (2015): 77-96

A NON-STUDENT DISTRIBUTION OF JACK-KNIFE RESIDUALS AND IDENTIFICATION OF OUTLIERS

G.S. David Sam Jayakumar and A. Sulthan
Jamal Institute of Management, Tiruchirappalli, India,
E Mail: samjaya77@gmail.com; Sulthan90@gmail.com

Received February 26, 2015
Modified November 14, 2015
Accepted November 28, 2015

Abstract

This paper proposes the exact distribution of Jack-Knife residual which is formally called as external studentized residual and used to evaluate the outliers in linear multiple regression analysis. The authors have proved that the Jackknife residuals do not follow student's t-distribution and they have explored the relationship among Jack-Knife residual, t-ratio and F-ratio and have expressed the derived density function of the residual in terms of series expression form. Moreover, the new form of the distribution is symmetric, first two moments of the distribution are derived and the authors have computed the critical points of Jack Knife residual at 5% and 1% level of significance and for varying sample sizes and predictors. Finally, the numerical example shows that the results extracted from the proposed approach and classical approach are similar even though the proposed distribution of the Jackknife residuals is different.

Key Words: Jack Knife Residuals, External Studentized Residual, Outliers, T-Ratio, F-Ratio, Moments, Critical Points.

AMS Classification: 62H10

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 standardization of a higher-degree statistic by another statistic of the same degree (Kendall and Stuart, 1973). In least-squares fitting it is important to understand the influence which an observed 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, 1978). The studentized residuals, t_t , (i.e. the residual divided by its standard error) have been recommended (Behnken and Draper (1972), Davies and Hutton (1975), Huber (1975)) 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, 1975). Cook (1977) 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 (1977), Cook and Weisberg (1982), Belsley et al. (1980) and Chatterjee and Hadi (1988) respectively. In non-parametric regression models, diagnostic results are quite rare. Among them, Eubank (1985), Silverman (1985), Thomas (1991), and Kim (1996) studied residuals, leverages, and several types of Cook's distance in smoothing splines, and Kim and Kim (1998), Kim et.al(2001) 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. (1980). 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 are concerned in the literature, the procedures were designed to detect the influence of observations on a specific regression result. However, Hadi (1992), 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 (2004) 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 outlying data points. Considering the above reviews, the authors have proposed the novel and exact distribution of Jack-Knife residual which indeed exactly identifies the outlying data points and is discussed in the subsequent sections.

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

The multiple linear regression model with random error is given by

$$Y = X\beta + e \tag{1}$$

where $\underset{(nX1)}{Y}$ is the vector of values of the dependent variable, $\underset{(nX(p+1))}{X}$ is a full column rank matrix of predictors, $\underset{(kX1)}{\beta}$ is the vector of beta co-efficients or partial regression co-efficients and $\underset{(nX1)}{e}$ is the residual vector following 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 follows approximately student's t distribution with $n-p-2$ degrees of freedom. Weisberg (1980) 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 always helps to evaluate the outliers in Y-space. Some authors like Cook (1977) and Hadi (1992) proposed measures to find the influential observations and potential outliers in the X as well as in the Y-space. Classically the general form of the external studentized residual or Jackknife residual $\left(R_i\right)$ of the i^{th} observation is given as

$$R_i = \frac{\widehat{e_i}}{\widehat{S}_{e(-i)}\sqrt{1-h_{ii}}} \tag{2}$$

Where $\widehat{e_i}$ is the estimated i^{th} regression residual, $\widehat{S}_{e(-i)}$ is the unbiased standard error of the estimated residuals without the i^{th} observation and $\left(h_{ii}\right)$ are 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 $\left(R_i\right)$ is compared with critical values of student's t-ratio for $n-p-2$ degrees of freedom and if the computed $\left(R_i\right)$ exceeds, then the observation is said to be outlier in the Y-space. Rewrite (2) in terms of the substitution of $\widehat{S}_{e(-i)} = S_e\sqrt{\frac{(n-p-1)-r_i^2}{n-p-2}}$ and the true standard error $\left(\sigma_e\right)$ of the residual as

$$R_i = \frac{\widehat{e_i}/\sigma_e}{\left(S_e/\sigma_e\right)\sqrt{\frac{(n-p-1)-r_i^2}{n-p-2}}\sqrt{1-h_{ii}}} \tag{3}$$

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

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

From (4), we know that the ratio $\left(z_i/\sqrt{\chi_{n-p-1}^2/n-p-1}\right)$ follows student's t-distribution with *n*-*p*-1 degrees of freedom. Similarly, Weisberg provided the monotonic transformation of internal studentized residuals $\left(r_i\right)$ in terms of student's t-ratio which follows t-distribution with *n-p*-2 degrees of freedom and is given as

$$t_i = r_i\sqrt{\frac{n-p-2}{n-p-1-r_i^2}} \tag{5}$$

From (5), we can write the squared internal studentized residuals $\left(r_i^2\right)$ in terms of F-ratio if student's t-ratio following *n-p*-2 degrees of freedom, then squared t-ratio follows F-distribution with $\left(1, n-p-2\right)$ degrees of freedom. Now, (5) can be written as

$$r_i^2 = \frac{(n-p-1)F_{i(1,n-p-2)}}{(n-p-2)+F_{i(1,n-p-2)}} \tag{6}$$

In the same manner, Belsley et al. (1980) proved when the set of predictors in a linear regression model follows multivariate normal distribution with (μ_X, Σ_X), then

$$\frac{(n-p)\left(h_{ii}-1/n\right)}{(p-1)(1-h_{ii})} \sim F_{(p-1,n-p)} \tag{7}$$

From (7) it follows F-distribution with $(p-1, n-p)$ degrees of freedom and it can be written in an alternative form as

$$h_{ii} = \frac{\left(((p-1)/(n-p))F_{i(p-1,n-p)}\right)+1/n}{1+((p-1)/(n-p))F_{i(p-1,n-p)}} \tag{8}$$

Now substituting (6) and (8) in (4), we get the form of Jack-Knife residual as

$$R_i = \frac{t_i}{\sqrt{\left(\frac{(n-p-1)-\left(\frac{(n-p-1)F_{i(1,n-p-2)}}{(n-p-2)+F_{i(1,n-p-2)}}\right)}{n-p-2}\right)\left(\frac{(n-1)/n}{1+\frac{p-1}{n-p}F_{i(p-1,n-p)}}\right)}} \tag{9}$$

From (5), it is the most important form identified and based on this, we proposed the exact distribution by utilizing the relationship among the Jack-Knife residual (R_i), t-ratio and the F-ratio's. From (9), the t and F--ratios are independent, because the computation of (t_i) involves the error term $e_i \sim N(0,\sigma_e^2)$, the $F_{i(1,n-p-2)}$ is the squared monotonic transformation of the internal studentized which involves the unstandardized true error term $e_i \sim N(0,\sigma_e^2)$, hat values (h_{ii}) and $F_{i(p-1,n-p)}$ is the transformation of hat values (h_{ii}) involving the set of predictors

$(H = X(X^{'}X)^{-1}X^{'})$. Therefore, from the property of least squares $E(eX) = 0$, $t_i, F_{i(1,n-p-2)}$ and $F_{i(p-1,n-p)}$ are also uncorrelated and independent. Using this assumption, we derived the exact distribution of $\left(R_i\right)$ by further modifying (9) as

$$R_i = \sqrt{\frac{n(n-p-2)}{(n-1)(n-p-1)}} \frac{t_i}{\sqrt{\dfrac{1}{1+\dfrac{1}{n-p-2}F_{i(1,n-p-2)}}}\sqrt{\dfrac{1}{1+\dfrac{p-1}{n-p}F_{i(p-1,n-p)}}}} \tag{10}$$

From (10), it can be simplified and $\left(R_i\right)$ is expressed in terms of independent t-ratio and beta variables θ_{1i} and θ_{2i} of the first kind by using the following facts

$$\theta_{1i} = \frac{1}{1+\dfrac{1}{n-p-2}F_{i(1,n-p-2)}} \sim \beta_1\left(\frac{n-p-2}{2}, \frac{1}{2}\right) \tag{11}$$

$$\theta_{2i} = \frac{1}{1+\dfrac{p-1}{n-p}F_{i(p-1),(n-p)}} \sim \beta_1\left(\frac{n-p}{2}, \frac{p-1}{2}\right) \tag{12}$$

Then, without loss of generality (10) can be written as

$$R_i = \sqrt{\frac{n(n-p-2)}{(n-1)(n-p-1)}} \frac{t_i}{\sqrt{\theta_{1i}\theta_{2i}}} \tag{13}$$

$$R_i = \alpha(p,n)\frac{t_i}{\sqrt{\theta_{1i}\theta_{2i}}} \tag{14}$$

Where $\alpha(p,n)$ is the normalizing constant which is a function of p, n and the final form of (14) can be given as

$$\frac{R_i}{\alpha(p,n)} = \frac{t_i}{\sqrt{\theta_{1i}\theta_{2i}}} = \psi_i \tag{15}$$

Based on the identified relationship from (15), the authors have derived the exact distribution of the Jack-Knife residual and is discussed in the next section.

3. Exact Distribution of Jackknife Residual

Using the technique of two-dimensional Jacobian of transformation, the joint probability density function of the t-ratio and the beta variable of Kind-1 namely θ_{1i} and θ_{2i} were transformed into density function of ψ_i and it is given as

$$f\left(\psi_i, u_{1i}, u_{2i}\right) = f\left(t_i, \theta_{1i}, \theta_{2i}\right)|J| \tag{16}$$

From (15), we know that t_i and θ_{1i}, θ_{2i} are independent. Then rewrite (16) as

$$f\left(\psi_i, u_{1i}, u_{2i}\right) = f\left(t_i\right) f\left(\theta_{1i}\right) f\left(\theta_{2i}\right) |J| \tag{17}$$

Using the change of variable technique, substitute $\theta_{1i} = u_{1i}$ and $\theta_{2i} = u_{2i}$ in (15) we get

$$\psi_i = \frac{t_i}{\sqrt{u_{1i} u_{2i}}} \tag{18}$$

Then partially differentiate (18), compute the Jacobian determinant and substitute in (17) as

$$f\left(\psi_i, u_{1i}, u_{2i}\right) = f\left(t_i\right) f\left(\theta_{1i}\right) f\left(\theta_{2i}\right) \left| \frac{\partial\left(t_i, \theta_{1i}, \theta_{2i}\right)}{\partial\left(\psi_i, u_{1i}, u_{2i}\right)} \right| \tag{19}$$

$$f\left(\psi_i, u_{1i}, u_{2i}\right) = f\left(t_i\right) f\left(\theta_{1i}\right) f\left(\theta_{2i}\right) \begin{vmatrix} \frac{\partial t_i}{\partial \psi_i} & \frac{\partial t_i}{\partial u_{1i}} & \frac{\partial t_i}{\partial u_{2i}} \\ \frac{\partial \theta_{1i}}{\partial \psi_i} & \frac{\partial \theta_{1i}}{\partial u_{1i}} & \frac{\partial \theta_{1i}}{\partial u_{2i}} \\ \frac{\partial \theta_{2i}}{\partial \psi_i} & \frac{\partial \theta_{2i}}{\partial u_{1i}} & \frac{\partial \theta_{2i}}{\partial u_{2i}} \end{vmatrix} \tag{20}$$

From (20), we know that t_i and θ_{1i}, θ_{2i} are independent, then the density function of the joint distribution of t_i and θ_{1i}, θ_{2i} is given as

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

$$\times \frac{1}{B\left(\frac{n-p}{2}, \frac{p-1}{2}\right)} \theta_{2i}^{\frac{n-p}{2}-1} (1-\theta_{2i})^{\frac{p-1}{2}-1} \tag{21}$$

where $-\infty < t_i < +\infty$, $0 \le \theta_{1i}, \theta_{2i} \le 1$, $n, p > 0$

and $$J = \frac{\partial(t_i,\theta_{1i},\theta_{2i})}{\partial(\psi_i,u_{1i},u_{2i})} = \begin{vmatrix} \sqrt{u_{1i}u_{2i}} & \frac{\psi_i\sqrt{u_{2i}}}{2\sqrt{u_{1i}}} & \frac{\psi_i\sqrt{u_{1i}}}{2\sqrt{u_{2i}}} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = \sqrt{u_{1i}u_{2i}} \tag{22}$$

Then substitute (21) and (22) in (20) in terms of the substitution of u_i, we get the joint distribution of ψ_i and u_{1i},u_{2i} as

$$f(\psi_i,u_{1i},u_{2i}) = \frac{1}{\sqrt{n-p-1}B\left(\frac{1}{2},\frac{n-p-1}{2}\right)}\left(1+\frac{\psi_i^2u_{1i}u_{2i}}{n-p-1}\right)^{-\left(\frac{n-p-1}{2}+\frac{1}{2}\right)} \times \frac{1}{B\left(\frac{n-p-2}{2},\frac{1}{2}\right)}u_{1i}^{\frac{n-p-2}{2}-1}(1-u_{1i})^{\frac{1}{2}-1}$$

$$\times \frac{1}{B\left(\frac{n-p}{2},\frac{p-1}{2}\right)}u_{2i}^{\frac{n-p}{2}-1}(1-u_{2i})^{\frac{p-1}{2}-1} \times |J|$$

$$f(\psi_i,u_{1i},u_{2i}) = \frac{1}{\sqrt{n-p-1}B\left(\frac{1}{2},\frac{n-p-1}{2}\right)B\left(\frac{n-p-2}{2},\frac{1}{2}\right)B\left(\frac{n-p}{2},\frac{p-1}{2}\right)}$$

$$\times\left(1+\frac{\psi_i^2u_{1i}u_{2i}}{n-p-1}\right)^{-\left(\frac{n-p-1}{2}+\frac{1}{2}\right)} u_{1i}^{\frac{n-p-2}{2}-1}(1-u_{1i})^{\frac{1}{2}-1} u_{2i}^{\frac{n-p}{2}-1}(1-u_{2i})^{\frac{p-1}{2}-1}\sqrt{u_{1i}u_{2i}} \tag{23}$$

where $-\infty < \psi_i < +\infty$, $0 \le u_{1i},u_{2i} \le 1$, $p,n > 0$ and $|J| = \sqrt{u_{1i}u_{2i}}$

Using Binomial series expansion rearrange (23) and integrate with respect to u_{1i},u_{2i}, we get the marginal distribution of ψ_i as

$$f(\psi_i) = \lambda(p,n)\sum_{k=0}^{\infty}\binom{-(n-p)/2}{k}$$

$$\times\left(\frac{\psi_i^{2k}}{(n-p-1)^k}\right)\int_0^1\int_0^1 u_{1i}^{\frac{n-p-1}{2}+k-1}(1-u_{1i})^{\frac{1}{2}-1}u_{2i}^{\frac{n-p+1}{2}+k-1}(1-u_{2i})^{\frac{p-1}{2}-1}du_{1i}du_{2i} \tag{24}$$

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

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

We know, from (24)

$$\int_0^1 u_{1i}^{\frac{n-p-1}{2}+k-1}\left(1-u_{1i}\right)^{\frac{1}{2}-1}du_{1i}=B\left(\frac{n-p-1}{2}+k,\frac{1}{2}\right) \tag{25}$$

$$\int_0^1 u_{2i}^{\frac{n-p+1}{2}+k-1}\left(1-u_{2i}\right)^{\frac{p-1}{2}-1}du_{2i}=B\left(\frac{n-p+1}{2}+k,\frac{p-1}{2}\right) \tag{26}$$

Then substitute (25), (26) in (24) and arrange the terms, we get the density function of ψ_i as

$$f(\psi_i)=\lambda(p,n)\sum_{k=0}^{\infty}\binom{-(n-p)/2}{k}\left(\frac{\psi_i^{2k}}{(n-p-1)^k}\right)B\left(\frac{n-p-1}{2}+k,\frac{1}{2}\right)B\left(\frac{n-p+1}{2}+k,\frac{p-1}{2}\right) \tag{27}$$

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

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

Using one-dimensional Jacobian of transformation in (15), differentiate and substitute in (27), we get the density function of the Jack-Knife residuals as

$$f(R_i;p,n)=\frac{\lambda(p,n)}{\alpha(p,n)}\sum_{k=0}^{\infty}\binom{-(n-p)/2}{k}B\left(\frac{n-p-1}{2}+k,\frac{1}{2}\right)B\left(\frac{n-p+1}{2}+k,\frac{p-1}{2}\right)\left(\frac{1}{(n-p-1)^k}\right)\left(\frac{R_i}{\alpha(p,n)}\right)^{2k} \tag{28}$$

where $-\infty<R_i<+\infty$, $n,p>0$, $n>p$ and $\alpha(p,n)=\sqrt{\frac{n(n-p-2)}{(n-1)(n-p-1)}}$

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

From (28), it is the density function of Jack-Knife residual $\left(R_i\right)$ which is symmetric in nature and involves the normalizing constants such as $B\left(\frac{n-p-1}{2}+k,\frac{1}{2}\right)$, $B\left(\frac{n-p+1}{2}+k,\frac{p-1}{2}\right)$, $\alpha(p,n)$, $\lambda(p,n)$ with two shape parameters (p,n),where B is the beta function, n is the sample size and p is the no. of predictors used in a multiple linear regression model respectively. In order to know the location and dispersion of Jack-Knife residual, the authors have derived the first two moments in

terms of mean, variance from (15) and are shown as follows.
Using (15), take expectation and substitute the moments of independent t-ratio and beta variables θ_{1i}, θ_{2i} of Kind-1, we get the first moment of R_i as

$$E(R_i) = 0 \tag{29}$$

From (29), if the moment of the residual is zero ($E(R_i) = 0$), then the second moment is equal to its variance. Hence, square (15) on both sides, then take expectation and substitute the appropriate second order moments of independent t-ratio and F-ratios, we get the variance of the R_i which is given as

$$V(R_i) = E(R_i^2) = V(t_i)E(\theta_{1i}^{-1})E(\theta_{2i}^{-1})$$

$$V(R_i) = E(R_i^2) = \frac{(n-3)(n-p-1)}{(n-p-2)(n-p-4)} \tag{30}$$

As a proposed approach, the authors adopted the test of significance approach to evaluate and identify the outliers in a sample. The approach is to derive the critical points of the Jack-Knife residual by using the following relationship from (9) and it is given as

$$R_{i(p,n)}(\alpha) = t_{i(n-p-1)}(\alpha)\sqrt{\frac{n(n-p-2)}{(n-1)(n-p-1)}\left(1+\frac{1}{n-p-2}F_{i(1,n-p-2)}(\alpha)\right)\left(1+\frac{p-1}{n-p}F_{i(p-1,n-p)}(\alpha)\right)} \tag{31}$$

From (31), for different combination of values of (p,n) and the significance probability $p\left(|R_i| > R_{i(p,n)}(\alpha)\right) = \alpha$, we have computed the critical points of Jack-Knife residual. If the sample size is very large $(n \to \infty)$, then the limiting distribution of R_i follows 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 exhibit the significant percentage points of the distribution of Jack-Knife residual for varying sample size*(n)* and predictors *(p)* at 5% and 1% significance (α) calculated based on (31).

n	***p***									
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
3	0.00	-	-	-	-	-	-	-	-	-
4	44.7	0.00	-	-	-	-	-	-	-	-
5	9.30	90.6	0.000	-	-	-	-	-	-	-
6	5.51	15.5	115.3	0.000	-	-	-	-	-	-
7	4.24	8.27	19.00	134.2	0.000	-	-	-	-	-
8	3.63	5.94	9.789	21.68	150.3	0.000	-	-	-	-
9	3.28	4.84	6.874	10.99	23.99	164.6	0.000	-	-	-
1	3.04	4.20	5.503	7.621	12.03	26.05	177.6	0.000	-	-
1	2.88	3.80	4.719	6.039	8.274	12.97	27.93	189.6	0.000	-
1	2.76	3.51	4.215	5.135	6.510	8.865	13.84	29.68	200.8	0.000
1	2.67	3.30	3.864	4.554	5.502	6.938	9.411	14.64	31.32	211.4
1	2.59	3.14	3.607	4.150	4.854	5.837	7.335	9.922	15.40	32.88
1	2.53	3.02	3.410	3.854	4.404	5.129	6.149	7.707	10.40	16.12
1	2.48	2.91	3.255	3.628	4.074	4.637	5.385	6.441	8.060	10.86
1	2.44	2.83	3.130	3.449	3.822	4.276	4.855	5.627	6.719	8.395
1	2.41	2.76	3.027	3.305	3.623	4.001	4.466	5.061	5.856	6.984
1	2.38	2.70	2.940	3.186	3.462	3.783	4.168	4.645	5.256	6.076
2	2.35	2.65	2.867	3.087	3.330	3.607	3.933	4.327	4.816	5.444
2	2.32	2.60	2.803	3.002	3.219	3.462	3.744	4.076	4.478	4.979
2	2.30	2.56	2.748	2.929	3.124	3.341	3.587	3.873	4.212	4.624
2	2.28	2.53	2.700	2.866	3.043	3.237	3.456	3.706	3.997	4.343
2	2.27	2.50	2.657	2.810	2.972	3.148	3.344	3.565	3.819	4.116
2	2.25	2.47	2.619	2.761	2.910	3.070	3.247	3.445	3.670	3.928
2	2.24	2.44	2.584	2.717	2.855	3.002	3.164	3.342	3.543	3.771
2	2.23	2.42	2.554	2.678	2.806	2.942	3.090	3.252	3.433	3.637
2	2.21	2.40	2.526	2.642	2.762	2.889	3.025	3.174	3.338	3.521
2	2.20	2.38	2.500	2.610	2.722	2.840	2.967	3.104	3.254	3.420
3	2.19	2.36	2.477	2.581	2.686	2.797	2.915	3.042	3.180	3.332
4	2.13	2.24	2.321	2.388	2.454	2.521	2.589	2.661	2.735	2.814
6	2.07	2.14	2.185	2.224	2.261	2.297	2.334	2.371	2.409	2.447
8	2.04	2.09	2.124	2.151	2.176	2.201	2.226	2.251	2.275	2.301
1	2.02	2.06	2.088	2.109	2.129	2.148	2.166	2.185	2.203	2.222
1	2.01	2.04	2.066	2.083	2.098	2.114	2.129	2.143	2.158	2.173
∞	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96

p-no.of predictors   n-Sample Size

Table-1: Significant two-tail percentage points of Jackknife residual at

$$p\left(\left|R_i\right| > R_{i(p,n)}\left(0.05\right)\right) = 0.05$$

n	*p*									
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
3	0.000	-	-	-	-	-	-	-	-	-
4	819.4	0.0000	-	-	-	-	-	-	-	-
5	52.10	2790.7	0.0000	-	-	-	-	-	-	-
6	19.85	128.14	3608.3	0.0000	-	-	-	-	-	-
7	12.19	40.375	159.19	4230.2	0.0000	-	-	-	-	-
8	9.103	21.899	48.681	183.21	4754.7	0.0000	-	-	-	-
9	7.494	14.977	25.804	55.182	203.69	5218.2	0.0000	-	-	-
10	6.525	11.556	17.330	28.887	60.771	221.91	5638.7	0.0000	-	-
11	5.883	9.5715	13.178	19.201	31.555	65.774	238.53	6026.8	0.0000	-
12	5.428	8.2945	10.785	14.474	20.828	33.955	70.357	253.92	6389.2	0.0000
13	5.091	7.4115	9.2530	11.758	15.606	22.298	36.161	74.616	268.34	6730.7
14	4.831	6.7680	8.1981	10.024	12.611	16.632	23.653	38.219	78.617	281.96
15	4.625	6.2798	7.4320	8.8329	10.702	13.386	17.580	24.920	40.155	82.404
16	4.458	5.8978	6.8527	7.9691	9.3921	11.319	14.105	18.469	26.115	41.992
17	4.320	5.5912	6.4004	7.3169	8.4433	9.9026	11.893	14.780	19.309	27.251
18	4.203	5.3399	6.0382	6.8084	7.7274	8.8769	10.377	12.433	15.418	20.109
19	4.104	5.1305	5.7420	6.4016	7.1698	8.1035	9.2810	10.825	12.944	16.027
20	4.019	4.9534	5.4956	6.0693	6.7239	7.5012	8.4544	9.6621	11.249	13.432
21	3.944	4.8017	5.2875	5.7931	6.3600	7.0199	7.8109	8.7857	10.024	11.655
22	3.879	4.6704	5.1095	5.5599	6.0575	6.6271	7.2968	8.1036	9.1011	10.371
23	3.821	4.5557	4.9556	5.3607	5.8024	6.3008	6.8773	7.5588	8.3825	9.4031
24	3.770	4.4546	4.8213	5.1886	5.5845	6.0257	6.5289	7.1142	7.8086	8.6498
25	3.724	4.3649	4.7031	5.0385	5.3963	5.7908	6.2352	6.7451	7.3404	8.0482
26	3.682	4.2848	4.5983	4.9065	5.2322	5.5879	5.9844	6.4338	6.9515	7.5574
27	3.644	4.2128	4.5048	4.7894	5.0879	5.4110	5.7678	6.1681	6.6237	7.1498
28	3.610	4.1477	4.4208	4.6850	4.9600	5.2555	5.5791	5.9387	6.3439	6.8062
29	3.578	4.0887	4.3450	4.5913	4.8460	5.1177	5.4131	5.7388	6.1023	6.5128
30	3.549	4.0348	4.2762	4.5067	4.7437	4.9949	5.2661	5.5630	5.8917	6.2596
40	3.353	3.6781	3.8289	3.9670	4.1034	4.2423	4.3861	4.5365	4.6951	4.8636
60	3.178	3.3723	3.4569	3.5316	3.6030	3.6733	3.7436	3.8148	3.8872	3.9613
80	3.097	3.2355	3.2940	3.3447	3.3924	3.4388	3.4846	3.5302	3.5759	3.6221
10	3.050	3.1580	3.2025	3.2409	3.2766	3.3110	3.3447	3.3781	3.4113	3.4446
12	3.020	3.1081	3.1441	3.1748	3.2033	3.2307	3.2573	3.2835	3.3095	3.3354
∞	2.58	2.58	2.58	2.58	2.58	2.58	2.58	2.58	2.58	2.58

p-no.of predictors ***n-Sample Siz***

Table-2: Significant two-tail percentage points of Jackknife residual at

$$p\left(\left|R_i\right| > R_{i(p,n)}(0.01)\right) = 0.01$$

4. Numerical Results and Discussion

In this section, the authors have shown a numerical study of evaluating the outliers based on the Jack-Knife residual of the i^{th} observation in a regression model. For this, the authors have fitted step-wise linear regression models with different sets of predictors in a brand equity study. The data in the study comprised of 18 different attributes about a car brand and the data were 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 was over, only 275 completed questionnaires were used for analysis. Using the step-wise regression results, 4 nested models were extracted from the regression procedure by using IBM SPSS version 22. For each model, the Jack-Knife residuals were computed. The comparison of proposed approach with the classical approach of identifying the outliers is discussed through the following tables.

Model	***p***	***df*** *(n-p-2)*	**Classical approach**			**Proposed approach**		
			Critical $t(0.05)$	(n) $\lvert {}^*R_i \rvert > t(0.05)$	**Outliers**	**Critical** $R(0.05)$	(n) $\lvert R_i \rvert > R(0.05)$	**Outliers**
1	1	272	1.96872	13	59, 70 ,71 ,25 ,57 ,18 ,20 ,3 ,197 ,1 ,122 ,34 , 51	1.9826	13	59, 70 ,71 ,25 ,57 ,18 ,20 ,3 ,197 ,1 ,122 ,34 , 51
2	2	271	1.96876	13	70,71,59,57,25,1, 20,3,197,18,122, 51,34	1.9967	13	70,71,59,57,25,1,20,3,197,18,122, 51,34
3	3	270	1.96879	14	70,71,59,57,25,1, 20,3,197,18,122, 51,34, 244	2.0047	14	70,71,59,57,25,1,20,3,197,18,122, 51,34, 244
4	4	269	1.96882	13	70,71,59,57,25,1, 20,3,197,18,122, 51,34	2.0116	13	70,71,59,57,25,1,20,3,197,18,122, 51,34

p-no.of predictors *n*=275 *df*-degrees of freedom

Table-3: Identification of Outliers based on Classical and Proposed approach at 5% Significance level

Model	***p***	***df*** *(n-p-2)*	**Classical approach**			**Proposed approach**		
			Critical $t(0.01)$	(n) $\lvert {}^*R_i \rvert > t(0.01)$	**Outliers**	**Critical** $R(0.01)$	(n) $\lvert R_i \rvert > R(0.01)$	**Outliers**
1	1	272	2.90292	11	70,71,59,57,25, 20,18,197,3,1,122	2.9385	11	70,71,59,57,25, 20,18,197,3,1,122
2	2	271	2.90301	10	70,71,59,57,25, 1,20,3,197,18	2.9747	10	70,71,59,57,25, 1,20,3,197,18
3	3	270	2.90310	10	70,71,59,57,25, 197,1,20,3,51	2.9891	9	70,71,59,57,25, 197,1,20,3
4	4	269	2.90319	10	70,71,59,57,25, 197,1,20,3,51	3.0011	10	70,71,59,57,25, 197,1,20,3,51

p-no.of predictors n=275 df-degrees of freedom

Table-4: Identification of Outliers based on Classical and Proposed approach at 1% Significance level

Table-3 and Table-4 clearly visualize the result of the identification of outliers based on classical and proposed approach. From the 4 fitted multiple regression models, the classical approach helps to identify 13 outliers in model-1, model-2 and model-4 at 5% level of significance respectively. Similarly it helps to identify 11 outliers in model-1, 10 outliers in model-2, model-3 and model-4 at 1% level of significance. On the other hand, based on the proposed distribution of Jack-Knife residual, the authors identified 13 outliers in model-1, model-2 and model-4 at 5% level of significance. Similarly, in model-4 the authors identified 14 outliers at 5% level of significance. Likewise, the authors identified 11 outliers in model-1, 10 outliers in model-2 and model-4 and 9 outliers in model-3 respectively. From the above discussion the authors explored that the proposed distribution of the Jack-Knife residuals helps the authors to fix a different calibration point but the number of outliers identified based on this calibration point are more or less similar when we compare it with the classical approach. This shows that the proposed distribution of the Jack-Knife residuals is different from the student's t distribution, but the results will be same. Hence, the proposed distribution of the Jack-Knife residuals can be used as proxy to detect the outliers in the Y-space in a multiple linear regression model. The following graphs visualize the outliers at 5% and 1% level based on both the approaches.

Identification of outliers based on Classical approach at 5% level

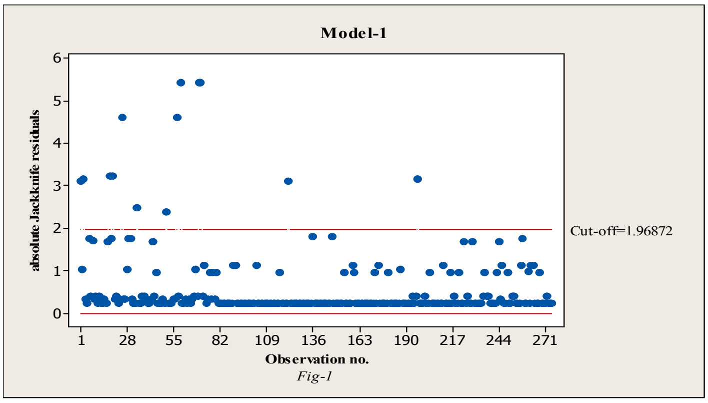

Fig-1

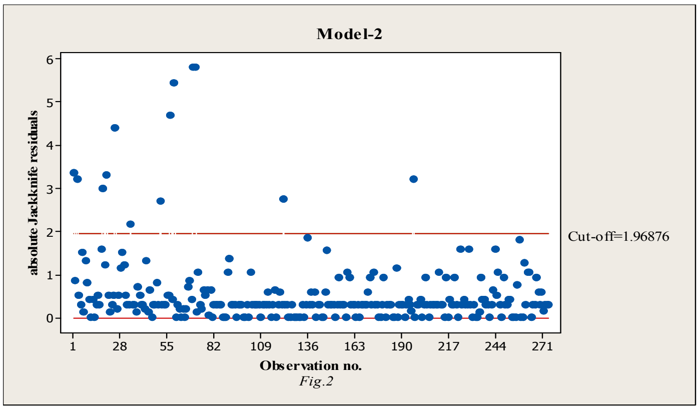

Fig.2

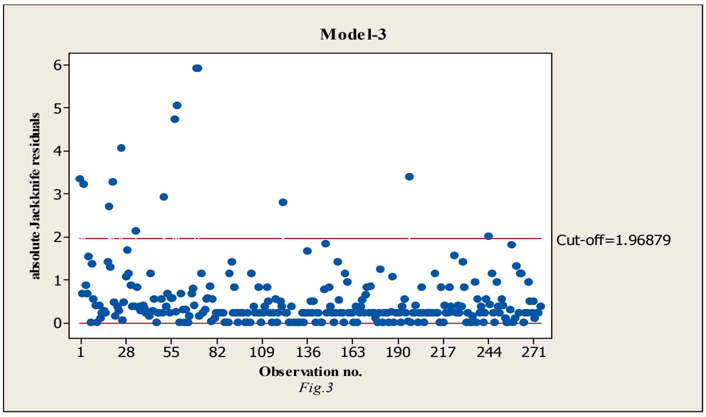

Fig.3

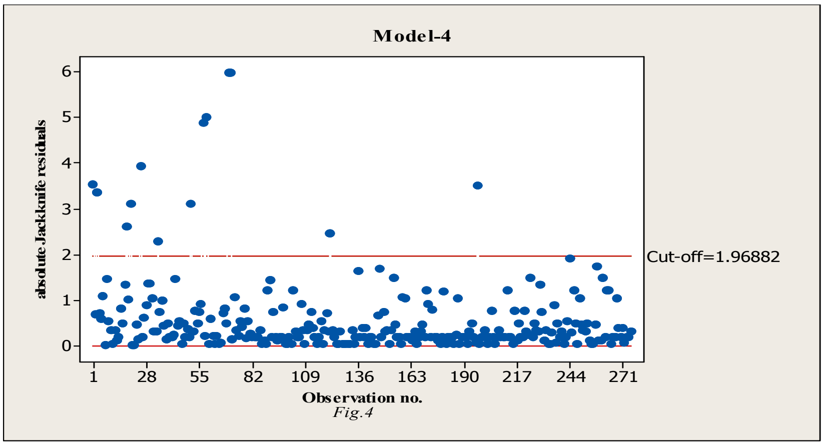

Fig.4

Identification of outliers based on Proposed approach at 5% level

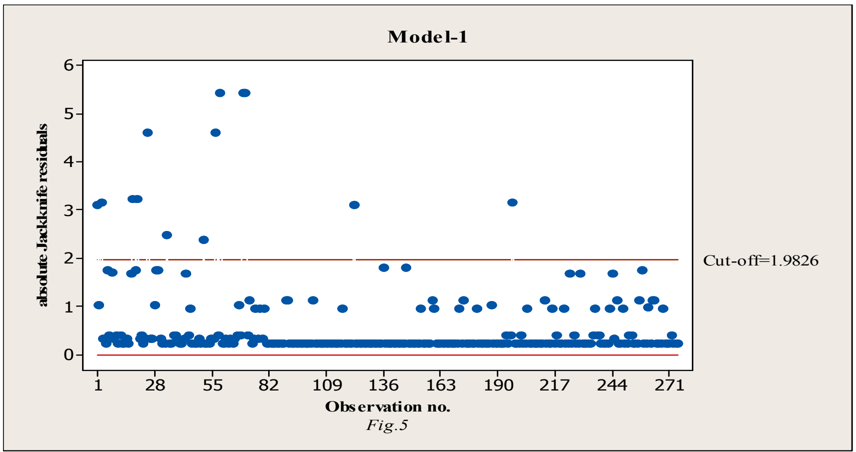

Fig.5

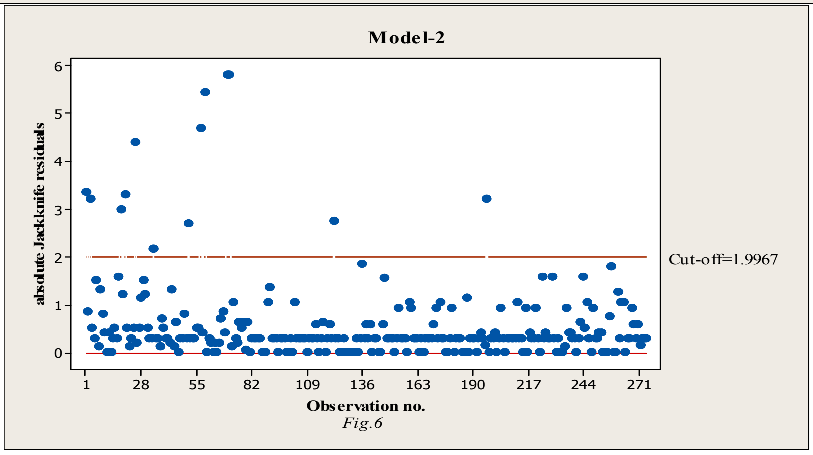

Fig.6

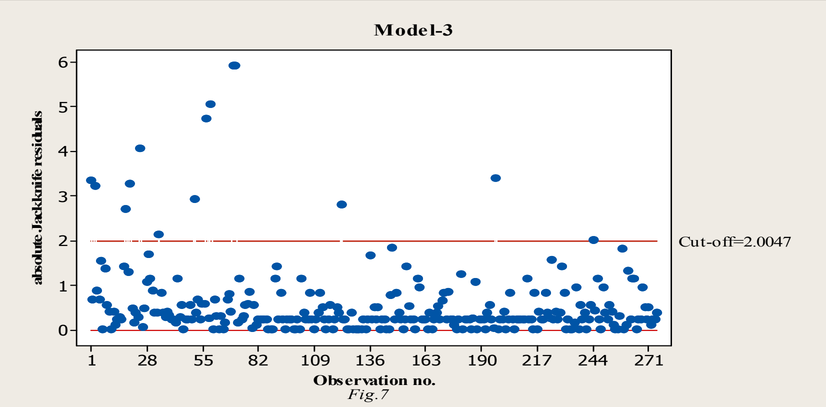

Fig.7

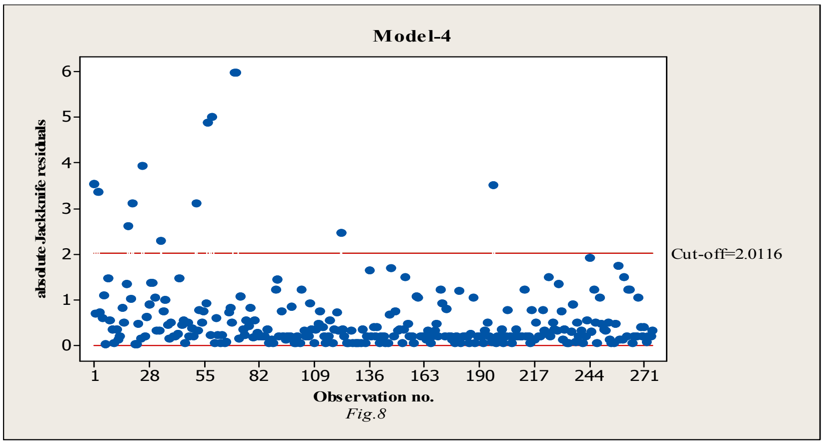

Fig.8

Identification of outliers based on Classical approach at 1% level

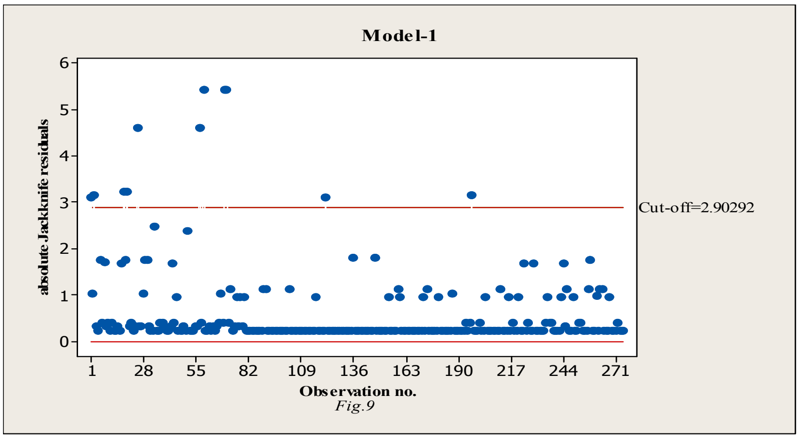

Fig.9

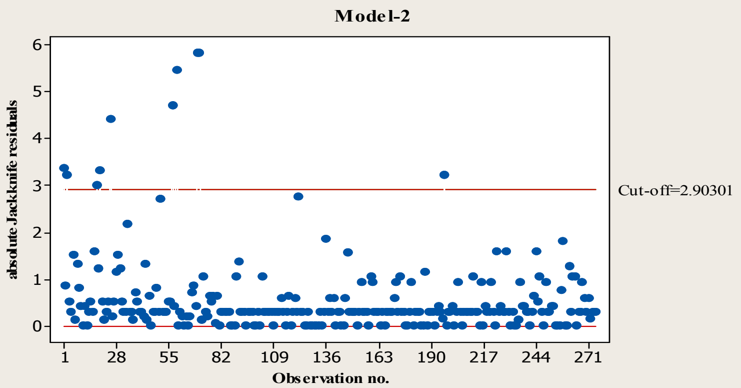

Fig.10

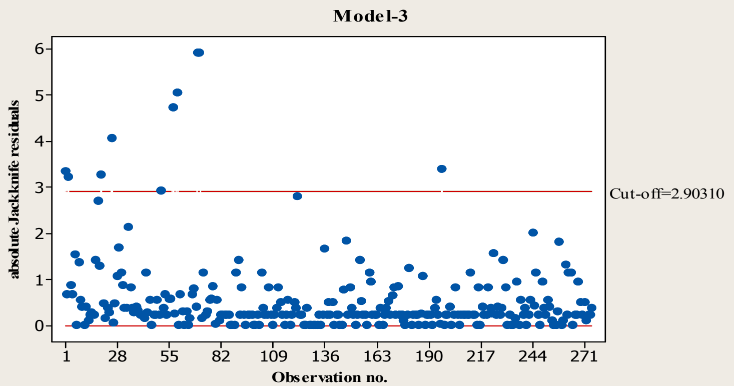

Fig.11

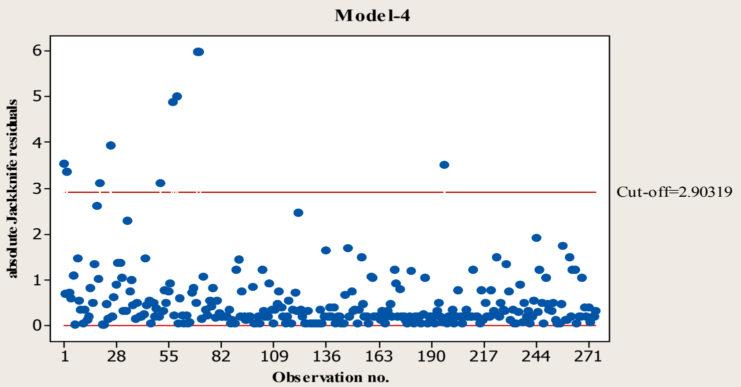

Fig.12

Identification of outliers based on Proposed approach at 1% level

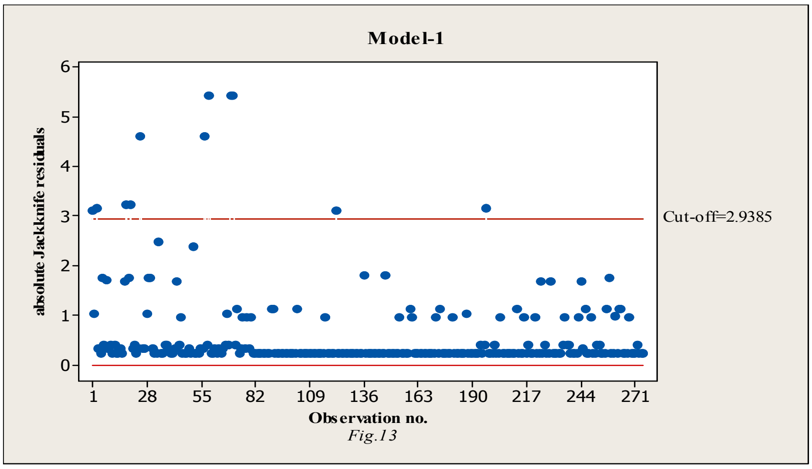

Fig.13

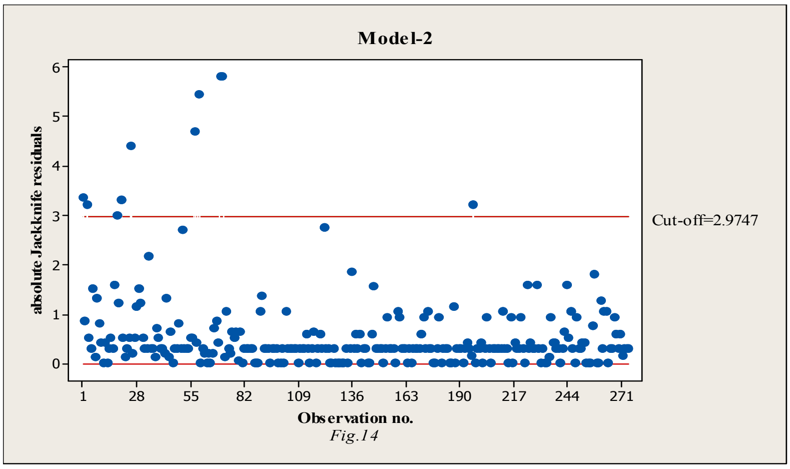

Fig.14

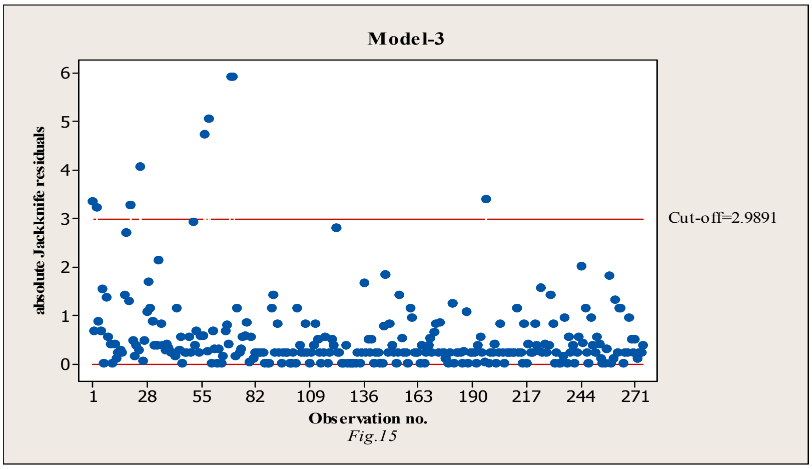

Fig.15

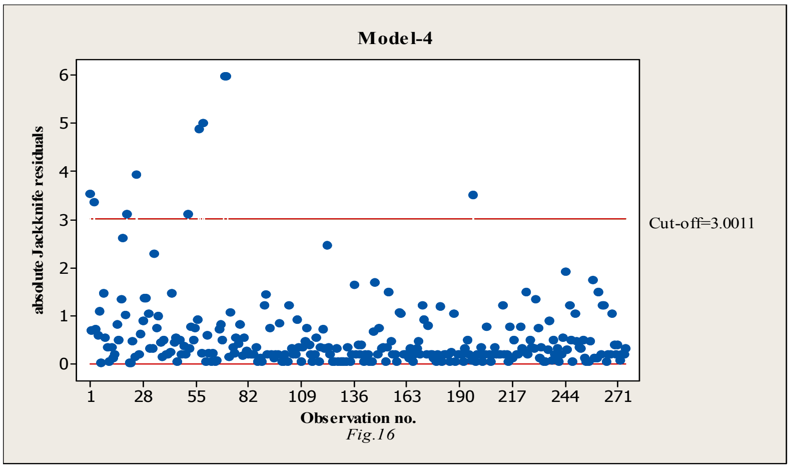

Fig.16

5. Conclusion

From the previous sections, the authors proposed the exact distribution of the Jack-Knife residuals which helps to evaluate the outliers in a multiple linear regression model. At first, the exact distribution of the Jack-Knife residual was derived and the authors visualized the density function in terms of series expression with shape parameters namely p and n. Moreover, the critical percentage points of Jack-Knife residuals at 5% and 1% levels of significance were also computed and are utilized to evaluate the outliers. Hence the authors conclude that the proposed distribution of the Jack-Knife residuals is non-student and we believe that it can be used as a proxy to identify the outliers in the functional data.

References

1. Ali S. Hadi (1992). A new measure of overall potential influence in linear regression, Computational Statistics & Data Analysis, 14(1), p. 1-27.
2. Behnken, D. W. and Draper, N. R. (1972). Residuals and their tariance patterns, Technometrics, 14, p. 101-111.
3. Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Wiley, New York.
4. Cook, R.D. (1977). Detection of influential observation in linear regression, Technometrics, 19, p. 15-18.
5. Chatterjee, S. and Hadi, A. S. (1988). Sensitivity Analysis in Linear Regression, New York: John Wiley and Sons.
6. Cook, R. D. and Weisberg, S. (1982). Residuals and Influence in Regression, Vol. 5, New York: Chapman and Hall.
7. Davies, R. B. and Hutton, B. (1975). The effects of errors in the independent variables in linear regression, Biometrika, 62, p. 383-391.
8. Díaz-García, J. A., and González-Farías, G. (2004). A note on the Cook's distance, Journal of Statistical Planning and Inference, 120(1), p. 119-136.
9. Eubank, R.L(1985). Diagnostics for smoothing splines, J. Roy. Statist. Soc. Ser. B, 47, p. 332–341.

10. Hoaglin, D.C, and Welsch, R.E. (1978). The Hat matrix in regression and ANOVA, The Amer. Statist., 32, p. 17-22.
11. Huber, P. J. (1975). Robustness and Designs, in A Survey of Statistical Design and Linear Models. North-Holland, Amsterdam.
12. Kendall, M.G., Stuart, A. (1973). The Advanced Theory of Statistics, Volume 2: Inference and Relationship, Griffin
13. Kim, C. (1996). Cook's distance in spline smoothing, Statist. Probab. Lett., 31, p. 139–144.
14. Kim, C. and Kim, W. (1998). Some diagnostics results in nonparametric density estimation, Comm. Statist., Theory Methods, 27, p. 291–303.
15. Kim, C., Lee, Y. and Park, B.U. (2001). Cook's distance in local polynomial regression, Statist. Probab. Lett., 54, p. 33–40.
16. Lund, R. E. (1975). Tables for an approximate test for outliers in linear models, Technometrics, 17, p. 473-476.
17. Silverman, B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion), J. Roy. Statist. Soc. Ser. B 47, p. 1–52.
18. Thomas, W (1991). Influence diagnostics for the cross-validated smoothing parameter in spline smoothing, J. Amer. Statist. Assoc., 86, p. 693–698.
19. Weisberg, S. (1980). Applied linear regression, New York: Wiley.