Hacettepe Journal of Mathematics and Statistics
Volume 48 (2) (2019), 536 – 551
Research Article

Exact distribution of Hadi's (H^2) influence measure and identification of potential outliers

G.S.David Sam Jayakumar*† and A. Sulthan‡

Abstract

This paper proposed an exact distribution of Hadi's influence measure that can be used to evaluate the potential outliers in a linear multiple regression analysis. The authors explored a relationship between the measure in terms of two independent F-ratios and they derived density function of the measure in a complicated series expression form with Gauss hyper-geometric function. Moreover, the first two moments of the distribution are derived in terms of Beta function and the authors computed the critical points of Hadi's measures at 5% and 1% significance level for different sample sizes and varying no. of predictors. Finally, the numerical example shows the identification of the potential outliers and the results extracted from the proposed approaches are more scientific, systematic and its exactness outperforms the Hadi's traditional approach.

Keywords: Hadi's measure, Potential outliers, Series expression form, Gauss hyper-geometric function, Moments, Beta function, Critical points.

Mathematics Subject Classification (2010): 62H10

Received : 14.12.2017 *Accepted* : 28.06.2018 *Doi* : 10.15672/HJMS.2018.614

*Assistant Professor, Jamal Institute of Management, Trichy 20, India Email: samjaya77@gmail.com

†Corresponding Author.

‡Research Scholar, Jamal Institute of Management, Trichy 20, India Email: contact@iamsulthan.in

1. Introduction and Related work

While fitting a regression model it is well understood that not all observations in a dataset play an equal role. Some observations have more impact than the others. Those observations which significantly influence the results of a regression analysis are called influential observations. Andrews and Pregibon [2] highlighted the need to find the outliers that matter. This means not all outliers need to be harmful in that they have an undue influence on the estimation of the parameters in the regression model. Hence, examining the residuals alone might not lead us to the detection of aberrant or unusual observations. Thus, other ways for finding influential observations are needed. Hoaglin and Welsch [9] discussed the importance of the projection matrix in linear regression, where the projection matrix is the matrix that projects onto the regression space. They argued that the diagonal elements of the projection matrix are important ingredients in influence analysis. The diagonal elements are referred to as leverages since they can be thought of as the amount of leverage concerning the response value on the corresponding predicted response value. Perhaps the most well-known influence measure was proposed by Cook [6], referred as Cook's distance. It is an influence measure used for assessing the influence of the observations on the estimated parameter vector in the linear regression model. It is widely used by practitioners for detecting influential observations. There are other influence measures to use in the linear regression analysis for assessing the influence of the observations on various results of the regression analysis. Such as, Andrews and Pregibon [2] derived a measure of the influence of an observation on the estimated parameters. This measure the AP statistic is based on the change in volume of confidence ellipsoids with and without a particular observation. Moreover, Belsley et al. [3] suggested an influence measure for assessing the influence of an observation on the variance of the estimated parameters in the linear regression model, known as COVRATIO. Besides the influence measures mentioned here there exist much more, see e.g. Chatterjee and Hadi [5] and Hadi [8] for excellent overviews of influence measures. Graphical investigation of data is a powerful tool in exploratory analysis. It can be used to examine relationships between variables and discover observations deviating from other. Hence, influential observations can also be detected using graphical tools. Mosteller and Tukey [11] introduced the added variable plot, which is used for graphically detecting observations that have a large influence on the parameter estimates. For details concerning the added variable plot, such as construction and properties, see, Belsley et al. [3], where the plot is referred to as the partial regression leverage plot, and Cook and Weisberg [7]. Other results on graphical tools in influence analysis are provided by Johnson and McCulloch [10]. It is important to note that the graphical tools used in influence analysis are not conclusive, but rather suggestive. From the previous discussions, we can see that the 1970's and the 1980's were the decades when most research results on influence analysis in linear regression came to see the light. However, influence analysis in linear regression is still an active research area. Nurunnabi et al. [12] proposed a modification of Cook's distance. This modification enables the identification of multiple influential observations. Furthermore, Beyaztas and Alin [4] used a combined Bootstrap and Jackknife algorithm to detect influential observations. In applied data analysis, there is an increasing availability of data sets containing a large number of variables. When such data is in the hands of the researcher sparse regression can be implemented, which is another field of research active today. In sparse regression, a penalty term on the regression parameters is added which shrinks the number of parameters. Common approaches to estimate the parameter in the sparse regression are, however, sensitive to influential observations and new methods are needed. Alfons et al. [1] and Park et al. [13] proposed robust

estimation methods, where influential observations are not harmful to the resulting estimates. Considering the above reviews, the authors proposed the exact distribution of Hadi's influence function $\left(H_i^2\right)$ which exactly identifies the influential data points and it is discussed in the subsequent sections.

2. Relationship between Hadi's $\left(H^2\right)$ influence measure and F-ratios

The multiple linear regression models with random error is given by

$$Y = X\beta + e \tag{2.1}$$

where $\underset{(nX1)}{Y}$ is the matrix of the dependent variable, $\underset{(kX1)}{\beta}$ is the vector of beta coefficients or partial regression coefficients and $\underset{(nX1)}{e}$ is the residual followed normal distribution N $(0,\sigma_e^2 I_n)$. From (2.1), 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 Hadi's $\left(H_i^2\right)$ influence measure is the interesting technique based on the simple fact that potentially influential observations are outliers in X-space, Y-space or both. The general form of the Hadi's influence measure of the i^{th} observation is given by

$$H_i^2 = \frac{(p+1)\,\hat{e}_i^2}{(1-h_{ii})\left(\hat{e}^t\,\hat{e} - \hat{e}_i^2\right)} + \frac{h_{ii}}{1-h_{ii}} \tag{2.2}$$

Where $\hat{e}_i^2$ is the vector of squared estimated residuals, p is the no.of predictors, $\hat{e}^t\,\hat{e}$ is the sum of the squared estimated residuals and h_{ii} is the hat values of i^{th} observation or diagonal elements of the hat matrix$(H = X(X^{'}X)^{-1}X^{'})$. This diagnostic measure is the sum of two components each of which has an interpretation. A large value for the first term indicates that the model has a poor fit (a large prediction error) and a large value for the second term indicates the presence of an outlier in the X-space. Similarly, Hadi pointed this diagnostic measure possess several desirable properties and it is also supplemented by a graphical display which shows the source of influence. He suggested,$\left(H_i^2\right)$ for observations more than a cut-off of $E\left(H_i^2\right)+c\sqrt{V\left(H_i^2\right)}$ which is treated as a potential outlier. Hadi's influence measure can also be written in an alternative form as

$$H_i^2 = \frac{p+1}{(1-h_{ii})\left(\left(\hat{e}^t\,\hat{e}\,/\,\hat{e}_i^2\right)-1\right)} + \frac{h_{ii}}{1-h_{ii}} \tag{2.3}$$

It is known the unbiased estimate of the true error variance is $s^2 = \hat{e}^t\,\hat{e}\,/n-p-1$ and substitute $\hat{e}^t\,\hat{e} = s^2\,(n-p-1)$ in (2.3) to get

$$H_i^2 = \frac{p+1}{\frac{n-p-1}{\hat{e}_i^2\,/s^2(1-h_{ii})} - (1-h_{ii})} + \frac{h_{ii}}{1-h_{ii}} \tag{2.4}$$

Rewrite (2.4), in terms of the internally studentized residual (r_i)which is equal to $\hat{e}_i\,/s\sqrt{1-h_{ii}}$ and it is given as

$$H_i^2 = \frac{p+1}{((n-p-1)\,/r_i^2) - (1-h_{ii})} + \frac{h_{ii}}{1-h_{ii}} \tag{2.5}$$

Though Hadi's influence measure is scientific and the yardstick used to detect the influential observation is not scientific and the authors believe it is based on the rule of thumb approach. Because $\left(H_i^2\right)$ is non-normally distributed and the usage of mean and variance in the cut-off $\left(E\left(H_i^2\right)+c\sqrt{V\left(H_i^2\right)}\right)$ is meaningless and illogic. Secondly, when using the cut-off, it is not recommended by the author to use a specific and fixed value for the constant (c). Finally, the usage of plots and graphs to identify the potential outliers and sources of influence leads to imprecision and ambiguity.In order to overcome this rule of thumb approach of identifying the influential observations, authors proposed the exact distribution for Hadi's influence measure and established a scientific yardstick to scrutinize the exact influential observations. For this, authors utilize the relationship among the Hadi's $\left(H_i^2\right)$, internally studentized residual(r_i) and hat elements(h_{ii}).The terms (r_i) and (h_{ii}) are independent because the computation of (r_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 r_i and h_{ii} are also uncorrelated and independent. Using this assumption, authors first determine the distribution of (r_i)based on the relationship given by Weisberg (1980) as

$$t_i = r_i\sqrt{\frac{n-p-2}{(n-p-1)-r_i^2}} \sim t_{(n-p-2)} \tag{2.6}$$

From (2.6) it follows student's t- distribution with $(n-p-2)$ degrees of freedom and it can be written in terms of the F-ratio as

$$r_i^2 = \frac{(n-p-1)t_i^2}{(n-p-2)+t_i^2}$$

$$r_i^2 = \frac{(n-p-1)F_{i(1,n-p-2)}}{(n-p-2)+F_{i(1,n-p-2)}} \tag{2.7}$$

From (2.7), if t_i follows student's t- distribution with $(n-p-2)$ degrees of freedom, then t_i^2 follows $F_{(1,n-p-2)}$ distribution with $(1, n-p-2)$ degrees of freedom. Similarly, authors identify the distribution of h_{ii} based on the relationship proposed by Belsey et al [3] and they showed when the set of predictors is multivariate normal with(μ_X, Σ_X), then

$$\frac{(n-p)\left(h_{ii}-1/n\right)}{(p-1)(1-h_{ii})} \sim F_{(p-1,n-p)} \tag{2.8}$$

From (2.8) it follows F-distribution with $(p-1, n-p)$ degrees of freedom and it can be written in an alternative form as

$$E\left(H_i^2\right) = (p+1)\left(\phi_1\left(p,n\right)\right) + \frac{n}{n-1}\left(\phi_2\left(p,n\right)\right) - 1 \tag{2.9}$$

In order to derive the exact distribution of $\left(H_i^2\right)$, substitute (2.7) and (2.9) in (2.5), authors get the Hadi's $\left(H_i^2\right)$measure in terms of the two independent F-ratios with $(1, n-p-2)$ and $(p-1, n-p)$ degrees of freedom respectively and the relationship is given as

$$H_i^2 = \frac{p+1}{\frac{(n-p-2)+F_{i(1,n-p-2)}}{F_{i(1,n-p-2)}} - \frac{(n-1)/n}{1+((p-1)/(n-p))F_{i(p-1,n-p)}}} + \frac{\left((p-1)/(n-p)\right)F_{i(p-1,n-p)}+1/n}{(n-1)/n} \tag{2.10}$$

$$H_i^2 = \frac{p+1}{1+(n-p-2)F_{i(n-p-2,1)} - \frac{(n-1)/n}{1+((p-1)/(n-p))F_{i(p-1,n-p)}}} + \frac{\left(1+((p-1)/(n-p))\,F_{i(p-1,n-p)}\right)-(n-1)/n}{(n-1)/n} \tag{2.11}$$

From (2.11), it can be further simplified and $\left(H_i^2\right)$ is expressed in terms of two independent beta variables namely θ_{1i} andθ_{2i} of the first kind by using the following facts

$$\frac{1}{1+(n-p-2)\,F_{i(n-p-2,1)}} = \theta_{1i} \sim \beta_1\left(\frac{1}{2}, \frac{n-p-2}{2}\right) \tag{2.12}$$

$$\frac{1}{1+((p-1)\,/\,(n-p))\,F_{i(p-1,n-p)}} = \theta_{2i} \sim \beta_1\left(\frac{n-p}{2}, \frac{p-1}{2}\right) \tag{2.13}$$

Then, without loss of generality (2.11) can be written as

$$H_i^2 = \frac{p+1}{(1/\theta_{1i}) - ((n-1)\,/n)\,\theta_{2i}} + \frac{(1/\theta_{2i}) - (n-1)\,/n}{(n-1)\,/n} \tag{2.14}$$

$$H_i^2 = \frac{(p+1)\,\theta_{1i}}{1-((n-1)\,/n)\,\theta_{1i}\theta_{2i}} + (n/n-1)\,(1/\theta_{2i}) - 1 \tag{2.15}$$

From (2.15), the authors showed the Hadi's influence measure in terms of $\theta_{1i} \sim \beta_1\left(\frac{1}{2}, \frac{n-p-2}{2}\right)$ and $\theta_{2i} \sim \beta_1\left(\frac{n-p}{2}, \frac{p-1}{2}\right)$ which followed beta distribution of first kind with two shape parameters p and n respectively. To avoid complexity further, the relationship from (2.15) modified as

$$\frac{n-1}{n}\left(1+H_i^2\right) = \left(\frac{1+p\,((n-1)\,/n)\,\theta_{1i}\theta_{2i}}{1-((n-1)\,/n)\,\theta_{1i}\theta_{2i}}\right)(1/\theta_{2i}) = \psi_i \tag{2.16}$$

Based on the identified relationship from (2.16), the authors derived the distribution of the Hadi's $\left(H_i^2\right)$ and it is discussed in the next section.

3. Exact Distribution of Hadi's (H_i^2)

Using the technique of two-dimensional Jacobian of transformation, the joint probability density function of the two Beta variables of kind-1 namelyθ_{1i},θ_{2i} were transformed into density function of ψ_i and it is given as

$$f\left(\psi_i, u_i\right) = f\left(\theta_{1i}, \theta_{2i}\right)|J| \tag{3.1}$$

From (3.1), It is known that$\theta_{1i}$and θ_{2i} are independent then rewrite (3.1) as

$$f\left(\psi_i, u_i\right) = f\left(\theta_{1i}\right) f\left(\theta_{2i}\right)|J| \tag{3.2}$$

Using the change of variable technique, substitute $\theta_{2i} = u_i$ in (2.16) to get

$$\theta_{1i} = \frac{\psi_i u_i - 1}{((n-1)\,/n)\,u_i\,(p+\psi_i u_i)} \tag{3.3}$$

Then partially differentiate (3.3) and compute the Jacobian determinant in (3.2) as

$$f\left(\psi_i, u_i\right) = f\left(\theta_{1i}\right) f\left(\theta_{2i}\right)\left|\frac{\partial\left(\theta_{1i}, \theta_{2i}\right)}{\partial(\psi_i, u_i)}\right| \tag{3.4}$$

$$f(\psi_i,u_i)=f(\theta_{1i})\,f(\theta_{2i})\begin{vmatrix}\frac{\partial\theta_{1i}}{\partial\psi_i} & \frac{\partial\theta_{1i}}{\partial u_i}\\ \frac{\partial\theta_{2i}}{\partial\psi_i} & \frac{\partial\theta_{2i}}{\partial u_i}\end{vmatrix} \tag{3.5}$$

From (3.5), It is known that the θ_{1i} and $\theta_{2i}$are independent, then the density function of the joint distribution of θ_{1i} and θ_{1i} is given as

$$f(\theta_{1i},\theta_{2i})=\frac{1}{\mathrm{B}\left(\frac{1}{2},\frac{n-p-2}{2}\right)}\theta_{1i}^{\frac{1}{2}-1}(1-\theta_{1i})^{\frac{n-p-2}{2}-1}\times\frac{1}{\mathrm{B}\left(\frac{n-p}{2},\frac{p-1}{2}\right)}\theta_{2i}^{\frac{n-p}{2}-1}(1-\theta_{2i})^{\frac{p-1}{2}-1} \tag{3.6}$$

where $0\le\theta_{1i},\theta_{2i}\le 1$,$n,p>0$
and

$$\begin{vmatrix}\frac{\partial\theta_{1i}}{\partial\psi_i} & \frac{\partial\theta_{1i}}{\partial u_i}\\ \frac{\partial\theta_{2i}}{\partial\psi_i} & \frac{\partial\theta_{2i}}{\partial u_i}\end{vmatrix}=\begin{vmatrix}\frac{p+1}{((n-1)/n)(p+\psi_i u_i)} & \frac{p-\psi_i^2u_i^2+2\psi_i u_i}{((n-1)/n)u_i^2(p+\psi_i u_i)^2}\\ 0 & 1\end{vmatrix}=\frac{p+1}{((n-1)/n)(p+\psi_i u_i)^2} \tag{3.7}$$

Then substitute (3.6) and (3.7) in (3.5) in terms of the substitution of u_i, to get the joint distribution of ψ_i and u_i as

$$f(\psi_i,u_i)=\frac{1}{\mathrm{B}\left(\frac{1}{2},\frac{n-p-2}{2}\right)}\left(\frac{\psi_i u_i-1}{\left(\frac{n-1}{n}\right)u_i(p+\psi_i u_i)}\right)^{\frac{1}{2}-1}\left(1-\frac{\psi_i u_i-1}{\left(\frac{n-1}{n}\right)u_i(p+\psi_i u_i)}\right)^{\frac{n-p-2}{2}-1}\times\frac{1}{\mathrm{B}\left(\frac{1}{2},\frac{n-p-2}{2}\right)}u_i^{\frac{n-p-2}{2}-1}(1-u_i)^{\frac{1}{2}-1}\times|J| \tag{3.8}$$

where $\frac{n-1}{n}\le\psi_i<\infty,0\le u_i\le 1$ and $|J|=\frac{p+1}{((n-1)/n)(p+\psi_i u_i)^2}$

Rearrange (3.8) and integrate with respect tou_i, to get the marginal distribution of ψ_i *as*

$$f(\psi_i;p,n)=\alpha(p,n)\left(\sum_{r=0}^{\frac{n-p-2}{2}-1}\sum_{s=0}^{r+\frac{1}{2}-1}\binom{\frac{n-p-2}{2}-1}{r}\binom{r+\frac{1}{2}-1}{s}\left(\tfrac{n}{n-1}\right)^{r+\frac{1}{2}-1}p^{-\left(r+\frac{3}{2}\right)}\times(-1)^{2r+s+\frac{1}{2}-1}\psi_i^s\int_0^1 u_i^{\frac{n-p-3}{2}+s-r-1}(1-u_i)^{\frac{p-1}{2}-1}\left(\tfrac{1}{1+(\psi_i/p)u_i}\right)^{r+\frac{3}{2}}du_i\right) \tag{3.9}$$

where $\frac{n-1}{n}\le\psi_i<\infty$ and $\alpha(p,n)=\frac{p+1}{\left(\frac{n-1}{n}\right)B\left(\frac{1}{2},\frac{n-p-2}{2}\right)B\left(\frac{n-p}{2},\frac{p-1}{2}\right)}$

It is known, from (3.9)

$$\int_0^1 u_i^{\frac{n-p-3}{2}+s-r-1}(1-u_i)^{\frac{p-1}{2}-1}\left(\frac{1}{1+(\psi_i/p)u_i}\right)^{r+\frac{3}{2}}du_i=\frac{\Gamma\left(\frac{p-1}{2}\right)}{(\psi_i/p)\Gamma\left(r+\frac{3}{2}\right)}\left(\Omega_1(\psi_i;p,n,r,s)+\Omega_2(\psi_i;p,n,r,s)\right) \tag{3.10}$$

where

$$\Omega_1(\psi_i;p,n,r,s)=(\psi_i/p)^{-\left(\frac{n-p-3}{2}+s-r\right)+1}\frac{\Gamma\left(\frac{n-p-3}{2}+s-r\right)\Gamma\left(2r+\frac{3}{2}-\left(\frac{n-p-3}{2}+s\right)\right)}{\Gamma\left(\frac{p-1}{2}\right)}$$

$$ {}_2F_1\left(\frac{n-p-3}{2}+s-r,\frac{p+1}{2};1-\left(r+\frac{3}{2}\right)+\frac{n-p-3}{2}+s-r;\frac{1}{\psi_i/p}\right) $$

$$ \Omega_2\left(\psi_i;p,n,r,s\right)=\left(\psi_i/p\right)^{-\left(r+\frac{3}{2}\right)+1}\frac{\Gamma\left(r+\frac{3}{2}\right)\Gamma\left(\frac{n-p}{2}+s-2r\right)}{\Gamma\left(\frac{p-1}{2}\right)} $$

$$ {}_2F_1\left(r+\frac{3}{2},1-\left(\frac{n-7}{2}+s-2r\right);1+r+\frac{3}{2}-\left(\frac{n-p-3}{2}+s-r\right);-\frac{1}{\psi_i/p}\right) $$

Then substitute (3.10) in (3.9) and arrange the terms, to get the density function of ψ_i in the series expression form as

$$ \begin{aligned} f\left(\psi_i;p,n\right)=\lambda\left(p,n\right)\Bigg(&\sum_{r=0}^{\frac{n-p-2}{2}-1}\sum_{s=0}^{r+\frac{1}{2}-1}\binom{\frac{n-p-2}{2}-1}{r}\binom{r+\frac{1}{2}-1}{s}\left(\tfrac{n}{n-1}\right)^{r+\frac{1}{2}-1}p^{-\left(r+\frac{3}{2}\right)}\\ &\times(-1)^{2r+s+\frac{1}{2}-1}\psi_i^{s-1}\tfrac{1}{\Gamma\left(r+\frac{3}{2}\right)}\left(\Omega_1\left(\psi_i;p,n,r,s\right)+\Omega_2\left(\psi_i;p,n,r,s\right)\right)\Bigg) \end{aligned} \tag{3.11} $$

where, $(n-1)/n\le\psi_i<\infty$,$n,p>0$, $n>p$ and $\lambda(p,n)=(p+1)\,\alpha(p,n)=\frac{(p+1)\Gamma\left(\frac{p-1}{2}\right)}{\left(\frac{n-1}{n}\right)B\left(\frac{1}{2},\frac{n-p-2}{2}\right)B\left(\frac{n-p}{2},\frac{p-1}{2}\right)}$ is the normalizing constant as a function of p *and* n. In order to derive the density function of Hadi's measure, the authors again utilize the relationship between ψ_i and $\left(H_i^2\right)$. It is known $H_i^2>0$, then $(n-1)/n\le\psi_i<\infty$.Hence from (2.16), using one-dimensional Jacobian of transformation, the density function of $\left(H_i^2\right)$ can be written as

$$ f\left(H_i^2\right)=f\left(\psi_i\right)|J| \tag{3.12} $$

$$ f\left(H_i^2\right)=f\left(\psi_i\right)\left|\frac{d\psi_i}{dH_i^2}\right| \tag{3.13} $$

Then substitute $\psi_i=\frac{n-1}{n}\left(1+H_i^2\right)$,$\frac{d\psi_i}{dH_i^2}=\frac{n-1}{n}$ and (3.11) in (3.13), to get the final form of the density function of $\left(H_i^2\right)$ as

$$ \begin{aligned} &f\left(H_i^2;p,n\right)=\\ &\varphi\left(p,n\right)\Bigg(\sum_{r=0}^{\frac{n-p-2}{2}-1}\sum_{s=0}^{r+\frac{1}{2}-1}\binom{\frac{n-p-2}{2}-1}{r}\binom{r+\frac{1}{2}-1}{s}\left(\Gamma\left(r+\tfrac{3}{2}\right)\right)^{-1}\left(\tfrac{n}{n-1}\right)^{r-s+\frac{1}{2}}\\ &p^{-\left(r+\frac{3}{2}\right)+1}(-1)^{2r+s+\frac{1}{2}-1}\left(1+H_i^2\right)^{s-1}\left(\Omega_1\left(H_i^2;p,n,r,s\right)+\Omega_2\left(H_i^2;p,n,r,s\right)\right)\Bigg) \end{aligned} \tag{3.14} $$

where,$0\le H_i^2<\infty$,$n,p>0$,$n>p$ and $\varphi(p,n)=\frac{n-1}{n}\lambda(p,n)=\frac{(p+1)\Gamma\left(\frac{p-1}{2}\right)}{B\left(\frac{1}{2},\frac{n-p-2}{2}\right)B\left(\frac{n-p}{2},\frac{p-1}{2}\right)}$

$$ \begin{aligned} \Omega_1\left(H_i^2;p,n,r,s\right)=&\frac{\Gamma\left(\frac{n-p-3}{2}+s-r\right)\Gamma\left(2r+\frac{3}{2}-\left(\frac{n-p-3}{2}+s\right)\right)}{\Gamma\left(\frac{p-1}{2}\right)}\left(\tfrac{n-1}{np}\left(1+H_i^2\right)\right)^{-\left(\frac{n-p-3}{2}+s-r\right)+1}\\ &\times{}_2F_1\left(\tfrac{n-p-3}{2}+s-r,\tfrac{p+1}{2};1-\left(r+\tfrac{3}{2}\right)+\tfrac{n-p-3}{2}+s-r;-\tfrac{1}{\frac{n-1}{np}\left(1+H_i^2\right)}\right) \end{aligned} $$

$$ \begin{aligned} \Omega_2\left(H_i^2;p,n,r,s\right)=&\frac{\Gamma\left(r+\frac{3}{2}\right)\Gamma\left(\frac{n-p}{2}+s-2r\right)}{\Gamma\left(\frac{n-7}{2}+s-2r\right)}\left(\tfrac{n-1}{np}\left(1+H_i^2\right)\right)^{-\left(r+\frac{3}{2}\right)+1}\\ &\times{}_2F_1\left(r+\tfrac{3}{2},1-\left(\tfrac{n-7}{2}+s-2r\right);1+r+\tfrac{3}{2}-\left(\tfrac{n-p-3}{2}+s-r\right);-\tfrac{1}{\frac{n-1}{np}\left(1+H_i^2\right)}\right) \end{aligned} $$

From (3.14), it is the density function of Hadi's $\left(H_i^2\right)$ influence measure which involves the following such as $\Omega_1\left(H_i^2;p,n,r,s\right)$,$\Omega_2\left(H_i^2;p,n,r,s\right)$are the auxiliary functions,${}_2F_1$ is the Gauss hypergeometric function and the normalizing constant $\lambda(p,n)$comprised of Beta and Gamma functions ($B(\frac{1}{2},\frac{n-p-2}{2})$,$B(\frac{p-1}{2},\frac{n-p}{2})$,$\Gamma\left(\frac{p-1}{2}\right)$) with two shape parameters (p,n),n is the sample size and p is the no. of predictors used in a multiple linear

regression model. In order to know the location and dispersion of Hadi's$\left(H_i^2\right)$, the authors derived the first two moments in terms of mean, variance from and it is shown as follows. Using (2.15), rewrite in terms of series expression form as

$$H_i^2 = (p+1)\left(\sum_{k=0}^{\infty}\left(\frac{n-1}{n}\right)^k \theta_{1i}^{k+1}\theta_{2i}^{k}\right) + \frac{n}{n-1}\left(\frac{1}{\theta_{2i}}\right) - 1 \tag{3.15}$$

Now take expectation and substitute the moments of two independent beta variables θ_{1i} and$\theta_{2i}$of kind-1, to get the first moment of $\left(H_i^2\right)$ as

$E\left(H_i^2\right) = (p+1)\left(\sum_{k=0}^{\infty}\left(\frac{n-1}{n}\right)^k E\left(\theta_{1i}^{k+1}\right) E\left(\theta_{2i}^{k}\right)\right) + \frac{n}{n-1}E\left(\frac{1}{\theta_{2i}}\right) - 1$

$$E\left(H_i^2\right) = (p+1)\left(\phi_1(p,n)\right) + \frac{n}{n-1}\left(\phi_2(p,n)\right) - 1 \tag{3.16}$$

Where $\phi_1(p,n) = \sum_{k=0}^{\infty}\left(\frac{n-1}{n}\right)^k \left(\frac{B\left(k+\frac{3}{2},\frac{n-p-2}{2}\right)}{B\left(\frac{1}{2},\frac{n-p-2}{2}\right)}\right)\left(\frac{B\left(\frac{n-p}{2}+k,\frac{p-1}{2}\right)}{B\left(\frac{n-p}{2},\frac{p-1}{2}\right)}\right)$

$\phi_2(p,n) = \frac{B\left(\frac{n-p-2}{2},\frac{p-1}{2}\right)}{B\left(\frac{n-p}{2},\frac{p-1}{2}\right)}$ and B is the beta function respectively.

From (2.15), rewrite and square both sides, then take expectation, to get the second moment of $\left(H_i^2\right)$ as

$$\left(H_i^2+1\right)^2 = (p+1)^2\theta_{1i}^2\left(\frac{1}{1-\left(\frac{n-1}{n}\right)\theta_{1i}\theta_{2i}}\right)^2 + \left(\frac{n}{n-1}\right)^2\left(\frac{1}{\theta_{2i}}\right)^2 + \frac{2n(p+1)}{n-1}\left(\frac{\theta_{1i}}{\theta_{2i}}\right)\frac{1}{1-\left(\frac{n-1}{n}\right)\theta_{1i}\theta_{2i}}$$

$$\left(H_i^2+1\right)^2 = (p+1)^2\left(\sum_{k=0}^{\infty}(k+1)\left(\frac{n-1}{n}\right)^k\theta_{1i}^{k+2}\theta_{2i}^{k}\right) + \left(\frac{n}{n-1}\right)^2\left(\frac{1}{\theta_{2i}}\right)^2 + \frac{2n(p+1)}{n-1}\left(\sum_{k=0}^{\infty}\left(\frac{n-1}{n}\right)^k\theta_{1i}^{k+1}\theta_{2i}^{k-1}\right)$$

$$E\left(H_i^2+1\right)^2 = (p+1)^2\left(\sum_{k=0}^{\infty}(k+1)\left(\frac{n-1}{n}\right)^k E\left(\theta_{1i}^{k+2}\right)E\left(\theta_{2i}^{k}\right)\right) + \left(\frac{n}{n-1}\right)^2 E\left(\frac{1}{\theta_{2i}}\right)^2 + \frac{2n(p+1)}{n-1}\left(\sum_{k=0}^{\infty}\left(\frac{n-1}{n}\right)^k E\left(\theta_{1i}^{k+1}\right)E\left(\theta_{2i}^{k-1}\right)\right)$$

$$E\left(H_i^2\right)^2 = (p+1)^2\left(\sum_{k=0}^{\infty}(k+1)\left(\frac{n-1}{n}\right)^k E\left(\theta_{1i}^{k+2}\right)E\left(\theta_{2i}^{k}\right)\right) + \left(\frac{n}{n-1}\right)^2 E\left(\frac{1}{\theta_{2i}}\right)^2 + \frac{2n(p+1)}{n-1}\left(\sum_{k=0}^{\infty}\left(\frac{n-1}{n}\right)^k E\left(\theta_{1i}^{k+1}\right)E\left(\theta_{2i}^{k-1}\right)\right) - 2E\left(H_i^2\right) - 1 \tag{3.17}$$

Therefore, It is known

$$V(H_i^2) = E\left(H_i^2\right)^2 - \left(E\left(H_i^2\right)\right)^2 \tag{3.18}$$

Then substitute (3.16) and (3.17) in (3.18), to get the variance of $\left(H_i^2\right)$ as

$$V\left(H_i^2\right) = (p+1)^2\left(\Phi_1(p,n)\right) + \left(\frac{n}{n-1}\right)^2\left(\Phi_2(p,n)\right) + \frac{2n(p+1)}{n-1}\left(\Phi_3(p,n)\right)$$

where

$$\Phi_1(p,n) = \left(\sum_{k=0}^{\infty}(k+1)\left(\frac{n-1}{n}\right)^k \frac{B\left(k+\frac{5}{2},\frac{n-p-2}{2}\right)B\left(\frac{n-p}{2}+k,\frac{p-1}{2}\right)}{B\left(\frac{1}{2},\frac{n-p-2}{2}\right)B\left(\frac{n-p}{2},\frac{p-1}{2}\right)}\right) - \left(\sum_{k=0}^{\infty}\left(\frac{n-1}{n}\right)^k \frac{B\left(k+\frac{3}{2},\frac{n-p-2}{2}\right)B\left(\frac{n-p}{2}+k,\frac{p-1}{2}\right)}{B\left(\frac{1}{2},\frac{n-p-2}{2}\right)B\left(\frac{n-p}{2},\frac{p-1}{2}\right)}\right)^2$$

$$\Phi_2(p,n) = \frac{B\left(\frac{n-p-4}{2}, \frac{p-1}{2}\right)}{B\left(\frac{n-p}{2}, \frac{p-1}{2}\right)} - \left(\frac{B\left(\frac{n-p-2}{2}, \frac{p-1}{2}\right)}{B\left(\frac{n-p}{2}, \frac{p-1}{2}\right)}\right)^2$$

$$\Phi_3(p,n) = \left(\sum_{k=0}^{\infty}\left(\frac{n-1}{n}\right)^k \frac{B\left(k+\frac{3}{2}, \frac{n-p-2}{2}\right)B\left(\frac{n-p-2}{2}+k, \frac{p-1}{2}\right)}{B\left(\frac{1}{2}, \frac{n-p-2}{2}\right)B\left(\frac{n-p}{2}, \frac{p-1}{2}\right)}\right) - \left(\frac{B\left(\frac{n-p-2}{2}, \frac{p-1}{2}\right)}{B\left(\frac{n-p}{2}, \frac{p-1}{2}\right)}\right)\left(\sum_{k=0}^{\infty}\left(\frac{n-1}{n}\right)^k \frac{B\left(k+\frac{3}{2}, \frac{n-p-2}{2}\right)B\left(\frac{n-p}{2}+k, \frac{p-1}{2}\right)}{B\left(\frac{1}{2}, \frac{n-p-2}{2}\right)B\left(\frac{n-p}{2}, \frac{p-1}{2}\right)}\right)$$

By using the mean and variance of Hadi's measure from (3.16) and (3.18), the authors established the upper control limit of $\left(H_i^2\right)$ for different combination of (p,n) by using (3.20).Therefore

$$UCL\left(H_i^2\right) = E\left(H_i^2\right) + \sqrt{V\left(H_i^2\right)} \tag{3.19}$$

$$UCL\left(H_i^2\right) = (p+1)\,\phi_1(p,n) + \frac{n}{n-1}\phi_2(p,n) - 1 + \sqrt{(p+1)^2\,\Phi_1(p,n) + \left(\frac{n}{n-1}\right)^2\Phi_2(p,n) + \frac{2n(p+1)}{n-1}\Phi_3(p,n)} \tag{3.20}$$

By using (3.19), as a first approach, the authors utilize the upper control limit as a cut-off to identify the influential observation in linear multiple regression models. The computed $\left(H_i^2\right)$ of any observation is greater than upper control limit, then the observation is said to be influential and it may be a potential outlier. As a second approach, the authors adopted the test of significance approach of evaluating and identifying the influential observations in a sample. The approach is to derive the critical points of the Hadi's $\left(H_i^2\right)$ measure by using the following relationship from (2.10) is given as

$$H^2_{i(p,n)}(\alpha) = \frac{p+1}{\frac{(n-p-2)+F_{i(1,n-p-2)}(\alpha)}{F_{i(1,n-p-2)}} - \frac{(n-1)/n}{1+((p-1)/(n-p))F_{i(p-1,n-p)}(\alpha)} + \frac{((p-1)/(n-p))\,F_{i(p-1,n-p)}(\alpha) + 1/n}{(n-1)/n}} \tag{3.21}$$

From (3.21) for a different combination of values of (p,n) and for the significance probability $p\left(H_i^2 > H^2_{i(p,n)}(\alpha)\right) = \alpha$, authors computed the critical points of Hadi's $\left(H_i^2\right)$ measure. By using the critical points, it is possible to test the significance of the influential observation computed from a multiple linear regression model. The following table-1 visualizes the upper control limit of the Hadi's $\left(H_i^2\right)$ measure computed from (3.20) and tables 2,3 exhibits the significant percentage points of the distribution of Hadi's $\left(H_i^2\right)$ measure for varying sample size*(n)* and no.of predictors *(p)* at 5% and 1% significance (α).

Table 1. Upper control limit of Hadi's$\left(H_i^2\right)$**for combinations of** (p, n)

n	***p***									
	1	2	3	4	5	6	7	8	9	10
5	3.8793	-	-	-	-	-	-	-	-	-
6	2.7869	4.7824	-	-	-	-	-	-	-	-
7	2.0960	4.1205	5.8544	-	-	-	-	-	-	-
8	1.6375	3.0649	5.3089	6.8240	-	-	-	-	-	-
9	1.3244	2.4222	3.9434	6.4396	7.75516	-	-	-	-	-
10	1.1016	1.9860	3.1307	4.7506	7.5498	8.7226	-	-	-	-
11	0.93798	1.6735	2.5131	3.8002	5.7432	8.6542	9.6091	-	-	-
12	0.81418	1.4407	2.2003	3.1624	4.4486	6.3327	9.7586	10.708	-	-
13	0.71781	1.2618	2.8756	2.7015	8.2252	5.0854	9.2888	10.864	13.987	-
14	0.64117	1.1207	1.6800	2.3529	3.1868	4.2587	5.7147	7.8927	11.973	12.787
15	0.57888	1.0069	3.4478	2.0806	9.4703	3.6615	16.541	6.3400	15.101	13.082
16	0.52736	0.91340	1.3518	1.8627	2.4698	3.2076	4.1290	5.3235	6.9626	9.4489
17	0.48413	0.83539	3.7858	1.6843	10.869	2.8508	18.092	4.5913	22.212	7.5833
18	0.44735	0.76937	1.1278	1.5361	2.0085	2.5631	3.2258	4.0347	5.0498	6.3734
19	0.41571	0.71281	4.6021	1.4112	16.095	2.3262	28.296	3.5961	46.585	5.5056
20	0.38821	0.66388	0.96602	1.3042	1.6880	2.1281	2.6393	3.2414	3.9631	4.8466
25	0.29153	0.49344	3.5089	0.94324	18.110	1.4852	42.721	2.1576	76.534	3.0180
30	0.23330	0.39220	0.55879	0.73673	0.92840	1.1359	1.3616	1.6083	1.8790	2.1778
40	0.16665	0.27779	0.39184	0.51107	0.63665	0.76936	0.91011	1.0597	1.2188	1.3888
60	0.10604	0.17523	0.24485	0.31616	0.38966	0.46577	0.54459	0.62640	0.71132	0.79969
80	0.077750	0.12797	0.17792	0.22875	0.28048	0.33352	0.38796	0.44375	0.50114	0.56014
100	0.061377	0.10078	0.13973	0.17908	0.21892	0.25970	0.30113	0.34346	0.38659	0.43076
120	0.050698	0.083074	0.11502	0.14716	0.17970	0.21257	0.24596	0.28005	0.31459	0.34982

p-no.of predictors n-Sample Size

Table 2. Significant two-tail percentage points of Hadi's

n	***p***									
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
3	6.5000	-	-	-	-	-	-	-	-	-
4	.3457	15.9119	-	-	-	-	-	-	-	-
5	.2550	4.4885	28.1667	-	-	-	-	-	-	-
6	.2031	2.5201	7.8663	39.8398	-	-	-	-	-	-
7	.1689	1.7131	4.2275	11.0202	51.2090	-	-	-	-	-
8	.1446	1.2866	2.7941	5.8051	14.0727	62.4021	-	-	-	-
9	.1264	1.0262	2.0582	3.7841	7.3268	17.0683	73.4846	-	-	-
10	.1123	.8518	1.6185	2.7595	4.7356	8.8176	20.0285	84.4930	-	-
11	.1010	.7273	1.3291	2.1536	3.4314	5.6662	10.2892	22.9647	95.4496	
12	.0918	.6341	1.1252	1.7579	2.6646	4.0872	6.5836	11.7478	25.8841	106.3683
13	.0842	.5618	.9744	1.4812	2.1664	3.1626	4.7331	7.4923	13.1973	28.7910
14	.0777	.5042	.8585	1.2776	1.8195	2.5639	3.6525	5.3722	8.3947	14.6399
15	.0721	.4572	.7669	1.1220	1.5652	2.1481	2.9544	4.1368	6.0066	9.2924
16	.0673	.4182	.6927	.9995	1.3714	1.8441	2.4706	3.3402	4.6172	6.6373
17	.0631	.3852	.6314	.9006	1.2193	1.6131	2.1176	2.7890	3.7226	5.0947
18	.0594	.3571	.5799	.8192	1.0968	1.4320	1.8498	2.3875	3.1045	4.1025
19	.0561	.3328	.5362	.7511	.9962	1.2865	1.6401	2.0831	2.6546	3.4177
20	.0531	.3115	.4985	.6933	.9122	1.1671	1.4719	1.8453	2.3141	2.9198
25	.0420	.2361	.3685	.4999	.6401	.7944	.9678	1.1654	1.3941	1.6624
30	.0348	.1900	.2921	.3904	.4921	.6006	.7181	.8473	.9905	1.1508
40	.0259	.1366	.2064	.2712	.3360	.4026	.4722	.5457	.6239	.7074
60	.0171	.0874	.1300	.1682	.2052	.2422	.2795	.3177	.3569	.3974
80	.0128	.0643	.0948	.1219	.1477	.1730	.1983	.2238	.2496	.2759
100	.0102	.0508	.0746	.0955	.1153	.1346	.1537	.1727	.1919	.2112
120	.0085	.0420	.0615	.0785	.0946	.1101	.1254	.1406	.1558	.1710
∞	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Table 3. Significant two-tail percentage points of Hadi's

n	p									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
3	6.5000	-	-	-	-	-	-	-	-	-
4	.3338	69.0471	-	-	-	-	-	-	-	-
5	.2502	14.4658	128.0323	-	-	-	-	-	-	-
6	.2001	6.5596	24.8542	183.7271	-	-	-	-	-	-
7	.1668	3.9604	10.6671	34.5340	237.7744	-	-	-	-	-
8	.1429	2.7611	6.2112	14.4528	43.8928	290.8798	-	-	-	-
9	.1251	2.0933	4.2220	8.2658	18.0998	53.0713	343.3963	-	-	-
10	.1112	1.6749	3.1419	5.5444	10.2376	21.6699	62.1368	395.5203	-	-
11	.1000	1.3909	2.4786	4.0844	6.8090	12.1642	25.1921	71.1262	447.3694	-
12	.0909	1.1867	2.0356	3.1965	4.9825	8.0420	14.0624	28.6819	80.0619	499.0186
13	.0834	1.0334	1.7213	2.6083	3.8785	5.8565	9.2553	15.9414	32.1486	88.9580
14	.0770	.9143	1.4879	2.1939	3.1509	4.5409	6.7154	10.4553	17.8068	35.5982
15	.0715	.8193	1.3084	1.8881	2.6406	3.6769	5.1910	7.5641	11.6457	19.6621
16	.0667	.7419	1.1664	1.6541	2.2654	3.0729	4.1926	5.8327	8.4054	12.8289
17	.0625	.6776	1.0514	1.4698	1.9793	2.6300	3.4961	4.7011	6.4685	9.2412
18	.0588	.6234	.9566	1.3213	1.7548	2.2932	2.9866	3.9132	5.2046	7.0997
19	.0556	.5771	.8771	1.1992	1.5742	2.0294	2.5999	3.3378	4.3259	5.7042
20	.0527	.5372	.8096	1.0972	1.4262	1.8178	2.2975	2.9016	3.6849	4.7352
30	.0345	.3166	.4551	.5880	.7260	.8740	1.0358	1.2151	1.4161	1.6437
40	.0256	.2241	.3156	.3998	.4838	.5703	.6608	.7567	.8591	.9692
60	.0170	.1413	.1953	.2432	.2892	.3349	.3811	.4283	.4767	.5268
80	.0127	.1032	.1413	.1746	.2060	.2368	.2673	.2980	.3290	.3606
100	.0101	.0812	.1107	.1361	.1600	.1830	.2058	.2284	.2510	.2738
120	.0084	.0670	.0910	.1116	.1307	.1491	.1672	.1851	.2029	.2207
∞	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

4. Numerical Results and Discussion

To evaluate the potential outliers based on Hadi's influence measure of the *ith* observation in a regression model in this section the authors showed the results of a numerical study. For this, the authors fitted Step-wise linear regression models with a different set of predictors in a Brand equity study. The study comprised of 18 different attributes about a car brand. The Step-wise regression results reveal 4 nested models were extracted from the regression procedure. For each model, the Hadi's ($H2$) were computed, and a comparison of proposed approaches I and II with the Hadi's traditional approach of identifying the potential outliers are visualized in the following tables.

Table 4. Identification of Potential Outliers, Comparative results of Hadi's approach and proposed approach-I

Model	p	**Hadi's traditional approach**								**Proposed approach-I**	
		*Cut-off $\left(H_i^2\right)=\left(E\left(H_i^2\right)+c\sqrt{V\left(H_i^2\right)}\right)$								**Cut-off $\left(H_i^2\right)=\left(E\left(H_i^2\right)+\sqrt{V\left(H_i^2\right)}\right)$	**n>B**
		$c=1$	**n > A**	$c=2$	**n > A**	$c=3$	**n > A**	$c=4$	**n > A**		
1	1	0.04271	17	0.07393	14	0.10514	7	0.13636	5	0.021590	31
2	2	0.06726	15	0.11531	12	0.16337	7	0.21142	5	0.035821	29
3	3	0.09110	15	0.15518	12	0.21926	5	0.28334	4	0.048238	33
4	4	0.11565	13	0.19635	12	0.27706	5	0.35776	4	0.061802	31

p-no.of predictors n=275 *A-Cut-off $\left(H_i^2\right)=\left(E\left(H_i^2\right)+c\sqrt{V\left(H_i^2\right)}\right)$ **B- Cut-off $\left(H_i^2\right)=\left(E\left(H_i^2\right)+\sqrt{V\left(H_i^2\right)}\right)$-refer(3.20)

Table 5. Identification of Potential Outliers, Comparative results of Hadi's approach and proposed approach-II

Model	p	**Hadi's traditional approach**								**Proposed approach-II**			
		***Cut-off** (H_i^2)=$\left(E\left(H_i^2\right)+c\sqrt{V\left(H_i^2\right)}\right)$								****Critical**(H_i^2) at 5% level	n>B	****Critical**(H_i^2) at 1% level	n>B
		$c=1$	$n>A$	$c=2$	$n>A$	$c=3$	$n>A$	$c=4$	$n>A$				
1	1	0.04271	17	0.07393	14	0.10514	7	0.13636	5	0.003679	78	0.003651	78
2	2	0.06726	15	0.11531	12	0.16337	7	0.21142	5	0.017942	50	0 .028388	35
3	3	0.09110	15	0.15518	12	0.21926	5	0.28334	4	0. 026061	58	0 .038219	43
4	4	0.11565	13	0.19635	12	0.27706	5	0.35776	4	0.033031	57	0.046480	47

p-no.of predictors n=275 *A-**Cut-off** $\left(H_i^2\right)$ **B-**Critical** $\left(H_i^2\right)$

Table 6. Identification of Potential Outliers, Comparative results of Proposed approach I and II

Model	p	**Proposed approach-I**		**Proposed approach-II**			
		***Cut-off** (H_i^2)=$\left(E\left(H_i^2\right)+c\sqrt{V\left(H_i^2\right)}\right)$	$n>A$	***Critical**(H_i^2) at 5% level	$n>B$	***Critical**(H_i^2) at 1% level	$n>B$
1	1	0.021590	31	0.003679	78	0.003651	78
2	2	0.035821	29	0.017942	50	0 .028388	35
3	3	0.048238	33	0. 026061	58	0 .038219	43
4	4	0.061802	31	0.033031	57	0.046480	47

p-no.of predictors n=275 *A-**Cut-off** $\left(H_i^2\right)$ *B-**Critical** $\left(H_i^2\right)$

Table 4 and 5 visualizes the comparative results of Hadi's traditional approach of evaluating the potential outliers with the proposed approached 1 and 2. Under Hadi's traditional approach, 4 nested multiple regression models are evaluated and the cut-offs' for different c values are shown in the table. As far as the fitted model-1 is a concern, the computed Hadi's influence measure for 17, 14, 7 and 5 observations were above the cut-off value and hence these observations are said to be potential outliers. Similarly, model-2 is concern 15,12,7 and 5 observations are finalized as potential outliers, in the same manner, in model-3, the calculated Hadi's influence measure for 15,12,5 and 4 observations was above the cut-off and hence these observations are said to be the potential outliers. Moreover, in model-4, 13, 12, 5 and 4 observations are treated as potential outliers because these observations exceeding the Cut-off. Under Hadi's approach at what value of c, an analyst can identify the potential outliers in the fitted models? For this question, the proposed approach-I has the answer. Under proposed approach-I, the cut-off was scientifically determined and in model-1, the calculated value of Hadi's influential measure for 31 observations are above the cut-off and in model-2 29 observations, in model-3, 33 observations and in model-4, 31 observations are exceeding the scientifically determined cut-off. Hence these observations are treated as potential outliers. Under the proposed approach-II, the authors utilize the test of significance approach to identify the potential outliers. As far as the model-1 is a concern, the computed values of Hadi's influential measure for 78 observations are greater than the critical Hadi's *H2* value at 5% significance level. Similarly, model-2, model-3, and model-4 are also evaluated and the authors identified 50, 58, 57 observations are potential outliers at 5% significance level. Likewise, 78, 35, 43 and 47 observations are treated as potential outliers at 1 % significance level for model-1, model-2, model-3 and model-4 respectively. Finally, among the three approaches to evaluate the outliers, the proposed approach-II is systematic and scientific when compared to Hadi's traditional approach and proposed approach-I,because the proposed approach II identified more number of outliers at different significance level and the cut-off critical Hadi's $\left(H_i^2\right)$ value is also scientifically determined from the distribution of Hadi's influence measure. Hence the authors observed, the proposed approach-2 outperforms the Hadi's traditional approach

and it will be the better when you compared it with the proposed approach-I. Finally, the comparative results emphasize the superiority of proposed approaches over the traditional approach and it is visualized through the graphical display from the following control charts.

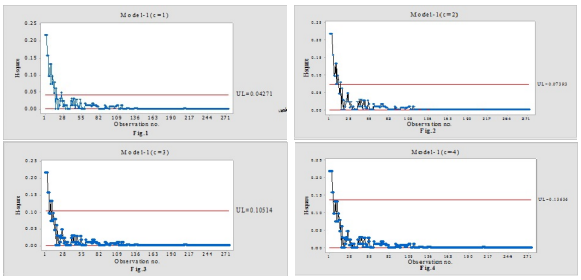

Figure 1. Control chart for fitted Model-1 shows the Identification of potential outliers based on Hadis approach

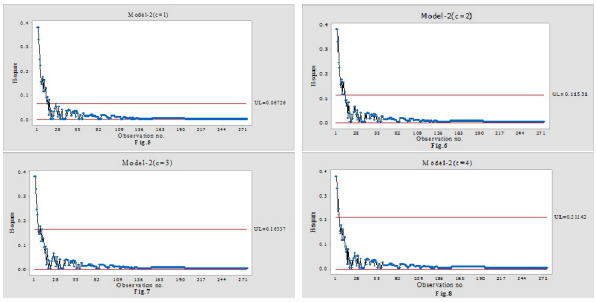

Figure 2. Control chart for fitted Model-2 shows the Identification of potential outliers based on Hadis approach

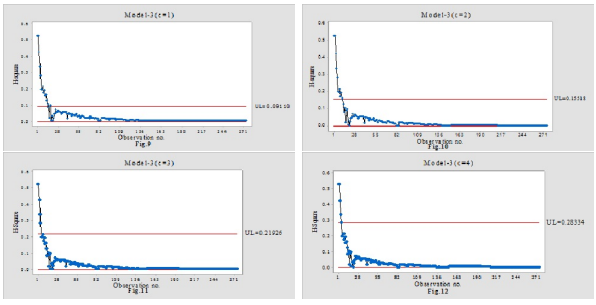

Figure 3. Control chart for fitted Model-3 shows the Identification of potential outliers based on Hadis approach

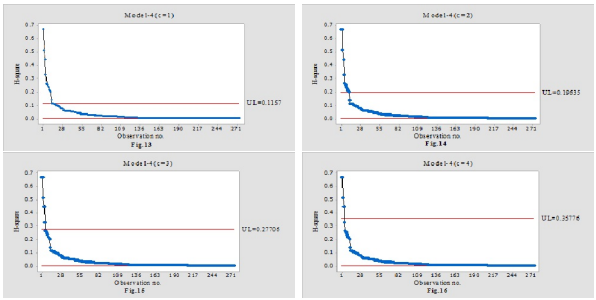

Figure 4. Control chart for fitted Model-4 shows the Identification of potential outliers based on Hadis approach

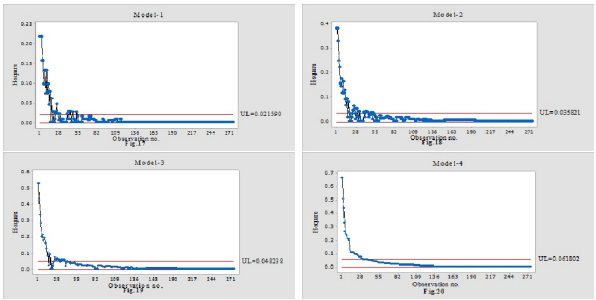

Figure 5. Control chart for each fitted model shows the Identification of potential outliers based on Proposed approach-I

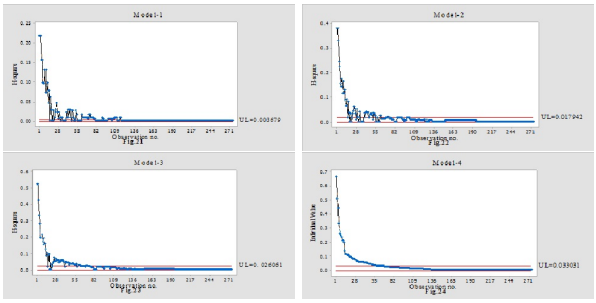

Figure 6. Control chart for each fitted model shows the Identification of potential outliers at 5% level based on proposed approach-II

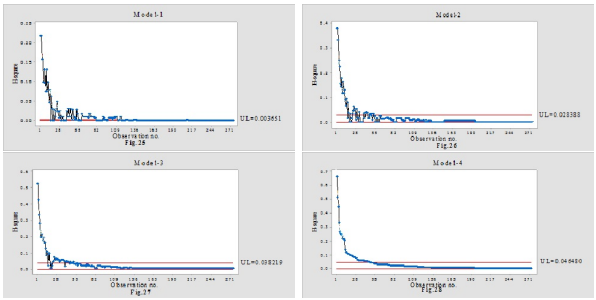

Figure 7. Control chart for each fitted model shows the Identification of potential outliers at 1% level based on proposed approach-II

5. Conclusion

From the previous sections, the authors proposed a scientific approach that is based on the test of significance for Hadi's influence measure to evaluate the potential outliers in a multiple linear regression model. At first, the exact distribution of the Hadi's $\left(H_i^2\right)$ was derived and the authors visualized the density function of *H2* in terms of complicated series expression form in terms of Gauss hypergeometric function and with two shape parameters namely p and n. Moreover, the authors computed the critical percentage points of $\left(H_i^2\right)$at 5 %, 1% level of significance and it is utilized to evaluate the potential outliers. Finally, the proposed approach II is more systematic and scientific because it is based on the test of significance and the results were superior when compared it with Hadi's traditional approach and proposed approach-I. Hence, the authors conclude, the proposed approach-II overrides the use of traditional approach, proposed approach-I and also it outperforms the traditional Hadi's approach in identifying more potential outliers in multiple regression models. Though Hadi's measure is used in the applied statistics for many years but authors found the absence of this technique in statistical software, limits the application of this efficient technique in the research. So the authors recommend the software developers and computational data analyst to include this valuable and pragmatic method in academic and commercial software in near future. Similarly, the authors believe that the scientific approach introduced in this study made Hadi's method a more significant tool in outlier detection as well as to the frequent users of linear multiple regression analysis.

References

[1] Alfons, A., Croux, C. & Gelper, S. *Sparse least trimmed squares regression for analyzing high-dimensional large data sets.* The Annals of Applied Statistics, 7, 226-248, 2013.

[2] Andrews, D. F., & Pregibon, D. *Finding the outliers that matter.* Journal of the Royal Statistical Society. Series B (Methodological), 85-93, 1978.

[3] Belsley, D. A., & Kuh, E. Welsch., RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Uiley Series in Probability and Mathematical Statistics, 1980.

[4] Beyaztas, U., & Alin, A. *Sufficient jackknife-after-bootstrap method for detection of influential observations in linear regression models.* Statistical Papers, 55(4), 1001-1018, 2014.

[5] Chatterjee, S., & Hadi, A. S. *Influential observations, high leverage points, and outliers in linear regression.* Statistical Science, 379-393, 1986.

[6] Cook, R. D. *Detection of influential observation in linear regression.* Technometrics, 19(1), 15-18, 1977.

[7] Cook, R. D., & Weisberg, S. *Criticism and influence analysis in regression.* Sociological methodology, 13(3), 313-361, 1982.

[8] Hadi, A. S. *Identifying multiple outliers in multivariate data.* Journal of the Royal Statistical Society. Series B (Methodological), 761-771, 1992.

[9] Hoaglin, D. C., & Welsch, R. E. *The hat matrix in regression and ANOVA*. The American Statistician, 32(1), 17-22, 1978.

[10] Johnson, B. W., & McCulloch, R. E. *AddedVariable Plots in Linear Regression.* Technometrics, 29(4), 427-433, 1987.

[11] Mosteller, F., & Tukey, J. W. *Data analysis and regression: a second course in statistics.* Addison-Wesley Series in Behavioral Science: Quantitative Methods, 1977.

[12] Nurunnabi, A. A. M., Hadi, A. S., & Imon, A. H. M. R. *Procedures for the identification of multiple influential observations in linear regression.*Journal of Applied Statistics, 41(6), 1315-1331, 2014.

[13] Park, H., Sakaori, F. & Konishi, S. *Robust sparse regression and tuning parameter selection via the efficient bootstrap information criteria.* Journal of Statistical Computation and Simulation, 84, 1596-1607, 2014.