Hacettepe Journal of Mathematics and Statistics
Volume 44 (1) (2015), 165 – 178

Exact distribution of Cook's distance and identification of influential observations

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

Abstract

This paper proposed the exact distribution of Cook's distance used to evaluate the influential observations in multiple linear regression analysis. The authors adopted the relationship proposed by Weisberg (1980), Belsey et al. (1980) and showed the derived density function of the cook's distance in terms of the series expression form. Moreover, the first two moments of the distribution are derived and the authors computed the critical points of Cook's distance at 5% and 1% significance level for different sample sizes based on no.of predictors. Finally, the numerical example shows the identification of the influential observations and the results extracted from the proposed approach is more scientific, systematic and it's exactness outperforms the traditional rule of thumb approach.

2000 AMS Classification: 62H10

Keywords: Cook's distance, influential observations, Series expression form, moments, critical points

Received 16/08/2013 : Accepted 26/02/2014 Doi : 10.15672/HJMS.201487459

1. Introduction and Related work

Cook's Distance (Di) is used for assessing 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), Belsey 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) proposed a type of Cook's distance in kernel density estimation. Later, Kim et al. (2001) suggested a type of Cook's distance

*Assistant Professor, Jamal Institute of Management, Tiruchirappalli - 620 020, India Email: samjaya77@gmail.com

†Research scholar, Jamal Institute of Management, Tiruchirappalli - 620 020, India Email:Sulthan90@gmail.com

in local polynomial regression. Recently, Diaz-Garcia and Gonzlez-Faras (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 outlier data points. Considering the above reviews, the authors proposed an alternative and scientific approach to identify the influential data points in multiple linear regression models and it is discussed in the subsequent sections.

2. Relationship between Cook's D and F-ratios

The multiple linear regression model with random error is given by

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

where $\underset{(nX1)}{Y}$ is the matrix of the dependent variable, $\underset{(kX1)}{\beta}$ is the vector of beta co-efficients or partial regression co-efficients and $\underset{(nX1)}{e}$ is the residual 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, cook's distance is the most frequently and interesting technique used to identify the influential observations in the Y as well as in the X-space in a multiple linear regression model. The general form of the Cook's distance of the *ith* observation is given by

$$D_i = \frac{1}{(p+1)\,\overset{\wedge}{\sigma}_e^2} \left(\overset{\wedge}{\beta} - \underset{(-i)}{\overset{\wedge}{\beta}}\right)^T X^T X \left(\overset{\wedge}{\beta} - \underset{(-i)}{\overset{\wedge}{\beta}}\right) \text{ - (2)}$$

where $\underset{(-i)}{\overset{\wedge}{\beta}}$ is the vector of estimated regression co-efficient with the *ith* observation deleted, p is the no.of predictors and $\overset{\wedge}{\sigma}_e^2$ is the residual error variance for the full data set. Removing the *ith* observation should keep $\underset{(-i)}{\overset{\wedge}{\beta}}$ close to $\overset{\wedge}{\beta}$ unless the *ith* observation is an outlier. Cook and Weisberg (1982) indicate that D_i of about 1, corresponding to distances between $\overset{\wedge}{\beta}$ and $\underset{(-i)}{\overset{\wedge}{\beta}}$ beyond a 50% confidence region would generally be considered large. Similarly, Bollen et al (1990) suggested, Cook's distance for observations more than a cut-off of $4/n-p$ which is treated as the traditional approach of evaluating the influential observations. Cook's distance (Cook and Weisberg (1982) p.118) can also be written in an alternative form as

$$D_i = \frac{r_i^2}{(p+1)}\left(\frac{h_{ii}}{1-h_{ii}}\right) \text{ - (3)}$$

Where from (3), r_i is the studentized residual which is equal to $\hat{e}_i \,/\, \overset{\wedge}{\sigma}_e \sqrt{1-h_{ii}}$ and h_{ii} is the hat element. Thus Cook's distance measures the joint influence on the case being an outlier on Y-space and in the space of the predictors (X-space). An influential observation in a multiple linear regression model may or may not be an outlier. In order to overcome the rule of thumb approach of evaluating and identifying the influential observation, we utilize the relationship among the cook's distance(D_i), 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, we first determine the distribution of (r_i)based on the relationship given by Weisberg (1980) as

$$t_i = r_i\sqrt{\frac{n-p-2}{(n-p-1)-r_i^2}} \sim t_{(n-p-2)} \text{ - (4)}$$

From (4) 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

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

$r_i^2 = \frac{(n-p-1)F_{i(1,n-p-2)}}{(n-p-2)+F_{i(1,n-p-2)}}$ - (5)

From (5), 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, we identify the distribution of (h_{ii})based on the relationship proposed by Belsey et al (1980) and they showed when the set of predictors is multivariate normal with(μ_X, Σ_X), then

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

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

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

In order to derive the exact distribution of (D_i),substitute (5) and (6) in (2), we get the Cook's D 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 by

$D_i = \frac{1}{(p+1)} \left(\frac{(n-p-1)F_{i(1,n-p-2)}}{(n-p-2)+F_{i(1,n-p-2)}} \right) \left(\frac{\big((p-1)F_{i(p-1,n-p)}/(n-p)\big)+1/n}{(n-1)/n} \right)$ - (8)

Based on the identified relationship from (8), the authors derived the distribution of the Cook's D-distance and it is discussed in the next section.

3. Exact Distribution of Cook's Distance

Using the technique of two-dimensional Jacobian of transformation, the joint probability density function of the two F-ratios namely $F_{i(1,n-p-2)}$,$F_{i(p-1,n-p)}$ with $(1, n-p-2)$ and $(p-1, n-p)$ degrees of freedom was transformed into density function of Cook's distance (D_i) and it is given as

$f(D_i, u_i) = f(F_{i(1,n-p-2)}, F_{i(p-1,n-p)})\,|J|$ - (9)

From (8), we know $F_{i(1,n-p-2)}$and $F_{i(p-1,n-p)}$ are independent then rewrite (9) as

$f(D_i, u_i) = f(F_{i(1,n-p-2)})f(F_{i(p-1,n-p)})\,|J|$ - (10)

Using the change of variable technique, substitute $F_{i(1,n-p-2)} = u_i$ in (8) we get

$F_{i(p-1,n-p)} = \frac{n-p}{p-1} \left(\frac{D_i((n-p-2)+u_i)((n-1)/n)}{((n-p-1)/(p+1))u_i} - 1/n \right)$ - (10a)

Then partially differentiate (10a) and compute the Jacobian determinant in (10) as

$f(D_i, u_i) = f(F_{i(1,n-p-2)})f(F_{i(p-1,n-p)}) \left| \frac{\partial(F_{i(1,n-p-2)}, F_{i(p-1,n-p)})}{\partial(D_i, u_i)} \right|$ - (11)

$f(D_i, u_i) = f(F_{i(1,n-p-2)})f(F_{i(p-1,n-p)}) \begin{vmatrix} \frac{\partial F_{i(1,n-p-2)}}{\partial D_i} & \frac{\partial F_{i(1,n-p-2)}}{\partial u_i} \\ \frac{\partial F_{i(p-1,n-p)}}{\partial D_i} & \frac{\partial F_{i(p-1,n-p)}}{\partial u_i} \end{vmatrix}$ - (12)

From (12), we know the F-ratios are independent, then the density function of the joint distribution of $F_{i(1,n-p-2)}$ and $F_{i(p-1,n-p)}$ are given as

$$f(F_{i(1,n-p-2)}, F_{i(p-1,n-p)}) = f(F_{i(1,n-p-2)})f(F_{i(p-1,n-p)})$$

$$f(F_{i(1,n-p-2)}, F_{i(p-1,n-p)}) = \left(\frac{(1/n-p-2)^{1/2}}{B(\frac{1}{2}, \frac{n-p-2}{2})} \left(F_{i(1,n-p-2)}\right)^{(1/2)-1} \left(1 + \frac{F_{i(1,n-p-2)}}{n-p-2}\right)^{-(\frac{1}{2}+\frac{n-p-2}{2})} \right) * \left(\frac{((p-1)/n-p)^{(p-1)/2}}{B(\frac{p-1}{2}, \frac{n-p}{2})} \left(F_{i(p-1,n-p)}\right)^{((p-1)/2)-1} (1 + \frac{p-1}{n-p} F_{i(p-1,n-p)})^{-(\frac{p-1}{2}+\frac{n-p}{2})} \right) \quad \text{- (13)}$$

where $0 \leq F_{i(1,n-p-2)}, F_{i(p-1,n-p)} \leq \infty$,$n, p > 0$

and $$\begin{vmatrix} \frac{\partial F_{i(1,n-p-2)}}{\partial D_i} & \frac{\partial F_{i(1,n-p-2)}}{\partial u_i} \\ \frac{\partial F_{i(p-1,n-p)}}{\partial D_i} & \frac{\partial F_{i(p-1,n-p)}}{\partial u_i} \end{vmatrix} = \begin{vmatrix} 0 & \frac{n-p}{p-1}\left(\frac{((n-p-2)+u_i)((n-1)/n)}{((n-p-1)/(p+1))u_i}\right) \\ 1 & -\frac{n-p}{p-1}\left(\frac{D_i(n-p-2)((n-1)/n)}{((n-p-1)/(p+1))u_i^2}\right) \end{vmatrix} \quad \text{- (14)}$$
$$= \frac{n-p}{p-1}\left(\frac{((n-p-2)+u_i)((n-1)/n)}{((n-p-1)/(p+1))u_i}\right)$$

Then substitute (13) and (14) in (12) in terms of the substitution of u_i we get the joint distribution of Cook's D and u_i as

$$f(D_i, u_i) = \left(\frac{(1/n-p-2)^{1/2}}{B(\frac{1}{2}, \frac{n-p-2}{2})} u_i^{(1/2)-1} (1 + \frac{u_i}{n-p-2})^{-(\frac{1}{2}+\frac{n-p-2}{2})} \right) * \left(\begin{array}{c} \frac{((p-1)/n-p)^{(p-1)/2}}{B(\frac{p-1}{2}, \frac{n-p}{2})} \left(\frac{n-p}{p-1} \left(\frac{D_i((n-p-2)+u_i)((n-1)/n)}{((n-p-1)/(p+1))u_i} - 1/n \right) \right)^{((p-1)/2)-1} \\ \left(1 + \left(\frac{D_i((n-p-2)+u_i)((n-1)/n)}{((n-p-1)/(p+1))u_i} - 1/n \right) \right)^{-(\frac{p-1}{2}+\frac{n-p}{2})} \end{array} \right) |J| \quad \text{- (15)}$$

where $0 \leq D_i \leq \infty, 0 \leq u_i \leq \infty$ and $|J| = \frac{n-p}{p-1}\left(\frac{((n-p-2)+u_i)((n-1)/n)}{((n-p-1)/(p+1))u_i}\right)$

Rearrange (15) and integrate with respect tou_i, we get the marginal distribution of D_i *as*

$$f(D_i, u_i) = \alpha(n,p) \sum_{q=0}^{(p-3)/2} \sum_{k=0}^{\infty} \left(\begin{array}{c} (p-3)/2 \\ q \end{array} \right) \left(\begin{array}{c} -(n-1)/2 \\ k \end{array} \right) \gamma^{((p-3)/2)-q} \lambda^k \int_0^{\infty} u_i^{q-(\frac{p-1}{2}+k)-1} (1 + \frac{u_i}{n-p-2})^{-(q-(\frac{p-1}{2}+k)+\frac{n-p}{2})} du_i \quad \text{-(16)}$$

where $0 \leq D_i \leq \infty$,

$$\alpha(n,p) = \frac{1}{B(\frac{1}{2}, \frac{n-p-2}{2}) B(\frac{p-1}{2}, \frac{n-p}{2})} \left(\frac{(n-p-2)^{1/2}(-1/n)^{(p-3)/2}}{((n-p-1)/(p+1))((n-1)/n))^{(\frac{n-3}{2})}} \right)$$

$$\lambda = \left(\frac{D_i(n-p-2)}{(n-p-1)/(p+1)} \right)$$

and $\gamma = -\left(\frac{D_i(n-p-2)(n-1)}{(n-p-1)/(p+1)}\right)$

Finally from (16), after the integration arranging the terms, we get the density of Cook's D distance as the form of series expression as

$$f(D_i;n,p)=\alpha(n,p)\sum_{q=0}^{(p-3)/2}\sum_{k=0}^{\infty}\left(\begin{array}{c}(p-3)/2\\ q\end{array}\right)\left(\begin{array}{c}-(n-1)/2\\ k\end{array}\right)$$
$$\beta(n,p,q,k)D_i^{((p-3)/2)-q+k}$$

- (17)

where, $0\leq D_i\leq\infty$,$n,p>0$, $n>p$

$$\alpha(n,p)=\frac{1}{B(\frac{1}{2},\frac{n-p-2}{2})B(\frac{p-1}{2},\frac{n-p}{2})}\left(\frac{(n-p-2)^{-1/2}(-1/n)^{(p-3)/2}}{((n-p-1)/(p+1))((n-1)/n))^{(n-3)/2}}\right)$$

$$\beta(n,p,q,k)=\left(\frac{1}{((n-p-1)/(p+1))}\right)^{k}\left(-\frac{(n-1)}{(n-p-1)/(p+1)}\right)^{((p-3)/2)-q}$$
$$B(q-(\frac{p-1}{2}+k),\frac{n-p-1}{2})$$

From (17), it is the density function of Cook's D distance which involves the normalizing constants such as$\alpha(n,p)$,$\beta(n,p,q,k)$ and$B(\frac{1}{2},\frac{n-p-2}{2})$,$B(\frac{p-1}{2},\frac{n-p}{2})$,$B(q-(\frac{p-1}{2}+k),\frac{n-p-1}{2})$ are the Beta functions respectively with two parameters (n,p), where n is the sample size and p is the no. of predictors used in the multiple linear regression model. In order to know the location and dispersion of Cook's D, the authors derived the first two moments in terms of mean, variance from (8) and it is given as follows.

$$D_i=$$
$$\frac{(n-p-1)}{(p+1)(n-1)}\left(\left(F_{i(1,n-p-2)}/(n-p-2)\right)\sum_{k=0}^{\infty}(-1)^{k}\left(F_{i(1,n-p-2)}/(n-p-2)\right)^{k}\right)$$
$$*\left(1+n(p-1)F_{i(p-1),(n-p)}/(n-p)\right)$$

$$D_i=\frac{(n-p-1)}{(p+1)(n-1)}\left(\sum_{k=0}^{\infty}(-1)^{k}\left(1/(n-p-2)\right)^{k+1}F_{i(1,n-p-2)}^{k+1}\right)$$
$$*\left(1+n(p-1)F_{i(p-1),(n-p)}/(n-p)\right)$$

- (18)

Therefore,

$$E(D)=\frac{(n-p-1)}{(p+1)(n-1)}\left(\sum_{k=0}^{\infty}(-1)^{k}\left(1/(n-p-2)\right)^{k+1}E(F_{(1,n-p-2)}^{k+1})\right)$$
$$*\left(1+n(p-1)E(F_{(p-1),(n-p)})/(n-p)\right)$$

$E(D)=\frac{(p(n-1)-2)(n-p-1)}{2(n-1)(p+1)(n-p+1)}$ - (19)

From (18), Squaring on both sides and take expectation, we get the second moment of the cook's D as

$$D_i^2 = \left(\frac{(n-p-1)}{(p+1)(n-1)}\right)^2 \left(\sum_{k=0}^{\infty}(-1)^k(k+1)\left(1/(n-p-2)\right)^{k+2}\left(F_{i(1,n-p-2)}\right)^{k+2}\right)$$
$$* \left(1 + (n(p-1)/(n-p))^2 F_{i(p-1),(n-p)}^2 + 2n(p-1)F_{i(p-1),(n-p)}/(n-p)\right)$$

$$E(D^2) = \left(\frac{(n-p-1)}{(p+1)(n-1)}\right)^2 \left(\sum_{k=0}^{\infty}(-1)^k(k+1)\left(1/(n-p-2)\right)^{k+2} E(F_{(1,n-p-2)}^{k+2})\right)$$
$$* \left(1 + (n(p-1)/(n-p))^2 E(F_{(p-1),(n-p)}^2) + 2n(p-1)E(F_{(p-1),(n-p)})/(n-p)\right)$$

$$E(D^2) = \frac{((n-p-1)/(p+1)(n-1))^2}{B(\frac{1}{2}, \frac{n-p-2}{2})}$$
$$* \left(\sum_{k=0}^{\infty}(-1)^k(k+1)B(\frac{1}{2}+k+2, \frac{n-p-2}{2} - (k+2))\right)$$
$$* \left(1 + \left(n^2(p^2-1)/(n-p-2)(n-p-4)\right) + (2n(p-1)/(n-p-2))\right)$$

- 20

Therefore, we know
$V(D) = E(D^2) - (E(D))^2$ - (21)
Substitute (19) and (20) in (21), we get
$V(D) = \frac{(n-p-1)}{(n-p+1)}\left(\frac{1}{(p+1)(n-1)}\right)^2 *$

$$\left(3\left(\frac{p^2+6p+n^2(p^2+2p-2)+n(2-2p^2-8p)+8}{(n-p-2)(n-p-4)}\right) - \frac{(n-p-1)}{(n-p+1)}\left(\frac{(p(n-1)-2)}{2}\right)^2\right)$$

Moreover, the authors adopted test of significance approach of evaluating and identifying the influential observations in a sample. The approach is to derive the critical points of the Cook's distance by using (8) for different values of (n, p) and the significance probability is given by by $p\left(D_i > D_{i(n,p)}(\alpha)\right) = \alpha$. Using the critical points, we can test the significance of the influential observation computed from a multiple linear regression model. The following tables 1 and 2 show the significance points of the distribution of Cook's D for varying sample size*(n)* and no.of predictors *(p)* at 5% and 1% significance (α).

Table 1 Significant two-tail values of the Distribution of Cook's D at 5% level of Significance
$(P\left(D_i > D_{i(0.05)}\right) = 0.05)$

n	*p*													
	2	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
3	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	**.33**	-	-	-	-	-	-	-	-	-	-	-	-	-
5	**.34**	**2.96**	-	-	-	-	-	-	-	-	-	-	-	-
6	**.31**	**2.27**	**3.90**	-	-	-	-	-	-	-	-	-	-	-
7	**.27**	**1.76**	**2.85**	**4.37**	-	-	-	-	-	-	-	-	-	-
8	**.24**	**1.41**	**2.15**	**3.14**	**4.65**	-	-	-	-	-	-	-	-	-
9	**.22**	**1.17**	**1.69**	**2.33**	**3.30**	**4.83**	-	-	-	-	-	-	-	-
10	**.20**	**.99**	**1.38**	**1.81**	**2.43**	**3.40**	**4.96**	-	-	-	-	-	-	-
11	**.18**	**.86**	**1.16**	**1.47**	**1.88**	**2.49**	**3.48**	**5.06**	-	-	-	-	-	-
12	**.16**	**.76**	**1.00**	**1.23**	**1.51**	**1.92**	**2.53**	**3.53**	**5.13**	-	-	-	-	-
13	**.15**	**.68**	**.87**	**1.05**	**1.26**	**1.54**	**1.94**	**2.56**	**3.57**	**5.19**	-	-	-	-
14	**.14**	**.61**	**.78**	**.92**	**1.08**	**1.28**	**1.56**	**1.96**	**2.59**	**3.60**	**5.24**	-	-	-
15	**.13**	**.56**	**.70**	**.81**	**.94**	**1.09**	**1.29**	**1.57**	**1.97**	**2.60**	**3.62**	**5.27**	-	-
16	**.12**	**.51**	**.63**	**.73**	**.83**	**.94**	**1.09**	**1.30**	**1.57**	**1.98**	**2.61**	**3.64**	**5.31**	-
17	**.12**	**.47**	**.58**	**.66**	**.74**	**.83**	**.95**	**1.10**	**1.30**	**1.58**	**1.99**	**2.62**	**3.66**	**5.33**
18	**.11**	**.44**	**.54**	**.60**	**.67**	**.75**	**.84**	**.95**	**1.10**	**1.30**	**1.58**	**1.99**	**2.63**	**3.67**
19	**.10**	**.41**	**.50**	**.55**	**.61**	**.67**	**.75**	**.84**	**.95**	**1.10**	**1.30**	**1.58**	**2.00**	**2.64**
20	**.10**	**.39**	**.46**	**.51**	**.56**	**.61**	**.67**	**.75**	**.84**	**.95**	**1.10**	**1.30**	**1.59**	**2.00**
21	**.09**	**.36**	**.43**	**.48**	**.52**	**.56**	**.61**	**.67**	**.75**	**.84**	**.95**	**1.10**	**1.30**	**1.59**
22	**.09**	**.34**	**.41**	**.45**	**.48**	**.52**	**.56**	**.61**	**.67**	**.74**	**.84**	**.95**	**1.10**	**1.30**
23	**.09**	**.33**	**.38**	**.42**	**.45**	**.48**	**.52**	**.56**	**.61**	**.67**	**.74**	**.83**	**.95**	**1.10**
24	**.08**	**.31**	**.36**	**.40**	**.42**	**.45**	**.48**	**.52**	**.56**	**.61**	**.67**	**.74**	**.83**	**.95**
25	**.08**	**.29**	**.34**	**.37**	**.40**	**.42**	**.45**	**.48**	**.52**	**.56**	**.61**	**.67**	**.74**	**.83**
26	**.08**	**.28**	**.33**	**.35**	**.38**	**.40**	**.42**	**.45**	**.48**	**.51**	**.56**	**.61**	**.67**	**.74**
27	**.07**	**.27**	**.31**	**.34**	**.36**	**.38**	**.40**	**.42**	**.45**	**.48**	**.51**	**.55**	**.60**	**.66**
28	**.07**	**.26**	**.30**	**.32**	**.34**	**.36**	**.37**	**.39**	**.42**	**.44**	**.47**	**.51**	**.55**	**.60**
29	**.07**	**.25**	**.29**	**.31**	**.32**	**.34**	**.35**	**.37**	**.39**	**.42**	**.44**	**.47**	**.51**	**.55**
30	**.07**	**.24**	**.27**	**.29**	**.31**	**.32**	**.34**	**.35**	**.37**	**.39**	**.41**	**.44**	**.47**	**.51**
40	**.05**	**.17**	**.20**	**.21**	**.21**	**.22**	**.22**	**.23**	**.24**	**.24**	**.25**	**.26**	**.27**	**.28**
50	**.04**	**.13**	**.15**	**.16**	**.16**	**.16**	**.17**	**.17**	**.17**	**.18**	**.18**	**.18**	**.19**	**.19**
60	**.03**	**.11**	**.12**	**.13**	**.13**	**.13**	**.13**	**.13**	**.14**	**.14**	**.14**	**.14**	**.14**	**.15**
70	**.03**	**.09**	**.10**	**.11**	**.11**	**.11**	**.11**	**.11**	**.11**	**.11**	**.11**	**.11**	**.12**	**.12**
80	**.02**	**.08**	**.09**	**.09**	**.09**	**.09**	**.09**	**.09**	**.10**	**.10**	**.10**	**.10**	**.10**	**.10**
90	**.02**	**.07**	**.08**	**.08**	**.08**	**.08**	**.08**	**.08**	**.08**	**.08**	**.08**	**.08**	**.08**	**.08**
100	**.02**	**.06**	**.07**	**.07**	**.07**	**.07**	**.07**	**.07**	**.07**	**.07**	**.07**	**.07**	**.07**	**.07**
∞	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2 Significant two-tail values of the Distribution of Cook's D at 1% level of Significance
$(P\left(D_i > D_{i(0.05)}\right) = 0.01)$

n	*p*													
	2	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
3	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	.33	-	-	-	-	-	-	-	-	-	-	-	-	-
5	.37	9.64	-	-	-	-	-	-	-	-	-	-	-	-
6	.37	6.43	12.42	-	-	-	-	-	-	-	-	-	-	-
7	.35	4.85	7.84	13.81	-	-	-	-	-	-	-	-	-	-
8	.33	3.87	5.71	8.50	14.63	-	-	-	-	-	-	-	-	-
9	.30	3.20	4.44	6.08	8.87	15.16	-	-	-	-	-	-	-	-
10	.28	2.72	3.60	4.66	6.27	9.10	15.53	-	-	-	-	-	-	-
11	.26	2.36	3.02	3.75	4.77	6.39	9.26	15.80	-	-	-	-	-	-
12	.25	2.08	2.59	3.12	3.81	4.83	6.46	9.37	16.01	-	-	-	-	-
13	.23	1.86	2.26	2.66	3.15	3.84	4.87	6.51	9.45	16.17	-	-	-	-
14	.22	1.68	2.01	2.31	2.67	3.16	3.85	4.89	6.55	9.52	16.30	-	-	-
15	.20	1.53	1.80	2.04	2.32	2.67	3.16	3.86	4.90	6.57	9.56	16.41	-	-
16	.19	1.41	1.63	1.82	2.04	2.31	2.67	3.16	3.86	4.91	6.59	9.60	16.49	-
17	.18	1.30	1.49	1.65	1.82	2.03	2.30	2.66	3.15	3.85	4.91	6.60	9.64	16.57
18	.17	1.21	1.38	1.50	1.64	1.81	2.02	2.29	2.65	3.14	3.85	4.91	6.61	9.66
19	.16	1.13	1.27	1.38	1.49	1.63	1.79	2.00	2.28	2.64	3.14	3.85	4.91	6.62
20	.16	1.06	1.19	1.28	1.37	1.48	1.61	1.78	1.99	2.26	2.63	3.13	3.84	4.91
21	.15	1.00	1.11	1.19	1.26	1.35	1.46	1.60	1.76	1.98	2.25	2.62	3.12	3.84
22	.14	.94	1.04	1.11	1.17	1.25	1.34	1.45	1.58	1.75	1.97	2.24	2.61	3.12
23	.14	.89	.98	1.04	1.09	1.16	1.23	1.32	1.43	1.57	1.74	1.96	2.24	2.61
24	.13	.85	.93	.98	1.02	1.08	1.14	1.22	1.31	1.42	1.56	1.73	1.95	2.23
25	.13	.81	.88	.92	.96	1.01	1.06	1.12	1.20	1.30	1.41	1.55	1.72	1.94
26	.12	.77	.84	.88	.91	.95	.99	1.05	1.11	1.19	1.28	1.40	1.54	1.71
27	.12	.74	.80	.83	.86	.89	.93	.98	1.03	1.10	1.18	1.27	1.39	1.53
28	.11	.71	.76	.79	.82	.84	.88	.92	.97	1.02	1.09	1.17	1.27	1.38
29	.11	.68	.73	.76	.78	.80	.83	.86	.91	.95	1.01	1.08	1.16	1.26
30	.11	.65	.70	.72	.74	.76	.79	.82	.85	.89	.94	1.00	1.07	1.15
40	.08	.47	.50	.50	.51	.51	.52	.52	.54	.55	.56	.58	.60	.62
50	.07	.37	.38	.39	.38	.38	.38	.39	.39	.39	.40	.40	.41	.42
60	.05	.30	.31	.31	.31	.31	.30	.30	.30	.31	.31	.31	.31	.32
70	.05	.26	.26	.26	.26	.26	.25	.25	.25	.25	.25	.25	.25	.25
80	.04	.22	.23	.23	.22	.22	.22	.21	.21	.21	.21	.21	.21	.21
90	.04	.20	.20	.20	.19	.19	.19	.19	.18	.18	.18	.18	.18	.18
100	.03	.18	.18	.18	.17	.17	.17	.16	.16	.16	.16	.16	.16	.16
∞	0	0	0	0	0	0	0	0	0	0	0	0	0	0

4. Numerical Results and Discussion

In this section, the authors show a numerical study of evaluating the influential observation based on cook's distance of the *ith* observation in a regression model. For this, the authors fitted Step-wise 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 Step-wise regression results reveals 4 models were extracted from the regression procedure by using IBM SPSS version 22. For each model, the cook's distance were computed and the identification of influential observations, comparison of proposed approaches with the traditional approach of identifying influential observations are visualized in the following table.3

Table 3

Model	p	Traditional approach		
		Cut-off $4/(n-p)$	No.of Influential observation (n)	Mean Cook's D of Influential observations
1	1	.014599	22	.0797472
2	2	.014652	20	0.074233
3	3	.014706	19	0.084601
4	4	.014760	24	0.062829
Proposed approach				
Model	p	5% Significance level		
		Critical Cook's D	No.of Influential observation (n)	Mean Cook's D of Influential observations
1	1	.00700	31	.0586684
2	2	.02288	15	0.093052
3	3	.02493	13	0.113835
4	4	.02528	15	0.088706
Model	p	1% Significance level		
		Critical Cook's D	No.of Influential observation (n)	Mean Cook's D of Influential observations
		.01203	22	.0797470
1	1	.06236	9	0.129777
2	2	.06297	10	0.13628
3	3	.06126	9	0.125272

p-no.of predictors n=275

Table-3 visualizes the results of the identification and evaluation process of the influential observation based on the cook's D distance in a multiple linear regression model. As far as the traditional approach is concern, the cut-off cook's D distances are 0.014599 for model-1, 0.014652 for model-2, 0.014706 for model -3 and 0.014760 for model-4 respectively. From model-1, we identified 22 observations are more than the prescribed cut-off followed by 20, 19, 24 observations from model-2 model-3 and model-4 respectively. This approach is traditional and if the analyst may change the cut-off then, it will give different results. As far as the proposed approach is concern, the authors identified the influential observations at 5% and 1% test of significance. As far as model 1 is concern 31 observations are said to be influential because the cook's D for these observations are more than the critical cook's D distance. Similarly 15 observations from model 2, 13 observations from model-3 and 15 observations from model-4 are also influential at 5% significance level. In the same manner, 22 observations are influential in model-1 at 1% level of significance followed by 9 observations from model-2 and 10 observations from model-3 and 9 observations from model-4 are more than the critical cook's D at 1% level of significance respectively. Another good evidence was also provided by the authors that is the mean cook's distance of the influential observations are higher than the critical cook's D for all the models at 5% and 1% significance level. This shows the identification of influential observation based on the test of significance gives different results when compared to the traditional approach we recommend the proposed approach is more scientific and it over rides the use of traditional approach in identifying influential observation in multiple regression model. The following control charts exhibits the results of Table 3 graphically.

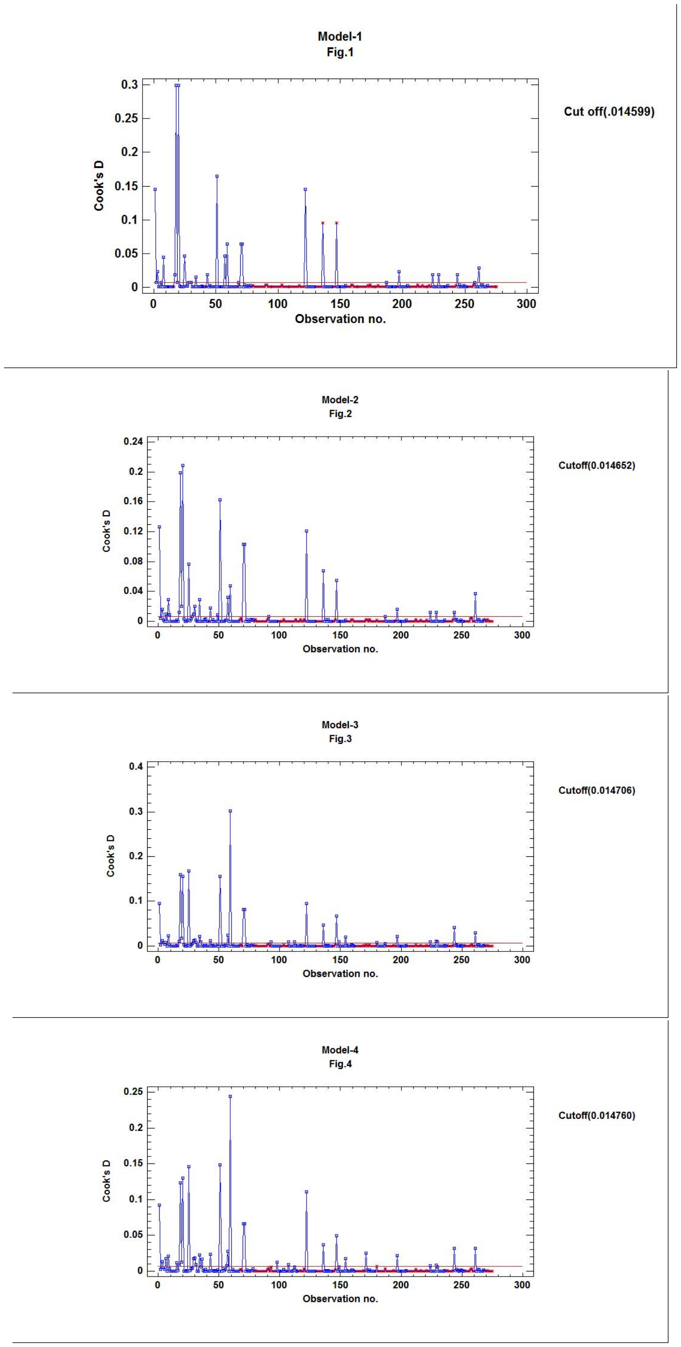

Figure 1. Control charts for each fitted model shows the identification of influential observations based on Traditional approach

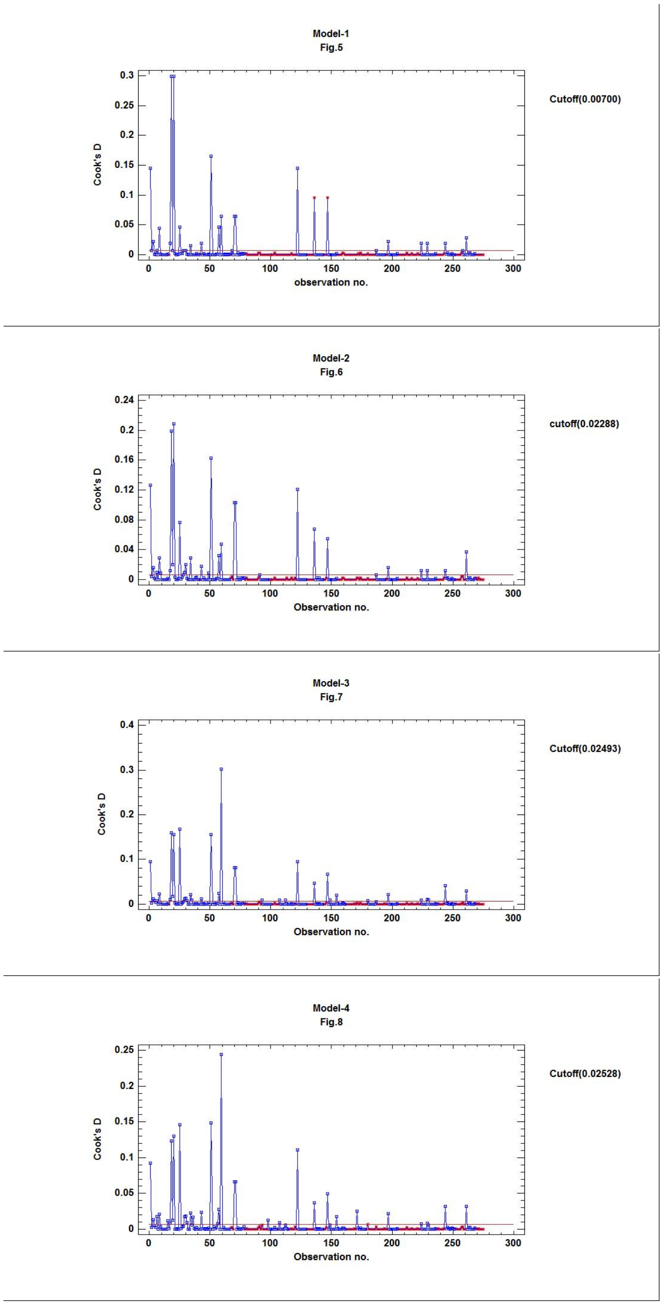

Figure 2. Control charts for each fitted model shows the identification of influential observations at 5% significance level proposed approach

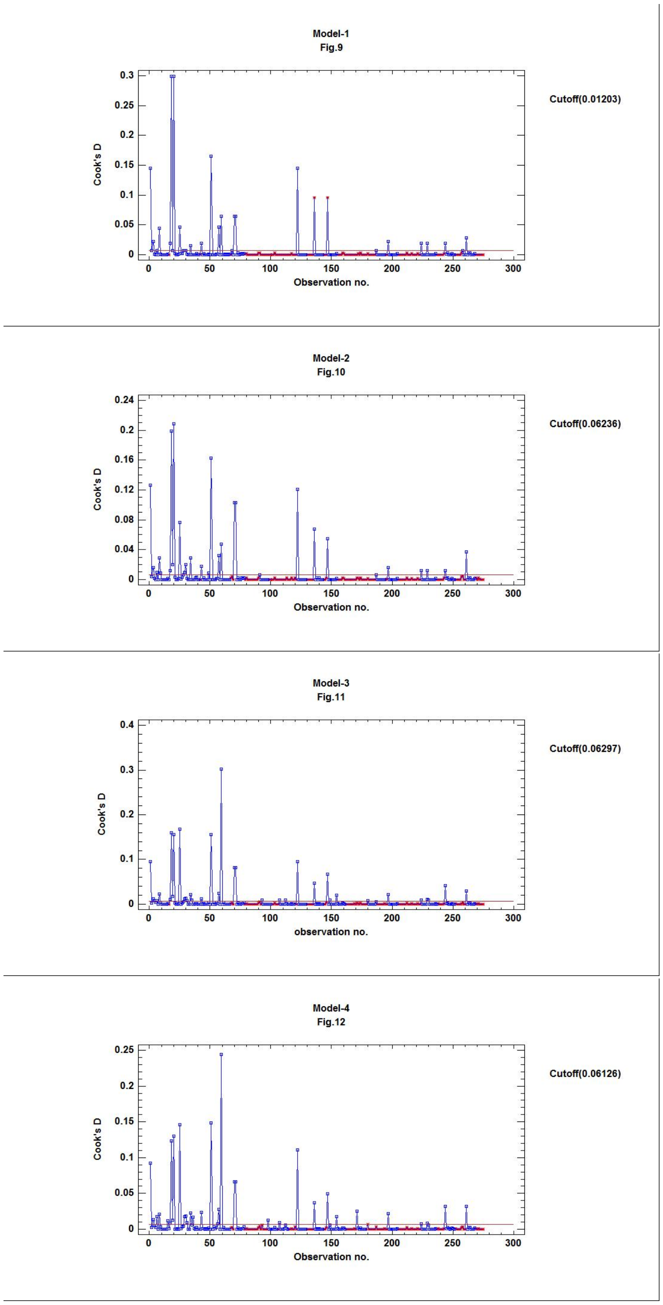

Figure 3. Control charts for each fitted model shows the identification of influential observations at 1% significance level based on proposed approach

5. Conclusion

From the previous sections, the authors proposed a scientific approach which is based on test of significance for identifying and evaluating the influential observation in a multiple linear regression model. At first, the exact distribution of the Cook's D distribution was derived and the authors proved, it followed a beta distribution with 2 shape parameters n and p and we expressed the density function of Cook's D in series expression form. Moreover, the authors computed the Critical points of Cook's D and it is utilized to evaluate the influential observations. Finally, the proposed approach which is more systematic and scientific method of identifying the influential observation because it is based on the test of significance and the results are different when compared it with traditional approach. So the authors found that the proposed approach over rides the use of traditional approach in identifying influential observation in multiple regression models.

References

[1] Belsey, D. A., Kuh, E., & Welsch, R. E. *Regression diagnostics: Identifying influential data and sources of collinearity.* (John Wiley1980).

[2] Bollen, K. A., & Jackman, R. W. Regression diagnostics: An expository treatment of outliers and influential cases. *Modern methods of data analysis*, 257-291, 1990.

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

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

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

[6] Diaz-Garcia, J. A., & Gonzlez-Faras, G. A note on the Cook's distance. *Journal of statistical planning and inference*, ***120***(1), 119-136, 2004.

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

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

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

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

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