Journal of Reliability and Statistical Studies; ISSN (Print): 0974-8024, (Online):2229-5666
Vol. 7, Issue 1 (2014): 61-78

EXACT DISTRIBUTION OF HAT VALUES AND IDENTIFICATION OF LEVERAGE POINTS

G.S. David Sam Jayakumar[1] and A. Sulthan[2]
Jamal Institute of Management, Tiruchirappalli – 620 020, India,
E Mail: [1]samjaya77@gmail.com, [2]Sulthan90@gmail.com

Received July 06, 2013
Modified May 30, 2014
Accepted June 02, 2014

Abstract

This paper proposed the exact distribution of centered hat values of the hat matrix of predictors in multiple linear regression analysis. The authors adopted the relationship proposed by Belsey et al. (1980) between the centered hat values and the F-ratio and we showed that the derived density function of the centered hat values followed Beta distribution $\beta(p-1, n-p)$ and it lies between $1/n \le h_{ii} \le 1$. Moreover, the first two moments of the distribution are derived and we established the upper and lower limits of the centered hat values. Moreover, the shape of the density function of hat values is also visualized and the authors computed the percentage points of centered hat values at 5% and 1% significance level for different sample sizes and predictors. Finally, the authors proposed two approaches. The first approach helps to identify the leverage points in multiple linear regression analysis in the X-space based on the test of significance and the second approach scrutinized the leverage points as well as the outliers. The proposed approaches were numerically illustrated and the results were compared the traditional approach.

Key Words: Centered Hat Values, Hat Matrix, Beta-Distribution, Moments, Leverage Points, Outliers, X-Space.

1. Introduction and Related work

The hat matrix is an important auxiliary quantity in regression theory and it is a standard measure of predictor influence. (Belsley *et al.* (1980) and Chatterjee and Hadi (1988)). Hoaglin and Welsch (1978) suggested, observations with h_{ii} >2p/n as high leverage points. A standard statistical measure of leverage is the size of the diagonal elements of the hat matrix, and many estimators use this quantity to detect and down weight the leverage values (Mallows (1975), Handschin *et al.* (1975) and Krasker and Welsch (1982)). Later, Dodge and Hadi (1999) presented graphs and bounds for the elements of hat matrix. In the work of Chave and Thomson (2003), a new bounded influence estimator is proposed that combines high asymptotic efficiency for normal data, high breakdown point behavior with contaminated data and computational simplicity for large data sets. The algorithm combines a standard *M*-estimator to down weight data corresponding to extreme regression residuals and removal of overly influential predictor values (leverage points) on the basis of the statistics of the hat matrix diagonal elements. Diaz-Garcia, J.A. and Gonzalez- Faras, G. (2004) investigated the Cook's D distance and extracted more properties to scrutinize the influential observation in linear regression. Moreover, Prendergast, L.A. (2005, 2006) studied the influential observations in the sliced inverse regression model and Huang, Y., Kuo, M. and Wang, T. (2007) proposed the perturbation influence functions and

visualized the local influence of observations in a sample. Finally M.A. Ullah1 and G.R. Pasha (2009) presented extensive literatures on the origin and developments of influence measures in regression analysis. Based on the reviews, the authors propose a more systematic and scientific approach to identify the leverage points and they discuss the characteristics of centered hat values and it's relationship with F-ratio, exact distribution, limits of hat values in the subsequent sections.

2. Relationship between centered hat values and F-ratio

Let the multiple linear regression model with random error can be given as

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

where $\underset{(nX1)}{Y}$ is the matrix of the dependent variable, $\underset{((p+1)X1)}{\beta}$ is the matrix of beta co-efficients or partial regression co-efficients and $\underset{(nX1)}{e}$ is the residual followed normal distribution $N\left(0, \sigma_e^2 I_n\right)$.From (1), the fitted model with estimates as

$$\hat{Y} = X\hat{\beta} \tag{2}$$

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{3}$$

Substitute $\hat{\beta}$ in (3), we get

$$\hat{Y} = X\left(X^T X\right)^{-1} X^T Y$$

$$\hat{Y} = HY \tag{4}$$

From (4), the estimated $\hat{Y}$ is predicted by the actual value of Y based on the projection matrix $H = X\left(X^T X\right)^{-1} X^T$, technically called as Hat matrix. Combine (2) and (4), we get a compact form of the disturbance in terms of the hat matrix and it is given as

$$e = Y - \hat{Y} \tag{5}$$

$$e = Y - X\hat{\beta}$$

$$e = Y - X\left(X^T X\right)^{-1} X^T Y$$

$$e = Y\left(I - X\left(X^T X\right)^{-1} X^T\right)$$

$$e = \left(I - H\right)Y \tag{6}$$

From (6), the regression disturbance is the product of actual value of *Y* and the residual operator $\left(I - H\right)$.Myers, Montgomery (1997) proved the magical properties of the residual operator matrix as idempotent and symmetric. Based on the properties they derived the variance-co-variance matrix of the disturbance as

$$\sum\nolimits_e = \sigma_e^2\left(I - H\right) \tag{7}$$

Where $\sum_e$ is the Variance-covariance matrix and σ_e^2 is the homoscedastic error variance of the linear regression model. The authors utilized the least squares estimates of the variance-covariance matrix of the disturbance and found the link between $\sum_e$

and $(I-H)$. From (7), the estimate of $\sum_e$ is given as

$$\widehat{\sum_e} = \widehat{\sigma_e^2}(I-H) \tag{8}$$

From (8), where

$$\widehat{\sum_e} = \begin{pmatrix} \widehat{\sigma_{e_1}^2} & \operatorname{cov}(\widehat{e}_1, \widehat{e}_2) & \cdots & \operatorname{cov}(\widehat{e}_1, \widehat{e}_n) \\ \operatorname{cov}(\widehat{e}_2, \widehat{e}_1) & \widehat{\sigma_{e_2}^2} & \cdots & \operatorname{cov}(\widehat{e}_2, \widehat{e}_n) \\ \cdots & \cdots & \cdots & \cdots \\ \operatorname{cov}(\widehat{e}_n, \widehat{e}_1) & \operatorname{cov}(\widehat{e}_n, \widehat{e}_2) & \cdots & \widehat{\sigma_{e_n}^2} \end{pmatrix} \tag{9}$$

$$H = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{pmatrix} \tag{10}$$

From (8), compare the diagonal elements of both sides, we get the estimated heteroscedastic error variance as

$$\widehat{\sigma_{e_i}^2} = \widehat{\sigma_e^2}(1-h_{ii}) \tag{11}$$

Where $\widehat{\sigma_{e_i}^2}, \widehat{\sigma_e^2}$ are the estimates of heteroscedastic, homoscedastic error variances and h_{ii} is the leading diagonal elements of the hat matrix, sometimes called as centered leverage values. If h_{ii} is close to 0, then the error variance of the i^{th} observation is equal to the homoscedastic variance, then the observation is said to be remote (outliers). In the same manner, if h_{ii} is close to 1, then the error variance of the i^{th} observation will be nearly 0 and the observation is a leverage point to the fitted regression equation. Many authors studied the hat matrix ($H = X\left(X^T X\right)^{-1} X^T$) and it's applications, but Belsley et al (1980) proposed a useful relationship between the centered hat values and F-ratio. They showed when the set of predictors followed a multivariate normal distribution with (μ_x, Σ_x) , then $\left((n-p)(h_{ii}-1/n)/(1-h_{ii})(p-1)\right) \sim F_{(p-1,n-p)}$ of i^{th} observation followed F-distribution with $(p-1, n-p)$ degrees of freedom respectively. Without loss of generality, the relationship can be written as

$$F_i = \frac{(n-p)(h_{ii}-1/n)}{(p-1)(1-h_{ii})} \sim F_{(p-1,n-p)} \tag{12}$$

From (12), solve it for h_{ii}, we get

$$h_{ii} = \frac{\left((p-1)/(n-p)\right)F_i + 1/n}{1+\left((p-1)/(n-p)\right)F_i} \tag{13}$$

From (13), the relationship between F_i and h_{ii} visualizes if $0 \le F_i < \infty$,then $(1/n) \le h_{ii} \le 1$ .So far, past researches emphasize $0 \le h_{ii} \le 1$,but based on the relationship from (13), we found h_{ii} lies between $1/n$ and 1.If the sample size is very large, that is $n \to \infty$,then the centered hat values will lies between 0 and 1.Based on the identified relationship from (13), the authors derived the distribution of the centered hat values and it is discussed in the next section.

3. Exact Distribution of Centered Hat Values

Using the technique of one-dimensional Jacobian transformation, the density function of F-ratio with $(p-1, n-p)$ degrees of freedom was transformed into density function of centered hat values for the i^{th} observation and it is given as

$$f\left(h_{ii}; p, n\right) = f\left(F_i; p-1, n-p\right) \left|J\right| \quad (14)$$

$$f\left(h_{ii}; p, n\right) = f\left(F_i; p-1, n-p\right) \left|\frac{dF_i}{dh_{ii}}\right| \quad (15)$$

From (15), we know the density function of F-ratio with $(p-1, n-p)$ degrees of freedom and the first derivative of the relationship from (15) are given as

$$f\left(F_i; p-1, n-p\right) = \frac{\left((p-1)/(n-p)\right)^{(p-1)/2}}{B\left(\frac{p-1}{2}, \frac{n-p}{2}\right)} F_i^{((p-1)/2)-1} \left(1+\frac{p-1}{n-p}F_i\right)^{-\left(\frac{p-1}{2}+\frac{n-p}{2}\right)} \quad (16)$$

where $0 \le F_i < \infty$, $n, p > 0$, $n > p$ and

$$\frac{dF_i}{dh_{ii}} = (n-p)(n-1)/n(p-1)(h_{ii}-1)^2 \quad (17)$$

Then substitute (16) and (17) in (15), we get the density function of centered hat values as

$$f\left(h_{ii}; p, n\right) = \frac{(n-1)/n(h_{ii}-1)^2}{B\left(\frac{p-1}{2}, \frac{n-p}{2}\right)} \left(\frac{h_{ii}-1/n}{1-h_{ii}}\right)^{((p-1)/2)-1} \left(1+\frac{h_{ii}-1/n}{1-h_{ii}}\right)^{-\left(\frac{p-1}{2}+\frac{n-p}{2}\right)} \quad (18)$$

where $(1/n) \le h_{ii} \le 1, n, p > 0, n > p$

From (18), it is the density function of centered hat values and it involves the beta function $B\left((p-1)/2, (n-p)/2\right)$ with two parameters (n, p), where n is the sample size and p is the no. of predictors in a multiple linear regression model. Moreover, the authors derived the first two moments of the distribution of centered hat values and it is given as follows.

$$E\left(h_{ii}\right) = \int_{1/n}^{1} h_{ii} f\left(h_{ii}; n, p\right) dh_{ii}$$

$$E\left(h_{ii}\right)=\int_{1/n}^{1} h_{ii}\frac{\left(n-1\right)/n\left(h_{ii}-1\right)^{2}}{B\left(\frac{p-1}{2},\frac{n-p}{2}\right)}\left(\frac{h_{ii}-1/n}{1-h_{ii}}\right)^{\left((p-1)/2\right)-1}\left(1+\frac{h_{ii}-1/n}{1-h_{ii}}\right)^{-\left(\frac{p-1}{2}+\frac{n-p}{2}\right)}dh_{ii}$$

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

$$E\left(h_{ii}\right)=p/n \tag{19}$$

$$E\left(h_{ii}^{2}\right)=\int_{1/n}^{1} h_{ii}^{2}\frac{\left(n-1\right)/n\left(h_{ii}-1\right)^{2}}{B\left(\frac{p-1}{2},\frac{n-p}{2}\right)}\left(\frac{h_{ii}-1/n}{1-h_{ii}}\right)^{\left((p-1)/2\right)-1}\left(1+\frac{h_{ii}-1/n}{1-h_{ii}}\right)^{-\left(\frac{p-1}{2}+\frac{n-p}{2}\right)}dh_{ii}$$

$$E\left(h_{ii}{}^{2}\right)=\frac{\left(\left(n-1\right)/n\right)^{2}}{B\left(\frac{p-1}{2},\frac{n-p}{2}\right)}\left(\sum_{k=0}^{\infty}\left(-1\right)^{k}\left(k+1\right)B\left(\frac{p-1}{2}+k,\frac{n-p}{2}-k\right)\right)+2\left(p/n\right)-1 \tag{20}$$

From (19) and (20), we know $E\left(h_{ii}\right)=p/n$ and $E\left(h_{ii}{}^{2}\right)=\frac{1}{n}\left(\frac{\left(n-1\right)\left(n-p+2\right)\left(n-p\right)}{n\left(n+1\right)}-\left(n-2p\right)\right)$ then the variance of the distribution was derived by substituting (19) and (20) in the following (21) we get

$$V\left(h_{ii}\right)=E\left(h_{ii}{}^{2}\right)-\left(E\left(h_{ii}\right)\right)^{2}$$

(21)

$$V\left(h_{ii}\right)=\frac{2\left(p-1\right)\left(n-p\right)}{n^{2}\left(n+1\right)} \tag{22}$$

$$\sigma\left(h_{ii}\right)=\sqrt{\frac{2\left(p-1\right)\left(n-p\right)}{n^{2}\left(n+1\right)}} \tag{23}$$

Moreover, the authors adopted two approaches of evaluating and identifying the leverage points in a sample. The first approach is to compute the critical points of the centered hat values by using the relationship between hat values and F-ratio from (13) is given as

$$h_{ii(n,p)}\left(\alpha\right)=\frac{\left(\left(p-1\right)/\left(n-p\right)\right)F_{i\left(p-1,n-p\right)}\left(\alpha\right)+1/n}{1+\left(\left(p-1\right)/\left(n-p\right)\right)F_{i\left(p-1,n-p\right)}\left(\alpha\right)} \tag{24}$$

From (24), for different values of (n, p) and for significance probability

$(\alpha = 0.05, 0.01)$, the percentage points were computed. The following table 1 and 2 shows the significance points of distribution of the hat values for varying sample size *(n)* and predictors *(p)* at 5% and 1% significance (α).using the percentage points, we can test the significance of the hat values computed from a multiple linear regression model to identify the leverage points. Similarly, the second approach is based on the control charts. The authors derived the lower and upper limits of the centered hat values by using the mean and standard deviation from (19) and (23) and both the limits are given as follows.

Lower limit of $h_{ii} = E(h_{ii}) - \sigma(h_{ii})$

$$\text{Lower limit of } h_{ii} = (p/n) - \sqrt{\frac{2(p-1)(n-p)}{n^2(n+1)}} \tag{25}$$

Upper limit of $h_{ii} = E(h_{ii}) + \sigma(h_{ii})$

$$\text{Upper limit of } h_{ii} = (p/n) + \sqrt{\frac{2(p-1)(n-p)}{n^2(n+1)}} \tag{26}$$

From (25) and (26), the authors computed the limits of the centered hat values for different combination of (n, p) and it is visualized in table no.3.Based on the limits, if the centered hat values of an observation is exceeding or above the upper limit, then the observation is said to be a leverage point in X-space. On the other hand, if it is below the lower limit, then the observation is said to be remote and technically it is an outlier in X-space. The following simulation graph shows the shape of the distribution of centered hat values for a fixed small sample size of 30 and for different values of p .From the curve of hat distribution, we observed for a fixed sample size of 30, it reaches the maximum probability if the no.of predictors used in a regression model will be more and the tail of the curves also touched the maximum hat value.

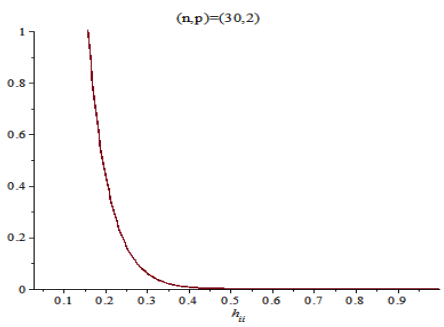

Fig.1

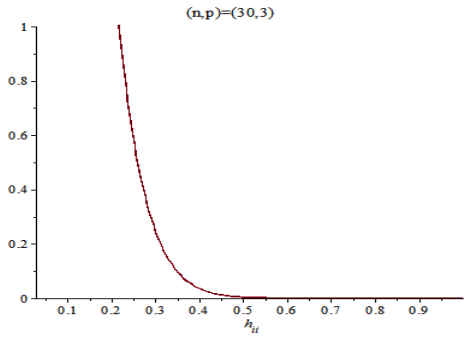

Fig.2

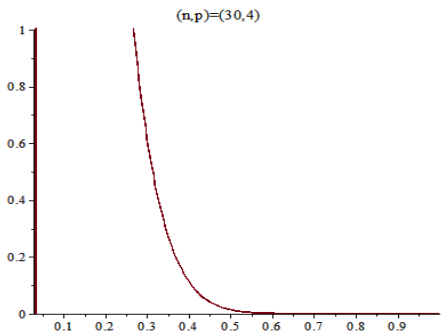

Fig.3

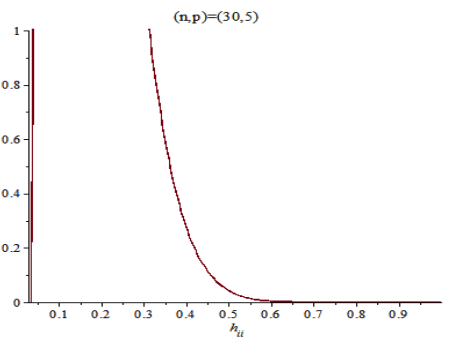

Fig.4

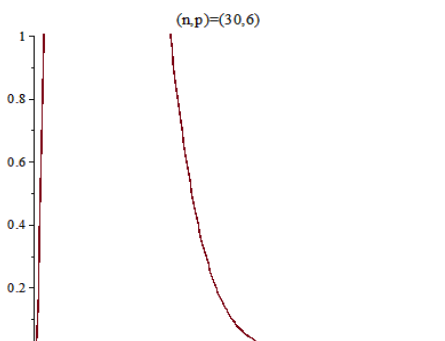

Fig.5

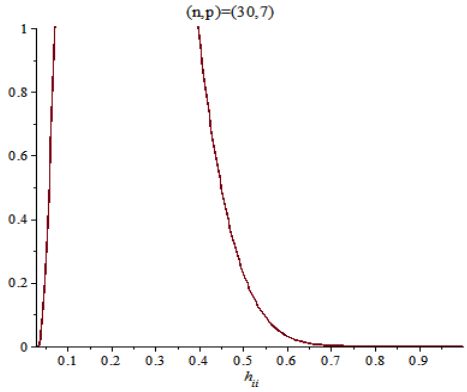

Fig.6

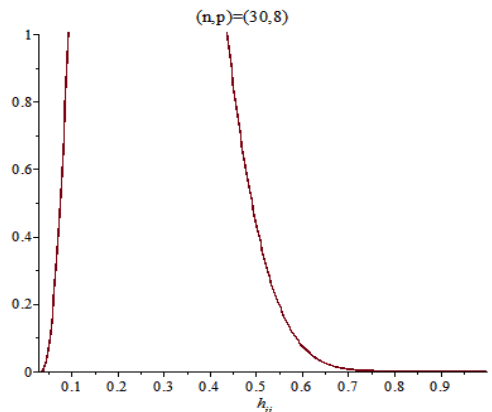

Fig.7

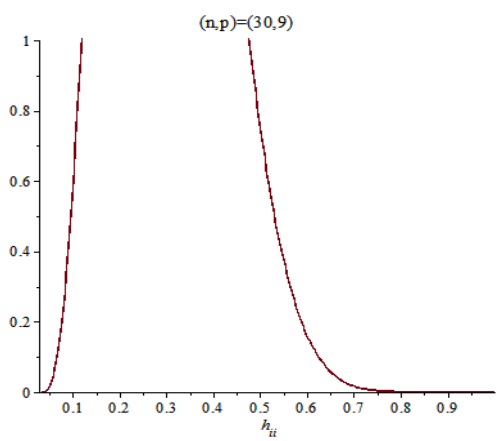

Fig.8

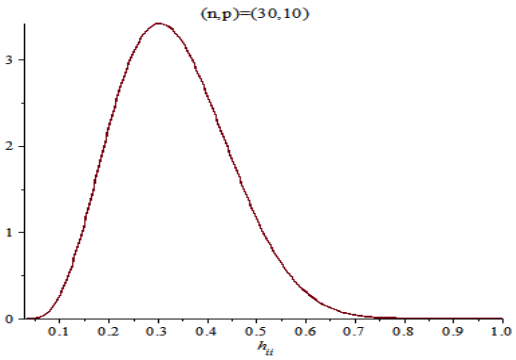

Fig.9

n	*p*									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
3	0.3333	.5531	-	-	-	-	-	-	-	-
4	0.2500	.3980	.5892	-	-	-	-	-	-	-
5	0.2000	.3127	.4337	.6264	-	-	-	-	-	-
6	0.1667	.2579	.3465	.4694	.6601	-	-	-	-	-
7	0.1429	.2196	.2895	.3791	.5033	.6894	-	-	-	-
8	0.1250	.1913	.2490	.3192	.4103	.5345	.7147	-	-	-
9	0.1111	.1695	.2186	.2763	.3478	.4397	.5629	.7365	-	-
10	0.1000	.1521	.1949	.2438	.3025	.3750	.4671	.5884	.7554	-
11	0.0909	.1380	.1759	.2183	.2680	.3276	.4006	.4924	.6114	.7719
12	0.0833	.1263	.1604	.1978	.2408	.2912	.3514	.4247	.5158	.6322
13	0.0769	.1164	.1473	.1808	.2187	.2624	.3134	.3740	.4472	.5373
14	0.0714	.1080	.1363	.1665	.2004	.2389	.2831	.3346	.3954	.4683
15	0.0667	.1007	.1267	.1544	.1850	.2193	.2583	.3030	.3548	.4156
16	0.0625	.0943	.1185	.1439	.1718	.2028	.2376	.2770	.3220	.3740
17	0.0588	.0887	.1112	.1348	.1604	.1886	.2200	.2552	.2950	.3402
18	0.0556	.0837	.1048	.1267	.1504	.1763	.2049	.2367	.2722	.3122
19	0.0526	.0793	.0991	.1196	.1416	.1656	.1918	.2208	.2528	.2886
20	0.0500	.0753	.0940	.1132	.1338	.1561	.1803	.2069	.2361	.2684
21	0.0476	.0717	.0894	.1075	.1268	.1476	.1701	.1946	.2215	.2509
22	0.0455	.0684	.0852	.1024	.1205	.1400	.1611	.1838	.2086	.2356
23	0.0435	.0654	.0814	.0977	.1149	.1332	.1529	.1741	.1971	.2221

24	0.0417	.0626	.0779	.0934	.1097	.1270	.1455	.1655	.1869	.2101
25	0.0400	.0601	.0747	.0895	.1050	.1214	.1389	.1576	.1777	.1993
26	0.0385	.0578	.0718	.0859	.1006	.1162	.1328	.1505	.1694	.1897
27	0.0370	.0556	.0691	.0826	.0966	.1115	.1272	.1440	.1618	.1809
28	0.0357	.0536	.0665	.0795	.0930	.1071	.1221	.1380	.1549	.1729
29	0.0345	.0518	.0642	.0766	.0895	.1031	.1174	.1325	.1486	.1656
30	0.0333	.0500	.0620	.0740	.0864	.0994	.1130	.1275	.1427	.1589
40	0.0250	.0413	.0547	.0675	.0802	.0930	.1059	.1191	.1325	.1464
60	0.0167	.0278	.0370	.0458	.0545	.0632	.0718	.0805	.0892	.0980
120	0.0083	.0139	.0186	.0230	.0273	.0316	.0358	.0400	.0441	.0483
∞	0	0	0	0	0	0	0	0	0	0

Table 1: Significant two-tail percentage points of hat values at $h_{ii(n,p)}\left(\alpha = 0.05\right)$

n	***p***									
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
3	0.3333	.9998	-	-	-	-	-	-	-	-
4	0.2500	.9996	.9999	-	-	-	-	-	-	-
5	0.2000	.9994	.9998	1	-	-	-	-	-	-
6	0.1667	.9992	.9998	.9999	1	-	-	-	-	-
7	0.1429	.9989	.9997	.9998	.9999	1	-	-	-	-
8	0.1250	.9987	.9996	.9998	.9999	.9999	1	-	-	-
9	0.1111	.9985	.9995	.9997	.9998	.9999	.9999	1	-	-
10	0.1000	.9982	.9994	.9997	.9998	.9999	.9999	1	1	-
11	0.0909	.9980	.9993	.9996	.9998	.9998	.9999	.9999	1	1
12	0.0833	.9977	.9992	.9995	.9997	.9998	.9999	.9999	.9999	1
13	0.0769	.9975	.9991	.9995	.9997	.9998	.9998	.9999	.9999	.9999
14	0.0714	.9973	.9990	.9994	.9996	.9997	.9998	.9999	.9999	.9999
15	0.0667	.9970	.9989	.9994	.9996	.9997	.9998	.9998	.9999	.9999
16	0.0625	.9968	.9988	.9993	.9995	.9997	.9998	.9998	.9999	.9999
17	0.0588	.9965	.9987	.9992	.9995	.9996	.9997	.9998	.9998	.9999
18	0.0556	.9963	.9986	.9992	.9995	.9996	.9997	.9998	.9998	.9999
19	0.0526	.9960	.9985	.9991	.9994	.9996	.9997	.9997	.9998	.9998

20	0.0500	.9958	.9984	.9991	.9994	.9995	.9996	.9997	.9998	.9998
21	0.0476	.9956	.9983	.9990	.9993	.9995	.9996	.9997	.9998	.9998
22	0.0455	.9953	.9982	.9989	.9993	.9995	.9996	.9997	.9997	.9998
23	0.0435	.9951	.9981	.9989	.9992	.9994	.9996	.9997	.9997	.9998
24	0.0417	.9948	.9980	.9988	.9992	.9994	.9995	.9996	.9997	.9998
25	0.0400	.9946	.9979	.9988	.9991	.9994	.9995	.9996	.9997	.9997
26	0.0385	.9943	.9978	.9987	.9991	.9993	.9995	.9996	.9997	.9997
27	0.0370	.9941	.9977	.9986	.9991	.9993	.9995	.9996	.9996	.9997
28	0.0357	.9939	.9976	.9986	.9990	.9993	.9994	.9995	.9996	.9997
29	0.0345	.9936	.9975	.9985	.9990	.9992	.9994	.9995	.9996	.9997
30	0.0333	.9934	.9974	.9985	.9989	.9992	.9994	.9995	.9996	.9996
40	0.0250	.2288	.3078	.3694	.4214	.4669	.5074	.5439	.5770	.6073
60	0.0167	.1306	.1738	.2090	.2402	.2689	.2957	.3210	.3452	.3683
120	0.0083	.0634	.0843	.1016	.1171	.1315	.1451	.1582	.1709	.1831
∞	0	0	0	0	0	0	0	0	0	0

Table 2: Significant two-tail percentage points of hat values at $h_{ii(n,p)}\left(\alpha = 0.01\right)$

n	*p*								
	1	*2*		*3*		*4*		*5*	
	LL/UL	*LL*	*UL*	*LL*	*UL*	*LL*	*UL*	*LL*	*UL*
3	0.333	0.431	0.902	-	-	-	-	-	-
4	0.250	0.276	0.724	0.526	0.974	-	-	-	-
5	0.200	0.200	0.600	0.369	0.831	0.600	1.000	-	-
6	0.167	0.155	0.512	0.282	0.718	0.448	0.885	0.655	1.000
7	0.143	0.126	0.445	0.227	0.631	0.357	0.786	0.512	0.916
8	0.125	0.106	0.394	0.189	0.561	0.296	0.704	0.421	0.829
9	0.111	0.091	0.354	0.161	0.505	0.252	0.637	0.357	0.754
10	0.100	0.079	0.321	0.140	0.460	0.219	0.581	0.309	0.691
11	0.091	0.070	0.293	0.124	0.421	0.194	0.534	0.273	0.636
12	0.083	0.063	0.270	0.111	0.389	0.173	0.493	0.244	0.590

13	0.077	0.057	0.250	0.101	0.361	0.157	0.459	0.220	0.549
14	0.071	0.053	0.233	0.092	0.337	0.143	0.429	0.201	0.514
15	0.067	0.048	0.218	0.085	0.315	0.131	0.402	0.184	0.482
16	0.063	0.045	0.205	0.078	0.297	0.121	0.379	0.170	0.455
17	0.059	0.042	0.194	0.073	0.280	0.113	0.358	0.158	0.430
18	0.056	0.039	0.183	0.068	0.265	0.105	0.339	0.148	0.408
19	0.053	0.037	0.174	0.064	0.252	0.099	0.322	0.139	0.388
20	0.050	0.035	0.165	0.060	0.240	0.093	0.307	0.130	0.370
21	0.048	0.033	0.158	0.057	0.229	0.088	0.293	0.123	0.353
22	0.045	0.031	0.151	0.054	0.219	0.083	0.280	0.117	0.338
23	0.043	0.029	0.144	0.051	0.210	0.079	0.269	0.111	0.324
24	0.042	0.028	0.139	0.049	0.201	0.075	0.258	0.106	0.311
25	0.040	0.027	0.133	0.046	0.194	0.072	0.248	0.101	0.299
26	0.038	0.026	0.128	0.044	0.186	0.069	0.239	0.096	0.288
27	0.037	0.025	0.124	0.043	0.180	0.066	0.230	0.092	0.278
28	0.036	0.024	0.119	0.041	0.173	0.063	0.222	0.089	0.269
29	0.034	0.023	0.115	0.039	0.168	0.061	0.215	0.085	0.260
30	0.033	0.022	0.111	0.038	0.162	0.059	0.208	0.082	0.251

n	*p*									
	6		*7*		*8*		*9*		*10*	
	LL	*UL*	*LL*	*UL*	*LL*	*UL*	*LL*	*UL*	*LL*	*UL*
3	-	-	-	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-	-	-	-
6	-	-	-	-	-	-	-	-	-	-
7	0.697	1.000	-	-	-	-	-	-	-	-
8	0.564	0.936	0.731	1.000	-	-	-	-	-	-
9	0.474	0.859	0.606	0.950	0.757	1.000	-	-	-	-
10	0.409	0.791	0.519	0.881	0.640	0.960	0.779	1.000	-	-
11	0.360	0.731	0.455	0.818	0.557	0.897	0.670	0.967	0.798	1.000

12	0.321	0.679	0.404	0.762	0.494	0.840	0.590	0.910	0.695	0.972
13	0.290	0.634	0.364	0.713	0.443	0.787	0.528	0.857	0.618	0.920
14	0.264	0.594	0.331	0.669	0.402	0.740	0.478	0.808	0.558	0.871
15	0.242	0.558	0.303	0.630	0.368	0.698	0.437	0.763	0.509	0.825
16	0.223	0.527	0.280	0.595	0.340	0.660	0.402	0.723	0.467	0.783
17	0.208	0.498	0.260	0.564	0.315	0.626	0.373	0.686	0.433	0.744
18	0.194	0.473	0.242	0.535	0.294	0.595	0.347	0.653	0.403	0.709
19	0.182	0.450	0.227	0.510	0.275	0.567	0.325	0.623	0.377	0.676
20	0.171	0.429	0.214	0.486	0.259	0.541	0.305	0.595	0.354	0.646
21	0.161	0.410	0.202	0.465	0.244	0.518	0.288	0.569	0.333	0.619
22	0.153	0.393	0.191	0.445	0.231	0.496	0.272	0.546	0.315	0.594
23	0.145	0.377	0.181	0.427	0.219	0.476	0.258	0.524	0.299	0.571
24	0.138	0.362	0.173	0.411	0.209	0.458	0.246	0.504	0.284	0.549
25	0.132	0.348	0.165	0.395	0.199	0.441	0.234	0.486	0.271	0.529
26	0.126	0.335	0.157	0.381	0.190	0.425	0.224	0.468	0.259	0.510
27	0.121	0.324	0.151	0.368	0.182	0.410	0.215	0.452	0.248	0.493
28	0.116	0.313	0.145	0.355	0.175	0.397	0.206	0.437	0.238	0.477
29	0.111	0.302	0.139	0.344	0.168	0.384	0.198	0.423	0.228	0.461
30	0.107	0.293	0.134	0.333	0.162	0.372	0.190	0.410	0.220	0.447

Table 3: Lower and upper limits of the centered hat values for combinations of (n, p)

4. Numerical Results and Discussion

In this section, we will investigate the discrimination between the traditional approach and the two proposed approaches of identifying the leverage points on the survey data collected from RSQ (Retail Services Quality) study. The data comprised of 20 different attributes about the retail stores and the data was collected from 275 customers. 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 following table shows the results extracted from the analysis by using SPSS version 20. At first, the authors used, stepwise multiple regression analysis by utilizing 19 independent variables and a dependent variable. The results of the stepwise regression analysis revealed, 4 fitted multiple linear regression models are significant with different set of predictors. For each model, the centered hat values were computed and the process of identifying the leverage points, comparative results of the proposed approaches with the traditional approaches are visualized in the following table.4

Model	*P*	Traditional approach			Proposed approach-I			
					Significance level			
					5%			1%
		Cut-off $(2p/n)$	Leverage points(n) > $(2p/n)$	Mean hat values of Leverage points	Critical hat values	Leverage points(n)	Mean hat values of Leverage points	Critical hat values
1	1	0.0072727	38	.0193428	0.00363	38	.0193428	0.00363

2	2	0.014545	21	0.039741	0.01758	21	.0446091	0.027603
3	3	0.021818	41	0.040602	0.02534	36	0.043004	0.036810
4	4	0.029091	43	0.05262	0.03191	42	0.05314	0.044413

Model	***p***	**Proposed approach-I**		**Proposed approach-II**					
		1% Significance level							
		Lever -age points (n)	**Mean hat values of Leverage points**	**UL**	**Leverage points(n) > UL**	**Mean hat values of Leverage points**	**LL**	**Outliers (n) <LL**	**Mean hat values of outliers**
1	1	38	0.193428	0.003636	38	0.003636	0.003636	237	0.001118
2	2	19	0.0412186	0.01238	41	0.0271215	0.002158	147	0.001498
3	3	21	0.0502705	0.01812	43	0.0396624	0.003689	122	0.002287
4	4	27	0.0607974	0.023371	49	0.0493099	0.005718	123	0.003515

p-no.of predictors *n*=275 LL-Lower limit UL-Upper limit

Table 4: Comparative results of proposed and traditional approaches

From the above tables, Table-4 visualizes the results of traditional approach and the proposed approach-I of evaluating the leverage points based on the centered hat values in a multiple linear regression model. As far as the fitted model-1 is concern, 38 observations found to be leverage points because the hat value of those observations are more than the cut-off of 0.0072727. Similarly, model-2, model-3 and model-4 are also having 21, 41 and 43 observations are leverage points respectively. This approach of evaluating the leverage points is a traditional one and experts revealed it is a rule of thumb but the proposed approach-I by the authors is scientific and it is based on the test of significance. In this approach the authors derived the percentage points of the centered hat values for each fitted multiple regression models at 5% and 1% level. As far as model-1 is concern 38 observations are said to be leverage points because the calculated hat values of these observations are greater than the critical values at 5% level of significance. Similarly, 21 observations in model-2, 36 observations in model-3 and 42 observations in model-4 are statistically proved as a leverage points at 5% level of significance. Moreover, 19 observations in model-2, 21 observations in model-3 and 27 observations in model-4 are found to be leverage points at 1% level of significance. Besides these the authors proposed another approach to evaluate the leverage as well as outlier points, because the proposed first approach only scrutinized the leverage points but fail to identify the remote or outliers in the X-space. The second approach is based

on the limits of the centered hat values. From Table-5 the authors computed the lower and upper limits of the centered hat values for each fitted regression models based on the no.of predictors and the sample size. As far as model-1 is concern the lower limit, upper limit and mean hat values of the leverage points are same because the fitted model-1 is having only one predictor. In this model, the hat values of 38 observations are above the upper limit and they are said to be leverage points, but the hat values of 237 observation are less than the lower limits. Hence these observations are treated as outliers. As far as model-2, 3 and 4 are concern, the hat values of 41, 43, 49 observations are above the respective upper limits such as 0.01238, 0.01812 and 0.023371 respectively. This shows, these observations are said to be leverage points in each fitted models. Similarly, the same models-2, 3 and 4 are concern, the hat values of 147, 122, 123 observations are below the respective lower limits such as 0.002158, 0.003689 and 0.005718 respectively. This shows, the hat value of these observations are close to 0 and they are said to be outliers. Finally from table-4 and table-5 the results of the proposed approach-I and II were shown along with the results of traditional approach. The identification of leverage points based on these two approaches are different when compared to the traditional approach. For example, the traditional approach emphasize model-3 is having 41 leverage points but the proposed approach-I critically claimed 36 observations are leverage points at 5% significance level and 21 observations are leverage points at 1% significance level respectively. This shows the proposed approach is discriminated from the traditional approach in identifying the leverage points. Similarly, the proposed approach-II is also different from the traditional approach, for example traditional approach emphasizes model-2 is having 21 leverage points but the proposed approach-II visualizes 41 observations are said to be leverage points and 147 observations are outliers in the same model. Hence, there is some discrimination between the two approaches in identifying the leverage points in a multiple linear regression model and the proposed approach-II helps us to identify the leverage as well as outliers. The following control charts visualize the results of the proposed approach-II.

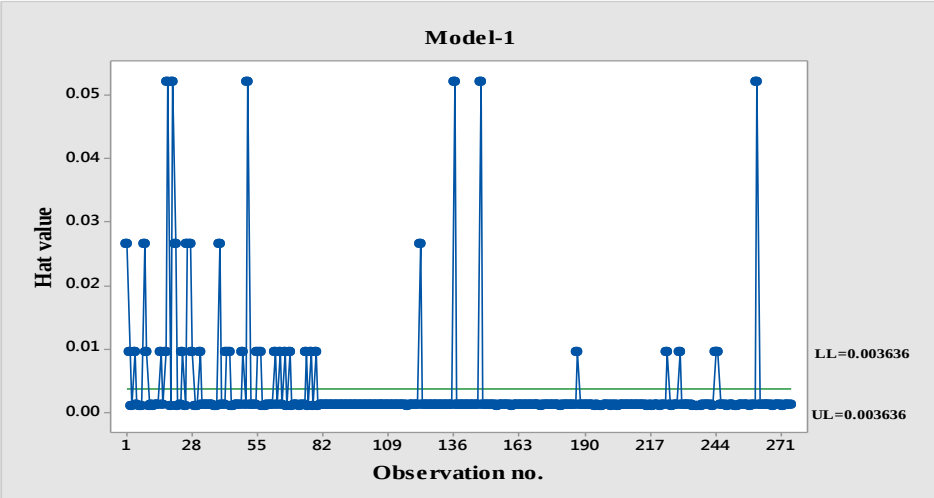

Fig.10

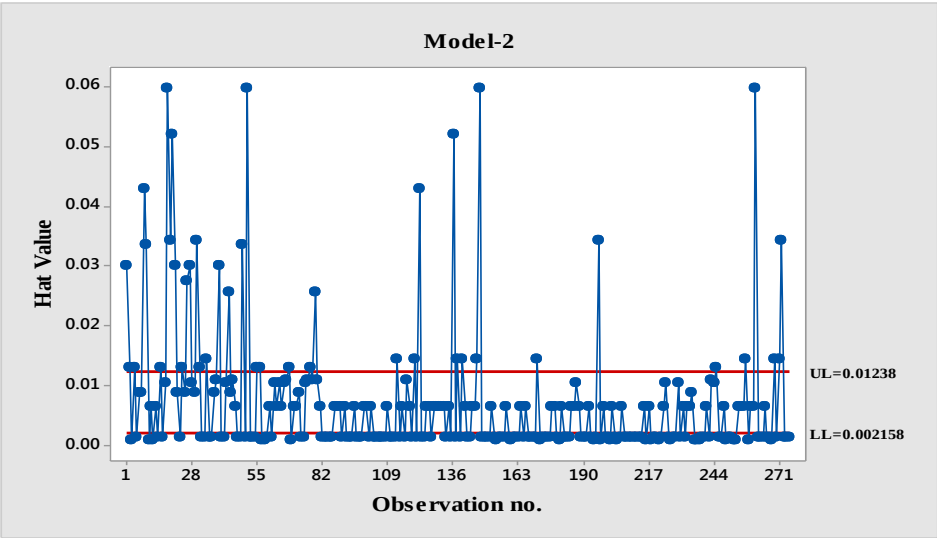

Fig.11

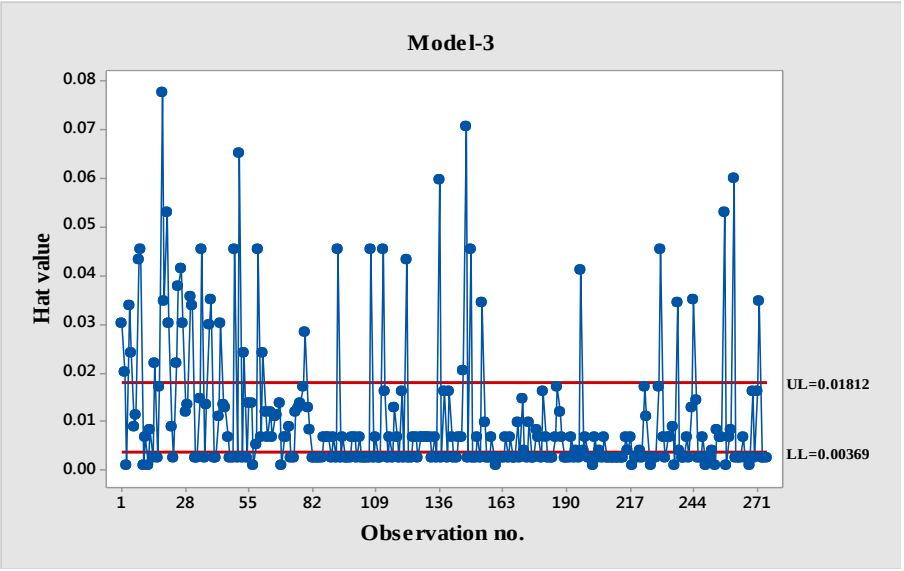

Fig.12

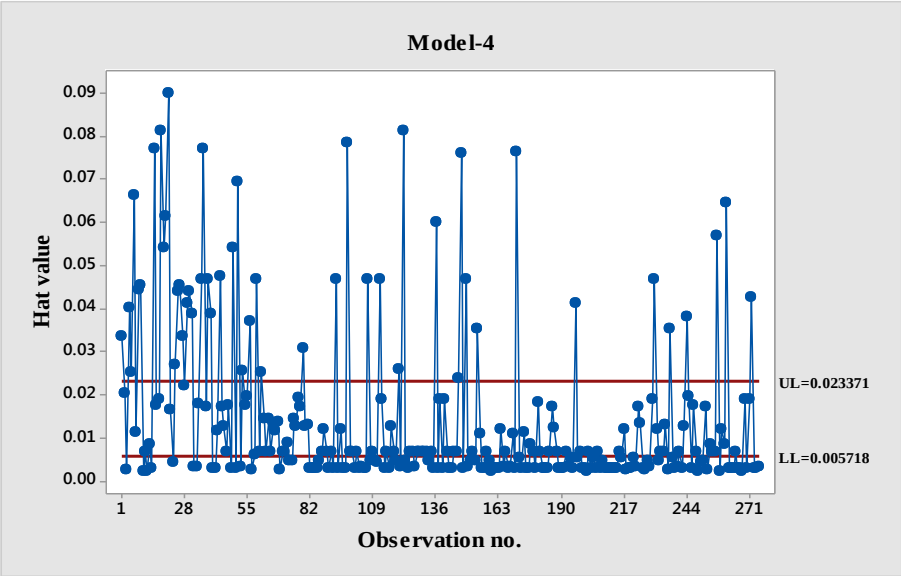

Fig.13

5. Conclusion

From the previous sections, the authors proposed two approaches of identifying and evaluating the leverage points in X-space in a multiple linear regression model. At first, the exact distribution of the leverage points was derived and the authors proved it followed a beta distribution with 2 shape parameters n and p. Similarly, the authors proved the leverage points are non-normally distributed and we found the hat values does not lies between 0 and 1 but it lies between 1/n and 1.If the sample size is very large, then the hat values in a multiple regression model lies between 0 and 1.Secondly, the proposed approach-I is a more systematic and scientific method of identifying the leverage points because it is based on the test of significance. The proposed approach-II is based on the limits of hat values and the authors treated if the hat values of an observation is beyond the upper limit is said to be a leverage point and if it is below the lower limit, then the observation is an outlier in X-space. Finally, the proposed approach-II is having an advantage over the Traditional approach and the proposed approach-I. This approach helps the statisticians to exactly identify the leverage as well as the outlier points in a multiple linear regression model. Moreover, the authors emphasizes that the hat values not only helps to identify the leverage points but also the outliers in X-space too. The hat values are non-normally distributed and using more rigorous parametric test based on normal distribution to test the differences in the means of the hat values are impossible.

References

1. Belsley, D. A., Kuh, E. and Welsch, R. E. (2005). Regression Diagnostics: Identifying Influential Data and Sources Of Collinearity, Wiley-Interscience., Vol. 571.
2. Chatterjee, S. and Hadi, A. S. (2009). Sensitivity Analysis in Linear Regression, Wiley, Vol. 327.
3. Chave, A. D. and Thomson, D. J. (2003). A bounded influence regression estimator based on the statistics of the hat matrix, J. Roy. Statist. Soc. C, 52(3), p. 307-322.
4. Diaz-Garcia, J.A. and Gonzalez- Faras, G. (2004). A note on the Cook's distance, J. Statist. Plan. Inference, 120, p. 119-136.
5. Dodge, Y. and Hadi, A. S. (1999). Simple graphs and bounds for the elements of the hat matrix, J. Appl.Statist., 26(7), p. 817-823.
6. Handschin, E., Schweppe, F. C., Kohlas, J. and Fiechter, A. (1975). Bad data analysis for power system state estimation. Power Apparatus and Systems, IEEE Transact, 94(2), p. 329-337.
7. Hoaglin, D. C. and Welsch, R. E. (1978). The hat matrix in regression and ANOVA, The Amer. Statist, 32(1), p. 17-22.
8. Huang, Y., Kuo, M. and Wang, T. (2007). Pair-perturbation influence functions and local influence in PCA, Comp. Statist. and Data Analysis, 51, p. 5886-5899
9. Krasker, W. S. and Welsch, R. E. (1982). Efficient bounded-influence regression estimation, J.Amer.Statist. Assoc., 77, p. 595-604.
10. Mallows, Colin L. (1975). On some topics in robustness, Unpublished memorandum, Bell Telephone Laboratories, Murray Hill, NJ.
11. Prendergast, L.A. (2005). Influence functions for sliced inverse regression. Scandinavian J. Statist., 32, p. 385-404.
12. Prendergast, L.A. (2006). Detecting influential observations in Sliced Inverse Regression analysis, Austral. and New Zealand J. Statist., 48, p. 285-304.
13. Pynnonen, Seppo (2010). Joint distribution of a linear transformation of OLS regression residuals with general spherical error distribution. Working Paper, Department of Mathematics and Statistics, University of Vaasa.
14. Ullah, M.A. and Pasha, G.R. (2009). The Origin and development of influence measures in regression, Pak. J. Statist, Vol. 25(3), p. 295-307.