Journal of Reliability and Statistical Studies; ISSN (Print): 0974-8024, (Online): 2229-5666
Vol. 9, Issue 1 (2016): 69-91

EXACT DISTRIBUTION OF SQUARED WELSCH-KUH DISTANCE AND IDENTIFICATION OF INFLUENTIAL OBSERVATIONS

G.S. David Sam Jayakumar[1] and A. Sulthan[2]
Jamal Institute of Management, Tiruchirappalli, India
E Mail: [1]samjaya77@gmail.com, [2]sulthan90@gmail.com.

Received March 22, 2015
Modified March 30, 2016
Accepted April 27, 2016

Abstract

This paper proposes the exact distribution of squared DFFITS alias squared Welsch-Kuh $\left(WK^2\right)$ distance measure used to evaluate the influential observations in a multiple linear regression analysis. The authors have explored the relationship between the WK^2 in terms of two independent F-ratio's and they have shown the derived density function of the WK^2 distance in a complicated series expression form involving Gauss hyper-geometric function with two shape parameters p and n. Moreover, the mean, variance of the distribution are derived in terms of the shape parameters and the authors have established the upper control limit of WK^2. Similarly, the critical points of squared Welsch-Kuh $\left(WK^2\right)$ distance measure are computed at 5% and 1% significance levesl for different sample sizes and varying no. of predictors. Finally, the numerical example shows the identification of the influential observations and the results extracted from the proposed approaches are more scientific, systematic and their exactness outperforms the Welsch-Kuh's traditional approach.

Key Words: Squared Welsch-Kuh Distance Measures, Influential Observation, Series Expression Form, Gauss Hyper-Geometric Function, Mean, Variance, Critical Points.

AMS Classification: 62H10

1. Introduction and Related work

The Studentized residuals and the plot of the residuals were considered the most appropriate statistical devices to detect potentially critical observations in the literature before the third quarter of the 20th century. Behnken and Draper (1972) have clarified that the estimated variance of the residuals includes pertinent information beyond that provided by plots of residuals or studentized residuals. Similarly, they discussed the variances of residuals in several more complicated designs. Hoaglin and Welsh (1978) expressed, projection matrix known as the hat matrix that contains this information and together with the studentized residuals, provides a means of identifying exceptional data points. Cook (1977) has been the first to establish a simple measure, D_i that incorporates information from the X-space and Y-space used for assessing the influential observations in regression models. The problem of outliers or influential data in the multiple or multivariate linear regression setting has been thoroughly discussed with reference to parametric regression models by the pioneers namely Cook (1977), Cook and Weisberg (1982), 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 & 2001) proposed a type of Cook's distance in kernel density estimation and in local polynomial regression. The phrase 'influence measures' has glimpsed a great surge of research interests. The developments of different measures are investigated to identify the influential observation from the early criteria of Cook's to the present and a definition about influence, which appears most suitable, is given by Belsey et al. (1980). Cook's statistical diagnostic measure is a simple, unifying and general approach for judging the local influence in statistical models. As far as the influence measures are concerned in the literature, the procedures were designed to detect the influence of observations on a specific regression result. However, Hadi (1992) proposed a diagnostic measure called Hadi's influence function to identify the overall potential influence which possesses several desirable properties that many of the frequently used diagnostics do not generally possess such as invariance to location and scale in the response variable and invariance to non-singular transformations of the explanatory variables. It is an additive function of measures of leverage and of residual error and it is monotonically increasing in the leverage values and in the squared residuals. Recently, Díaz-García and González-Farías (2004) modified the classical Cook's distance with generalized Mahalanobis distance in the context of multivariate elliptical linear regression models and they also established the exact distribution for identification of outlier data points. Considering the above reviews, the authors have proposed the exact distribution of Squared Welsch-Kuh distance $\left(WK^2\right)$ to exactly identify the influential data points and is discussed in the subsequent sections.

2. Relationship between Squared Welsch-Kuh distance $\left(WK^2\right)$ and F-ratios

The multiple linear regression model with random error is given by

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

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, but the DFFITS is the interesting technique based on the simple fact that the impact of the i^{th} on the predicted value can be measured by scaling the change in prediction at x_i ,when the i^{th} observations is omitted , i.e.

$$\frac{\left|\widehat{y_i} - \widehat{y_{(i)}}\right|}{\sigma\sqrt{h_{ii}}} = \frac{\left|x_i\left(\widehat{\beta} - \widehat{\beta_{(i)}}\right)\right|}{\sigma\sqrt{h_{ii}}} \tag{2}$$

Welsch and Kuh (1977), Welsch and Peters (1978) and Belsley, Kuh and Welsch

(1980) suggested using $\sigma^2_{(i)}$ an estimate of σ^2 and called (2) as DFFITS. For simplicity, they refer (2) by Welsch-Kuh distance $\left(WK_i\right)$,

$$WK_i = \frac{\left| x_i \left(\widehat{\beta} - \widehat{\beta_{(i)}} \right) \right|}{\widehat{\sigma}(i)\sqrt{h_{ii}}} = \left|R_i\right| \sqrt{\frac{h_{ii}}{1-h_{ii}}} \tag{3}$$

Where $\left|R_i\right|$ is the absolute externally studentized residual, '*n*' is the sample size, and h_{ii} is the hat value of i^{th} observation or diagonal element of the hat matrix $(H = X(X'X)^{-1}X')$. Welsch (1980) suggested WK_i as a diagnostic tool and $2\sqrt{\left(p+1\right)/n}$ as a calibration point for observations. The value of WK_i for observations exceeding this calibration point which is treated as influential observation and seems reasonable to nominate points for special attention, Welsch-Kuh distance measure can also be written in a squared alternative form as

$$WK_i^2 = R_i^2 \frac{h_{ii}}{1-h_{ii}} \tag{4}$$

Though the measure is scientific and the criterion $2\sqrt{\left(p+1\right)/n}$ used to detect the influential observation is not scientific and the authors believe that it is based on rule of thumb approach. In order to overcome this rule of thumb approach, authors made an attempt to make this approach more scientific by fixing meaning full criterion as calibration point. To identify the exact influential observations, we propose the exact distribution for squared Welsch-Kuh distance measure. For this, we utilize the relationship among the squared Welsch-Kuh distance $\left(WK_i^2\right)$, externally studentized residual $\left(R_i\right)$ 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 involve the set of predictors $(H = X(X'X)^{-1}X')$. Therefore, from the property of least squares if $E(eX) = 0$, then R_i and h_{ii} are also uncorrelated and independent. Using this assumption, we already know that the externally studentized residual $\left(R_i\right)$ exactly follows t-distribution with *n-p-2* degrees of freedom and it's squared form is given as

$$R_i^2 = \frac{\widehat{e_i}^2}{s^2_{e(-i)}\left(1-h_{ii}\right)} \sim F_{(1,n-p-2)} \tag{5}$$

From (5), it is the squared form of the externally studentized residual and it follows F-distribution with (1, *n-p-2*) degrees of freedom. Similarly, we identify the distribution of h_{ii} based on the relationship proposed by Belsley et al. (1980) who have shown that if the set of predictors follows multivariate normal distribution with (μ_X, Σ_X), then

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

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

$$h_{ii} = \frac{\left(\frac{p-1}{n-p}F_{i(p-1,n-p)}\right)+1/n}{1+\frac{p-1}{n-p}F_{i(p-1,n-p)}} \tag{6a}$$

In order to derive the exact distribution of squared Welsch-Kuh distance, without loss of generality substituting (5) and (6a) in (4), we get WK_i^2 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

$$WK_i^2 = \frac{n}{n-1}\left(\frac{p-1}{n-p}F_{i(p-1,n-p)}+\frac{1}{n}\right)F_{i(1,n-p-2)} \tag{7}$$

$$WK_i^2 = \frac{n(n-p-2)}{n-1}\left(\frac{p-1}{n-p}F_{i(p-1,n-p)}+\frac{1}{n}\right)\left(\frac{1}{n-p-2}F_{i(1,n-p-2)}\right) \tag{8}$$

From (8), it can be further simplified and WK_i^2 is expressed in terms of two independent beta variables of kind-2 namely θ_{1i} and θ_{2i} by using the following facts

$$\frac{p-1}{n-p}F_{i(p-1,n-p)} = \theta_{1i} \sim \beta_2\left(\frac{p-1}{2}, \frac{n-p}{2}\right) \tag{9}$$

$$\frac{1}{n-p-2}F_{i(1,n-p-2)} = \theta_{2i} \sim \beta_2\left(\frac{1}{2}, \frac{n-p-2}{2}\right) \tag{10}$$

Then, without loss of generality (8) can be written as

$$WK_i^2 = \frac{n(n-p-2)}{n-1}\left(\theta_{1i}+\frac{1}{n}\right)\theta_{2i} \tag{11}$$

$$WK_i^2 = \frac{n-p-2}{n-1}(n\theta_{1i}+1)\theta_{2i} \tag{12}$$

$$WK_i^2 = \alpha(p,n)(n\theta_{1i}+1)\theta_{2i} \tag{13}$$

From (13), the authors have shown the squared Welsch-Kuh distance measure in terms of $\theta_{1i} \sim \beta_2\left(\frac{p-1}{2}, \frac{n-p}{2}\right)$ and $\theta_{2i} \sim \beta_2\left(\frac{1}{2}, \frac{n-p-2}{2}\right)$ which followed beta distribution of kind-2 with two shape parameters *p*, *n* and $\alpha(p,n) = (n-p-2)/n-1$ is a normalizing function which involves the shape parameters respectively. Based on the identified relationship from (13), the authors have derived the distribution of the squared Welsch-Kuh distance which discussed in the next section.

3. Exact Distribution of Squared Welsch-Kuh distance

Using the technique of two-dimensional Jacobian of transformation, the joint probability density function of the two beta variables of kind-2 namely θ_{1i} and θ_{2i} were transformed into density function of new random variables WK_i^2 and u_i. It is given as

$$f\left(WK_i^2, u_i\right) = f\left(\theta_{1i}, \theta_{2i}\right)|J| \tag{14}$$

From (14), we know θ_{1i} and θ_{2i} are independent then rewrite (14) as

$$f\left(WK_i^2, u_i\right) = f\left(\theta_{1i}\right) f\left(\theta_{2i}\right)|J| \tag{15}$$

Using the change of variable technique, substitute $\theta_{2i} = u_i$ in (13) we get

$$\theta_{1i} = \frac{1}{n}\left(\frac{WK_i^2}{\alpha(p,n)u_i} - 1\right) \tag{16}$$

Then partially differentiate (16), compute the Jacobian determinant and rewrite (15) as

$$f\left(WK_i^2, u_i\right) = f\left(\theta_{1i}\right) f\left(\theta_{2i}\right)\left|\frac{\partial\left(\theta_{1i}, \theta_{2i}\right)}{\partial\left(WK_i^2, u_i\right)}\right| \tag{17}$$

$$f\left(WK_i^2, u_i\right) = f\left(\theta_{1i}\right) f\left(\theta_{2i}\right)\begin{vmatrix} \dfrac{\partial\theta_{1i}}{\partial WK_i^2} & \dfrac{\partial\theta_{1i}}{\partial u_i} \\ \dfrac{\partial\theta_{2i}}{\partial WK_i^2} & \dfrac{\partial\theta_{2i}}{\partial u_i} \end{vmatrix} \tag{18}$$

From (15), we know that θ_{1i} and θ_{2i} are independent and then the density function of the joint distribution of θ_{1i} and θ_{1i} is given as

$$f(\theta_{1i}, \theta_{2i}) = \frac{1}{B\left(\frac{p-1}{2}, \frac{n-p}{2}\right)}\theta_{1i}^{\frac{p-1}{2}-1}(1+\theta_{1i})^{-\left(\frac{p-1}{2}+\frac{n-p}{2}\right)} \times \frac{1}{B\left(\frac{1}{2}, \frac{n-p-2}{2}\right)}\theta_{2i}^{\frac{1}{2}-1}(1+\theta_{2i})^{-\left(\frac{1}{2}+\frac{n-p-2}{2}\right)} \tag{19}$$

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

and

$$\begin{vmatrix} \dfrac{\partial\theta_{1i}}{\partial WK_i^2} & \dfrac{\partial\theta_{1i}}{\partial u_i} \\ \dfrac{\partial\theta_{2i}}{\partial WK_i^2} & \dfrac{\partial\theta_{2i}}{\partial u_i} \end{vmatrix} = \begin{vmatrix} \dfrac{1}{n\alpha(p,n)u_i} & -\dfrac{(WK_i)^2}{n\alpha(p,n)u_i^2} \\ 0 & 1 \end{vmatrix} = \frac{1}{n\alpha(p,n)u_i} \tag{20}$$

Then substituting (19) and (20) in (18) in terms of the substitution of u_i, we get the joint distribution of WK_i^2 and u_i as

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

$$\times\frac{1}{B\left(\frac{1}{2},\frac{n-p-2}{2}\right)}u_i^{\frac{1}{2}-1}\left(1+u_i\right)^{-\left(\frac{1}{2}+\frac{n-p-2}{2}\right)}\times|J| \tag{21}$$

where $0\le WK_i^2<\infty, 0\le u_i<\infty$ and $|J|=\dfrac{1}{n\alpha(p,n)u_i}$

Using Binomial series expansion, rearrange (21), and integrate with respect to u_i, we get the marginal distribution of WK_i^2 as

$$f\left(WK_i^2\right)=\phi(p,n)\sum_{k=0}^{\infty}\sum_{r=0}^{\frac{p-3}{2}+k}\binom{-(n-1)/2}{k}\binom{\frac{p-3}{2}+k}{r}\left(\frac{1}{n}\right)^k(-1)^r\left(\alpha(p,n)\right)^{\frac{p-3}{2}+k-r}\left(WK_i^2\right)^{\frac{p-3}{2}+k-r}.$$

$$\int_0^\infty u_i^{r-\left(\frac{p-2}{2}+k\right)-1}\left(1+u_i\right)^{-\left(r-\left(\frac{p-2}{2}+k\right)+\left(\frac{n-p-1}{2}-r-\left(\frac{p-2}{2}+k\right)\right)\right)}du_i \tag{22}$$

where $0\le WK_i^2<\infty$, $n,p>0$

We know, from (22)

$$\int_0^\infty u_i^{r-\left(\frac{p-2}{2}+k\right)-1}\left(1+u_i\right)^{-\left(r-\left(\frac{p-2}{2}+k\right)+\left(\frac{n-2p+1}{2}-(k+r)\right)\right)}du_i=B\left(r-\left(\frac{p-2}{2}+k\right),\frac{n-2p+1}{2}-(k+r)\right) \tag{23}$$

Then substitute (23) in (22) and arrange the terms, we get the density function of WK_i^2 in the series expression form as

$$f\left(WK_i^2;p,n\right)=\phi(p,n)\sum_{k=0}^{\infty}\sum_{r=0}^{\frac{p-3}{2}+k}\binom{-(n-1)/2}{k}\binom{\frac{p-3}{2}+k}{r}\left(\frac{1}{n}\right)^k(-1)^r\left(\frac{WK_i^2}{\alpha(p,n)}\right)^{\frac{p-3}{2}+k-r}.$$

$$B\left(r-\left(\frac{p-2}{2}+k\right),\frac{n-2p+1}{2}-(k+r)\right) \tag{24}$$

Further (24) is reduced by expanding the series with respect to 'r' and we get

$$f\left(WK_i^2;p,n\right)=\phi\left(p,n\right)\sum_{k=0}^{\infty}\binom{-(n-1)/2}{k}\left(\frac{1}{n}\right)^k\left(\frac{WK_i^2}{\alpha\left(p,n\right)}\right)^{\frac{p-3}{2}+k}B\left(\frac{-p}{2}+1-k,\frac{n+1}{2}-\left(p+k\right)\right)$$
$$\cdot{}_2F_1\left(\frac{-p}{2}+1-k,\frac{-p+3}{2}-k;\frac{-n+1}{2}+p+k;-\frac{\alpha\left(p,n\right)}{WK_i^2}\right)$$

(25)

where, $0\le WK_i^2<\infty$, $n,p>0$, $n>p$,

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

From (25), it is the density function of squared Welsch-Kuh distance measure which involves the functions such as ${}_2F_1$ is the Gauss hyper-geometric function and the normalizing constants are $\alpha(p,n)$ and $\phi(p,n)$ comprised of two Beta functions namely $B\left(\frac{p-1}{2},\frac{n-p}{2}\right)$, $B\left(\frac{1}{2},\frac{n-p-2}{2}\right)$ 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 respectively. In order to know the location and dispersion of squared Welsch-Kuh distance, the authors derived the mean, variance from (13) and it is shown as follows. Using (13), taking expectation, we get

$$E\left(WK_i^2\right)=\frac{n-p-2}{n-1}\left(nE\left(\theta_{1i}\right)+1\right)E\left(\theta_{2i}\right) \tag{26}$$

Then, substituting the mean of two independent beta variables θ_{1i} and θ_{2i} of kind-2 in (26), we get the mean of $\left(WK_i^2\right)$ as

$$E\left(WK_i^2\right)=\frac{p\left(n-1\right)-2}{\left(n-p-4\right)\left(n-1\right)} \tag{27}$$

Similarly, compute the difference between (13) and (26), then square it and take expectations, we get

$$V\left(WK_i^2\right)=\left(\frac{n-p-2}{n-1}\right)^2\left(n^2\left(E\left(\theta_{1i}^2\right)E\left(\theta_{2i}^2\right)-\left(E\left(\theta_{1i}\right)E\left(\theta_{2i}\right)\right)^2\right)+V\left(\theta_{2i}\right)\left(1+2nE\left(\theta_{1i}\right)\right)\right)$$

(28)

Then, substitute the appropriate moments of beta variables θ_{1i} and θ_{2i} of kind-2 in (28), we get the Variance of $\left(WK_i^2\right)$ as

$$V\left(WK_i^2\right)=\left(\frac{n-p-2}{n-1}\right)^2\left(n^2\left(\varphi_1\left(p,n\right)-\varphi_2\left(p,n\right)\right)+\varphi_3\left(p,n\right)\right) \tag{29}$$

Where

$$\varphi_1(p,n)=\frac{3\left(p^2-1\right)}{(n-p-4)^2(n-p-2)(n-p-6)}$$

$$\varphi_2(p,n)=\left(\frac{p-1}{(n-p-2)(n-p-4)}\right)^2$$

$$\varphi_3(p,n)=\frac{2(n-p-3)(2np-n-p-2)}{(n-p-4)^2(n-p-2)(n-p-6)}$$

Moreover, by using the mean and variance of squared Welsch-Kuh distance measure, the authors established the upper control limit of $\left(WK_i^2\right)$ for i^{th} observation based on different combination of (p,n) and it is given as

$$UCL\left(WK_i^2\right)=E\left(WK_i^2\right)+\sqrt{V\left(WK_i^2\right)} \tag{30}$$

Then substitute (27) and (29) in (30), we get

$$UCL\left(WK_i^{\,2}\right)=\frac{p(n-1)-2}{(n-p-4)(n-1)}+\frac{n-p-2}{n-1}\sqrt{n^2\left(\phi_1(p,n)-\phi_2(p,n)\right)+\phi_3(p,n)} \tag{31}$$

where $n-p>6$

By using (31), as a first approach, the authors utilize the upper control limit as a cut-off to identify the influential observation in a multiple linear regression model. The computed $\left(WK_i^2\right)$ of any observation exceeds the upper control limit, then the observation is treated as influential. Secondly, 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 squared Welsch-Kuh distance measure by utilizing the following relationship from (7) and it is given as

$$WK^2_{i(p,n)}(\alpha)=\frac{n}{n-1}\left(\frac{p-1}{n-p}F_{i(p-1,n-p)}(\alpha)+\frac{1}{n}\right)F_{i(1,n-p-2)}(\alpha) \tag{32}$$

where $n-p>2$

From (32) for different combinations of values of (p,n) and based on the significance probability $p\left(WK_i^2>WK^2_{i(p,n)}(\alpha)\right)=\alpha$, we computed the critical points of squared Welsch-Kuh distance measure. By using the critical points, we can test the significance of the influential observation in a multiple linear regression model. The following Table-1 visualizes the upper control limit of the squared Welsch-Kuh distance measure computed from (31) and Tables 2, 3 exhibit the significant two tail percentage points of the distribution of WK_i^2 measure for varying sample size *(n)* and no.of predictors *(p)* at 5% and 1% significance (α).

n	***p***				
	1	2	3	4	5
8	1.0462	-	-	-	-
9	0.67662		-	-	-
10	0.51112	2.3285	6.0708	-	-
11	0.41395	1.6594	3.5767	8.0755	-
12	0.34921	1.2887	2.5379	4.7496	10.023
13	0.30264	1.0530	1.9640	3.3647	5.8899
14	0.26738	0.88994	1.6001	2.6002	4.1692
15	0.23969	0.77052	1.3491	2.1158	3.2196
16	0.21736	0.67929	1.1656	1.7821	2.6183
17	0.19889	0.60732	1.0257	1.5385	2.2041
18	0.18337	0.54913	0.91570	1.3529	1.9019
19	0.17015	0.50112	0.82684	1.2069	1.6717
20	0.15873	0.46083	0.75369	1.0892	1.4909
21	0.14878	0.42653	0.69232	0.9922	1.3450
22	0.14001	0.39697	0.64021	0.9111	1.2250
23	0.13223	0.37127	0.59535	0.8421	1.1245
24	0.12528	0.34867	0.55635	0.7829	1.0391
25	0.11903	0.32869	0.52216	0.7314	0.96580
26	0.11338	0.31085	0.49187	0.6862	0.90208
27	0.10824	0.29486	0.46492	0.6463	0.84621
28	0.10356	0.28043	0.44078	0.6107	0.79684
29	0.099262	0.26735	0.41902	0.5788	0.75289
30	0.095310	0.39930	0.39930	0.5502	0.71349
40	0.068226	0.17672	0.27148	0.3677	0.46832
60	0.043548	0.10936	0.16548	0.2210	0.27742
80	0.031993	0.079184	0.11902	0.1579	0.19706
100	0.025286	0.062065	0.092923	0.1229	0.15279
120	0.020904	0.051030	0.076216	0.1007	0.12476

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

n	***p***				
	6	7	8	9	10
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-
11	-	-	-	-	-
12	-	-	-	-	-
13	11.939	-	-	-	-
14	7.0125	13.835	-	-	-
15	4.9617	8.1249	15.718	-	-
16	3.8300	5.7466	9.2293	17.592	-
17	3.1135	4.4347	6.5270	10.329	19.459
18	2.6201	3.6042	5.0356	7.3036	11.424
19	2.2599	3.0321	4.0919	5.6340	8.0771
20	1.9861	2.6151	3.4420	4.5776	6.2303
21	1.7708	2.2977	2.9681	3.8500	5.0614
22	1.5972	2.0483	2.6075	3.3195	4.2568
23	1.4544	1.8473	2.3242	2.9160	3.6699
24	1.3349	1.6819	2.0960	2.5991	3.2235
25	1.2334	1.5434	1.9081	2.3435	2.8729
26	1.1462	1.4260	1.7509	2.1332	2.5904
27	1.0704	1.3250	1.6175	1.9576	2.3580
28	1.0040	1.2374	1.5029	1.8083	2.1635
29	0.94534	1.1605	1.4033	1.6801	1.9984
30	0.89307	1.0925	1.3161	1.5686	1.8566
40	0.57494	0.68871	0.81076	0.94219	1.0843
60	0.33551	0.39566	0.45811	0.52316	0.59099
80	0.23684	0.27749	0.31917	0.36200	0.40606
100	0.18300	0.21366	0.24489	0.27674	0.30925
120	0.14911	0.17371	0.19865	0.22398	0.24972

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

Table 1: Upper control limit of squared Welsch-Kuh $\left(WK_i^2\right)$ distance for combinations of $\left(p,n\right)$

n	*P*									
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
4	53.8	-	-	-	-	-	-	-	-	-
5	4.62	721.	-	-	-	-	-	-	-	-
6	2.02	46.5	1266.	-	-	-	-	-	-	-
7	1.28	17.3	78.07	1774.	-	-	-	-	-	-
8	.944	9.89	28.23	107.2	2266.	-	-	-	-	-
9	.748	6.76	15.83	38.24	135.3	2748.	-	-	-	-
1	.621	5.08	10.67	21.22	47.86	162.9	3225.	-	-	-
1	.531	4.05	7.940	14.20	26.39	57.27	190.1	3698.	-	-
1	.465	3.36	6.278	10.50	17.57	31.44	66.55	217.0	4168.	-
1	.413	2.86	5.170	8.265	12.94	20.85	36.41	75.74	243.8	4637.
1	.372	2.49	4.383	6.779	10.15	15.32	24.08	41.33	84.86	270.5
1	.339	2.20	3.799	5.729	8.306	11.98	17.65	27.28	46.21	93.94
1	.311	1.98	3.348	4.951	7.003	9.786	13.78	19.96	30.45	51.07
1	.287	1.79	2.990	4.353	6.040	8.238	11.23	15.56	22.24	33.60
1	.267	1.63	2.700	3.881	5.302	7.094	9.447	12.67	17.32	24.51
1	.249	1.50	2.461	3.498	4.719	6.219	8.126	10.63	14.08	19.07
2	.234	1.39	2.259	3.182	4.248	5.529	7.116	9.142	11.81	15.49
2	.220	1.30	2.088	2.918	3.860	4.972	6.320	7.999	10.14	12.98
2	.208	1.21	1.940	2.693	3.536	4.514	5.679	7.099	8.872	11.14
2	.197	1.14	1.812	2.500	3.261	4.131	5.152	6.375	7.869	9.737
2	.188	1.07	1.699	2.332	3.024	3.807	4.712	5.780	7.062	8.631
2	.179	1.01	1.600	2.185	2.819	3.528	4.340	5.283	6.399	7.742
2	.171	.965	1.511	2.055	2.639	3.287	4.020	4.863	5.846	7.013
2	.163	.918	1.431	1.940	2.481	3.076	3.743	4.502	5.379	6.404
2	.157	.875	1.360	1.836	2.340	2.890	3.501	4.191	4.978	5.890
2	.150	.836	1.295	1.743	2.214	2.725	3.288	3.918	4.632	5.449
3	.145	.800	1.236	1.659	2.101	2.577	3.099	3.678	4.329	5.068
4	.105	.560	.8487	1.117	1.387	1.666	1.960	2.271	2.603	2.961
6	.068	.349	.5208	.6746	.8239	.9729	1.124	1.278	1.438	1.602
8	.050	.254	.3754	.4827	.5852	.6861	.7868	.8883	.9912	1.095
1	.039	.199	.2935	.3757	.4537	.5297	.6049	.6801	.7557	.8320
1	.033	.164	.2409	.3075	.3704	.4313	.4913	.5509	.6105	.6703
∞	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Table 2: Significant two-tail percentage points of squared Welsch-Kuh $\left(WK_i^2\right)$ distance at $p\left(WK_i^2 > WK_{i(p,n)}^2(0.05)\right) = 0.05$

n	***p***									
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
4	1350	-	-	-	-	-	-	-	-	-
5	24.6	5861	-	-	-	-	-	-	-	-
6	6.82	646.1	10070	-	-	-	-	-	-	-
7	3.53	135.1	1050.	13993	-	-	-	-	-	-
8	2.32	59	211.8	1423.	17785	-	-	-	-	-
9	1.71	34.03	89.49	281.9	1782.	21504	-	-	-	-
1	1.36	23.02	51.07	117.5	349.2	2134.	25178	-	-	-
1	1.12	17.03	34.06	66.40	144.3	414.9	2481.	28820	-	-
1	.960	13.36	24.92	43.93	81.00	170.4	479.7	2825.	32441	-
1	.837	10.91	19.37	31.94	53.30	95.21	196.1	543.8	3166.	36046
1	.742	9.183	15.71	24.69	38.58	62.41	109.1	221.6	607.4	3506.
1	.666	7.902	13.14	19.94	29.72	45.02	71.34	122.9	246.8	670.7
1	.604	6.921	11.25	16.61	23.92	34.59	51.34	80.16	136.6	271.9
1	.553	6.148	9.809	14.17	19.88	27.77	39.35	57.56	88.90	150.2
1	.510	5.524	8.679	12.32	16.92	23.03	31.54	44.05	63.73	97.58
1	.473	5.011	7.772	10.88	14.68	19.57	26.11	35.25	48.70	69.85
2	.442	4.582	7.029	9.722	12.94	16.95	22.16	29.14	38.91	53.30
2	.414	4.219	6.411	8.776	11.54	14.92	19.17	24.70	32.13	42.55
2	.389	3.908	5.889	7.990	10.40	13.29	16.85	21.35	27.21	35.10
2	.368	3.638	5.443	7.328	9.463	11.97	14.99	18.74	23.49	29.69
2	.348	3.403	5.057	6.764	8.669	10.87	13.49	16.67	20.61	25.62
2	.331	3.195	4.721	6.277	7.993	9.953	12.24	14.98	18.32	22.46
2	.315	3.011	4.426	5.854	7.411	9.169	11.20	13.59	16.45	19.95
2	.300	2.846	4.165	5.482	6.904	8.495	10.31	12.42	14.91	17.91
2	.287	2.698	3.932	5.153	6.461	7.909	9.547	11.43	13.62	16.22
2	.275	2.565	3.723	4.861	6.069	7.396	8.883	10.57	12.53	14.81
3	.264	2.444	3.535	4.599	5.721	6.943	8.303	9.837	11.59	13.61
4	.189	1.657	2.341	2.975	3.614	4.276	4.975	5.722	6.527	7.400
6	.120	1.004	1.390	1.733	2.064	2.394	2.728	3.070	3.423	3.789
8	.088	.7202	.9872	1.220	1.441	1.657	1.872	2.089	2.308	2.532
1	.069	.5610	.7650	.9409	1.106	1.266	1.424	1.581	1.739	1.898
1	.057	.4594	.6243	.7655	.8972	1.024	1.148	1.271	1.394	1.517
∞	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Table 3: Significant two-tail percentage points of Squared Welsch-Kuh $\left(WK_i^2\right)$ **distance at** $p\left(WK_i^2 > WK_{i(p,n)}^2\left(0.01\right)\right) = 0.01$

4. Numerical Results and Discussion

In this section, the authors have shown a numerical study of evaluating the influential observation based on squared Welsch-Kuh distance of the i^{th} observation in a regression model. For this, the authors have fitted step-wise linear regression models with different set of predictors in a brand equity study. The data in the study comprising of 19 different attributes about a car brand were collected from 275 car users. A well-structured questionnaire was prepared and distributed to 300 customers and the questions were anchored at five point Likert scale from 1 to 5. After the data collection was over, only 275 completed questionnaires were used for analysis. The step-wise regression results revealing 4 nested models were extracted from the regression procedure by using IBM SPSS version 22. For each model, the Welsch-Kuh (WK) and squared Welsch-Kuh (WK^2) distances were computed, the comparative results of proposed approaches I and II with the traditional Welsch-Kuh's distance approach of identifying the influential observations are visualized in the following Tables 4 and 5.

Model	*p*	**Traditional Welsch-Kuh's approach**		**Proposed approach-I**	
		Cut-off $(WK_i)=2\sqrt{(p+1)/n}$	***n*** > Cut-off (WK_i)	*UCL (WK_i^2)	***n*** > UCL (WK_i^2)
1	1	0.17056	12	0.0089248	25
2	2	0.20889	17	0.021462	24
3	3	0.24121	17	0.031842	29
4	4	0.26968	22	0.041762	28

p-no.of predictors n=275 *UCL $(WK_i^2)=E(WK_i^2)+\sqrt{V(WK_i^2)}$ - refer (31)

Table-4: Identification of influential observations, Comparative results of Traditional Welsch-Kuh's approach and proposed approach-I

Model	***p***	**Traditional Welsch-Kuh's approach**		**Proposed approach-II**			
		Cut-off $(WK_i)=2\sqrt{(p+1)/n}$	***n*** > Cut-off (WK_i)	**Critical** $WK_i^2(0.05$	***n*** > $WK_i^2(0.05)$	**Critical** $WK_i^2(0.01)$	***n*** > $WK_i^2(0.01)$
1	1	0.17056	12	0.01415	19	0.02456	17
2	2	20889	17	0.06937	13	0.19102	7
3	3	0.24121	17	0.10079	13	0.25719	8
4	4	0.26968	22	0.12775	13	0.31279	7

p-no.of predictors n=275

Table-5: Identification of influential observations, Comparative results of Traditional Welsch-Kuh's approach and proposed approach-II

Table 4 and 5 visualizes the comparative results of Welsch-Kuh's traditional approach of evaluating the influential observations with the proposed approach I and II. Firstly four nested multiple regression models are fitted and the cut-off distances based on Welsch-Kuh's traditional approach are shown in the tables. As far as the fitted model-1 is concerned, the computed Welsch-kuh's distance measure for 12 observations where above the cut-off distance and hence these observations are said to be influential. Similarly in model 2 and model 3, 17 observations are finalized as influential and in the same manner, in model 4, the calculated Welsch-Kuh distance measure for 22 observations are above the cut-off and hence these observations are said to be influential. Under proposed approach I, the cut-off was scientifically determined and in model 1, the calculated value of squared Welsch-Kuh distance measure for 25 observations is above the cut-off and in model-2, 24 observations, in model-3, 29 observations and in model-4, 28 observations are exceeding the scientifically determined upper control limit. Hence these observations are treated as influential observations. Under the proposed approach-II, the authors adopted the test of significance approach to identify the influential observations. As far as the model 1 is concerned, the computed values of squared Welsch-Kuh distance measure for 19 observations is greater than the critical (WK^2) value at 5% significance level and in model 2, model 3and model 4, the authors identified 13 observations in each fitted model as influential at 5% significance level. Likewise 17, 7, 8 and 7 observations are treated as influential at 1 % significance level in model 1, model 2, model 3 and model 4 respectively. Finally, among the three approaches, the proposed approach-I identified more influential observations when compared to Welsch-kuh's traditional approach and proposed approach II. On the other hand, the proposed approach II is systematic and scientific when compared to Welsch-Kuh's traditional approach and proposed approach-I, because the cut-off critical (WK^2) at different significance levels is scientifically determined from the distribution of squared Welsch-Kuh's distance measure. Hence the authors observed, the proposed approach-I and II outperforms the Welsch-Kuh's traditional approach in identifying influential observations and 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.

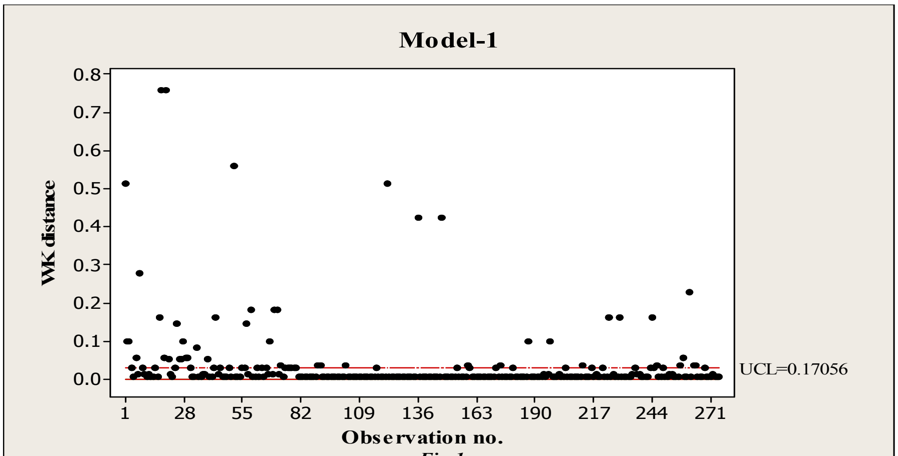

Model-1
WK distance
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
UCL=0.17056
1 28 55 82 109 136 163 190 217 244 271
Observation no.

Fig.1

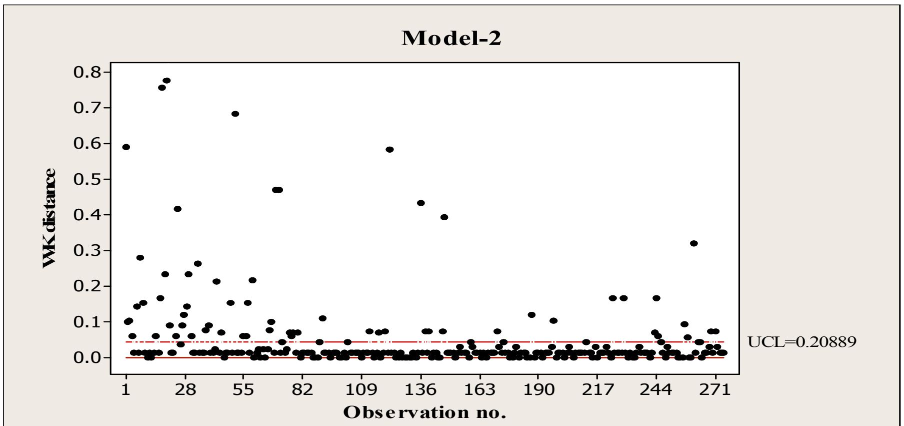

Model-2
WK distance
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
UCL=0.20889
1 28 55 82 109 136 163 190 217 244 271
Observation no.

Fig.2

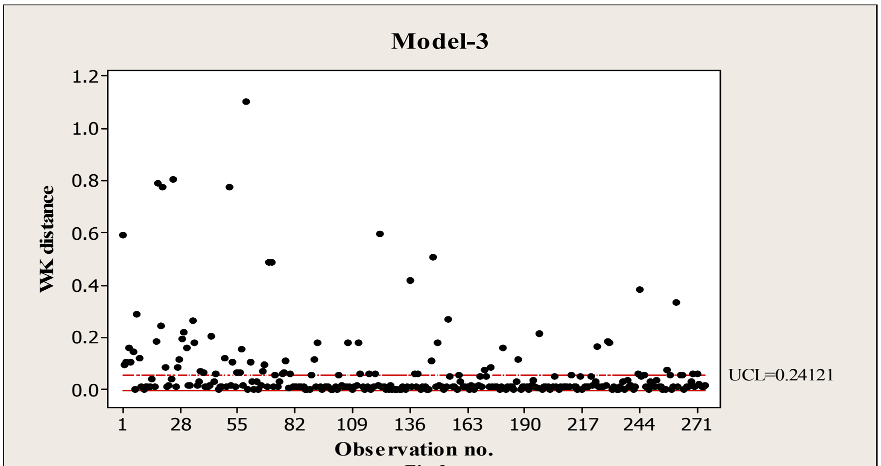


Fig.3

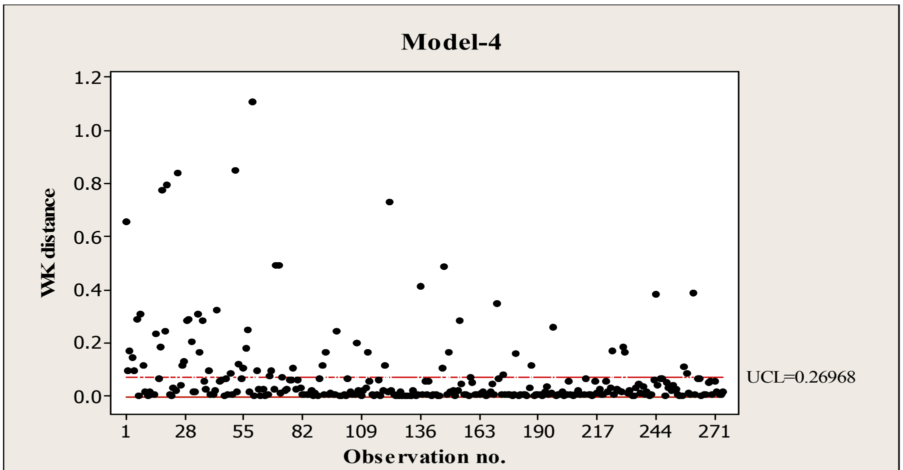


Fig.4

Control charts for fitted models show the identification of influential observations based on proposed approach-I

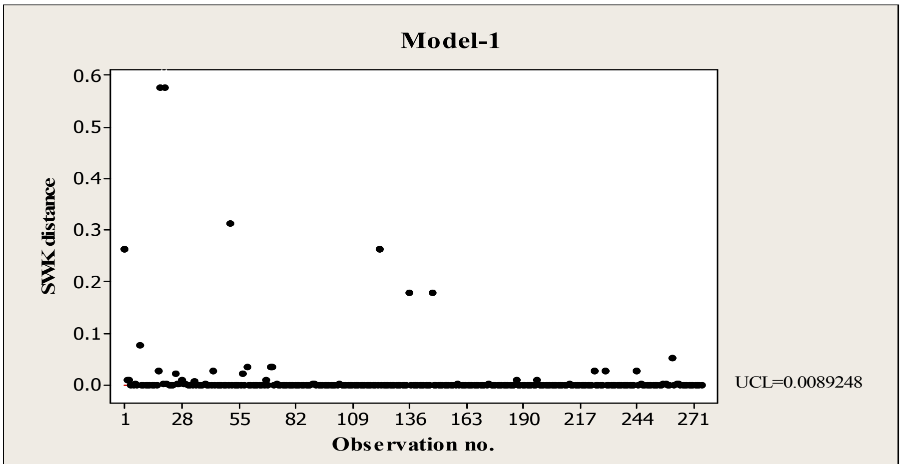

Model-1
SWK distance
0.6
0.5
0.4
0.3
0.2
0.1
0.0
UCL=0.0089248
1 28 55 82 109 136 163 190 217 244 271
Observation no.

Fig.5

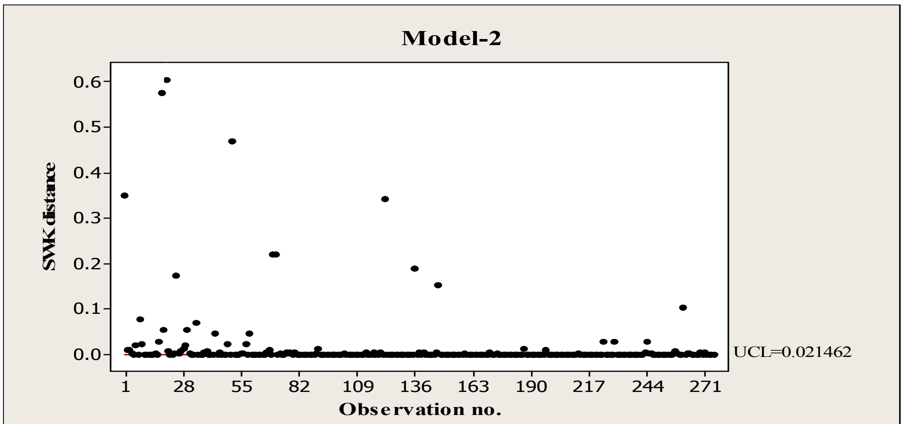

Model-2
SWK distance
0.6
0.5
0.4
0.3
0.2
0.1
0.0
UCL=0.021462
1 28 55 82 109 136 163 190 217 244 271
Observation no.

Fig.6

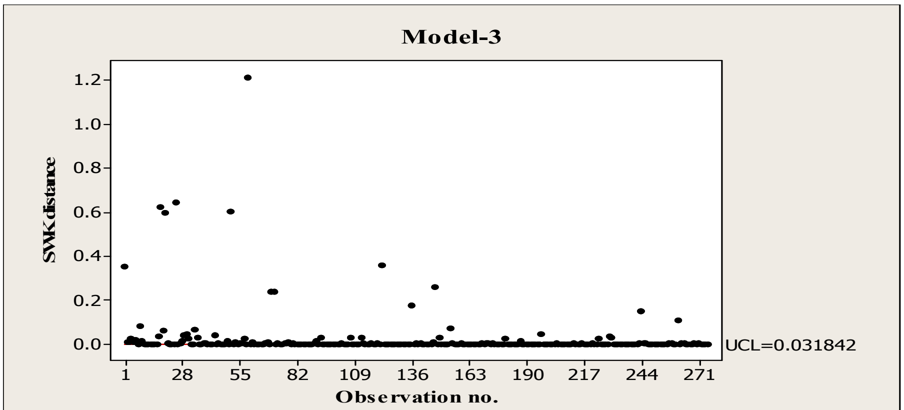


Fig.7

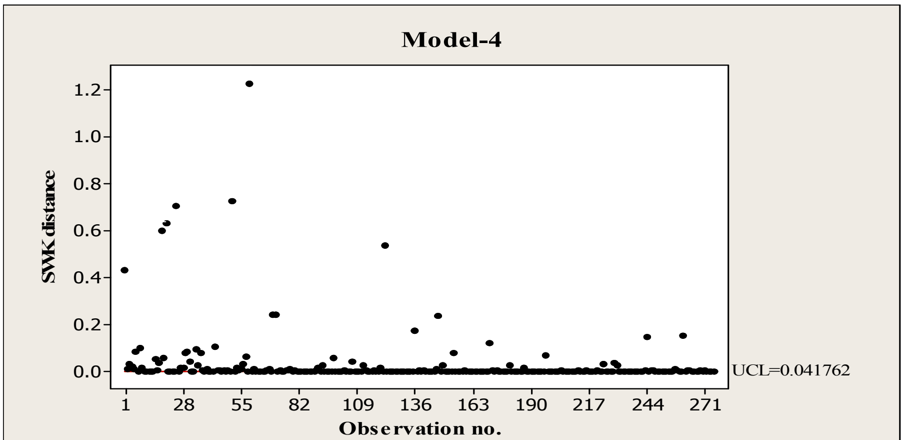


Fig.8

Control charts for fitted models show the identification of influential observations at 5% level based on proposed approach-II

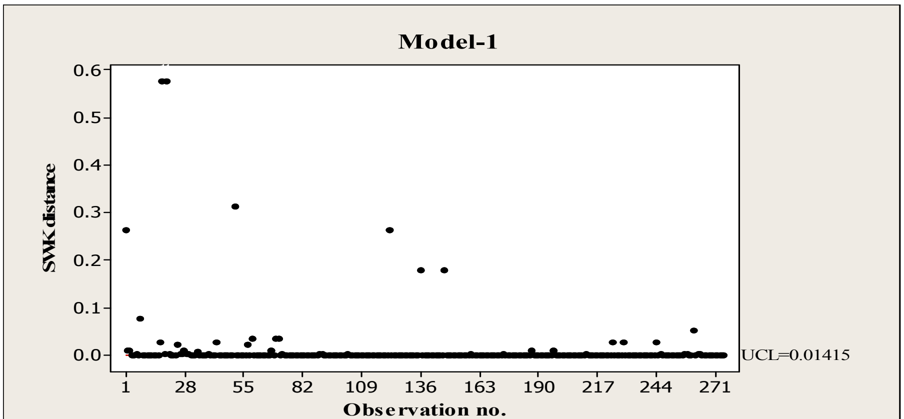
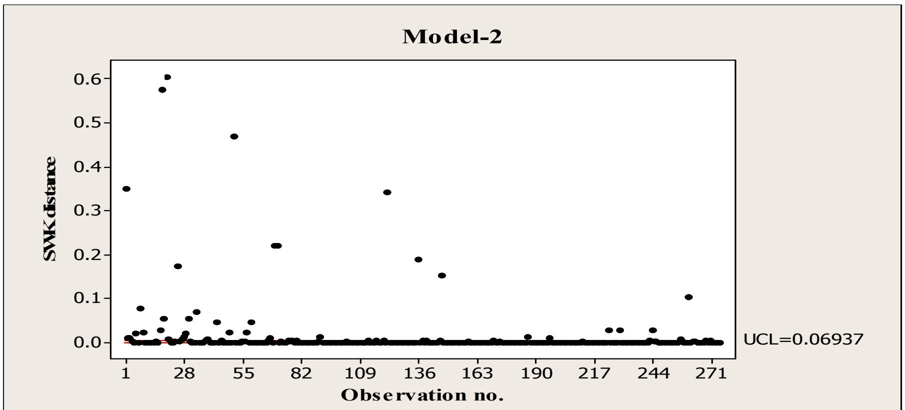
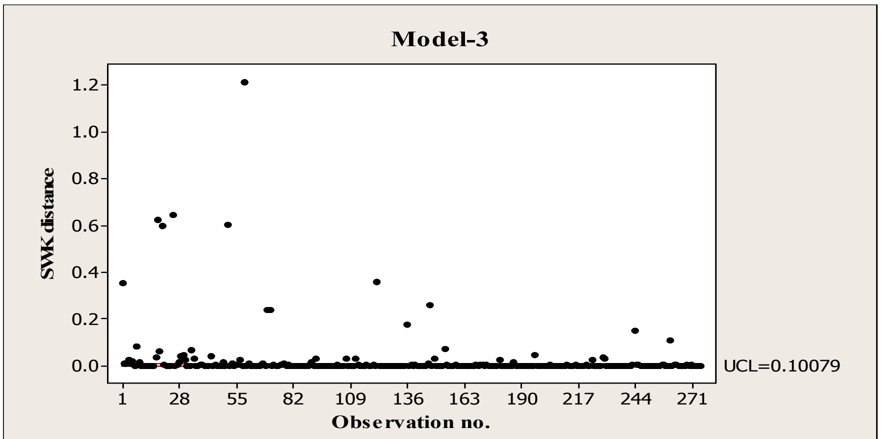


Fig.9


Model-2
SWK distance
0.6
0.5
0.4
0.3
0.2
0.1
0.0
UCL=0.06937
1
28
55
82
109
136
163
190
217
244
271
Observation no.
Fig.10


Model-3
SWK distance
1.2
1.0
0.8
0.6
0.4
0.2
0.0
UCL=0.10079
1
28
55
82
109
136
163
190
217
244
271
Observation no.
Fig.11

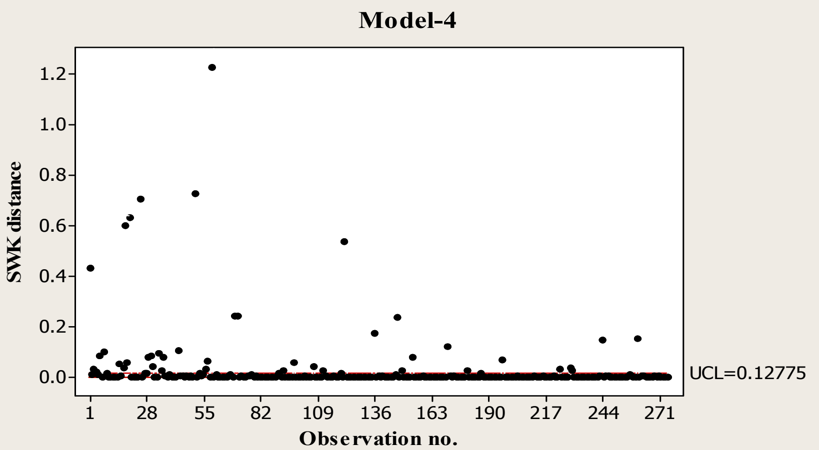


Fig.12

Control charts for fitted models show the identification of influential observations at 1% level based on proposed approach-II

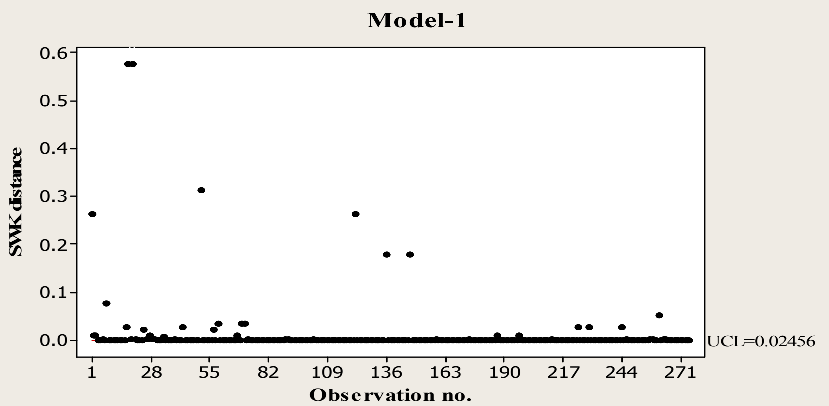


Fig.13

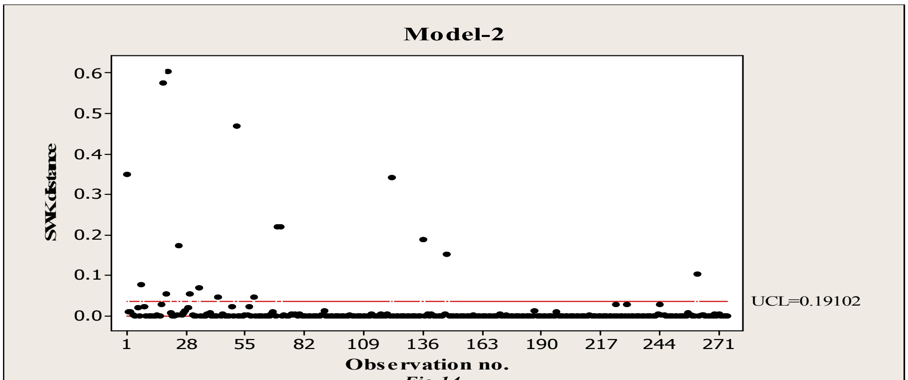


Fig.14

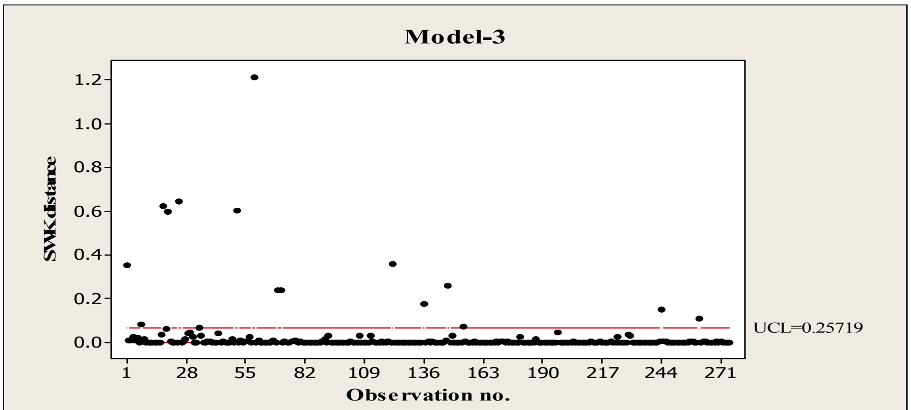


Fig.15

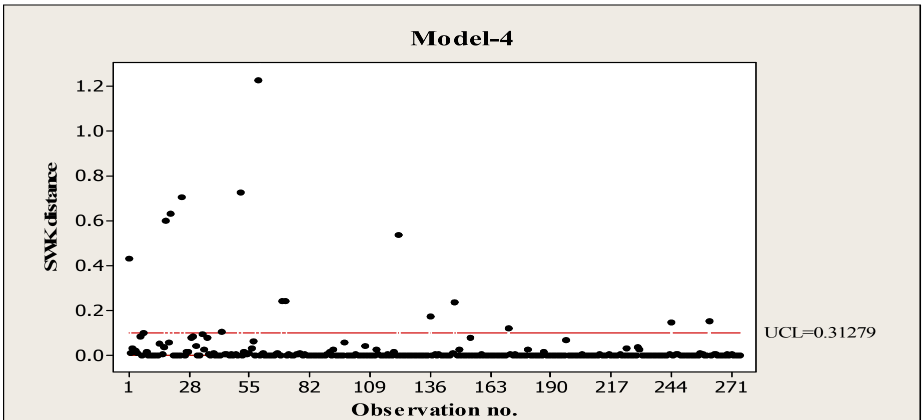


Fig.16

5. Conclusion

The authors have proposed a scientific approach which is based on the test of significance for squared Welsch-Kuh's distance measure to evaluate the influential observations in a multiple linear regression model. At first, the exact distribution of the squared Welsch-Kuh distance is derived and the authors have visualized the density function of WK^2 in terms of complicated series expression form and Gauss hyper-geometric function with two shape parameters namely p and n. Moreover, the authors have established the upper control limit of WK^2 by using the mean, variance of the distribution and the observations exceeding the UCL are identified as influential. Similarly, significant two-tail percentage points of WK^2 at 5 % and 1% level of significance are also computed and are utilized to evaluate the influential observations. The proposed approach-I identifies more influential observations than the traditional approach and the proposed approach II is systematic and scientific because it is based on the test of significance and the results are superior when compared it with Welsch-Kuh's traditional approach. Hence, based on the evidences, the authors conclude that the proposed approaches I and II override the use of traditional approach and they outperform the traditional Welsch-Kuh's approach in the process of exact identification of influential observations in multiple regression models.

References

1. Ali, S. Hadi (1992). A new measure of overall potential influence in linear regression, Computational Statistics & Data Analysis, 14(1), p. 1-27.
2. Behnken, D. W. and Draper, N. R. (1972). Residuals and Their Variance Patterns, Technometrics, 14, p. 101-111.
3. Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Wiley, New York.
4. Bollen, K. A. and Jackman, R. W. (1990). Regression diagnostics: An expository treatment of outliers and influential cases, Chapter 6, Modern Methods of Data Analysis, p. 257-291.
5. Cook, R.D. (1977). Detection of influential observation in linear regression. Technometrics, 19, p. 15-18.
6. Chatterjee, S. and Hadi, A. S. (1988). Sensitivity Analysis in Linear Regression, New York: John Wiley and Sons.
7. Cook, R. D. and Weisberg, S. (1982). Residuals and influence in regression (Vol. 5). New York: Chapman and Hall.
8. Díaz-García, J. A., and González-Farías, G. (2004). A note on the Cook's distance, Journal of Statistical Planning and Inference, 120(1), p. 119-136.
9. Eubank, R.L (1985). Diagnostics for smoothing splines, J. Roy. Statist. Soc., Ser. B, 47, p. 332–341.
10. Hoaglin, D.C. and Welsch, R.E. (1978). The Hat matrix in regression and ANOVA, The Amer. Statist., 32, p. 17-22.
11. Kim, C. (1996). Cook's distance in spline smoothing, Statist. Probab. Lett., 31, p.139–144.
12. Kim, C. and Kim, W. (1998). Some diagnostics results in nonparametric density estimation, Comm. Statist. Theory Methods, 27, p. 291–303.
13. Kim, C., Lee, Y. and Park, B.U. (2001). Cook's distance in local polynomial regression, Statist. Probab. Lett., 54, p. 33–40.

14. Silverman, B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion), J. Roy. Statist. Soc., Ser. B, 47, p. 1–52.
15. Thomas, W. (1991). Influence diagnostics for the cross-validated smoothing parameter in spline smoothing, J. Amer. Statist. Assoc., 86, p. 693–698.
16. Weisberg, S. (1980). Applied linear regression. New York: Wiley.
17. Welsch, R. E., and Kuh, E. (1977). Technical Report 923-77: Linear Regression Diagnostics, Cambridge, MA: Sloan School of Management, Massachusetts Institute of Technology.
18. Welsch, R. E. and Peters, S.C. (1978). Finding influential subsets of data in regression models, Proc. Eleventh Inference Symp, Comput.Sci. Statist, p. 240-244.