Assistant Professor Arezoo BAGHERI, PhD
E-mail:abagheri_000@yahoo.com
National Population Studies & Comprehensive Management Institute

SAMPLE SIZE IMPACTS ON HIGH LEVERAGE COLLINEARITY-ENHANCING OBSERVATIONS

Abstract: *the latest known source of multicollinearity, a nonorthogonality of two or more explanatory variables in multiple regression models, is high leverage points. Interpreting a fitted regression model may become impossible by the influential impacts of multicollinearity. In this paper, we attempt to investigate the impact of different sample sizes as one of the main causing factors of high leverage points to be collinearity-influential observations in non-collinear data. To do so, the influence of changing sample size on High Leverage Collinearity-Influential Measure (HLCIM) and Condition Number (CN) was studied. According to the simulation results, by increasing the percentage of high leverage points for each magnitude of contamination and fixed sample size and also by increasing magnitude of contamination for each percentage of high leverage point and fixed sample size, the CN and the absolute HLCIM values increase. The simulation results have been confirmed by a well-known real data set.*

Keywords: *High Leverage Collinearity-Influential Measure (HLCIM); High leverage collinearity-enhancing observations; condition number; Diagnostic Robust Generalized Potential (DRGP) method.*

JEL Classification: C15, C39, C63

1. Introduction

Kamruzzaman and Imon (2002) have announced high leverage points as the latest known source of multicollinearity. Multicollinearity is a near-linear dependency of two or more explanatory variables in multiple regression models which may have significant influential impacts on regression analysis. High leverage points may change a non-collinear data set to be collinear and vice versa (Bagheri, 2011). Bagheri, 2011 has called high leverage points which cause multicollinearity as high leverage collinearity-enhancing observations and those points which decrease multicollinearity are referred as high leverage collinearity-reducing observations. It is important mentioning that collinearity-enhancing observations are crucial points in non-collinear data sets for changing the pattern of multicollinearity. Habshah and Bagheri (2015) has

introduced a robust diagnostic measures based on minimum covariance determination approach for identification of these points and Bagheri and Habshah (2015) has defined diagnostic plots for their detection.

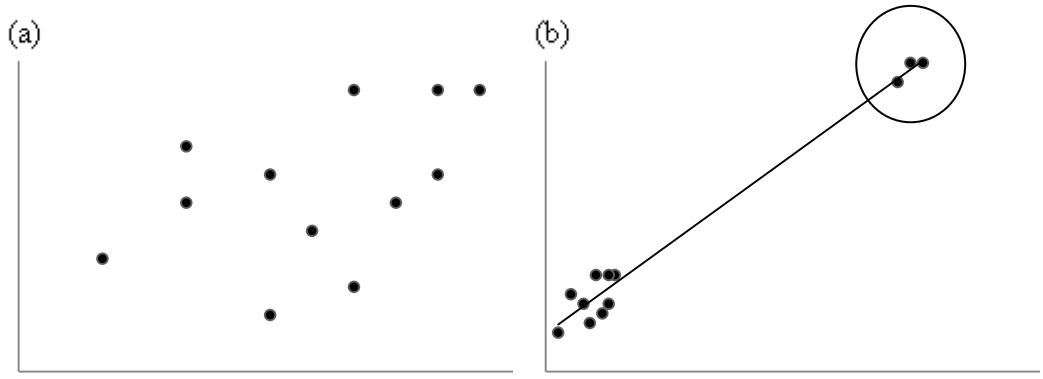
Figure 1 illustrates the influence of collinearity-enhancing observations on non-collinear Artificial data set which has taken from Bagheri (2011). Table 1 is presented this Artificial data set. This figure part (a) presents a non-collinear data set. According to this figure, there is no collinearity in the data set. However, by modifying the data set to have a collinearity-enhancing observation, it is obvious that this data set changed to be collinear date set (Figure 1 part (b)).

Table 1.Artificial data set

Non-collinear Data Set		**Modified Data With High Leverage Collinearity-Enhancing Observations**	
x_1	x_2	x_1	x_2
4	10	4	10
5	7	5	7
3	2	28	30
4	3	29	32
2	8	2	8
5.5	10	5.5	10
4.5	6	4.5	6
3.5	5	3.5	5
3	7	3	7
5	10	5	10
1	4	1	4
2	6	30	32

Habshah et al. (2011) investigated the effect of some factors for changing high leverage points to be collinearity-influential observations in the non-collinear data set. These factors can be listed as percentage of high leverage points, Magnitude of Contamination (MC) and the position of high leverage points. To achieve their aims, they proposed High Leverage Collinearity-Influential Measure (HLCIM). However, they didn't consider the influence of changing the sample size which was the limitation of their study. In addition, Bagheri et al. (2010) has studied the influence of these factors on high leverage points to be collinearity-influential observations in the collinear data set.

Figure 1. (a) Original Artificial Data Set (2) (Non-collinear), (b) Modified Artificial Data Set (2) with a Collinearity-enhancing Observation (Bagheri, 2011).



In this paper, Monte Carlo simulation studies have been performed to study the influence of sample size as one of the most important factors in causing high leverage points to bring multicollinearity in non-collinear data sets. Insight is gained only by simulation experience and by real data set. This paper is organized as follows. High leverage and High Leverage Collinearity-Influential Measures are introduced in Section 2. The effect of high leverage collinearity-influential observations in HLCIM and Condition Number (CN) of X matrix on a well-known non-collinear set is investigated in Section 3. The Monte Carlo simulation studies have been performed in Section 4. Section 5 consists of a brief conclusion of the results.

2. High leverage and High Leverage Collinearity-Influential Measures

There are different high leverage diagnostic measures such as Three-Sigma edit rule (Maronna et al. 2006), hat matrix (see Kutner et al. 2005), Mahalanobis Distance (Rousseeuw and Leroy, 1983), Robust Mahalanobis Distance (Rousseeuw, 1985), Potential measure (Hadi, 1992), Generalized potential measure (Imon, 2002).

The commonly used diagnostics fail to identify high leverage points correctly when a group of high leverage points is present in a data set due to the masking and/or swamping effects (Rousseeuw and Leroy, 1987).

Let's define a multiple regression model as:

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

where Y is an ($n\times1$) vector of dependent variables, X is an ($n\times p$) matrix of explanatory variables, β is a ($p\times1$) vector of unknown finite parameters to be estimated and ε is an ($n\times1$) vector of random errors. Rousseeuw (1985) introduced Robust Mahalanobis Distance (RMD), based on the Minimum Volume Ellipsoid (MVE) (for more details, one can refer to Rousseeuw ,1983, 1984), RMD(MVE), as:

$$\mathrm{RMD}_{\mathrm{i}}(\mathrm{MVE}) = \sqrt{(\mathbf{X} - T_R(\mathbf{X}))' C_R(\mathbf{X})^{-1} (\mathbf{X} - T_R(\mathbf{X}))} \qquad i = 1, 2, \ldots, n \qquad (2)$$

where $T_R(\mathbf{X})$ and $C_R(\mathbf{X})$ are robust locations and shape estimates of the MVE, respectively. The robust alternative diagnostic methods such as RMD(MVE) can detect the high leverage points correctly but they have a tendency to identify too many low leverage points as high leverages which is not also desired. Habshah et al. (2009) attempt to make a compromise between these two approaches. They proposed an adaptive method where the suspected high leverage points are identified by robust methods and then after diagnostic checking, the low leverage points (if any) are put back into the estimation data set. Following the idea of Imon (2002) in developing Generalized Potentials, they proposed the diagnostic robust generalized potential (DRGP) based on MVE. DRGP(MVE) procedure can be summarized as follows.

Step1: Calculate RMD(MVE) in Equation (2) and consider a non-parametric cutoff point to determine the suspected high leverage points as:

$$\text{median}\,(\mathrm{RMD}_{\mathrm{i}}(\mathrm{MVE})) + c\,\mathrm{MAD}(\mathrm{RMD}_{\mathrm{i}}(\mathrm{MVE})) \qquad i = 1, 2, \ldots, n \quad (3)$$

where c is constant value equals to 2 or 3. Hence, the members of group D are observations whose RMD(MVE) values exceeded the cutoff point in Equation (3).

Step 2: Compute GP, p_{ii} ,as follows to see whether all members of the deletion set have potentially high leverages or not:

First denote remaining cases in the data set as R and suspected high leverage points as D. if hat matrix defined based on a group of deleted cases indexed by D, we can define:

$$h_{ii}^{(-D)} = \mathbf{x_i}'(\mathbf{X_R}'\mathbf{X_R})^{-1}\boldsymbol{x_i} \qquad i = 1, 2, \ldots, n \quad . \qquad (4)$$

It should be noted that $h_{ii}^{(-D)}$ is the i^{th} diagonal elements of $\mathbf{X}(\mathbf{X_R}'\mathbf{X_R})^{-1}\mathbf{X}'$ matrix. Imon (2002) also introduced GP as:

$$p_{ii} = \begin{cases} h_{ii}^{(-D)} & for\ i \in D \\ \dfrac{h_{ii}^{(-D)}}{1-h_{ii}^{(-D)}} & for\ i \in R \end{cases} \tag{5}$$

Step 3: If all members of the *D* set are greater than:

$$\text{median}(p_{ii}) + c\,\text{MAD}(p_{ii}) \tag{6}$$

,we declare them as high leverage points. Otherwise, those observations are put back into the estimation subset *R*.

The generalized potential values based on the final deletion set will be called DRGP(MVE) namely as p_{ii}^{*} and the points which are detected will be finally declared as high leverage points. Leverage measures based on DRGP(MVE) have proven to be very effective in the identification of multiple high leverage points (Habshah et al., 2009).

In the multiple regression model, multicollinearity maybe be defined as linear dependences of the columns of *X* matrix. There are different multicollinearity diagnostic methods such as Variance Inflation Factor (VIF) by Marquardt (1970) and Condition Number (CN) by Belsley et al. (1980). For comprehensive details about these diagnostics tools, one can referee to Montgomery et al. (2001).

Singular-value decomposition of ($n\times p$) *X* matrix is identified by Belsley et al. (1980) as:

$$X = UDV' \tag{7}$$

where *U* ,*V* , and *D* are ($n\times p$), *(p×p)*,and ($p\times$p) matrices. *U* is the matrix which columns are the eigenvectors associated with the *p* non-zero eigen values of $X'X$ and $U'U = I$. The matrix of eigenvectors of $X'X$ is *V, and* $V'V = I$, and *D* is a diagonal matrix with non-negative diagonal elements μ_j (*j*=1,2,…, *p*) which is called singular-values of *X*. Belsley et al. (1980) also defined the CI of the *X* matrix as:

$$\boldsymbol{k_j} = \frac{\mu_{\max}}{\mu_j} \qquad j=1,\ldots,p \tag{8}$$

where $\mu_1, \mu_2, \ldots, \mu_p$ are the singular values of *X* matrix. It is noticeable that the largest value of μ_j can be defined as CN of *X* matrix. There are Belsley (1991)'s rule of thumb for indicating the degree of multicollinearity from CN value which has been accepted as the standard in application in the literature. Belsley (1991) recommended that CN

values of *X* matrix between 10 and 30 is indicated as moderate multicollinearity while the values more than 30 resulted as severe multicollinearity.

Collinearity-influential observations are those observations which change the multicollinearity pattern whether create or hide it in a data set (Hadi, 1988, Sengupta and Behimasankaram, 1997; Gross, 2003). Hadi (1988) introduced a collinearity-influential measure as follows:

$$\delta_i = \frac{k_{(i)} - k}{k} \qquad i = 1, 2, \ldots, n \tag{9}$$

where k is the Condition Number of $\boldsymbol{X}$ matrix and $k_{(i)}$ is the Condition Number of $\boldsymbol{X}$ matrix when the i^{th} row of *X* matrix has been deleted. Hadi's measure has the lack of symmetry which is due to the additive change in the Condition Number of **X**. To overcome the weakness of Hadi's measure, Sengupta and Behimasankaram (1997) introduced the following collinearity-influential measure as:

$$l_i = \log(\frac{k_{(i)}}{k}) \qquad i = 1, 2, \ldots, n \tag{10}$$

It is important mentioning that a large negative or positive value of δ_i and l_i indicates that the i^{th} observation is a collinearity-enhancing or collinearity-reducing observation. Habshah et al. (2011) proposed a new measure to study the influence of high leverage points in the multicollinearity pattern of a data based on the idea of Sengupta and Bhimasankaram in proposing l_i as collineaity-influential measure. If *D* is the group of high leverage collinearity-influential observations, we define High Leverage Collinearity-Influential Measure (HLCIM) as follows:

$$HLCIM = \log(\frac{k_{(D)}}{k}) \tag{11}$$

where $k_{(D)}$ is the Condition Number of *X* matrix when *D* rows of $\boldsymbol{X}$ matrix have been deleted. The HLCIM shows whether these leverage points can cause multicollinearity or not in the data set. Habshah et al. (2011) similar to Sengupta and Bhimasankaram's measure have defined cutoff points for HLCIM. If $\log(\frac{\boldsymbol{k}_{(D)}}{\boldsymbol{k}}) \prec 0$ then the *D* group of high leverage points is referred as high leverage collinearity-enhancing observations. Otherwise, the deletion of the *D* group high leverage points may increase the degree of multicollinearity. Thus, in this situation these high leverage points are referred as collinearity-reducing observations.

3. Real data set

In this section, we will consider a real non-collinear data set that shows how changing the sample size can change the influence of high leverage points to be collinearity-enhancing observations. Commercial Properties data set which has taken from Kutner et al. (2005) is a three-predictor data set contains 81 observations. This data set contains nineteen high leverage points (observations 1, 2, 3, 6, 7, 8, 17, 21, 26, 29, 37, 45, 53, 54, 58, 61, 62, 72, and 79) while none of these leverages cause multicollinearity(Habshah et al. ,2010, Bagheri et al., 2011, Bagheri, 2011). To investigate the influence of changing sample size in changing the multicollinearity pattern of this data set, two different sample sizes with first 40 and first 60 observations of this data set has been chosen. Table 2 presents the multicollinearity diagnostics for these two original sample sizes data set. According to this table, these two selected sample sizes are non-collinear. To make multicollinearity in this data set, 5 and 10 percent of last observations of these two sample sizes have been fixed by large values of 100 and 200. HLCIM and CN of these two modified sample sizes are presented in Table 3. The result of this table indicates that by fixing sample size and Magnitude of Contamination (MC) and increasing the percentage of contamination, CN and the absolute value of HLCIM both increase. Furthermore, the same result can be drawn, when we increase MC and fix sample size and percentage of high leverage points. However, when the sample size is increased while MC and percentage of high leverage points are fixed, the value of CN decreases and the absolute value of HLCIM increases.

Table 2. Multicollinearity diagnostics for different sample sizes of Commercial Properties data set

Diagnostics	***n***	**1**	**2**	**3**
Pearson correlation coefficient	40	r_{12}= 0.1830	r_{13}= -0.1766	r_{14}= -0.3720
	60	r_{12}= 0.2373	r_{13}= -0.1867	r_{14}= -0.3134
VIF > 5	40	1.0495	1.1800	1.1772
	60	1.0756	1.1512	1.1256
Condition index of X matrix > 10	40	1	1.3128	1.5461
	60	1	1.3462	1.4835

Table 3. HLCIM and CN values for the modified Different sample sizes of Commercial Properties data set

α	***n*=40, *MC*=100**		α	***n*=40, *MC=200***	
	HLCIM	**CN**		**HLCIM**	**CN**
5	-1.0949	19.6792	5	-1.4194	41.5364
10	-1.2643	28.2542	10	-1.5837	59.5891
α	***n*=60, *MC*=100**		α	***n*=60, *MC*=200**	
	HLCIM	**CN**		**HLCIM**	**CN**
5	-1.1088	18.7702	5	-1.4327	39.5687
10	-1.2607	26.1666	10	-1.5884	55.0485

4. Simulation study

The objective of this simulation study is to investigate the effect of sample size and different magnitude, contamination, and percentage of high leverage points on HLCIM and CN. In this simulation study, we considered different sample sizes that varied from 20, 60, 100, and 300 and different Magnitude of Contamination (MC) values equal to 10 and 20. The results of simulation study of Habshah et al. (2011) for fixed sample size (equal to 100) reveal that when contamination exists in only X_1, the contamination did not cause multicollinearity problem for the data set. However, when contamination exists in X_1 and X_2 and X_1, X_2 and X_3 multicollinearity will be present in the data set. Hence, in studying the influence of different sample sizes on HLCIM and CN values, we only consider contamination which exists in all three explanatory variables. Table 4 shows the HLCIM and the CN values for different sample sizes and different percentages and different magnitudes of high leverage collinearity-enhancing observations. The values of CN for X matrix without high leverage points for sample sizes 20, 60, 100, and 300 are equal to 1.5978, 1.2752, 1.2102, and 1.1091, respectively.

These results indicate that the simulated data sets are non-collinear. However, the values of CN for contaminated data reveal the presence of multicollinearity evident by CN and HLCIM values which become large and negative large, respectively (Table 4). By increasing the percentage of high leverage points for each MC and fixed *n* and also by increasing MC for each percentage of high leverage point and fixed *n*, the CN and the absolute HLCIM values increase. However, by increasing the *n* for fixed MC and fixed percentage of high leverage pints, the CN values slightly decrease.

Table 4. HLCIM and CN values for diffident sample sizes and different percentages and magnitudes of high leverage collinearity-influential observations

α	***n*=20, *MC*=10**		α	***n*=20, *MC*=20**	
	HLCIM	**CN**		**HLCIM**	**CN**
5	-0.9661	13.7911	5	-1.2660	27.8221
10	-1.1103	19.6207	10	-1.4162	39.3272
15	-1.1938	24.4832	15	-1.4976	49.0967
20	-1.2478	27.8097	20	-1.5501	56.0013
25	-1.2923	31.5767	25	-1.5964	64.3129
α	***n*=60, *MC*=10**		α	***n*=60, *MC*=20**	
	HLCIM	**CN**		**HLCIM**	**CN**
5	-1.0120	12.7763	5	-1.3137	25.6585
10	-1.1592	18.1459	10	-1.4622	36.3971
15	-1.2485	22.2854	15	-1.5483	44.5819
20	-1.30974	25.8682	20	-1.6073	51.7914
25	-1.3532	29.0820	25	-1.6531	57.6371
α	***n*=100, *MC*=10**		α	***n*=100, *MC*=20**	
	HLCIM	**CN**		**HLCIM**	**CN**
5	-1.0253	12.5505	5	-1.3285	25.1005
10	-1.1760	17.7956	10	-1.4741	35.4610
15	-1.2616	21.8333	15	-1.5626	43.6404
20	-1.3218	25.2510	20	-1.6231	50.2790
25	-1.3720	28.3189	25	-1.6679	56.3383
α	***n*=300, *MC*=10**		α	***n*=300, *MC*=20**	
	HLCIM	**CN**		**HLCIM**	**CN**
5	-1.0431	12.1559	5	-1.3442	24.2507
10	-1.1938	17.1748	10	-1.4931	34.3476
15	-1.2796	21.0600	15	-1.5795	42.1525
20	-1.3417	24.3259	20	-1.6433	48.7147
25	-1.3903	27.3057	25	-1.6919	54.6307

5.Conclusion

Multicollinearity consequences are obvious to the regression analysis. Some of these drawbacks are producing unstable and inconsistent parameters estimates, and

insignificant regression coefficients, where in fact it is significant. There are lots of referable works which devoted to this area. However, little work has been explored when high leverage points are the cause of multicollinearity in the data set. Thus, it is very essential to investigate the factors which cause any high leverage points to change to collinearity-influential observation. The main focus of this paper was to study the effect of changing the sample size on multicollinearity pattern of non-collinear. Monte Carlo simulations were carried out to achieve this aim. The simulation results indicated that the sample size has a principal impact on changing the multicollinearity pattern of non-collinear data sets.

REFERENCES

[1] **Bagheri A. and Habshah M. (2015),** ***Diagnostic Plot for the Identification of High Leverage Collinearity-influential Observations*****;** *Statistics and Operations Research Transactions*, 39 (1) January-June 2015, 51-70;

[2] **Bagheri, A. (2011),** ***Robust Estimation Methods And Robust Multicollinearity Diagnostics For Multiple Regression Model in the Presence of High Leverage Collinearity-Influential Observations*****;** Unpublished doctorial thesis. University Putra Malysia;

[3] **Bagheri, A., Habshah, M. and Imon, A.H.M.R. (2010),** ***The Effect of Collinearity- Influential Observations on Collinear Data Set: A Monte Carlo Simulation Study.*** *Journal of Applied Sciences. 10(18): 2086-2093;*

[4] **Bagheri, A., Habshah, M. and Imon, A.H.M.R. (2011),** ***A Novel Collinearity-Influential Observation Diagnostic Measure Based on a Group Deletion Approach.*** *Journal of Communication and Statistics, Simulation and Computation(accepted)*;

[5] **Belsley, D.A. (1991),** ***Conditioning Diagnostics: Collinearity and Weak Data in Regression*****.** *New York: Wiley*;

[6] **Belsley, D.A. (1991),** ***Conditioning Diagnostics: Collinearity and Weak Data in Regression*****.** *New York: Wiley*;

[7]**Belsley, D.A., Kuh, E. and Welsch, R.E. (1980),** ***Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*****.** *New York :Wiley;*

[8] **Gross, J. (2003),** ***Linear Regression*****.** 1th edition. New York*: Springer-Verlag*;

[9] **Habshah M. and Bagheri1 A., (2015),** ***Robust Multicollinearity Diagnostic Measures Based On Minimum Covariance Determination Approach*****;** *Economic Computation and Economic Cybernetics Studies and Research*, No. 4/2013, *ASE Publishing, Bucharest;*

[10] **Habshah, M., Bagheri, A. and Imon, A.H.M.R. (2010),** ***The Application of Robust Multicollinearity Diagnostic Method Based on Robust Coefficient Determination to a Non-Collinear Data.*** *Journal of Applied Sciences*. 10(18): 611-619;
[11]**Habshah, M., Bagheri, A. and Imon, A.H.M.R. (2011*), High Leverage Collinearity- Enhancing Observation and Its Effect on Multicollinearity Pattern; Monte Carlo Simulation Study*.** Sains Malaysiana (In press);
[12]**Habshah, M., Norazan, M.R. and Imon, A.H.M.R. (2009*), The Performance of Diagnostic-Robust Generalized Potentials for the Identification of Multiple High Leverage Points in Linear Regression.*** *Journal of Applied Statistics.* 36(5): 507-520;
[13]**Hadi, A. S. (1992),** ***A New Measure of Overall Potential Influence in Linear Regression.*** *Computational and Statistical Data Analysis*. 14:1-27;
[14]**Hadi, A.S. (1988),** ***Diagnosing Collinearіy-influential Observations*.** *Computational Statistics and Data Analysis*. 7:143-159;
[15]**Imon, A.H.M.R. (2002),** ***Identifying Multiple High Leverage Points in Linear Regression.*** *Journal of Statistical Studies*. *Special Volume in Honour of Professor Mir Masoom Ali*. 3: 207–218;
[16]**Kamruzzaman, MD. and Imon, A.H.M.R. (2002),** ***High Leverage Point: Another Source of Multicollinearity***. Pakistanian Journal of Statistics. 18:435-448;
[17]**Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005),** ***Applied Linear Regression Models*.** 5th edition. New York*: MacGRAW-Hill;*
[18]**Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006),** ***Robust Statistics Theory and Methods*.** New York: *Willy and sons;*
[19]**Marquardt, D.W. (1970),** ***Generalized Inverses, Ridge Regression, Biased Linear Estimation and Nonlinear Estimation*.** *Technometrics*.12: 591-612;
[20]**Montgomery, D. C., Peck, E. A. and Viving, G.G. (2001),** ***Introduction to Linear Regression Analysis*.** 3rd edition. New York: *John Wiley and sons;*
[21]**Rousseeuw, P. J. (1983),** ***Multivariate Estimation with High Breakdown Point*.** *Mathematical Statistics and Applications*. Vol (B): 283-297;
[22]**Rousseeuw, P. J. (1984),** ***Least Median of Squares Regression*.** *Journal of the American Statistical Association.* 79: 871–880;
[23]**Rousseeuw, P.J. (1985),** ***Multivariate Estimation with High Breakdown Point.*** *Mathematical and Statistical Applications*. B: 283-297;
[24]**Rousseeuw, P.J. and Leroy, A.M. (1987),** ***Robust Regression and Outlier Detection*.** New York: *Wiley;*
[25]**Sengupta, D. and Bhimasankaram, P. (1997),*****On the Roles of Observations in Collinearіy in the Linear Model.*** *Journal of American Statistical Association*. 92:1024-1032.