https://doi.org/10.56093/JISAS.V79I3.4
Available online at http://isas.org.in/isa/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 79(3) 2025 239–245

Development of Influential Matrix for Detection of Outliers in Presence of Masking in Linear Regression using Survey Data

Raju Kumar, Ankur Biswas, Deepak Singh and Tauqueer Ahmad

ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Received 08 October 2025; Revised 06 November 2025; Accepted 11 December 2025

SUMMARY

Identification of outliers and influential points is crucial in linear regression, particularly when dealing with survey data. Earlier works by Li and Valliant (2011, 2015) proposed diagnostic measures for detecting single influential observations in both unclustered and clustered survey data. However, these methods face challenges when multiple influential subsets exist, primarily due to the phenomenon of masking, where the presence of multiple outliers conceals their influence. In contrast, substantial advancements for non-survey data have been made, notably by Peña and Yohai (1995) and Lawrance (1995), who introduced diagnostic procedures capable of detecting influential subsets under masking conditions. Peña and Yohai (1995) developed an influence matrix defined as the matrix of uncentered covariances representing the effect of deleting each observation on the entire dataset, with normalization to include univariate Cook's statistics on the diagonal. Candidate outliers are identified by examining the eigenvectors associated with the non-null eigenvalues of this matrix. Building on this concept, we have developed an influence matrix tailored for survey data, enabling the detection of influential subsets even in the presence of masking. The proposed diagnostic is demonstrated using a real-life survey dataset, highlighting how masked outliers can significantly affect both ordinary and survey-weighted least squares estimation.

Keywords: Eigenvectors; Extended Conditional Cook-statistic; Influence matrix; Outlier; Masking; Complex survey data; Linear regression.

1. INTRODUCTION

The basic random sampling scheme i.e. simple random sampling (SRS) available in the literature is not used in the majority of the major population surveys conducted by social scientists. Instead, they depend on a complex sampling design in which sampling units have varying probabilities of being selected. To use this type of data to develop population estimates, sampling weights must be used. However, these sampling weights should be used for estimating regression equations. The pseudo maximum likelihood (PML) method is frequently used in the literature to analyze complex survey data using linear regression models (e.g., Binder, 1983; Skinner *et al.,* 1989). Outliers are commonly present in every field where data collection and statistical analysis are involved. In sample surveys, outlying data values are frequently encountered. Outliers can occur in the data at any time during the survey. Outliers can have a significant impact on survey data analysis. The presence of outlier(s) in the sample could completely influence the survey estimate, resulting in wrong inference. Since weighted values are used to estimate regression coefficient in survey-weighted regression, the definition of the diagnostic measure must account for both the observed values and the weights. Cook (1977), Cook and Weisberg (1982), Neter *et al.* (1996), and Weisberg (1999) have all studied diagnostics for detecting influential points in standard linear regression models. They developed diagnostic tools for non-survey data in linear regression models. Li and Valliant (2011, 2015) developed diagnostics tools for detecting outliers in survey data. They extended certain common regression diagnostics for unclustered and clustered survey data. The problem of identifying outliers or influential observations becomes more difficult if the data set contains more

Corresponding author: Raju Kumar
E-mail address: raju.iasri@gmail.com

than one outlier, which is likely to be the case in most data sets. This is due to the masking/swapping effect. In the presence of masking in linear regression, Pena and Yohai (1995) introduced a technique for detecting influential subsets. They developed a method that makes use of the eigenstructure of an influence matrix. The influence matrix is a matrix of uncentered covariances of the effect of deleting each observation on the entire data set, normalized to include the univariate Cook-statistics on the diagonal. There are some multi-row deletion techniques developed by researchers like Belsley *et al.* (1980) discussed MDFFIT which is an extension of Cook's D. Atkinson and Riani (2000, 2004) developed the forward search method that is capable of identifying groups of outliers. Li and Valliant (2011) adapted the forward search to survey data. Kumar *et al.* (2024) enhanced Lawrance's (1995) methodology by incorporating survey weighting and design features. This advancement enables us to compute conditional Cook statistics, thereby determining masking factors for linear regression in survey data. The eigenstructure of the influence matrix is shown to be a useful tool for identifying influential subsets. In this paper, we have developed a method to identify outliers in survey-weighted linear regression in presence of masking. We formed an influential matrix in which elements of this matrix are derived from Extended Cook-distance. By computing the eigenvectors corresponding to the non-null eigenvalues of the influential matrix a set of candidate outliers are identified.

2. PROPOSED INFLUENCE MATRIX

Consider a fixed effect linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi},\ \boldsymbol{\xi} \sim N(\mathbf{0}, \sigma^2\mathbf{V}) \tag{1}$$

where $\mathbf{Y} = \left(Y_1, Y_{2,}..., Y_n\right)^T$ is an n x1 vector of response where $\mathrm{Y_i}$ is a response variable for unit i, $\mathbf{X} = \left(\mathbf{X}_1, \mathbf{X}_{2,}..., \mathbf{X}_n\right)^T$, $\mathbf{X}_\mathrm{i}$ is a p-vector of fixed covariates, $\boldsymbol{\beta}$ is a fixed but unknown parameter, $\boldsymbol{\xi} = \left(\xi_1, \xi_{2,}..., \xi_n\right)^T$ and $\mathbf{V} = diag(v_i)$ is an $\mathrm{n \times n}$ diagonal matrix.

Equation (1) can be written as follows

$$\mathrm{Y_i} = \mathbf{X}_i^{T'}\boldsymbol{\beta} + \xi_\mathrm{i}, \tag{2}$$

The model-based likelihood for $\boldsymbol{\beta}$ is

$L(\boldsymbol{\beta}) = \prod_{i \in s} f(Y_i; \mathbf{X}_i, \boldsymbol{\beta}, \sigma^2 v_i)$, where s is the set of sample units and $f(Y_i \mathbf{X}_i, \boldsymbol{\beta}, \sigma^2 v_i)$ is the normal density with mean $\mathbf{X}_i'\boldsymbol{\beta}$ and variance $\sigma^2 v_i$.

The PML estimate of $\boldsymbol{\beta}$ is the solution of the set of equations

$$\sum_{i \in s} w_i \frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \tag{3}$$

where w_i is a survey weight for unit i. Survey weights, which in probability samples are usually inversely proportional to inclusion probabilities, are used in the PMLE to account for an informative design in which the sample distribution of the Y's is likely to differ from that of the finite population. Equation (3) can be simplified as

$$\mathbf{X}^T\mathbf{W}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \tag{4}$$

where $\mathbf{W} = diag(w_1, w_2, ..., w_n)$ and $\mathbf{V} = diag(v_1, v_2, ..., v_n)$.

After solving equation (3), we get $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{W}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}^{-1}\mathbf{Y}$ (Survey weighted estimator). After putting **V**=**I**, the survey-weighted (SW) estimator will consequently take the form of a weighted least squares estimator $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$ (weighted least squares estimator) when survey weights are accounted for regression.

Let $\hat{\boldsymbol{\beta}}_{(i)}$ be a weighted least squares estimator when an i^{th} data point is deleted. The effect of this point on $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y} - \left(\mathbf{X}_{(i)}^T\mathbf{W}_{(i)}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_{(i)}^T\mathbf{W}_{(i)}\mathbf{Y}_{(i)}$$

$$= \frac{\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}_i e_i w_i}{1 - h_{ii}} = \frac{\mathbf{A}^{-1}\mathbf{X}_i e_i w_i}{1 - h_{ii}}, \text{ where}$$

$\mathbf{A} = \mathbf{X}^T\mathbf{W}\mathbf{X}$ and Leverage on the diagonal of the hat matrix are $h_{ii} = w_i\mathbf{X}_i^T\mathbf{A}^{-1}\mathbf{X}_i$ (see Li and Valliant (2006, 2011, 2015)).

The extended Cook's statistic is used for individual influence assessment. However, to detect masking and

multiple influential units, we adopt the eigenstructure-based approach introduced by Peña and Yohai (1995) in linear regression and further applied to block design for incomplete multiresponse experiments by Kumar & Bhar (2023).

Now, h_{ij} is the ij^{th} element of the hat matrix. Suppose $\hat{y}_{j(i)}$ denote the new fitted value of the observation j when the i^{th} data point is delated.

$$\hat{y}_j - \hat{y}_j(i) = \frac{h_{ij}e_i}{1-h_{ii}}$$

where $h_{ij} = w_i \mathbf{x}_i^{'} \mathbf{A}^{-1} \mathbf{x}_j$ and $h_{ii} = w_i \mathbf{x}_i^{'} \mathbf{A}^{-1} \mathbf{x}_i$

Now, we define $\mathbf{f}_i = \hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)} = \frac{\mathbf{h}_i e_i}{1-h_{ii}}$, where $\mathbf{h}_i$ is the i^{th} column of Hat Matrix.

Extended Cook statistic for detection of i^{th} observation can be written as $ED_i = \mathbf{f}_i^{'}\mathbf{f}_i / p\hat{\sigma}^2$. Two observations i and j have effects on the survey-weighted regression coefficient when a scalar λ is such that $\mathbf{f}_i \approx \lambda \mathbf{f}_j$. Pena and Yohai (1995) showed that all types of masking do not imply proportional effects. To detect the possible sets of observations having similar or opposite effects on the fit, a covariance matrix can be used.

Let **F** the n x n matrix F= ($\mathbf{f}_1$,...,$\mathbf{f}_n$) whose columns are the vectors $\mathbf{f}_i$. we define the $n \times n$ influence matrix **M** as

$$\mathbf{M} = \frac{1}{p\hat{\sigma}^2}\mathbf{F}_i'\mathbf{F}_i \qquad (5)$$

The diagonal elements of the **M** matrix can be written as $m_{ij} = \frac{e_i e_j \tilde{h}_i}{(1-h_{ii})(1-h_{ii})}$, where, $\tilde{h}_i = \left(h_{ii}h_{ij} + h_{jj}h_{ji}\right)$. Here, we can observe those diagonal elements of **M** which are individual Extended Cook-statistics for i^{th} observation vector and off diagonals are products of two DFBETAs of two observation vectors.

3. PROCEDURE FOR DETECTING INFLUENTIAL SETS

To obtain the eigenvectors corresponding to the largest non-zero eigenvalues of the influence matrix **M**, in each eigenvector, we search for the sets of coordinates with relatively large values and the same sign. The search is done in the following way:

Assume that e denotes the eigenvector, which corresponds to a large eigenvalue. Order the eigenvector v's coordinates to obtain $e_{(1)} \leq e_{(2)} \leq \ldots e_{(n)}$.

Calculate the ratio $a_j = e_j / e_{j-1}$ for $j=n, \ldots, n\text{-}c_1$ and $b_i = e_i / e_{i+1}$ for $i=1, \ldots, c_2$. The constants c_1 and c_2 are determined by the coordinates and are smaller than n/2.

Search the first j_0 for which $|a_j| > t$ and first i_0 such that $|b_i| > t$, If i_0>1 and/or j_0>1, then the set i_0=($i_{(1)}$, $i_{(2)}$, ..., $i_{(j_{0-1})}$) and j_0=($i_{(n)}$, $i_{(n-1)}$, ..., $i_{(n-i_{0-1})}$) as candidate outlier vectors. The constants c_1 and $c_2$are generally chosen based on the breakdown point. They serve as cutoff thresholds applied to to identify sharp breaks that indicate influential groups of observations. The parameter t sets a maximum limit on the size of the subset considered influential, preventing over-identification.

4. ILLUSTRATION

The MU284 population dataset (Särndal *et al.,* 1992), which includes variables appropriate for linear regression analysis, was used. The data set includes 284 Swedish municipalities that are divided into 50 groups based on a cluster indicator (a cluster consisting of a set of neighbouring municipalities).There are 5 to 9 municipalities in each cluster. The interest model was to regress revenue from municipal taxes in 1985 [RMT 85, in millions of kronor] on S82: the total number of seats in municipal employees in 1984, the number of municipal employees in 1984 (ME84), and REV84: Real estate values based on a 1984 assessment (in millions of kronor).

For demonstration purposes,a sample of size 81 was drawn using single-stage sampling with probability proportional to size with replacement. P85: The population (in thousands) was used as a measure for size. R software was used to conduct all the analysis. R

Table 1. OLS and SW parameter estimate with outliers

Independent variable	OLS Estimation				SW Estimation			
	Coefficient	SE	t	Pr(>|t|)	Coefficient	SE	t	Pr(>|t|)
Intercept	17.035	13.440	1.267	0.209	22.150	13.645	1.623	0.108
S82	-1.475	1.048	-1.181	0.241	-1.817	1.055	-1.723	0.089
ME84	0.145	0.001	107.21	0.000	0.150	0.003	313.890	0.000
REV84	-0.004	0.001	-3.214	0.001	-0.008	0.001	-11.292	0.000

Table 2. OLS and SW parameter estimate after deleting detected influential observation using Extended Cook's Distance statistics

Independent variable	OLS Estimation				SW Estimation			
	No. of units deleted=5				No. of units deleted=7			
	Coefficient	SE	t	Pr(>|t|)	Coefficient	SE	t	Pr(>|t|)
Intercept	28.061	8.462	3.316	0.002	22.970	9.730	2.361	0.021
S82	-3.230	0.817	-3.594	0.000	-2.062	0.760	-2.714	0.008
ME84	0.144	0.001	134.882	0.000	0.150	0.000	456.910	0.000
REV84	0.002	0.002	-3.036	0.003	0.002	0.002	1.206	0.232

programme was written to compute the eigenvalues and eigenvectors. Survey weights have been incorporated for analysis in the Survey weighted parameter estimate of linear regression and the data set has been analysed by incorporating weights into the PML estimation approach (Skinner *et al.,* 1989; Chamber and Skinner, 2003).

In the latter scenario, the modified Cook's distance method was used to diagnose a single influential point. For both OLS and SW least squares. Estimation of coefficient and standard error are computed both before and after deletion of influential points. Following that, the identification of outliers in the presence of masking was considered. When outliers are present in the data, the estimated coefficient and standard error of OLS and SW least squares are reported in Table 1.

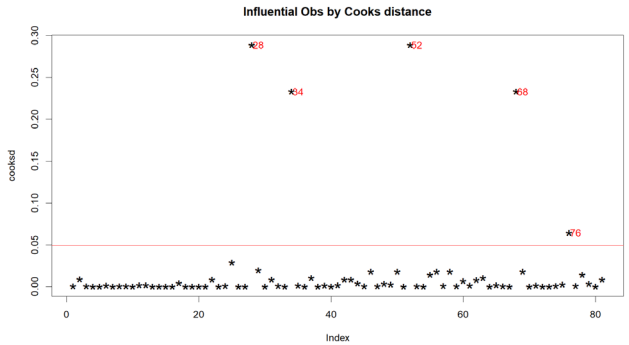

Fig. 1. Influential observations linear regression

From Table 1, we observed that Intercept and CS82 are non-significant in fitting the OLS and survey-weighted linear regression model. Fig. 1 shows influential points of linear regression and observations 28, 34, 52, 68 and 76 are identified as influential observations.

Fig. 2 shows influential points of weighted linear regression for survey data. observations 18, 21, 26, 47, 53, 82 and 85 are identified as influential observations.

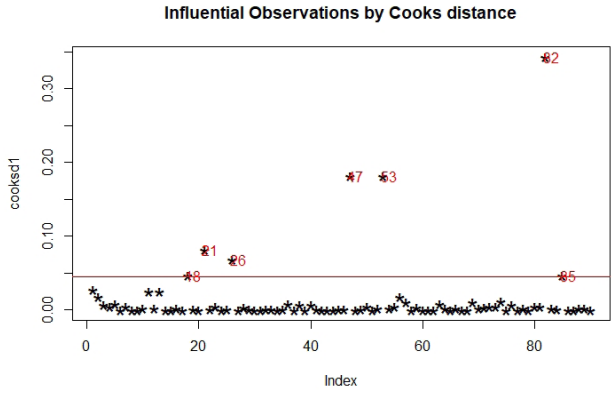

Fig. 2. Influential observations for survey-weighted linear regression

After deletion of influential observation again regression analysis was performed. Table 2 reports the estimated coefficient and their standard error after the deletion of influential observations.

From Table 2, it is observed that after the deletion of outliers, all coefficients become significant, except for REV84 in the case of the survey-weighted (SW) estimation, where it becomes non-significant. The standard errors of the coefficients are reduced in all

cases after the deletion, in both estimation methods. Subsequently, the strategy proposed by Pena and Yohai (1995) for linear regression, along with the developed approach for survey-weighted regression, was applied to identify groups of influential observations. In the case of ordinary linear regression, no observations were detected as outliers indicative of masking. However, for the survey-weighted regression, outliers were identified in the presence of masking and these results are reported here. The first three positive eigenvalues of **M** were $\lambda_1 = 1.524$, $\lambda_2 = 0.8985$, $\lambda_3 = 0.042176$. For other eigenvalues are less than 0.04. We have chosen $c_1 = c_2 = 36$ and $k = 2.0$. Here, value of $c1$ and $c2$ has been adjusted empirically to account for the variability of survey weights leverage to balance sensitivity and robustness in detecting masked effectively.

In Table 3 the coordinates of eigenvectors corresponding to the first two eigenvalues are ordered.

Table 3. Ordered co-ordinates of eigenvectors

λ_1=1.524				λ_2=0.8985			
Orderd co-ordenate No	Orderd Co-ordinate value	a_j	b_i	Orderd co-ordenate No.	Co-ordinate value	a_j	b_i
81	0.999	3.874		39	0.389	1.565	
62	0.258	1.351		24	0.249	1.273	
15	0.191	1.069		64	0.195	1.309	
36	0.179	1.013		31	0.149	1.005	
5	0.176	1.000		10	0.149	1.029	
4	0.176	1.041		58	0.144	1.060	
48	0.169	1.000		56	0.136	1.028	
39	0.169	1.138		51	0.132	1.050	
72	0.149	1.482		81	0.126	1.080	
64	0.116	1.153		36	0.118	1.010	
61	0.100	1.028		6	0.117	1.003	
74	0.098	1.000		29	0.116	1.234	
14	0.098	1.464		77	0.094	1.194	
8	0.067	1.136		54	0.079	1.054	
18	0.059	1.023		41	0.075	1.045	
76	0.057	1.000		3	0.072	1.010	
43	0.057	1.618		34	0.071	1.047	
32	0.035	1.700		75	0.068	1.026	
65	0.021	1.156		8	0.066	1.038	
51	0.018	1.156		23	0.064	1.005	
78	0.016	1.100		4	0.063	1.318	
38	0.014	1.171		79	0.048	1.235	
12	0.012	1.000		45	0.039	1.014	
9	0.012	1.197		43	0.038	1.032	
63	0.037	1.050					
12	0.035	1.227					
25	0.029	1.319					
40	0.022	1.054					
76	0.021	1.480					
19	0.014	1.268					
1	0.011	1.158					
35	0.010	1.001					
62	0.010	1.032					
46	0.009	1.040					
69	0.009	1.961					
13	0.005						
7	0.004						
17	0.000						
15	-0.002						
38	-0.003						
66	-0.005						
72	-0.010						
57	-0.013						
16	-0.014						
44	-0.015						
48	-0.016		1.038				
65	-0.016		1.092				
22	-0.020		1.010				
28	-0.021		1.220				
9	-0.022		1.048				
21	-0.031		1.048				
14	-0.033		1.462				
73	-0.034		1.029				
61	-0.046		1.356				
33	-0.049		1.054				
42	-0.055		1.140				
11	-0.057		1.032				
59	-0.059		1.025				
37	-0.059		1.008				
60	-0.059		1.004				
68	-0.067		1.124				
2	-0.072		1.076				
30	-0.076		1.059				
70	-0.079		1.045				
18	-0.084		1.058				
55	-0.093		1.108				
78	-0.098		1.054				
53	-0.103		1.051				

λ_1=1.524				λ_2=0.8985			
Orderd co-ordenate No	Orderd Co-ordinate value	a_j	b_i	Orderd co-ordenate No.	Co-ordinate value	a_j	b_i
3	-0.043		1.222	32	-0.111		1.077
60	-0.043		1.000	67	-0.114		1.029
26	-0.043		1.000	74	-0.117		1.025
70	-0.043		1.000	20	-0.118		1.008
13	-0.043		1.000	80	-0.124		1.052
19	-0.043		1.000	26	-0.154		1.240
30	-0.050		1.000	52	-0.165		1.071
68	-0.052		1.166	27	-0.178		1.079
50	-0.095		1.044	47	-0.192		1.077
34	-0.110		1.827	71	-0.195		1.018
52	-0.138		1.161	5	-0.263		1.348
53	-0.338		**2.446**	49	-0.274		1.042
41	-0.938		**2.773**	50	-0.327		1.193

The a_j and b_j values are then obtained. Table 3 shows that the first eigenvalue identifies two candidate outliers with b_i>2, whereas the second eigenvalue identifies no candidate outliers. Other eigenvalues' coordinates are not shown in the table because they do not identify any outliers.

The interesting note here is that observation number 53 was detected as an outlier when using the Extended Cook-statistic for detecting a single outlier, but observation number 41 was also identified as influential when using the developed method for identifying outliers in the presence of masking. As a result, eight outliers where observation number 41 was masked by observation number 53 were identified as outliers. After eliminating these observations, the regression analysis was re-conducted.

Table 4. SW parameter estimate after deleting detected influential observation

Independent variable	SW Estimation			
	No. of units deleted=7+1=8			
	Coefficient	SE	t	Pr(>|t|)
Intercept	-10.981	3.781	-2.904	0.005
CS82	2.295	0.348	6.589	0.000
ME84	0.124	0.002	69.25	0.000
REV84	-0.008	0.000	-16.011	0.000

Table 4 displays that standard errors decrease substantially after deleting observations and regression coefficient like CS82 becomes non-significant to significant. Hence, there is a change in inference.

5. CONCLUSION

We found that conclusions drawn from complex survey data can be misleading if outliers are present in the data set and the presence of masking makes it more difficult to identify outliers. It is important to identify influential observations in survey-weighted regression. The developed approach to detect outliers in the presence of masking performs well in solving this problem.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to the anonymous reviewers for providing suggestions and comments which led to the significant improvement in the manuscript.

Statement of Funding

No funding related to this work was received.

Statement of Potential Conflicts of Interest

The author confirms no potential conflicts of interest.

REFERENCES

Atkinson, A.C. and Riani, M. (2000). *Robust Diagnostic Regression Analysis*. New York: Springer- Verlag.

Atkinson, A.C. and Riani, M. (2004). The forward search and data visualization. *Computational Statistics*, **19**, 29-54.

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics.* New York: JohnWiley & Sons.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51,** 279-292.

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

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. London: Chapman & Hall Ltd.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376-382.

Kumar, R., Biswas, A., Singh, D. and Ahmad, T. (2024). Detection of outliers in survey–weighted linear regression. *Mathematical Population Studies*, **31(3),** 147-164.

Kumar, R. and Bhar, L.M. (2023). Procedure for the identification of multiple influential observations in block design for incomplete multi-response experiments in presence of masking.

Communications in Statistics-Simulation and Computation, **52(5),** 2167-2176.

Lawrance, A.J. (1995). Deletion influence and masking in regression**.** *Journal of the Royal Statistical Society* B, **57(1**), 181-189.

Li, J. and Valliant, R. (2006). Influence analysis in linear regression with sampling weights. Proceedings of the Section on Survey Research Methods, American Statistical Association,3330-3337. Retrieved from https://www.asasrms.org/Proceedings/y2006/Files/JSM2006-000435.pdf

Li, J. and Valliant, R. (2011). Detecting Groups of Influential Observations in Linear Regression Using Survey Data—Adapting the Forward Search Method. *Pakistan Journal of Statistics*, **27(4),** 507-528.

Li, J. and Valliant, R. (2011). Linear regression influence diagnostics for unclustered survey data. *Journal of Official Statistics*, **27,** 99-119.

Li, J. and Valliant, R. (2015). Linear regression diagnostics in cluster samples. *Journal of Official Statistics*, **71**, 61-75.

Neter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W. (1996). *Applied Linear Statistical Models*. McGraw Hill/Irwin Series: Operations and Decision Sciences (Fourth edition).

Pena, D. and Yohai, V.J. (1995). The detection of influential subsets in linear regression by using an influence matrix. *Journal of Royal Statistical Society B*, 57, 145-156.

Pfeffermann, D. and Holmes, D.J. (1985). Robustness considerations in the choice of method of inference for the regression analysis of survey data. *Journal of the Royal Statistical Society A,* **148**, 268-278.

Ren, R. and Chambers, R. (2001). *Studies on outlier robust estimators for finite populations*. Euredit project report.

Särndal, C.E., Swensson, B. and Wretman, J. (1992), Model Assisted Survey Sampling, Springer Verlag, New York.

Scholl, J.R. (1997). *Matrix Analysis for Statistics.* New York: John Wiley.

Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). *Analysis of Complex Surveys*. New York: John Wiley.

Weisberg, S. (2005), *Applied Linear Regression*, New York: John Wiley.