Journal of Applied Economic Sciences

A NEW METHOD OF ROBUST LINEAR REGRESSION ANALYSIS: SOME MONTE CARLO EXPERIMENTS

Sudhanshu Kumar **MISHRA**
Department of Economics, North-Eastern Hill University,
Shillong, Meghalaya, **India**
mishrasknehu@yahoo.com

Abstract:

This paper has elaborated upon the deleterious effects of outliers and corruption of dataset on estimation of linear regression coefficients by the Ordinary Least Squares method. Motivated to ameliorate the estimation procedure, it introduces the robust regression estimators based on Campbell's robust covariance estimation method. It investigates into two possibilities: first, when the weights are obtained strictly as suggested by Campbell and secondly, when weights are assigned in view of the Hampel's median absolute deviation measure of dispersion. Both types of weights are obtained iteratively and using those weights, two different types of weighted least squares procedures have been proposed. These procedures are applied to detect outliers in and estimate regression coefficients from some widely used datasets such as stackloss, water salinity, Hawkins-Bradu-Kass, Hertzsprung-Russell Star and pilot-point datasets. It has been observed that Campbell-II in particular detects the outlier data points quite well. Subsequently, some Monte Carlo experiments have been carried out to assess the properties of these estimators whose findings indicate that for larger number and size of outliers, the Campbell-II procedure outperforms the Campbell-I procedure. Unless perturbations introduced to the dataset are numerous and very large in magnitude, the estimated coefficients are also nearly unbiased.

Keywords: Robust regression, Campbell's robust covariance, outliers, Monte Carlo Experiment, Median absolute Deviation

JEL classification: C13, C14, C63, C15, C01

1. Introduction

The outliers in a dataset are the points in a minority that are highly unlikely to belong to the population from which the other points (i.e. inliers), which are in a majority, have been drawn. Alternatively, the outliers exhibit a pattern or characteristics that are alien or non-conformal to those of the inliers. Stated differently, if a majority of data points, $p_i \in p$, lie in a range (a, b), then a minority of data points, $q_j \in q$, far exterior to (a, b), are outliers in the data set $D \approx p \cup q$. The said range that divides D into p and q is often fuzzy since the definition of 'far exterior' cannot be exact. The points in the 'near exterior', which belong neither to p nor to q are in the indeterminate zone and to consider them the outliers or the inliers often needs some criterion, often ad hoc or presumptive in nature.

In any case, outliers in a data set pull the measures of central tendency towards themselves and also inflate the measures of dispersion leading to biased and inefficient estimators. The pulled measures of location and inflated measures of dispersion often lead to masking of the outliers. A single prominent outlier can mask other relatively less prominent outliers and thus may cause delusion and evade their detection by a cursory inspection.

2. Linear Regression Analysis

On many occasions we desire to explain changes in a dependent variable (Y) as a response to changes in (a single or multiple) explanatory variables (X) and we hypothesize that the relationship between Y and X is linear. That is to say that the data set is described as $Y = b_1X_1 + b_2X_2 + ... + b_mX_m$ or, in another sense, $Y = \frac{\partial Y}{\partial X_1}X_1 + \frac{\partial Y}{\partial X_2}X_2 + ... + \frac{\partial Y}{\partial X_m}X_m$. We obtain a data set $Y(n,1)$ and $X(n,m)$ such that $n \geq m$. This dataset may be presented as a system of n equations in m unknowns or, in matrix representation, $Y = Xb$. If $n > m$ and the equations are inconsistent among themselves, no b will exactly satisfy the relationship $Y = Xb$, but a residual, $e(n)$, will make up $Y = Xb + e$. From this, we

have $X^{-g}Y = X^{-g}Xb + X^{-g}e$, where X^{-g} is the generalized inverse of X. Since $X^{-g} = (X'X)^{-1}X'$ such that $X^{-g}X = (X'X)^{-1}X'X = I$, we have $(X'X)^{-1}X'Y = b + (X'X)^{-1}X'e$. We assume X and e to be uncorrelated such that $X'e = 0$ whence we obtain $\hat{b} = (X'X)^{-1}X'Y$. This procedure of estimation of b is known as the method of (ordinary) least squares or the OLS.

The method of ordinary least squares is very powerful, but at the same time it is very sensitive to contamination in Y or X and the nature of e as well the relationship between X and e. As for the residuals (e), it is required that each e_i should have zero mean and constant (non-zero) standard deviation, or $\mathrm{E}(e_i) = 0;\ E(e_i^2) = \sigma^2 \neq 0,$ where E(g) is the (statistical) expectation of (g). It is also necessary that $\mathrm{E}(e_L e_T) = 0,$ where e_L and e_T are leading and trailing points, which is relevant only if the data points obey some order such as one in the time series. Together, these requirements are summarized to state that $E(ee') = \sigma^2 I$. As to X and its relationship with e, it is necessary that $\mathrm{E}(X'e) = 0$. Normally, X should be fixed or non-stochastic. If these conditions are satisfied, the OLS provides the BLUE or best linear unbiased estimator (of the parameters, b). These requirements are collectively called as the Gauss-Markov conditions [Plackett, (1950); Theil, (1971)]. It may be noted that the OLS to be BLUE does not require e_i to be normally or even identically distributed.

Aitken (1935), who was perhaps the first statistician to present the method of the ordinary least squares in matrix notations, extended the OLS to his Generalized Least Squares (GLS) to take care of the cases when $\mathrm{E}(ee') = \Omega \neq \sigma^2 I$. The GLS-estimated b (to be denoted by b_{GLS}) is obtained as $b_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$. Since Ω (and hence Ω^{-1} too) is a symmetric positive definite matrix, we may factorize $\Omega^{-1} = \omega'\omega$, whence $b_{GLS} = [(\omega X)'(\omega X)]^{-1}(\omega X)'(\omega Y)$. In this sense, the GLS is a weighted least squares, where ω is the weight matrix. In the OLS we have $\omega = (1/\sigma)I$. Aitken showed that the GLS estimators are BLUE. In particular, when the off-diagonal elements of Ω are all zero, we have $\omega_{ii} = 1/\sigma_i$ for all $i = 1, n$ and $\omega_{ij} = 0$ for $i \neq j;\ i, j = 1, n.$

3. The Case of Contaminated Datasets

In spite of meeting all the conditions mentioned above, contamination of the dataset makes the OLS an unsatisfactory method of estimation. This fact can be demonstrated by a simple example.

Table 1. Generated Data Set to Demonstrate the Effect of Mutilation by Introduction of Outliers

Original (Generated) Data Set								Mutilated Data Set							
Sl	Y	X_1	X_2	Sl	Y	X_1	X_2	Sl	Y	X_1	X_2	Sl	Y	X_1	X_2
1	176.1168	10.1683	10.3990	9	181.7555	10.9547	10.7604	1	176.1168	10.1683	10.3990	9	181.7555	**10.9547**	10.7604
2	170.4097	10.8740	10.0387	10	194.5983	11.5369	11.5459	2	170.4097	10.8740	10.0387	10	194.5983	11.5369	11.5459
3	222.4446	11.6551	13.2853	11	179.4799	12.8106	10.5921	3	222.4446	11.6551	13.2853	11	179.4799	12.8106	10.5921
4	209.2376	12.1879	12.4598	12	178.4198	12.0033	10.5327	4	209.2376	12.1879	12.4598	12	178.4198	12.0033	10.5327
5	193.3530	12.4701	11.4621	13	191.7982	11.4577	11.3721	5	193.3530	12.4701	11.4621	13	191.7982	11.4577	11.3721
6	192.5315	9.7968	11.4266	14	191.1635	11.3358	11.3349	6	192.5315	9.7968	11.4266	14	191.1635	11.3358	11.3349
7	164.5448	11.3915	9.6676	15	200.7671	11.1565	11.9338	7	164.5448	11.3915	9.6676	15	200.7671	11.1565	11.9338
8	164.8064	11.2341	11.6864	16	171.5413	11.5555	10.1067	8	164.8064	11.2341	**11.6864**	16	171.5413	11.5555	10.1067

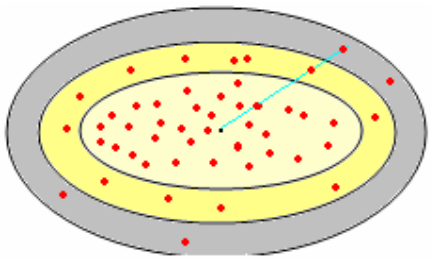
The dataset presented in Table-1 (left panel) has been generated such that $Y = 8.7 + 0.1X_1 + 16X_2 + e,$ where e is a very small disturbance. The ordinary least squares estimation of parameters from this data set gives $\hat{Y} = 8.74057 + 0.09872X_1 + 15.99787X_2$ which is very close to the generator equation. Next, we have mutilated $X_{8,2}$ and $X_{9,1}$ only slightly, which cannot possibly be detected by a mere eye inspection (right panel). Once again we apply the ordinary least squares estimation, which gives $\hat{Y} = 11.52752 + 0.98960X_1 + 14.65721X_2$. It may be noted that there is a tenfold increase in the magnitude of the coefficient associated with X_1. The value of R^2 has dropped down from 0.999998 to 0.760542. The moral of this story is clear: presence of outliers and corruption of

only a few data points can sizably distort the estimated values of some or all parameters of the regression equation.

4. Detection of Contaminated or Outlier Data Points

If the contaminated or outlier data points can be detected, something can be done to eliminate them from the dataset or to abate their influence on the estimated regression coefficients. In particular, such data points can be assigned a relatively lower (even zero) weights vis-à-vis the inlier data points and a weighted least squares approach to estimation can be employed.

Figure 1: The Mahalanobis Distance



Mahalanobis (1936) defined his generalized distance, $d=[\{Y-\mathrm{E}(Y)\}'S^{-1}\{Y-\mathrm{E}(Y)\}]^{1/2}$, where the symbol S stands for the covariance matrix of Y. This distance is a measure of deviation of a (multivariate) data point from its center. If this distance is larger than a presumed value, the data point may be considered as an outlier. This measure is formally very similar to the device Ω used by Aitken in developing his Generalized Least Squares.

5. Campbell's Robust Covariance Matrix

Using the Mahalanobis distance as a measure of deviation from center, Campbell (1980) obtained a robust covariance matrix. Campbell's method is an iterative method that obtains the $m-$element vector of weighted (arithmetic) mean, $\bar{x}$, and weighted variance-covariance matrix, $S(m,m)$, in the following manner. Initially, all weights, $\omega_i; i=1,n$ are considered to be equal, $1/n$, and the sum of weights, $\sum_{i=1}^{n}\omega_i=1$. Further, we define $d_0=\sqrt{m}+b_1/\sqrt{2};\ b_1=2,\ b_2=1.25$. Then we obtain

$$\bar{x}=\sum_{i=1}^{n}\omega_i x_i/\sum_{i=1}^{n}\omega_i$$

$$S=\sum_{i=1}^{n}\omega_i^2(x_i-\bar{x})'(x_i-\bar{x})/\left[\sum_{i=1}^{n}\omega_i^2-1\right]$$

$$d_i=\left\{(x_i-\bar{x})S^{-1}(x_i-\bar{x})'\right\}^{1/2};i=1,n$$

$$\omega_i=\omega(d_i)/d_i; i=1,n: \omega(d_i)=d_i \text{ if } d_i\le d_0 \text{ else } \omega(d_i)=d_0\exp[-0.5(d_i-d_0)^2/b_2^2].$$

If S is ill-conditioned for ordinary inversion, a generalized or the Moore-Penrose inverse [Theil, (1971)] of S or S^+ may be used for S^{-1} and if $d_i=0$ or $d_i\approx 0$ then $\omega_i=1$. We will call it the Campbell-I procedure to obtain a robust covariance matrix.

6. Use of Hampel's Median Absolute Deviation

Hampel *et al*. (1986) defined the median of absolute deviations (from median) as a measure of scale, $s_H^*(x_a)=\underset{i}{median}\,|x_{ia}-\underset{i}{median}(x_{ia})|$ and $s_H=s_H^*/0.6745$, which is a very robust measure of deviation. Using s_H, we may assign weights to different data points. If we heuristically assign the weight $\omega_i=1$ for $d_i-s_H(d)\le d_i<d_i+s_H(d)$, $\omega_i=(1/2)^2$ for $d_i-2s_H(d)\le d_i<d_i-s_H(d)$ as well as $d_i+2s_H(d)\ge d_i>d_i+s_H(d)$ and so on, and use Campbell's iterative method incorporating these weights, we may obtain a robust covariance matrix. Although not suggested so by Campbell (1980) himself, we will, however, obtain ω in this manner and call the resulting procedure as the Campbell-II method to obtain a robust covariance matrix.

7. Two Algorithms for Robust Regression Analysis

Let $Z=[Y \mid X]$. First, we obtain a robust covariance matrix $S=S(Z)$. In the process, we also obtain $\omega_i;\ i=1,n.$ With ω we construct a matrix $W_{n,n}$ such that $w_{ij}=\omega_i$ for $i=j$ else $w_{ij}=0;\ i,j=1,n.$ Then, using this weight matrix we obtain the robust regression estimator, $b_c=[(WX)'(WX)]^{-1}(WX)'(WY).$ In obtaining $S(Z)$ we may use the Campbell-I or the Campbell-II procedure and accordingly, we get two different b_c both of which are notably robust against data contamination and outliers.

8. Performance of Robust Regression Algorithms on Some Test Datasets

Many investigators in robust statistics [e.g. Andrews, (1974); Rupert and Carrol, (19809; Rousseeuw and Leroy, (1987); Kashyap and Maiyuran, (19939, etc.] have tested their methods on certain specific datasets that contain outliers. In particular, the datasets used by Rousseeuw and Leroy (1987) are available on http://www.uni-koeln.de/themen/Statistik/data/rousseeuw. Those datasets provide a good and widely accepted test bed for robust regression analysis. Among those the "stackloss datasets" [Brownlee, (1965)], water salinity dataset [Rupert and Carrol, (1980)], Hawkins-Bradu-Kass dataset [Hawkins *et al*., (1984)], the Hertzsprung-Russell star dataset [analysed by Rousseeuw and Leroy, (1987)], and the Pilot-Point dataset [Daniel and Wood, (1971)] have been used here to test the performance of the presently proposed methods of robust regression. Other test datasets also could be used, but we consider that exercise unnecessary.

8.1. The Stackloss Dataset

The dataset describes the operation of a plant for the oxidation of ammonia to nitric acid. The stackloss (y) is a function of the rate (x_1), temperature (x_2) and acid concentration (x_3). The dataset has 21 observations or cases. The dataset is reproduced in the Table 2.

Table 2. Stackloss Dataset [Brownlee, (1965); Rousseeuw and Leroy, (1987)]

sl	y	x_1	x_2	x_3	sl	y	x_1	x_2	x_3	sl	y	x_1	x_2	x_3
1	42	80	27	89	8	20	62	24	93	15	8	50	18	89
2	37	80	27	88	9	15	58	23	87	16	7	50	18	86
3	37	75	25	90	10	14	58	18	80	17	8	50	19	72
4	28	62	24	87	11	14	58	18	89	18	8	50	19	79
5	18	62	22	87	12	13	58	17	88	19	9	50	20	80
6	18	62	23	87	13	11	58	18	82	20	15	56	20	82
7	19	62	24	93	14	12	58	19	93	21	15	70	20	91

It is widely acclaimed that the data points (1, 3, 4, 21) and possibly the point (2) are outliers. While the points (1, 3, 4, 21) are considered outliers, Kashyap and Maiyuran (1993) estimate the parameters as (-37.65, 0.80, 0.577, -0.067) of which the first is the y-intercept and the subsequent three are the coefficients associated with x_1, x_2 and x_3 respectively.

We applied Campbell-I robust estimator on the data, but it did not detect any outlier and therefore the estimated coefficients were the OLS estimates (-39.92, 0.716, 1.295, -0.152) only. However, Campbell-II detected the points (1, 2, 3, 4, 21) as clear outliers and the points (13, 17) as very mild outliers. The estimated coefficients were (-32.47, 0.852, 0.451, -0.132).

8.2. The Water Salinity Dataset

The water salinity (i.e., its salt concentration) dataset (Rupert&Carrol, 1980) comprises data on water salinity (y) as the dependent variable and lagged salinity (x_1), trend (x_2) and river discharge in North Carolina's Pamlico Sound (x_3) as the explanatory variables. The dataset has 28 points (Table 3).

In this dataset, Rousseeuw and Leroy's method detects the points (5, 16, 23, 24) as outliers whereas Rupert and Carrol's method detects (1, 11, 13, 15, 16, 17) as outliers. Kashyap and Maiyuran's method detects (5, 8, 15, 16, 17) as outliers for which the coefficients are (22.30, 0.724, -0.279, -0.786).

Table 3. Water Salinity Dataset [Rupert and Carrol, (1980); Rousseeuw and Leroy, (1987)]

sl	y	x_1	x_2	x_3	sl	y	x_1	x_2	x_3	sl	y	x_1	x_2	x_3	sl	y	x_1	x_2	x_3
1	7.6	8.2	4	23.005	8	8.2	8.3	5	21.862	15	10.4	13.3	0	23.927	22	14.1	13.6	5	21.005
2	7.7	7.6	5	23.873	9	13.2	10.1	0	22.274	16	10.5	10.4	1	33.443	23	13.5	15	0	25.865
3	4.3	4.6	0	26.417	10	12.6	13.2	1	23.830	17	7.7	10.5	2	24.859	24	11.5	13.5	1	26.290
4	5.9	4.3	1	24.868	11	10.4	12.6	2	25.144	18	9.5	7.7	3	22.686	25	12.0	11.5	2	22.932
5	5.0	5.9	2	29.895	12	10.8	10.4	3	22.430	19	12.0	10	0	21.789	26	13.0	12	3	21.313
6	6.5	5	3	24.200	13	13.1	10.8	4	21.785	20	12.6	12	1	22.041	27	14.1	13	4	20.769
7	8.3	6.5	4	23.215	14	12.3	13.1	5	22.380	21	13.6	12.1	4	21.033	28	15.1	14.1	5	21.393

The Campbell-I detects the points (5, 16) as outliers and yield the estimates of regression equation as (20.63 0.708 -0.202 -0.725). On the other hand, Campbell-II detects the points (5, 16) as clear outliers, points (23, 24) as severe outliers and points (9, 12, 15, 18, 19, 25) as very mild outliers. The estimated regression coefficients are (21.98, 0.722 , -0.276, -0.783).

8.3. Hawkins-Bradu-Kass Dataset

This dataset was artificially generated by Hawkins *et al.* (1984) and consists of 75 points of four variables, y, x_1, x_2 and x_3. It is widely held that the dataset has ten extreme outliers and four other points which obey the regression model, but are located away from other inliers [Kashyap and Maiyuran, (1993)].

Table 4. Hawkins-Bradu-Kass Dataset [Hawkins, Bradu & Kass, (1984); Rousseeuw and Leroy, (1987)]

sl	y	x_1	x_2	x_3	sl	y	x_1	x_2	x_3	sl	y	x_1	x_2	x_3
1	9.7	10.1	19.6	28.3	26	-0.8	0.9	3.3	2.5	51	0.7	2.3	1.5	0.4
2	10.1	9.5	20.5	28.9	27	-0.7	3.3	2.5	2.9	52	-0.5	3.3	0.6	1.2
3	10.3	10.7	20.2	31.0	28	0.3	1.8	0.8	2.0	53	0.7	0.3	0.4	3.3
4	9.5	9.9	21.5	31.7	29	0.3	1.2	0.9	0.8	54	0.7	1.1	3.0	0.3
5	10.0	10.3	21.1	31.1	30	-0.3	1.2	0.7	3.4	55	0.0	0.5	2.4	0.9
6	10.0	10.8	20.4	29.2	31	0.0	3.1	1.4	1.0	56	0.1	1.8	3.2	0.9
7	10.8	10.5	20.9	29.1	32	-0.4	0.5	2.4	0.3	57	0.7	1.8	0.7	0.7
8	10.3	9.9	19.6	28.8	33	-0.6	1.5	3.1	1.5	58	-0.1	2.4	3.4	1.5
9	9.6	9.7	20.7	31.0	34	-0.7	0.4	0.0	0.7	59	-0.3	1.6	2.1	3.0
10	9.9	9.3	19.7	30.3	35	0.3	3.1	2.4	3.0	60	-0.9	0.3	1.5	3.3
11	-0.2	11.0	24.0	35.0	36	-1.0	1.1	2.2	2.7	61	-0.3	0.4	3.4	3.0
12	-0.4	12.0	23.0	37.0	37	-0.6	0.1	3.0	2.6	62	0.6	0.9	0.1	0.3
13	0.7	12.0	26.0	34.0	38	0.9	1.5	1.2	0.2	63	-0.3	1.1	2.7	0.2
14	0.1	11.0	34.0	34.0	39	-0.7	2.1	0.0	1.2	64	-0.5	2.8	3.0	2.9
15	-0.4	3.4	2.9	2.1	40	-0.5	0.5	2.0	1.2	65	0.6	2.0	0.7	2.7
16	0.6	3.1	2.2	0.3	41	-0.1	3.4	1.6	2.9	66	-0.9	0.2	1.8	0.8
17	-0.2	0.0	1.6	0.2	42	-0.7	0.3	1.0	2.7	67	-0.7	1.6	2.0	1.2
18	0.0	2.3	1.6	2.0	43	0.6	0.1	3.3	0.9	68	0.6	0.1	0.0	1.1
19	0.1	0.8	2.9	1.6	44	-0.7	1.8	0.5	3.2	69	0.2	2.0	0.6	0.3
20	0.4	3.1	3.4	2.2	45	-0.5	1.9	0.1	0.6	70	0.7	1.0	2.2	2.9
21	0.9	2.6	2.2	1.9	46	-0.4	1.8	0.5	3.0	71	0.2	2.2	2.5	2.3
22	0.3	0.4	3.2	1.9	47	-0.9	3.0	0.1	0.8	72	-0.2	0.6	2.0	1.5
23	-0.8	2.0	2.3	0.8	48	0.1	3.1	1.6	3.0	73	0.4	0.3	1.7	2.2
24	0.7	1.3	2.3	0.5	49	0.9	3.1	2.5	1.9	74	-0.9	0.0	2.2	1.6
25	-0.3	1.0	0.0	0.4	50	-0.4	2.1	2.8	2.9	75	0.2	0.3	0.4	2.6

We have applied Campbell-I and Campbell-II methods to detect the outliers and estimate the coefficients of robust regression. Campbell-I detects the points (11, 12, 13, 14) as outliers and the estimated coefficients are (-0.828, 0.156, 0.106, 0.226). Campbell-II detects the points (1 through 14) as clear outliers, points (18, 53, 71, 72) as strong outliers and the points (19, 28, 29, 40, 47, 50, 55, 59, 67, 68) as very mild outliers. The estimated regression coefficients by the Campbell-II method are (-0.775, 0.1625, 0.1812, 0.06517). The OLS estimates of coefficients are (-0.38755, 0. 239185, -

0.334548, -0.383341). A comparison of the Campbell-II and the OLS estimates of regression coefficients show the damage done by the outliers.

8.4. The Hertzsprung-Russell Star Dataset

This data set was introduced by Rousseeuw and Leroy (1987). It has 47 points in two variables, logarithm of the light intensity of the star as the dependent variable (y) and logarithm of the effective temperature at the surface of the star as the explanatory variable (x_1). It has four very strong outliers, the so-called giant stars, represented by the points (11, 20, 30, 34).

Table 5. Hertzsprung-Russell Star Dataset [Rousseeuw and Leroy, (1987)]

sl	y	x_1	sl	y	x_1	sl	y	x_1	sl	y	x_1	sl	y	x_1
1	4.37	5.23	11	3.49	5.73	21	4.29	4.38	31	4.38	4.42	41	4.38	4.62
2	4.56	5.74	12	4.43	5.45	22	4.29	4.22	32	4.56	5.10	42	4.45	5.06
3	4.26	4.93	13	4.48	5.42	23	4.42	4.42	33	4.45	5.22	43	4.50	5.34
4	4.56	5.74	14	4.01	4.05	24	4.49	4.85	34	3.49	6.29	44	4.45	5.34
5	4.30	5.19	15	4.29	4.26	25	4.38	5.02	35	4.23	4.34	45	4.55	5.54
6	4.46	5.46	16	4.42	4.58	26	4.42	4.66	36	4.62	5.62	46	4.45	4.98
7	3.84	4.65	17	4.23	3.94	27	4.29	4.66	37	4.53	5.10	47	4.42	4.50
8	4.57	5.27	18	4.42	4.18	28	4.38	4.90	38	4.45	5.22			
9	4.26	5.57	19	4.23	4.18	29	4.22	4.39	39	4.53	5.18			
10	4.37	5.12	20	3.49	5.89	30	3.48	6.05	40	4.43	5.57			

The Campbell-I method detects the points (11, 20, 30, 34) as clear outliers, the point (7) as a strong outlier, and points (9, 14) as very mild outliers. The regression coefficients are (3.7789, 0.126). The Campbell-II detects the points (7, 9, 11, 14, 20, 30, 34) as clear outliers and the points (3, 5, 18, 25, 28, 33, 38, 41, 42, 43, 46) as very mild outliers. The estimated regression equations are (3.7415, 0.13688). Against this, the OLS estimates of the coefficients are (4.847, -0.1071). The OLS estimates indicate that light intensity decreases as the temperature increases, which is obviously misleading. The robust regression coefficient, however, is positive.

8.5. The Pilot-Plant Dataset

Daniel and Wood (1971) provide the dataset of 20 points in two variables, where the dependent variable (y) is the acid content determined by titration, and the explanatory variable (x_1) is the organic acid content determined by extraction and weighing.

Table 6. Pilot Point Dataset [Daniel and Wood, (1971); Rousseeuw and Leroy, (1987)]

sl	y	x_1	sl	y	x_1	sl	y	x_1	sl	y	x_1	sl	y	x_1
1	76	123	5	55	57	9	41	16	13	88	159	17	89	169
2	70	109	6	48	37	10	43	28	14	58	75	18	88	167
3	55	62	7	50	44	11	82	138	15	64	88	19	84	149
4	71	104	8	66	100	12	68	105	16	88	164	20	88	167

The Campbell-I method does not detect any outlier in this data and hence the estimated regression coefficients, (35.4583, 0.3216) are the OLS estimates. However, the Campbell-II method detects a single very strong outlier point (11), four strong outlier points (4, 10, 13, 15) and three very mild outlier points (2, 8, 14). None of the points is a clear outlier. The estimated regression coefficients are (36.190, 0.3137).

These tests indicate that in detecting the outliers (and yielding the estimates of robust regression coefficients), the Campbell-I method is rather blunt and the Campbell-II is very sensitive. Where the outliers are not much deviant from the center, the Campbell-I fails to detect them. But Campbell-II detects very mild outliers too, occasionally signaling false positive.

9. Some Monte Carlo Experiments

We generated artificially a forty points 'base data' on three variables (x_1, x_2 and x_3), obtained $y = 80 - 16x_1 + 12x_2 - 2x_3$, and add a very small error to it to meet the requirements of regression analysis.

We present the 'base data' in Table 7. These data have no outliers and the OLS regression coefficients are (80.0071, -16.0001, 12.0001, -1.9998).

Table 7. Generated (Artificial) Base Dataset for Introducing Outliers in Monte Carlo Experiments

sl	y	x_1	x_2	x_3	sl	y	x_1	x_2	x_3	sl	y	x_1	x_2	x_3
1	325.62554	5.48761	28.73814	5.75294	15	-320.93381	38.27473	25.78603	49.00109	29	-216.70787	46.50270	44.03694	40.58253
2	216.23692	4.39939	17.35204	0.79249	16	380.45138	-4.53886	18.59881	-2.30804	30	65.97138	13.22553	21.07927	27.69959
3	61.22945	23.83300	35.86611	33.93566	17	240.77406	3.03338	23.03262	33.57479	31	100.21909	17.65419	24.29101	-5.58615
4	-503.35919	41.23092	11.65128	31.77223	18	33.14312	25.33601	31.93026	12.38928	32	76.79387	21.43170	30.03208	10.35714
5	235.15369	16.11028	37.32482	17.48946	19	-230.16532	33.78115	26.40625	43.28304	33	-372.67139	43.93101	28.55601	46.25669
6	-25.00594	22.25174	25.10806	25.09546	20	226.89812	7.18313	22.76687	5.72171	34	522.41150	-4.43900	32.89357	11.66127
7	176.81192	9.90141	25.04717	22.64996	21	115.94871	0.69938	9.38464	32.74268	35	280.68193	-6.29805	7.42427	-5.37605
8	-233.23841	19.83920	4.29028	23.65183	22	319.34732	12.12583	40.15470	24.25351	36	-398.03503	32.25189	7.91189	28.43924
9	-73.33875	8.76262	3.63445	28.40374	23	272.84371	6.58704	28.01780	19.00688	37	-373.70179	39.99472	20.60245	30.50913
10	154.46532	1.84157	11.41580	16.51787	24	-23.69083	17.15236	19.89292	33.96668	38	-270.06547	24.93147	4.11388	0.26708
11	-55.74443	30.61515	33.49451	23.86843	25	-375.92013	34.26759	10.90453	19.25040	39	195.19782	15.80573	33.94128	19.57595
12	-396.83923	30.81785	5.96886	27.70361	26	-78.48044	33.31807	32.40769	7.14970	40	-82.85239	27.95279	27.01828	19.91053
13	-337.50365	40.83092	21.78991	12.82849	27	-405.38187	32.61658	7.07663	24.22386	$y = 80 - 16x_1 + 12x_2 - 2x_3$				
14	-9.86483	29.19798	35.25464	22.88996	28	425.10924	11.33292	48.91941	30.30439					

9.1. Experiment-1

We add one quantum of a random size between (-10, -5) and (5, 10) to equi-probably randomly chosen point of every variable (including y). We do this exercise 200 times and find mean, $\bar{b}$, standard deviation, $s(b)$, and root-mean-square, Rms, for each coefficient (having 200 replicates). Estimation is done by Campbell-I and Campbell-II methods. Then we change the number of perturbation quanta to be made to each variable to 2, 5 and 10 keeping other parameters of the experiment constant. The results are presented in Table 8.

Table 8. Results of the Monte Carlo Experiments for Perturbations between (-10, -5) and (5, 10)

NO	EM	$\bar{b}_0$	$\bar{b}_1$	$\bar{b}_2$	$\bar{b}_3$	$s(b_0)$	$s(b_1)$	$s(b_2)$	$s(b_3)$	Rms_0	Rms_1	Rms_2	Rms_3
1	C1	79.9653	-15.9975	12.0000	-2.0000	0.4637	0.0146	0.0171	0.0147	0.4650	0.0148	0.0171	0.0147
	C2	79.9484	-15.9982	12.0000	-1.9987	1.0798	0.0282	0.0388	0.0257	1.0811	0.0283	0.0388	0.0258
2	C1	80.0780	-15.9944	11.9882	-1.9991	1.9785	0.0395	0.0825	0.0452	1.9800	0.0399	0.0833	0.0452
	C2	79.8596	-15.9927	11.9992	-1.9993	1.4753	0.0426	0.0490	0.0405	1.4820	0.0432	0.0490	0.0405
5	C1	83.8509	-15.7211	11.6521	-2.0359	16.6481	0.5297	0.5722	0.5203	17.0877	0.5986	0.6697	0.5216
	C2	80.7855	-15.9914	11.9688	-2.0015	4.6661	0.1116	0.1441	0.1247	4.7317	0.1120	0.1474	0.1247
10	C1	83.1449	-15.2665	11.3806	-2.1560	24.2318	0.7452	0.8149	0.7094	24.4350	1.0456	1.0236	0.7263
	C2	83.4905	-15.7827	11.6428	-2.0122	28.4410	0.8036	0.8973	0.8302	28.6544	0.8324	0.9658	0.8303

Note: NO=No. of perturbations per variable; EM= Estimation Method; C1=Campbell-I method; C2=Campbell-II method.

9.2. Experiment-2

Next, we repeat the experiment with change in the size of perturbation quanta, but keeping everything else as elaborated in Experiment 1. The perturbation quanta now lie in a larger range of (-25, -20) and (20, 25). The results are presented in Table 9.

Table 9. Results of the Monte Carlo Experiments for Perturbations between (-25, -20) and (20, 25)

NO	EM	$\bar{b}_0$	$\bar{b}_1$	$\bar{b}_2$	$\bar{b}_3$	$s(b_0)$	$s(b_1)$	$s(b_2)$	$s(b_3)$	Rms_0	Rms_1	Rms_2	Rms_3
1	C1	79.9745	-15.9756	11.9848	-2.0048	1.4511	0.0483	0.0595	0.0597	1.4513	0.0541	0.0615	0.0599
	C2	79.9635	-15.9800	11.9854	-2.0015	3.0999	0.0856	0.1173	0.0938	3.1001	0.0880	0.1182	0.0939
2	C1	80.3301	-15.9319	11.9346	-2.0095	5.9550	0.2491	0.1728	0.1467	5.9641	0.2583	0.1847	0.1470
	C2	79.9951	-15.9494	11.9633	-2.0092	4.8627	0.1564	0.1658	0.1419	4.8627	0.1644	0.1698	0.1422
5	C1	102.2277	-13.3051	9.1346	-2.4067	42.5440	1.3607	1.3525	1.3570	48.0007	3.0190	3.1686	1.4167
	C2	81.6860	-15.9585	11.7975	-1.8904	11.9805	0.3729	0.4931	0.4217	12.0986	0.3752	0.5330	0.4357
10	C1	100.8889	-10.9714	7.2610	-2.5784	53.1775	1.4913	1.6434	1.4639	57.1331	5.2451	5.0159	1.5740

	C2	108.7593	-13.4929	8.5766	-2.1912	75.5747	2.4112	2.3043	2.1935	80.8618	3.4784	4.1267	2.2018

Note: NO=No. of perturbations per variable; EM= Estimation Method; C1=Campbell-I method; C2=Campbell-II method.

9.3. Experiment-3

Once again we repeat the experiment with further changes in the size of perturbation quanta, but keeping everything else as elaborated in Experiment 1. The perturbation quanta now lie in a still larger range of (-100, -50) and (50, 100). The results are presented in Table 10.

Table 10. Results of the Monte Carlo Experiments for Perturbations between (-100, -50) and (50, 100)

NO	EM	$\bar{b}_0$	$\bar{b}_1$	$\bar{b}_2$	$\bar{b}_3$	$s(b_0)$	$s(b_1)$	$s(b_2)$	$s(b_3)$	Rms_0	Rms_1	Rms_2	Rms_3
1	C1	80.9852	-15.7714	11.8010	-2.0469	7.3542	0.2993	0.3634	0.3857	7.4198	0.3766	0.4144	0.3885
	C2	80.4523	-15.7762	11.8439	-2.0677	11.5777	0.3792	0.4672	0.4666	11.5865	0.4403	0.4926	0.4715
2	C1	84.0103	-15.5509	11.5696	-2.1508	16.6481	0.7617	0.7323	0.7673	17.1243	0.8843	0.8495	0.7820
	C2	82.3807	-15.5821	11.6483	-2.1339	18.3389	0.7976	0.7269	0.7900	18.4927	0.9004	0.8075	0.8012
5	C1	114.2855	-10.9362	5.8761	-1.7315	92.7163	5.2181	3.9374	1.8410	98.8524	7.2713	7.2804	1.8605
	C2	89.6809	-15.1442	11.0092	-2.1745	40.8676	1.6265	1.6630	1.3783	41.9986	1.8379	1.9357	1.3893
10	C1	53.8971	-3.7093	1.7535	-1.2798	65.2209	2.6154	1.7382	1.4089	70.2505	12.5658	10.3929	1.5823
	C2	97.7946	-8.0965	3.9146	-1.7399	116.3931	5.2697	3.4541	2.1912	117.7455	9.4992	8.7922	2.2065

Note: NO=No. of outliers per variable; EM= Estimation Method; C1=Campbell-I method; C2=Campbell-II method.

9.4. Observations

For small perturbations both Campbell-I and Campbell-II do perform well, but if the number of perturbations is smaller, the Campbell-I performs better. This edge of Campbell-I over Campbell-II is lost with an increase in the number of perturbations. Secondly, as the size as well as the number of perturbations increase, the robust estimators by both the methods tend to become biased as reflected in the increasing difference between $s(b)$ and Rms values. It may be noted that for unbiased ness $s(b) = Rms$. It has been empirically observed that ten perturbations per variable amount to corruption of about 35 percent points in the dataset. Further, considering the size/magnitude of independent variables (x_1, x_2 and x_3) that lie between (-10, 50), a perturbation lying between (-100, -50) or (50, 100) is quite large. Such perturbations can always induce biases in the estimated coefficients. As it is observed, when the perturbations per variable is up to five in number, Campbell-II produces very good results even when the size of perturbations is large.

10. Concluding Remarks

In this paper we have elaborated upon the deleterious effects of outliers and corruption of dataset on estimation of linear regression coefficients by the Ordinary Least Squares method. Motivated to ameliorate the estimation procedure, we have introduced the robust regression estimators based on Campbell's robust covariance estimation method. We have investigated into two possibilities: first, when the weights are obtained strictly as suggested by Campbell and secondly, when weights are assigned in view of the Hampel's median absolute deviation measure of dispersion. Both types of weights are obtained iteratively. Using these two types of weights, two different types of weighted least squares procedures have been proposed. These procedures are applied to detect outliers in and estimate regression coefficients from some widely used datasets such as stack loss, water salinity, Hawkins - Bradu-Kass, Hertzsprung-Russell Star and pilot-point datasets. It has been observed that Campbell-II in particular detects the outlier data points quite well (although occasionally signaling false positive too as very mild outliers). Subsequently, some Monte Carlo experiments have been carried out to assess the properties of these estimators. Findings of these experiments indicate that for larger number and size of outliers, the Campbell-II procedure outperforms the Campbell-I procedure. Unless perturbations introduced to the dataset are sizably numerous and very large in magnitude, the estimated coefficients by the Campbell-II method are also nearly unbiased.

11. References:

[1] Aitken, A.C., (1935), *On Least Squares and Linear Combinations of Observations*, in: *Proceedings of the Royal Society of Edinburgh*, 55: 42 – 48.

[2] Andrews, D.F., (1974), *A Robust Method for Multiple Linear Regression*, in: T*echnometrics,* 16: 523 – 531.

[3] Brownlee, K.A., (1965), *Statistical Theory and Methodology in Science and Engineering,* Wiley, New York.

[4] Campbell, N.A., (1980), *Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation*, in: *Applied Statistics*, 29 (3): 231 – 237.

[5] Daniel, C. and Wood, F.S., (1971), *Fitting Equations to Data*, Wiley, New York.

[6] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A., (1986), *Robust Statistics:The Approach Based on Influence Functions,* Wiley, New York.

[7] Hawkins, D.M., Bradu, D., and Kass, G.V., (1984), *Location of Several Outliers in Multiple Regression Using Elemental Sets*, in: *Technomenics,* 26: 197 – 208.

[8] Kashyap, R.L., and Maiyuran, S., (1993), *Robust Regression and Outlier Set Estimation using Likelihood Reasoning*, in: *Electrical and Computer Engineering ECE Technical Reports*, TR-EE 93-8, Purdue University School of Electrical Engineering. http://docs.lib.purdue.edu/ecetr/33/

[9] Mahalanobis, P.C., (1936), *On the Generalized Distance in Statistics*, in: *Proceedings of the National Institute of Science of India,* 12: 49 – 55.

[10] Plackett, R.L., (1950), *Some Theorems in Least Squares*, in: *Biometrika,* 37: 149 – 157.

[11] Rousseeuw, P.J., and Leroy, A.M., (1987), *Robust Regression and Outlier Detection,* Wiley. New York.

[12] Rupert, D., and Carrol, R.J., (1980), *Trimmed Least Squares Estimation in the Linear Model*, in: *Journal of American Statistical Association, 75:* 828 – 838.

[13] Theil, H., (1971), *Principles of Econometrics*, Wiley, New York.

Note: A Fortran Computer Program for both of the proposed methods is available from the author on request. Contact: mishrasknehu@yahoo.com