ED 411 271 TM 027 241

AUTHOR Lin, Chowhong; Davenport, Ernest C., Jr.
TITLE A Weighted Least Squares Approach To Robustify Least Squares Estimates.
PUB DATE 1997-03-00
NOTE 38p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Estimation (Mathematics); *Least Squares Statistics; Regression (Statistics); *Robustness (Statistics)
IDENTIFIERS *Weighting (Statistical)

ABSTRACT

This study developed a robust linear regression technique based on the idea of weighted least squares. In this technique, a subsample of the full data of interest is drawn, based on a measure of distance, and an initial set of regression coefficients is calculated. The rest of the data points are then taken into the subsample, one after another, and a weighted least squares procedure is performed each time a new data point is brought in, until all data points are included. The weighted average and standard errors of the regression coefficients from all iterations are calculated and compared with those from ordinary least squares and two other robust techniques. It is shown that the technique developed in this paper performs better than the least absolute deviation approach and at the same level with the least median of squares (LMS) approach. The simplicity of this approach, however, seems to justify its use over the LMS. (Contains 5 figures, 10 tables, and 8 references.) (Author/SLD)

A Weighted Least Squares Approach to Robustify Least Squares Estimates

by

Chowhong Lin[1]

Ernest C. Davenport, Jr.

University of Minnesota

A Weighted Least Squares Approach to Robustify Least Squares Estimates

ABSTRACT

This study develops a robust linear regression technique based on the idea of the Weighted Least Squares. In this technique, a subsample of the full data of interest is drawn, based on a measure of distance, and an initial set of regression coefficients are calculated. The rest of the data points is then taken into the subsample, one after another, and a weighted least squares procedure is perform each time a new data point is brought in, until all data points are included. The weighted average and standard errors of the regression coefficients from all iterations are calculated and compared with those from OLS, and two other robust techniques. It is shown that the technique developed in this article performs better than LAD and at the same level with LMS. The simplicity of this technique, however, seems to justify its use over the LMS.

One of the assumptions of an Ordinary Least Squares regression is the homogeneity of error variances. When this assumption is violated, error variances within all levels of the same independent variable(s) are not equal. In such a situation, differential weight, with weight being the inverse of the error variances, is often assigned to each case and then a Weighted Least Squares regression is performed.

Let Y_i be the value of the dependent variable for the i*th* observation, X_{ij} be the j*th* dependent variable of the i*th* observation. Further, let b_j be the regression coefficient for the j*th* independent variable, and e_i be the error for the i*th* observation. Then the regression equations can be expressed in the following manner:

$$Y_i = \sum_j X_{ij} b_j + e_i \qquad (1a)$$

where *i : 1, 2, 3,........,n;*

j : 1, 2, 3,.......,p;

or equivalently,

$$\hat{Y}_i = \sum_j X_{ij} b_j \qquad (1b)$$

where $\hat{Y}_i = Y_i - e_i$ is the predicted value for the i*th* observation.

In an Ordinary Least Squares procedure, the regression coefficients are solved from the following equation, in matrix form:

$$\beta_{j\times 1} = (X'X)^{-1}_{j\times j}(X'Y)_{j\times 1} \qquad (2)$$

The regression coefficients (elements in the *j by 1* matrix) solved from equation 2 minimize the sum of squared errors, $\sum (Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i$ is the predicted value for the ith observation.

In a Weighted Least Squares approach, however, the regression equations are expressed as:

$$w_i Y_i = \sum_j w_i X_{ij} b_j + w_i e_i \qquad (3a)$$

where *i : 1, 2, 3,........,n;*

j : 1, 2, 3,.......,p;

w_i*: weight assigned to the* i*th observation.*

or

$$w_i \hat{Y}_i = \sum_j w_i X_{ij} b_j$$

The weighted regression coefficients are solved from the following equation, again in matrix form:

$$\beta_{j\times 1} = (X'wX)^{-1}_{j\times j}(X'wY)_{j\times 1} \qquad (4)$$

where *w* is an *i by i* diagonal matrix with *w(i,i)* being the weight for the i*th* observation.

The regression coefficients from equation (4) minimize the sum of *weighted* squared errors: $\sum w_i (Y_i - \hat{Y}_i)^2$.

The Weighted Least Squares (or WLS) approach thus can reduce any particular data point's influence on the regression line (or surface) by assigning a smaller weight to it, or can increase any particular point's influence by adding more weight to it. (An Ordinary Least Square Approach, however, assigns unit weight to each observation.) In fact, many of the robust regression techniques, in which the influence of some extreme points are reduced to a certain degree, have utilized this idea and found satisfactory ways of determining the appropriate weights.

The Determination of Weights

Almost all of the parametric robust regression procedures determine weights by consideration of some *distances*, or *errors*. One gets an initial estimates of the regression coefficients, obtains the errors, and determines the weight based on some function (usually the inverse) of the errors. One then iterates the regression procedure in a weighted manner until the errors converges to a certain small value.

This procedure looks simple, but many issues need to be resolved to reach a satisfactory final solution. A major consideration is about the initial estimates. What should be the initial estimates? How would one calculate them? If the Ordinary Least Squares solution is used to serve as the initial estimates, then it is not robust to start with. It might be desirable to start the iteration by using some robust estimates, such as the Least Absolute Deviation solution.

One then has to decide the means of calculation of errors. They could just be the differences between the predicted values and the observed values for each observation. However, how to make use of these errors poses another question. In finding appropriate weights, Huber's M-estimators are obtained by standardizing the errors by a robust estimate of the standard error, and by giving differential weights based on the standardized errors. For example, one of the M-estimators gives weights to observations on the following manner:

$$\begin{aligned} w_i &= 1 \qquad , \text{if } |r_i| \leq 1.5; \qquad (5) \\ &= \frac{1.5}{|r_i|} \qquad , \text{otherwise.} \end{aligned}$$

where r_i is the "standardized residual for the *ith* observation.

There are other different schemes of assigning weights. Refer to Holland and welsch (1977) for some summary information.

This family of the M-estimators suffers from a low breakdown. In fact, Rousseeuw and Leroy (1987) have pointed out that the M-estimators have as low as a 0% breakdown (p.149), which means that even a single outlier can arbitrarily change the regression line/surface.

The programming effort involved in obtaining the M-estimators also contributes to the infrequent use among researchers in education. Many of the statistical packages, such as SPSS, SAS, and MINITAB, do not have a built-in routine to accomplish a robust regression procedure.

Other procedures have also been developed, such as the Least Median of Square procedure (Rousseeuw and Leroy, 1987), and the Minimum Volume Ellipsoid method (Rousseeuw and Zomeren, 1990). An MVE approach utilizes the variance-covariance matrix to define an ellipse (for bivariate datasets) or ellipsoid (for more than two dimensional datasets) that is the smallest in volume among those that contain at least 50% of the actual data point. Hawkins (1993) has proposed a feasible set algorithm (FSA) for solving efficiently the regression coefficients of an MVE solution. His method, however, requires knowledge on linear programming, which is unfamiliar to many of the data analysts in an educational setting.

The Iterative Expanding Reweighted Least Squares Regression Procedure

This paper blends the idea of defining an ellipse/ellipsoid and the idea of a Weighted Least Square approach, and proposes an Iterative Expanding Reweighted Least Square Regression technique to robustify a linear regression procedure. Let *X* be the combined matrix of both the independent variables and the dependent variable. Let X_i be the i*th* row vector (observation). And let X_{ij} be the j*th* variable in that row, $\overline{X}_j$ be the mean vector of the j*th* column (variable). Finally, let S^2 be the variance-covariance matrix. An Ellipsoid can then be expressed in the following way:

$$\left(X_{ij} - \overline{X}\right)'\left(S^2\right)^{-1}\left(X_{ij} - \overline{X}_{ij}\right) \leq \mathrm{C}, \qquad (6)$$

It turns out that the left-hand side of the above inequality is the "distance" between a particular data point to the centroid $(\overline{X}_1, \overline{X}_2, \overline{X}_3, \ldots\ldots \overline{X}_p, \overline{Y})$ of an ellipsoid, taking into account the correlation among variables. Choosing a specific C value will form an Ellipsoid enclosing points whose distances from the centroid are no greater than C.

Apparently, every data point has a distance value. If points are sorted, with ascending distance, then the first 50% (with the lowest distance measures) of the full data set should be a good representation of the whole set of data. (In their research, Rousseeuw and Leroy have indicated a similar one-step improvement over the Ordinary Least Squares by assigning unite weight to points within the ellipsoid defined by equation (6) and zero weight to the rest of the points, and do an ordinary regression.) An initial regression surface can be obtained by using this subsample of the 50% data points. After the initial regression surface is obtained, errors (the difference between the observed value and the predicted value for the dependent variable) can be calculated; and points are again sorted by the absolute amount of errors, with an ascending order. The first 50% of the data points (with the lowest absolute errors) plus the next lowest, forms the starting set of data points that can be used to obtain the regression coefficients at iteration one.

It should be noted that since points are sorted based on some measure of errors, those that with a smaller error should have more weight than those with larger errors. The weight for each point is defined as:

$$\begin{aligned} w_i &= 1, \quad \text{if the point is in the lower half with smaller errors;} \qquad (7) \\ &= \frac{1}{(1+|error|)}, \quad \text{otherwise.} \end{aligned}$$

(Note that the error in (7) is the initial error, and weights are defined once only, after the first iteration which includes n/2 data points.)

Then a weighted least squares is performed with the *(n/2 + 1)* data points [or *(n+ 1)/2* if the number of data points in odd], and regression coefficients stored. The program will search for the next data point on the list and add it into the *(n/2 + 1)* points, with weights being assigned by (7). (However, weights in (7) are rescaled each time so that for each individual iteration, the sum of weights of included data points will be the number of data points included.) The observed regression coefficients are again stored. This program then goes to the next point, adds it in the model, and recalculates and stores the regression coefficients. Therefore, there will be *n/2* [or *(n/2+ 1)*] sets of regression coefficients. The first set contains either *(n/2+ 1)* or *(n+1)/2* data points, the final set contains *n* data points.

After all sets of regression coefficients are obtained, the "overall" regression coefficients are calculated as the weighted average of the previously obtained regression coefficients, with the weight being the number of data points used at each iteration. Therefore, the first set of the regression coefficients will have a weight of *(n/2+ 1)* or *(n+1)/2*, and the final set will have a weight of *n*.

If we let b_j be the "overall" weighted average (for the jth independent variable) of the *n/2* or *n/2+ 1* sets of regression coefficients; w_{ik}' the rescaled weight at the k*th* iteration for the i*th* data point, and b_{jk} be the estimate of the j*th* independent variable at the k*th* iteration, then we can write:

$$b_j = \frac{\sum w_i' b_{jk}}{\sum w_i'}, \text{ where} \qquad (8a)$$

$$w_i' = \frac{w_{ik} \times n}{\sum w_{ik}}, \text{ updated at each iteration.} \qquad (8b)$$

The standard error of the regression coefficients can also be obtained via the weighted manner:

$$S.E._{\cdot b_j} = \sqrt{\frac{\sum n_k b_{jk}^2 - \frac{\sum (n_k b_{jk})^2}{\sum n_k}}{\sum n_k}} \tag{9}$$

where n_k is the sample size at the k*th* iteration,

and b_{jk} as defined in (8a).

The model standard error can be obtained by the following equation, using the overall average regression coefficients and the weights at the final (the n/2 *th*) iteration:

$$S.E._{\cdot \text{model}} = \sqrt{\frac{\sum w_i^{'} (Y_i - \hat{Y}_i)^2}{\sum w_i^{'} - p}}, \text{ with no intercept; or} \tag{10a}$$

$$S.E._{\cdot \text{model}} = \sqrt{\frac{\sum w_i^{'} (Y_i - \hat{Y}_i)^2}{\sum w_i^{'} - p - 1}}, \text{ with an intercept.} \tag{10b}$$

Algorithm for an Iterative Expanding Reweighted Least Squares

Step 1. Input the data matrix, which includes the dependent and independent variables.

Step 2. Calculate the distance measure by using equation

$$\text{Distance} = (X_{ij} - \overline{X})^{'} (S^2)^{-1} (X_{ij} - \overline{X}_{ij}) \tag{11}$$

Step 3. Sort points on the ascending order on the distance.

Step 4. Select half of the points with the shortest distance measures as the base and perform an Ordinary Least Squares regression.

Step 5. Calculate the unstandardized errors for all data points based on the regression coefficients obtained in step 4.

Step 6. Sort errors from Step 5 on the ascending order of their magnitudes, irrespective of the signs.

Step 7. Calculate weight associated with each data point via equation (7). The first 50% of the data points will have an initial weight of 1.

Step 8. Select the first $n/2+ 1$ or $(n+ 1)/2+ 1$ points, depending on number of observations one has. This is the start-up working dataset.

Step 9. Rescale the weight so that the sum of weight is th number of data points at this iteration.

Step 10. Perform a weighted least squares regression and store the regression coefficients.

Step 11. If all data points are in the working dataset, go to Step 13. Otherwise, add the point that is next on the list (with the next smallest error) to the working dataset.

Step 12. Go to Step 9.

Step 13. Calculate the weighted mean, as well as the standard error, for each regression coefficient.

Step 14. Calculate the model standard error via equation (10), using the rescaled weight.

Step 15. Output regression coefficients and standard error and the model standard error.

Step 16. Exit.

Empirical Runs of the Iterative Expanding Reweighted Least Squares Regression Procedure

Five different datasets are used to run the IERLS procedure, and results are compared with those obtained via the Ordinary Least Squares, the Least Median of Squares, and the Least Absolute Deviation approaches. The five datasets are: the salinity data (Rupport and Carroll, 1980), the air quality data as used by Rousseeuw and Leroy (1984), the stackloss data (Brownlee, 1965), the Coleman data (Mosteller and Tukey, 1977) and an artificial data set from Hawkins, Bradu and Kass (1984).

The solutions from LMS can be found at Rousseeuw and Leroy (1984). The OLS is performed with an SPSS routine. The other two programs, for the reweighted and the LAD, are written in the FORTRAN-77 code and are available from the authors. Several subroutines in IMSL are used to complete the IEWLS program. The RLAD subroutine is used to calculate the Least Absolute Deviation solutions for a linear regression. It is shown that this proposed reweighted procedure does have the desired ability to detect multivariate outliers where the OLS and LAD both perform less satisfactorily.

The Stackloss data:

The Least Medium of Squares identifies cases 1, 2, 3, 4 and case 21 as outliers. The Iterative Expanding Reweighted Least Weighted Squares identifies cases 1, 3, 4 and case 21 as outliers. The above two identify outliers as those whose standardized (Residual./Scale) residual is larger than 2.5 in magnitude. Judged somewhat subjectively, by the authors, on the magnitude of the unstandardized errors, the Least Absolute Deviation identifies cases 1, 3, 4, 21 as outliers. As shown in table 6, an OLS procedure is unable to detect any of the outliers identified by the above robust techniques.

The Salinity data:

The Least Median of Squares identifies cases 5, 8, 16, 23 and case 24 as outliers. The Iterative Expanding Reweighted Least Weighted Squares identifies cases 1, 5, 8, 9, 15, 16 and case 17 as outliers. The Least Absolute Deviation identifies cases 16 as an outlier. However, for comparison purposes 6 other cases with largest residuals are also undelined (see table 7). Table 7 also indicates that the OLS procedure is nonrobust with regard to this data set. It identifies no data points as outliers.

The Air Quality data:

The Least Median of Squares identifies case 30 as the single outlier. The Iterative Expanding Reweighted Least Squares has identified cases 1, 4, 9, 21 and 30 as outliers, with case 30 being much more extreme than the other 4 points. LAD suggests case 30 as the only outlier. The OLS procedure has case 30 as the outlier, but with a much less standardized residual than that obtained by the IERLS procedure.

The Hawkins et.al., data:

The Least Median of Squares and the Iterative Expanding Reweighted Least Squares both successfully identify the ten outliers in this artificial data set. The LAD, however, fits poorly for this data set. It fits case 5, an outlier, perfectly; yet cases 11, 12, 13 and 14 are identified as outliers. However, it also fails to identify the first 10 cases. This is an example of the "masking" effect, in which outlier cluster together so any single one of them, taken separately, does not appear to be an outlier.

The Coleman data:

The Least Median of Squares has cases 3, 17 and 18 as outliers. The Iterative Expanding Reweighted Least Squares identifies cases 3 and 18 as outliers. The Least Absolute Deviation identifies cases 3 and 18. Again, the OLS procedure fails to detect any outliers from this data set.

Discussion

OLS performs the worst with respect to the identification of outliers. This is to be expected since the quantity it minimizes gives undue weight to outlying observations. The LMS and IERLS perform more consistently with each other across different data sets, although there is minor inconsistency between them. The performance of an LAD procedure is somewhat unreliable. In the Hawkins et.al., data, it even gives a perfect fit to an oitlier, which is undesirable.

Also, the standard errors from an IERLS are much less than an LMS procedure. The standard errors from an LAD or an LMS procedure are not listed, but it is believed that the reweighted procedure will yield the lowest standard errors, due to the way weights are assigned.

While the LMS is also well recognized as one of the standard robust regression technique, it requires more programmming background. However, this reweighted procedure only requires that researchers have the knowledge of an WLS solution. Therefore, it seems to suggest that more consistent estimates are likely obtained by using the IERLS procedure.

Recommendations

Several directions of further investigations may be of interest with respect to this new robust regression procedure.

First, a more comprehensive simulation study is needed to determine how this procedure will fare in various contaminating conditions. Also, the breakdown point of this procedure is yet to be determined.

Second, it seems reasonable to apply this reweighting scheme to other parametric procedures. For example, to calculate the correlation coefficients among several variables, it is possible to select first a small set of points based on the criteria mentioned in this paper, then expand and calculate the correlation coefficient at each iteration in a weighted manner, and finally calculate the weighted mean of these correlation coefficients as the "robust" correlation coefficient.

Third, it would be instructive to calculate the ordinary regression coefficients without those data points identified by the IERLS as outliers. it is expected that the "trimmed" data will yield regression coefficients not much different from the "untrimmed" ones.

Fourth, the sampling distribution of the regression coefficients obtained by IERLS is unknown. This paper gives a preliminary estimate of the standard errors, both for the coefficients and for the model. It seems that the standard errors obtained this way are very liberal.

References

Brownlee, K. A. (1965). Statistical theory and methodology in science and engineering. (2nd ed.). New York: John Wiley and Sons.

Hawkins, D.M., Bradu, D., & Kass, G. V. (1984). Location of several outliers in multiple regression data using elemental sets. Technometrics, 26, 197-208.

Hawkins, D. M. (1993). The feasible set algorithm for least median of squares regression. Computational statistics and data analysis, 16, 81-101.

Holland, P. W., & Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. Communications in statistics, theory and methodology, A6(9), 813-827.

Tatsuoka, M. M. (1971). Multivariate data analysis; techniques for educational and psychological research. New York: John Wiley and Sons.

Rousseeuw, P. J., & Leroy, A. M. (1987). Robust regression and outlier detection. New York: John Wiley and Sons.

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association, 85, 633-639.

Ruppert, D., & Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. Journal of the American Statistical Association, 75, 828-838.

Footnotes

[1] Correspondence concerning this article should be sent to:

Chowhong Lin
Department of Educational Psychology
University of Minnesota
324 Burton
Minneapolis, MN 55455

Table 1. Regression Coefficients and Standard Errors for the Stackloss data*:

	OLS		IERLS	
b0	-39.92	(11.90)	-33.30	(1.57)
b1	.72	(.14)	.75	(.03)
b2	1.30	(.37)	.52	(.07)
b3	-.15	(.16)	-.07	(.02)
model		(3.24)		(1.72)

	LMS	LAD
b0	-34.5	-39.69
b1	.71	.83
b2	.36	.57
b3	0.00	-.006

*: Standard errors for LMS and LAD are unavailable.

Table 2. Regression Coefficients and Standard Errors for the Salinity data*:

	OLS	IERLS
b0	9.59 (3.13)	19.78 (.47)
b1	.78 (.09)	.72 (.01)
b2	-.03 (.16)	-.16 (.02)
b3	-.30 (.11)	-.69 (.02)
model	(1.33)	(.48)

	LMS	LAD
b0	36.7	14.21
b1	.36	.74
b2	-.07	-.11
b3	-1.30	-.46

*: Standard errors for LMS and LAD are unavailable.

Table 3. Regression Coefficients and Standard Errors for the Air Quality data^:

	OLS		IERLS	
b0	-69.66	(45.94)	-70.12	(5.31)
b1	-.02	(.04)	.004	(.002)
b2	-2.19	(1.15)	2.27	(.31)
b3	1.84	(.67)	.97	(.10)
model		(18.33)		(7.01)

	LMS		LAD
b0	-37.52	(28.95)	-46.77
b1	.01	(.02)	-.002
b2	-.75	(.71)	-.79
b3	.99	(.43)	1.18

^: Standard errors for LMS and LAD are unavailable.

Table 4. Regression Coefficients and Standard Errors for the Hawkins et.al., data[*,**]:

	OLS	IERLS
b0	-.39 (.42)	-.13 (.15)
b1	.24 (.26)	.18 (.03)
b2	-.33 (.16)	-.05 (.04)
b3	.38 (.13)	-.08 (.08)
model	(2.25)	(1.59)

	LAD
b0	-.88
b1	.10
b2	.15
b3	.22

[*] : Standard errors for LAM and LAD are unavailable.

[**] : Regression Coefficients for LMS are unavailable.

Table 5. Regression Coefficients and Standard Errors for the Coleman data*:

	OLS		IERLS	
b0	19.95	(13.63)	20.68	(4.28)
b1	-1.79	(1.23)	-1.96	(.15)
b2	.44	(.05)	.05	(.02)
b3	.56	(.09)	.65	(.01)
b4	1.11	(.43)	1.24	(.04)
b5	-1.81	(2.03)	-2.56	(.70)
model		(2.07)		(1.18)

	LMS	LAD
b0	29.75	29.21
b1	-1.20	-1.73
b2	.08	-.006
b3	.66	.67
b4	1.10	1.12
b5	-3.90	-3.55

* : Standard errors for LAM and LAD are unavailable.

Table 6. Residuals associatd with various regression fits, Stackloss data*

ID	NZ	ZRE_1	LAD	LMS
1	3.637	.997	5.06	>7
2	.678	-.591	.00	>3
3	3.799	1.405	5.43	>6
4	4.702	1.757	7.64	>7
5	-.377	-.528	-1.22	
6	-.739	-.927	-1.79	
7	-.178	-.737	-1.00	
8	.403	-.428	.00	
9	-.654	-.969	-1.46	
10	.175	.391	-.02	
11	.689	.813	.53	
12	.414	.857	.04	
13	-1.452	-.440	-2.90	
14	-.604	-.016	-1.80	
15	.860	.728	1.18	
16	.108	.279	.00	
17	-.474	-.469	-.43	
18	-.074	-.140	.00	
19	.202	-.184	.49	
20	1.059	.435	1.62	
21	-4.819	-2.232	-9.48	<-5

*: NZ is the "standardized residual" from the iterative procedure. ZRE_1 is the standarized residual from an OLS run. LAD, however, is the raw score difference between the actual values and predicted values for the dependent variable. Values underlined are those that are likely to be outliers. For the LMS residuals (standardized), only those identified as outliers are shown.

Table 7. Residuals associatd with various regression fits, Salinity data*

ID	NZ	ZRE_1	LAD	LMS
1	-3.188	-1.107	-1.70	
2	-.565	-.470	-.65	
3	-1.182	-.805	-1.23	
4	.812	.249	.00	
5	3.710	-.229	.33	>8
6	.704	.181	.00	
7	1.156	.458	.35	
8	-3.345	-.949	-1.59	<-4
9	3.455	1.753	1.71	
10	-.073	-.144	-.36	
11	-1.668	-1.137	-1.40	
12	-.909	-.134	-.50	
13	2.735	1.238	1.32	
14	-1.318	-.556	-.80	
15	-5.021	-1.854	-2.70	
16	12.996	2.045	4.01	>20
17	-4.540	-2.003	-2.68	
18	.818	.523	.31	
19	.418	.802	.36	
20	-.709	.160	-.29	
21	.774	.687	.51	
22	-.188	.200	.00	
23	1.573	-.087	.03	>6
24	.544	-.601	-.55	>6
25	.323	.218	.00	
26	-.256	.338	.00	
27	.093	.479	.22	
28	1.687	.746	.81	

*: NZ is the "standardized residual" from the iterative procedure. ZRE_1 is the standarized residual from an OLS run. LAD, however, is the raw score difference between the actual values and predicted values for the dependent variable. Values underlined are those that are likely to be outliers. For the LMS residuals (standardized), only those identified as outliers are shown.

Table 8. Residuals associatd with various regression fits, Air Quality data*

ID	NZ	ZRE_1	LAD	LMS
1	4.517	.865	20.47	
2	2.136	-.404	4.03	
3	-2.730	-1.338	-18.65	
4	.210	.209	.87	
7	1.179	-.178	.00	
8	.277	.648	7.00	
9	-3.135	.527	-1.45	
12	-.697	-.869	-10.91	
13	-.833	-.870	-12.72	
14	-1.126	-.718	-10.75	
15	.439	.592	6.67	
16	-.670	-.192	-5.47	
17	1.760	.734	12.53	
18	-2.399	.669	.00	
19	1.004	.269	5.78	
20	-.310	-.628	-7.87	
21	-1.273	-.904	-14.36	
22	-3.725	-.663	-15.09	
23	-1.153	-.926	-13.71	
24	2.269	.936	16.20	
28	.078	-.228	.00	
29	.331	.125	8.01	
30	12.891	3.023	73.03	3.08
31	1.791	-.679	.00	

*: NZ is the "standardized residual" from the iterative procedure. ZRE_1 is the standarized residual from an OLS run. LAD, however, is the raw score difference between the actual values and predicted values for the dependent variable. Values underlined are those that are likely to be outliers. For the LMS residuals (standardized), only those identified as outliers are shown.

Table 9. Residuals associatd with various regression fits, Hawkins et.al., data*

ID	NZ	ZRE_1	LAD	LMS
1	7.038	1.495	.55	>14
2	7.417	1.768	.74	>14
3	7.498	1.329	.42	>14
4	7.131	1.234	-.52	>14
5	7.388	1.354	.00	>14
6	7.216	1.518	.47	>14
7	7.766	1.995	1.24	>14
8	7.464	1.698	1.06	>14
9	7.187	1.200	-.26	>14
10	7.356	1.348	.38	>14
11	1.160	-3.478	-11.55	
12	.984	-4.158	-12.12	
13	1.628	-2.725	-10.83	
14	1.618	-1.711	-12.55	
15	-.362	-.298	-.73	
16	.193	.378	.79	
17	.019	.287	.40	
18	-.031	-.176	-.01	
19	.226	.288	.12	
20	.197	.145	.00	
21	.517	.296	.79	
22	.422	.414	.25	
23	-.536	-.194	-.63	
24	.475	.598	1.00	
25	-.199	-.133	.40	
26	-.296	-.215	-1.04	
27	-.514	-.614	-1.13	
28	.190	-.106	.46	
29	.204	.176	.76	
30	-.055	-.558	-.37	
31	-.177	-.123	.16	
32	-.134	.245	.01	
33	-.294	-.052	-.65	
34	-.368	-.297	-.01	
35	.141	-.180	-.12	
36	-.471	-.520	-1.14	
37	-.084	-.100	-.74	
38	.528	.554	1.42	
39	-.538	-.564	-.28	
40	-.166	-.010	-.23	
41	-.176	-.490	-.41	
42	-.228	-.476	-.58	
43	.600	.762	.78	
44	-.391	-.796	-.75	
45	-.415	-.338	.06	

ID	NZ	ZRE_1	LAD	LMS
46	-.212	-.629	-.41	
47	-.784	-.668	-.49	
48	-.011	-.386	-.20	
49	.469	.286	.70	
50	-.178	-.307	-.77	
51	.330	.390	1.05	
52	-.531	-.517	-.28	
53	.664	-.044	.78	>2.5
54	.510	.756	.96	
55	.147	.321	.28	
56	.087	.342	.13	
57	.376	.276	1.16	
58	-.072	.117	-.28	
59	-.075	-.328	-.53	
60	-.309	-.593	-.98	
61	.103	-.008	-.62	
62	.378	.308	1.32	
63	-.135	.285	.03	
64	-.315	-.398	-.96	
65	.387	-.127	.61	
66	-.410	-.117	-.48	
67	-.418	-.217	-.53	
68	.505	.245	1.24	
69	.016	.085	.74	
70	.622	.212	.53	
71	.150	.005	.00	
72	.026	.062	.00	
73	.463	.199	.52	
74	-.335	-.171	-.70	
75	.315	-.148	.43	

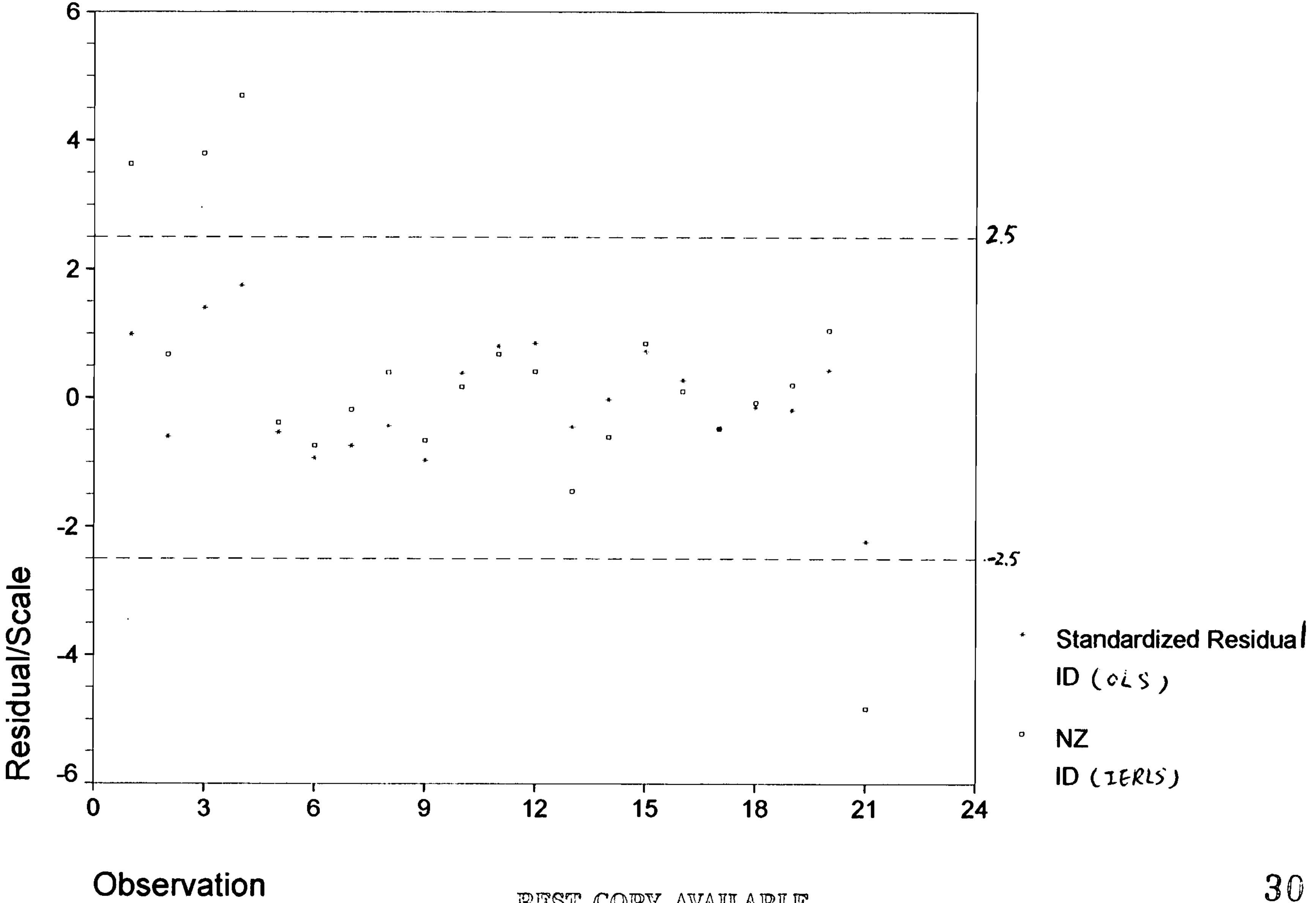
*: NZ is the “standardized residual″ from the iterative procedure. ZRE_1 is the standarized residual from an OLS run. LAD, however, is the raw score difference between the actual values and predicted values for the dependent variable. Values underlined are those that are likely to be outliers. For the LMS residuals (standardized), only those identified as outliers are shown.

Table 10. Residuals associatd with various regression fits, Coleman data*

ID	NZ	ZRE_1	LAD	LMS
1	.078	.168	.00	
2	.661	-.169	.00	
3	-3.800	-1.904	-3.98	-4.04
4	-.939	-.228	-.58	
5	.555	.376	.52	
6	.203	-.041	.00	
7	-.002	.347	.23	
8	-.136	-.209	-.45	
9	-.093	.301	.00	
10	.058	.101	.00	
11	-.431	-1.064	-.76	
12	1.527	.842	.88	
13	-.928	-.506	-1.44	
14	.132	-.167	.31	
15	.068	-.858	.00	
16	.837	.625	1.30	
17	-.939	-.694	-1.18	-2.82
18	5.627	2.411	6.80	5.76
19	.425	.541	1.34	
20	-.294	.127	-.03	

*: NZ is the "standardized residual" from the iterative procedure. ZRE_1 is the standarized residual from an OLS run. LAD, however, is the raw score difference between the actual values and predicted values for the dependent variable. Values underlined are those that are likely to be outliers. For the LMS residuals (standardized), only those identified as outliers are shown.

Figure 1.



BEST COPY AVAILABLE

Figure 2

Residual Comparison

Salinity data

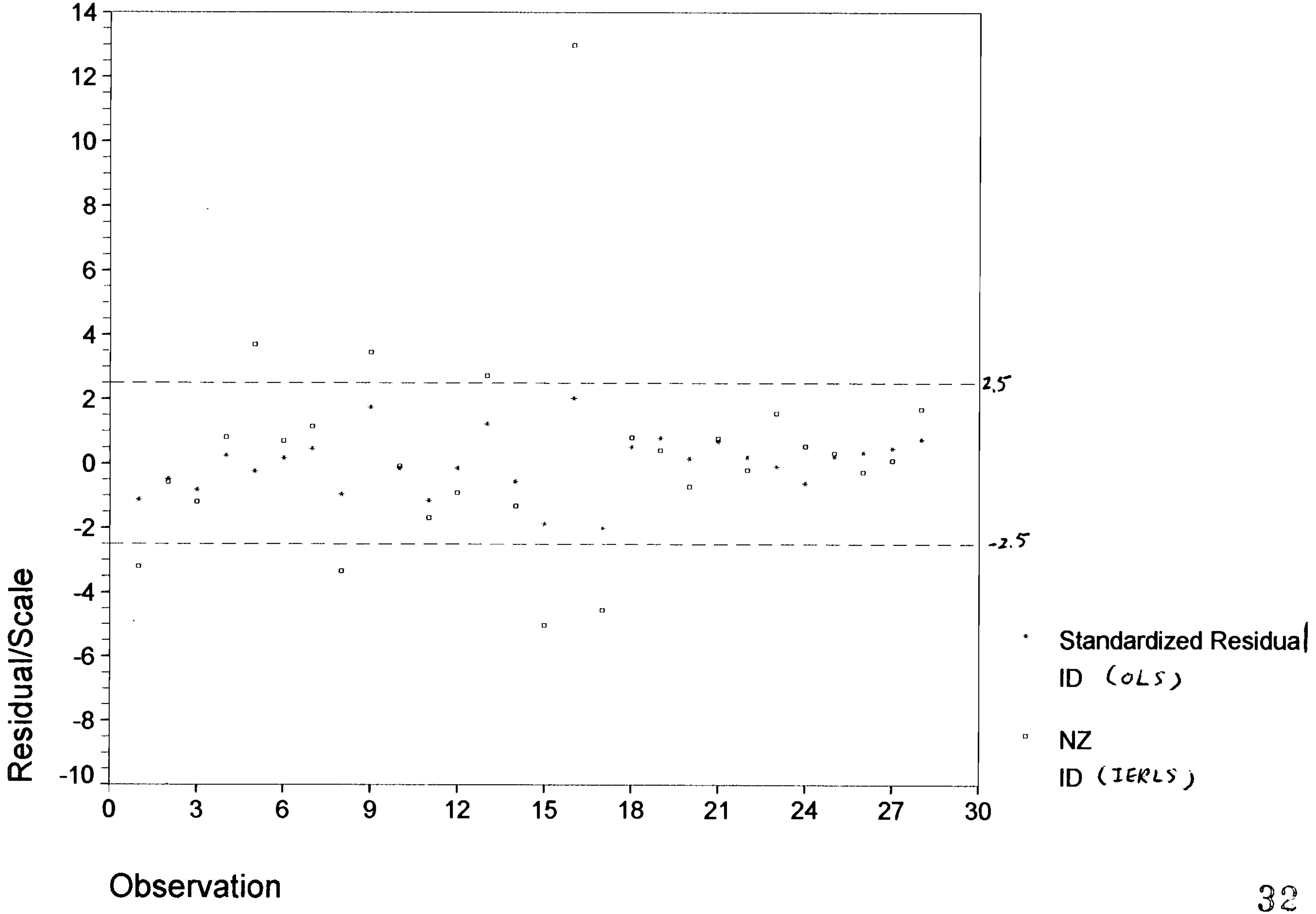


Figure 3.

Residual Comparison

Air Quality data

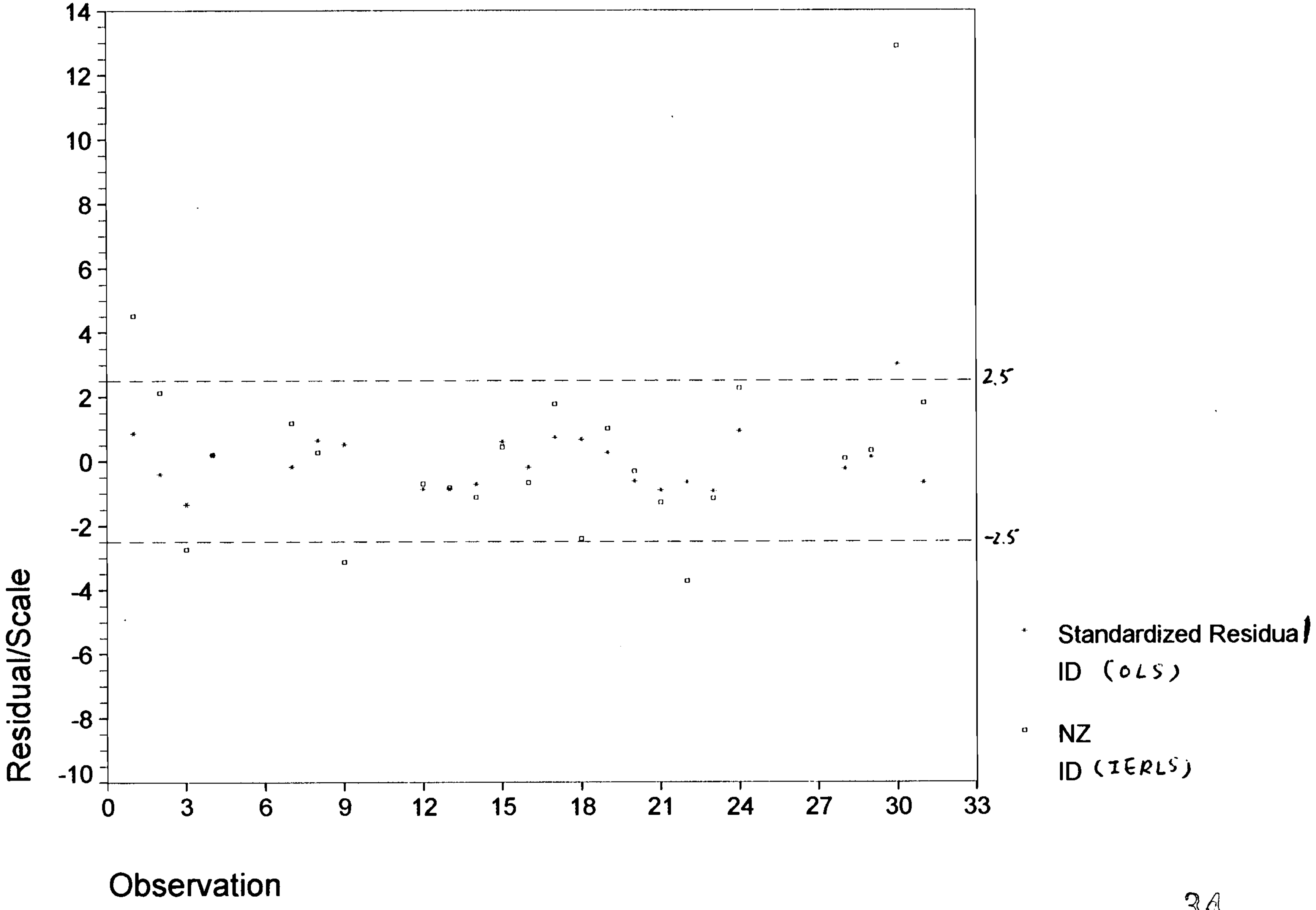


Figure 4.

Residual Comparison

Hawkins et.al., data

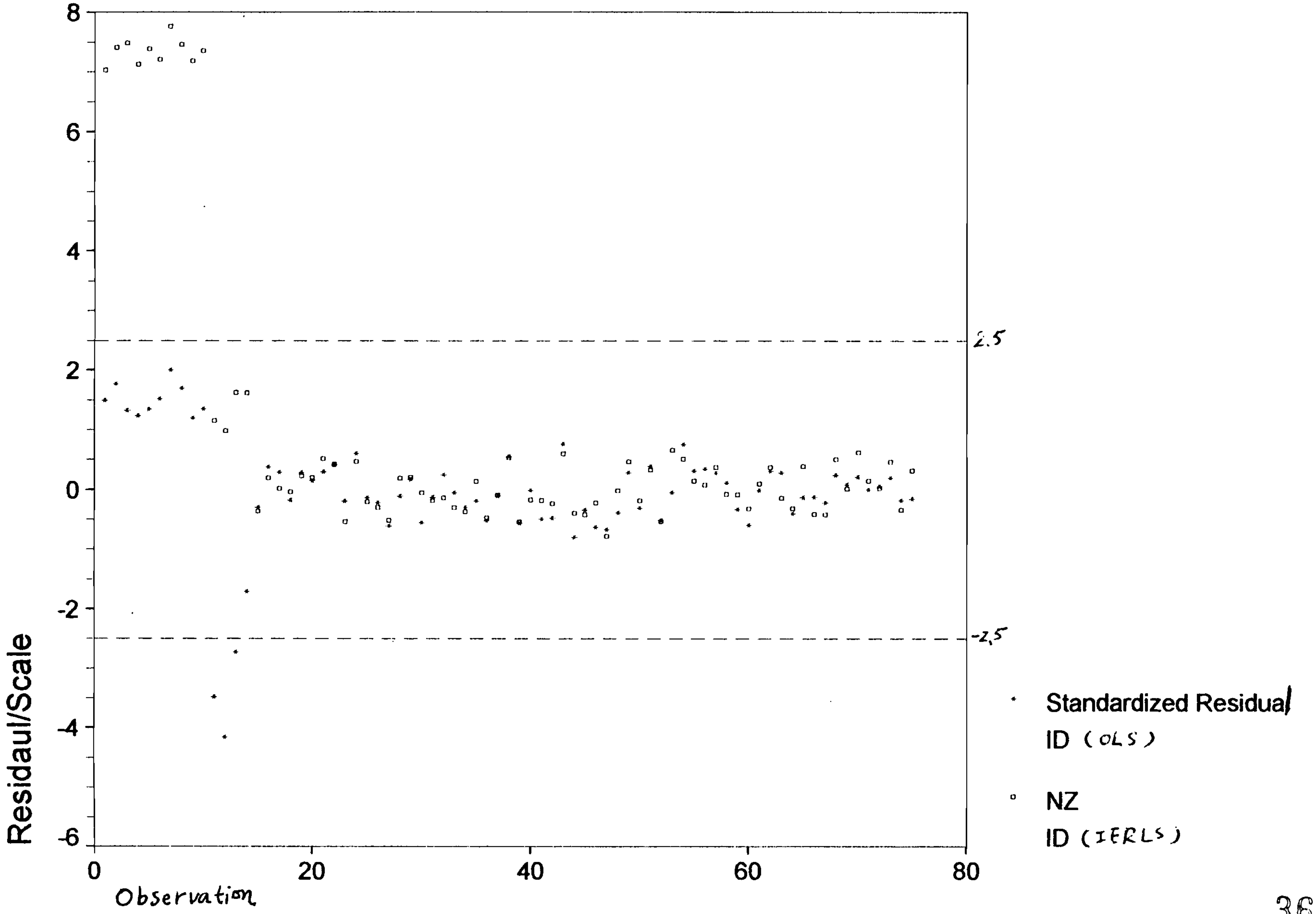
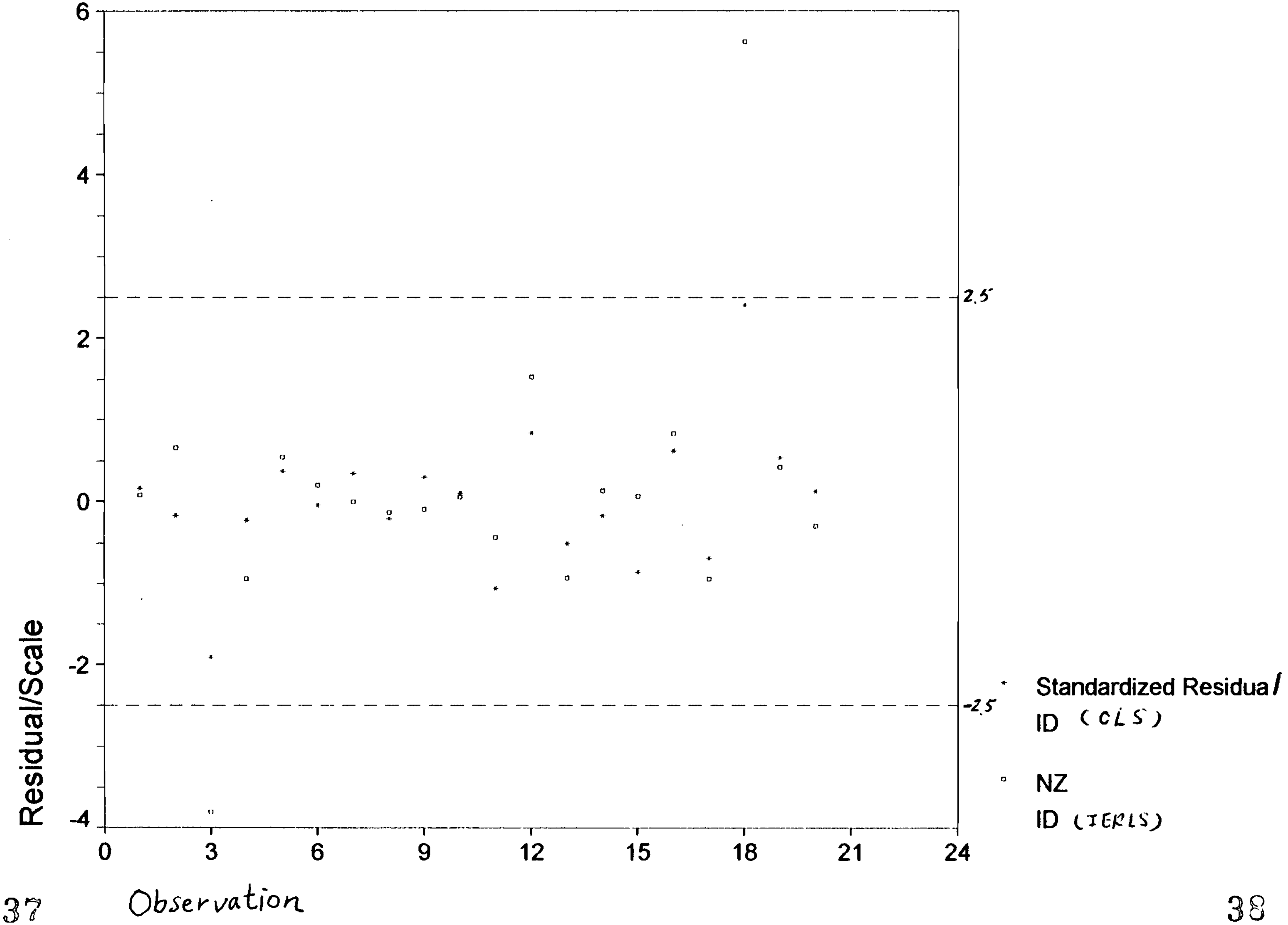


Figure 5.

Residual Comparison

Coleman data



AREA 1997

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

TM027241

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: A weighted least squares Approach to robustify Least Squares Estimates

Author(s): Chowhong Lin, Ernest C. Davenport

Corporate Source:

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☑ **Check here** Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction	**Sample sticker to be affixed to document** "PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)." **Level 1**	**Sample sticker to be affixed to document** "PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)." **Level 2**	□ **or here** Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature:	Position:
Printed Name: CHOWHONG LIN	Organization: University of Minnesota, Mpls
Address:	Telephone Number: (612) 624-1896
	Date: 4/1/97

OVER

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064
202 319-5120

February 21, 1997

Dear AERA Presenter,

Congratulations on being a presenter at AERA[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a printed copy of your presentation.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our processing of your paper at http://ericae2.educ.cua.edu.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (523)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

AERA 1997/ERIC Acquisitions
The Catholic University of America
O'Boyle Hall, Room 210
Washington, DC 20064

This year ERIC/AE is making a **Searchable Conference Program** available on the AERA web page (http://aera.net). Check it out!

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

[1] If you are an AERA chair or discussant, please save this form for future use.