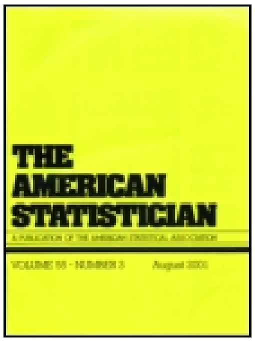

THE AMERICAN STATISTICIAN

The American Statistician

Taylor & Francis
Taylor & Francis Group

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://amstat.tandfonline.com/loi/utas20

The Use of Recursive Residuals in Checking Model Fit in Linear Regression

Jacqueline S. Galpin & Douglas M. Hawkins

To cite this article: Jacqueline S. Galpin & Douglas M. Hawkins (1984) The Use of Recursive Residuals in Checking Model Fit in Linear Regression, The American Statistician, 38:2, 94-105

To link to this article: https://doi.org/10.1080/00031305.1984.10483175

Published online: 30 Mar 2012.

The Use of Recursive Residuals in Checking Model Fit in Linear Regression

JACQUELINE S. GALPIN and DOUGLAS M. HAWKINS*

Recursive residuals are independently and identically distributed and, unlike ordinary residuals, do not have the problem of deficiencies in one part of the data being smeared over all the residuals. In addition, recursive residuals may be interpreted as showing the effect of successively deleting observations from the data set. We propose the use of the normal probability plot and the cumulative sum plots of the recursive residuals, and of the square roots of the absolute values of the recursive residuals to check the model assumptions of normality and homoscedasticity, and other aspects of model misfits such as change of regime, outliers, and omitted predictors, in place of plots based on ordinary residuals. A further advantage of recursive residuals is that they are open to formal statistical testing, so that these plots can be automated and in fact produced only when a model misfit has been detected.

KEY WORDS: Recursive residuals; Normal probability plot; Cumulative sum plots; Multiple regression.

1. INTRODUCTION

We consider the usual regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{Y}$ is the $(n \times 1)$ vector of observations on the dependent variable; $X(n \times p)$ contains the n observations on the p predictors (including the intercept term); $\beta(p \times 1)$ is the regression vector to be estimated; and $\epsilon(n \times 1)$ is the vector of error terms, which are assumed to be iid.

The usual least squares estimate of the regression vector is then given by

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

When such a regression model is fitted to a set of data, certain implicit assumptions are made. These include the assumption that the same regression model applies to the whole data set, that is, that there is no change of regime over the data (and also that there are no outliers). It is also assumed that all important variables are included in the predictor set. If any testing of the model is to be done (as is usually the case), the assumption of normality of the errors is also invoked.

The regression residuals, $e_i = Y_i - \hat{Y}_i$, i $= 1, \ldots, n$ (where $\hat{\mathbf{Y}} = \mathbf{Xb}$), are often used to check some of these model assumptions and to check for lack of fit of the regression model. These ordinary residuals do, however, have some major defects: they are not independent, and in general they do not all have the same variance. (If $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, then $\mathbf{e} \sim N(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'))$.) This results in the effect of structural change, for example, being smeared over all the residuals. It also implies that the distribution of the residuals is dependent on the particular design matrix under consideration. Since the set of ordinary residuals is a scale mixture of normal distributions, the normal probability plot of ordinary residuals will, if the influences of the data points vary greatly, indicate that the data are nonnormal. Ordinary residuals are also highly susceptible to masking and swamping problems with multiple outliers (see Cook and Weisberg 1982).

Recursive residuals have frequently been suggested for testing model fit and model assumptions (see Brown, Durbin, and Evans 1975; Phillips and Harvey 1974; Riddell 1980; and Hawkins 1980). Recursive residuals are a linear transformation of ordinary residuals, such that they are identically and independently distributed. Under the assumption that the data are normal, model defects are thus not smeared across all recursive residuals. Since they are all of the same scale, indications of nonnormality from normal probability plots really do indicate nonnormality (in the absence of other model defects that can mimic nonnormality) and not differences in the influences of the observations. Recursive residuals are also not subject as much to masking and swamping as ordinary residuals. Although it is possible for outliers occasionally to remain unidentified, their presence can still be detected by the normal probability plot failing to pass through the origin. This is due to another advantage that recursive residuals have over ordinary residuals—they are not constrained to sum to zero. Recursive residuals also allow us to test for a change of regime, something for which ordinary residuals are not well suited.

The fact that the recursive residuals follow independent identical $N(0, \sigma^2)$ distributions under the full model has the important implication that exact tests may be applied to them. This removes the element of subjectivity that can cloud the interpretation of ordinary residual plots. It further implies that omnibus tests may be applied automatically to the set of recursive residuals by a regression routine, with detailed results and plots only being printed out for the user's attention if a significant model misfit is detected. This is explored more fully in Section 3.4.

The recursive residuals are defined in Section 2, where a simple interpretation of them is also given. Three plots of the recursive residuals are defined in

*Jacqueline S. Galpin is a Chief Research Officer and Douglas M. Hawkins is a Senior Consultant at the National Research Institute for Mathematical Sciences of the Council for Scientific and Industrial Research, P.O. Box 395, Pretoria, South Africa.

Section 3, where the effects of various model defects on these plots are discussed. The plots are the normal probability plot, the cumulative sum (cusum) plot of the recursive residuals, and the cusum plot of the square roots of the recursive residuals. The model defects discussed are nonnormality of the errors, heteroscedasticity and change of scale of the errors, outliers, change of regime, omitted variables, and serial correlation of the errors. The results are summarized in Section 4, and an example is discussed in Section 5.

We have used these recursive residual plots as well as ordinary residual plots for some time and have found many cases in which we obtained valuable information from the recursive residual plots, but not from the ordinary residual plots, but we have not come across any cases where the reverse holds.

2. DEFINITION OF THE RECURSIVE RESIDUALS

Let $\mathbf{b}_r$ be the least squares regression vector based on the first r observations. Let $\mathbf{X}$ and $\mathbf{Y}$ be partitioned accordingly, so that $\mathbf{X}' = (\mathbf{x}_1, \dots, \mathbf{x}_r)$ and $\mathbf{Y}_r' = (y_1, \dots, y_r)$. Assume that $\mathbf{X}_r'\mathbf{X}_r$ is nonsingular. Then $\mathbf{b}_r = (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{Y}_r$, and the recursive residuals are defined (Brown, Durbin, and Evans 1975) as

$$w_r = \frac{y_r - \mathbf{x}_r'\mathbf{b}_{r-1}}{\sqrt{1 + \mathbf{x}_r'(\mathbf{X}_{r-1}'\mathbf{X}_{r-1})^{-1}\mathbf{x}_r}}, \qquad r = p+1, \dots, n.$$

It is well known that if $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ then $\mathbf{w} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{n-p})$.

The recursive residuals can be interpreted as follows. For a data set of n points, discard the last data point and fit the model to the first $n-1$ points, thus obtaining $\mathbf{b}_{n-1}$. The recursive residual is then defined as the standardized residual of the last observation from the new line, the standardization making the variance σ^2. Now discard the second to last point as well, and fit the regression model to the first $n-2$ points. The standardized residual of this second to last point from the new line is then w_{n-1}. Continue omitting points in this way, obtaining $w_n, w_{n-1}, \dots, w_{n-p}$. It is obvious that only $n-p$ recursive residuals can be calculated, as at least p points are needed for the fitting of a p-parameter regression. Brown, Durbin, and Evans suggest using the first p data points as the base for the regression (as described above), but this is by no means the only sensible choice. The last p points (or, in fact, any arbitrarily selected p points) can also be used.

The calculation of the recursive residuals may appear to be a time-consuming operation, involving the fitting of $n-p$ regressions, but the use of the well-known updating formulas (Plackett 1950; Bartlett 1951) allows this to be done in a very economical manner. These updating formulas are

$$(\mathbf{X}_r'\mathbf{X}_r)^{-1} = (\mathbf{X}_{r-1}'\mathbf{X}_{r-1})^{-1} - \mathbf{dd}'/(1 + \mathbf{x}_r'\mathbf{d}),$$

$$\text{where } \mathbf{d} = (\mathbf{X}_{r-1}'\mathbf{X}_{r-1})^{-1}\mathbf{x}_r$$

$$\mathbf{b}_r = \mathbf{b}_{r-1} + (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{x}_r(\mathbf{y}_r - \mathbf{x}_r'\mathbf{b}_{r-1})$$

$$S_r = S_{r-1} + w_r^2, \qquad r = p+1, \dots, n,$$

where S_r is the residual sum of squares based on r observations.

As an indication of the time required for these operations, a recent regression problem involving 25 variables and 1,068 observations used (on a CDC CYBER 174) 3.9 central processor seconds to calculate the recursive residuals, 3.8 seconds to perform a stepwise regression, and 14 seconds to obtain the means, standard deviations, and correlation matrix. Timings for smaller problems have had similar ratios.

It is the interpretation of the recursive residuals as showing the effect of successively deleting points from the data set, in addition to their property of independence, that makes their use so attractive. In this respect, they have an advantage over Theil's (1965) BLUS (best linear unbiased scalar) residuals, which cannot be so interpreted. Because of this interpretation, recursive residuals are very flexible, and they can be used directly to detect changes in the regression equation over the observations. They can also be used to indicate possible outliers that have a great effect on the regression line. (As emphasized by Andrews and Pregibon 1978, Hoaglin and Welsch 1978, and Cook 1977,1979, the detection and examination of outliers with high leverage are the object of the outlier detection procedures, since such outliers have a catastrophic effect on the regression. Outliers that are not high leverage points may be retained in the data set without changing the regression equation greatly.) In considering the detection of outliers, Belsley, Kuh, and Welsch (1980) state that the use of recursive residuals appears to have greater power to detect departures from the null hypothesis than some other methods. These writers, however, consider the use of recursive residuals only when the data vary over time, while it is our contention that recursive residuals are useful for all data sets, and particularly for those where the data can be ordered in some natural way.

3. GRAPHICAL PLOTS USING RECURSIVE RESIDUALS

We discuss three plots of these recursive residuals.

3.1 The Normal Probability Plot of the Recursive Residuals

On the assumption that the errors are identically and independently distributed as $N(0, \sigma^2)$, the recursive residuals are also independently and identically distributed as $N(0, \sigma^2)$. Thus they may be plotted on normal probability paper, and the assumption of normality may be checked. If all assumptions of the model are satisfied, then the normal probability plot should show a straight line through the origin such as that in Figure 1a.

If the errors come from a heavy-tailed distribution, the normal probability plot will have an S shape (Fig. 1b), while if they come from a short-tailed distribution, the normal probability plot will have an inverted S shape (Fig. 1c). If the error terms have a skew distribution, the normal probability plot will be a combination of these plots (Fig. 1d).

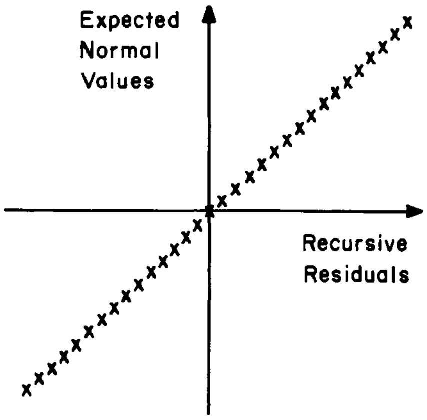

Figure 1a

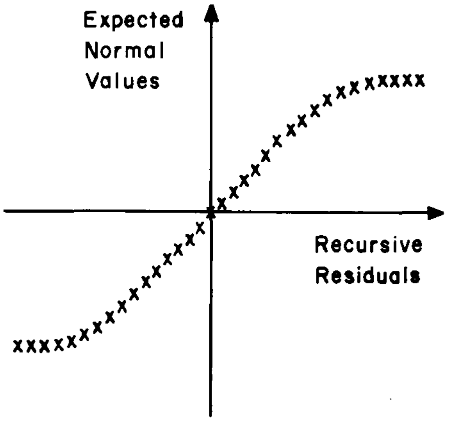

Figure 1b

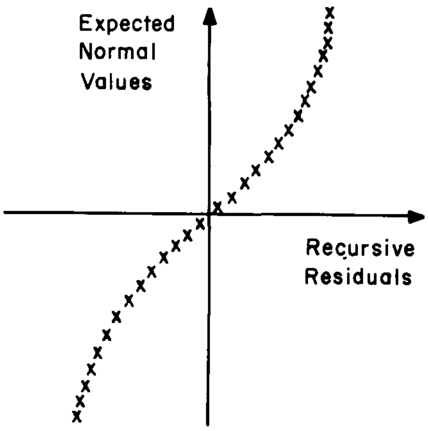

Figure 1c

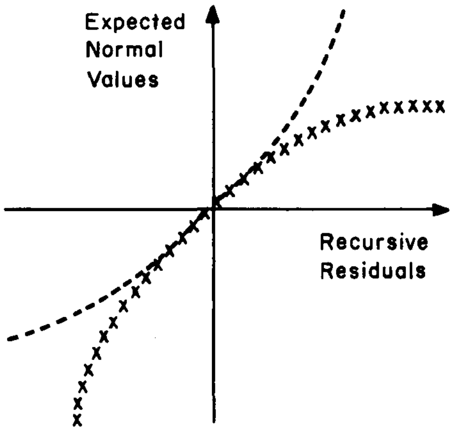

Figure 1d

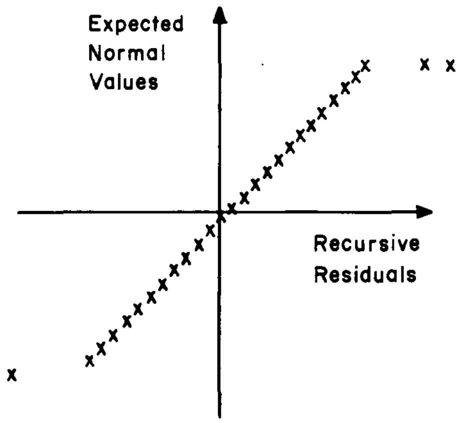

Figure 1e

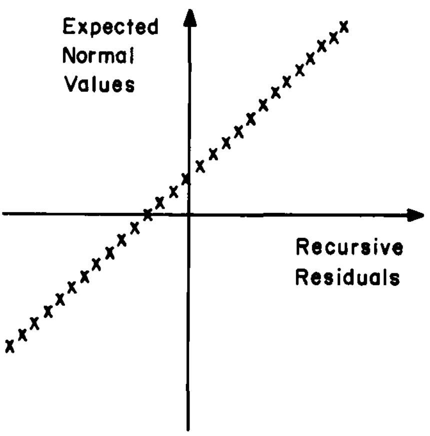

Figure 1f

If the normal probability plot is obtained by plotting the ordered recursive residuals against the order statistic medians from the $N(0, 1)$ distribution, the straightness of the line obtained can be measured using the normal probability plot correlation coefficient test proposed by Filliben (1975). Note that the recursive residuals should not be centered in the calculation of this correlation coefficient, since one also wishes to test whether the line passes through the origin.

It should be noted that several other model defects, such as heteroscedasticity, can mimic a heavy-tailed distribution, so that a nonstraight line on a normal probability plot does not necessarily imply nonnormality of the recursive residuals, and thus of the data.

The occurrence of outliers might also be considered as mimicking a heavy-tailed distribution. However, consider the effect on the recursive residuals of an outlier with a positive displacement. While included in the data set, this outlier will draw the regression function toward itself, and the recursive residuals will accordingly tend to be negative. Once the outlier has been removed from the data set, however, the recursive residuals will once again become $N(0, \sigma^2)$. Thus the pattern of the recursive residuals with a single positive outlier is for the initial recursive residuals to be biased towards negative values, for the recursive residual corresponding to the outlier to be large and positive, and for subsequent recursive residuals to be random.

If the positive outlier is among the base set, and so is never removed, then all the recursive residuals will have a negative bias.

The substance of these remarks applies also to multiple outliers, and mutatis mutandis to negative outliers.

Thus, if the outliers occur among the observations for which recursive residuals have been calculated, the normal probability plot will be most likely to show a straight line through the origin with a few points off the line at one (or both) ends of the line, and not a curved line (see Fig. 1e). If the outliers occur among the base points, they will cause a permanent distortion of the line, which will then not pass through the origin (see Fig. 1f). Such a straight line not passing through the origin shows that the mean of these recursive residuals is not zero, and can also indicate a model misfit, such as may be due to an omitted variable. In order to check whether this type of plot has been caused by outliers, the recursive residuals may be calculated using a different base (one that allows recursive residuals to be calculated for the points previously used as base), and the plot may be redrawn. The outliers (if any) should now be among the points for which recursive residuals have been calculated, and the plot should then pass through the origin. If it does not, the problem is unlikely to be due to outliers, but rather to an omitted term.

Thus a normal probability plot showing a more or less straight line with a few aberrant points is one indication of possible outliers. Since the recursive residuals correspond to actual observations, points on the plot that are suspect can be identified from a listing of the recursive residuals and checked for coding errors, measurement

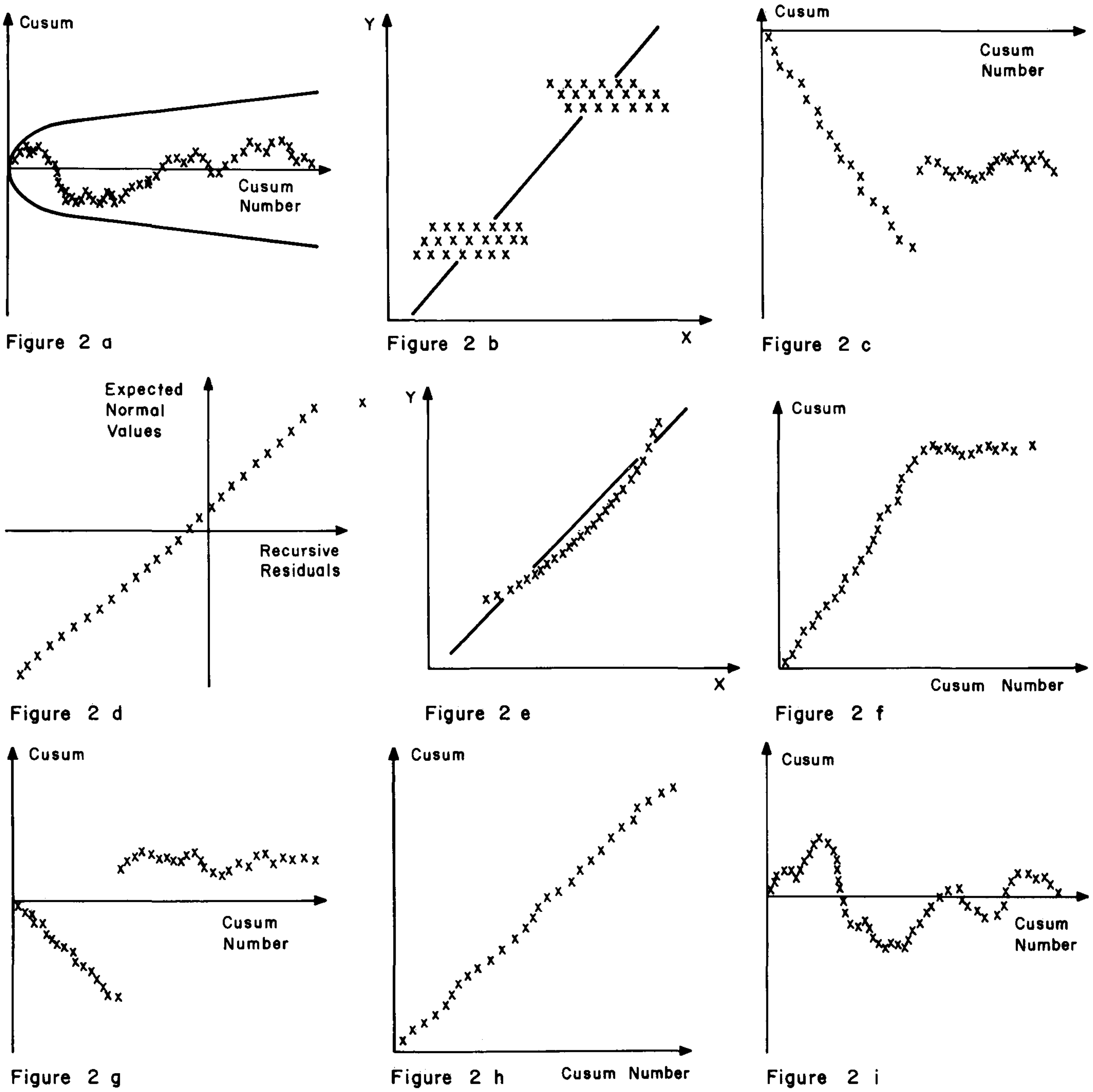

Figure 2 a

Figure 2 b

Figure 2 c

Figure 2 d

Figure 2 e

Figure 2 f

Figure 2 g

Figure 2 h

Figure 2 i

errors, or other possible causes of outliers. Such an indication of possible outliers should be followed up by formal outlier testing and possible deletion, for which optimal methods exist (Hawkins 1980, pp. 92–99). After the deletion of outliers, the other regression assumptions, must, of course, be rechecked.

We mentioned that there are several possible reasons for a curved normal probability plot. Fortunately, however, these defects reflect differently on the two cusum plots, which will now be discussed.

3.2 Plot of the Cusums of the Recursive Residuals

The plot of the cusums of the recursive residuals (standardized by the estimated standard deviation of the data) was first proposed by Brown, Durbin, and Evans (1975). The plot considered here is a plot against i of the ith cusum, defined by

$$\text{sum}_i = \sum_{j=1}^{i} w_j / sd, \qquad i = 1, \ldots, n-p,$$

where w_j is the $(n+1-j)$th recursive residual and is based on $(n-j)$ observations.

If all the regression assumptions are satisfied, this plot should show a random walk within a parabolic envelope about the origin, since the expectation of these recursive residuals is zero (see Fig. 2a). This plot should also show such a random walk when the only deviation is nonnormality. As mentioned by Brown, Durbin, and Evans, it is preferable to use the recursive residuals rather than the ordinary residuals to detect a change of model, since the recursive residuals behave exactly as under the null hypothesis, until a change occurs.

In the case of a change of regime of the regression, such as shown in the rather simplistic sketch in Fig. 2b, the cusum plot in Fig. 2c will show a sudden downsurge, since the deletion of points will result in a long string of negative recursive residuals. This will be followed by a jump in the cusum that corresponds to the large positive recursive residual obtained when the last point of the second set is omitted. The horizontal random segment indicates that the current line is correct, so that the mean of the recursive residuals is again zero. (In Fig. 2b it is assumed that the leftmost points are used as the base in the calculation of the recursive residuals and that the points are deleted from right to left, so that the rightmost point corresponds to the first recursive residual.) The normal probability plot corresponding to this cusum will be similar to that illustrated in Figure 2d, which shows the preponderance of negative recursive residuals, as well as the large positive recursive residual.

The effect of an omitted variable is more difficult to specify, as the variable may take many forms. If the omitted variable is a quadratic term such as is shown in Figure 2e, all but a few recursive residuals will be positive, since the slope of the line will steadily decrease as points are omitted. The recursive residuals will then have, in contrast to the ordinary residuals, a nonzero mean. While it may be possible to detect a quadratic trend in the plot of residuals versus predicted values, this trend may be rather confused, and it may easily be overlooked, particularly if the data are fairly noisy. The cusum plot of the recursive residuals will, in contrast, show a straight-line segment (followed by a short random segment), as in Figure 2f.

The effect of outliers will have a somewhat different plot. Following the discussion in Section 3.1, if the outliers are among the calculated recursive residuals, the cusum plot (Fig. 2g) will show a sudden blip (corresponding to the large recursive residual), preceded by a downward drift. (The corresponding normal probability plot will show a straight line through the origin with a few points off the line, as is shown in Fig. 1e.) If the outlier is among the base observations, the cusum will show merely a nonhorizontal straight line segment (see Fig. 2h), indicating that the mean of the recursive residuals is nonzero. (This cusum plot corresponds to a normal probability plot that does not pass through the origin.)

The effect of a change of scale of the regression over the observations will be merely to increase (or decrease) the variance of the random variation (see Fig. 2i) and may be difficult to detect. Such a change of scale is detected more efficiently by a second cusum plot.

3.3 Plot of the Cusums of the Square Roots of the Absolute Values of the Standardized Recursive Residuals

It has been shown by Hawkins (1981) that if $X \sim N(0,\sigma^2)$, $Y = (|X|/\sigma)^{1/2}$ is distributed very nearly normally with mean .82218 and standard deviation .34914, and that the distribution is comparatively robust for heavy-tailed departures from normality. Since the recursive residuals are distributed as $N(0,\sigma^2)$ on the null hypothesis of no change of scale over the observations, the cusums

$$\sum_{j=1}^{i} \{(|w_j|/sd)^{1/2} - .82218\}/.34914$$

may be used to check this null hypothesis. In the situation where all the model assumptions are satisfied, this plot should show a random walk about the X-axis (see Fig. 3a). Such a random walk will also be evident in the case of nonnormality.

It may be noted that these cusum plots form a bridge-like process. The terminal point of this second cusum plot also provides some evidence about the tail weight of the distribution of the ϵ_i. Defining $z_i = |w_i/\sigma|^{1/2}$ we find that, if the errors are normally distributed, the z_i are approximately distributed as $N(.82218, (.34914)^2)$. If the ϵ_i follow a heavy-tailed distribution, so does z_i, and likewise for light-tailed distributions. Now

$$sd^2 = \sigma^2 \sum_{i=1}^{n-p} z_i^4/(n-p),$$

the fourth sample moment of z_i. This has an expectation that increases with the kurtosis of the z_i. Thus a heavy-tailed distribution of the ϵ_i leads to the normalization of $|w_i|$ by a factor that is too large, and hence causes a slow downward drift in the cusum. Conversely, a light-tailed error distribution will lead to a slow upward drift.

If there has been a change of scale over the observations, this plot will show some straight-line segments, for example a decreasing trend followed by an upsurge (see Fig. 3b) if the variance increases over the observations (and the first p data points are used as base). It should be noted that a rise on this plot corresponds to a recursive residual of magnitude greater than .676sd, while a dip corresponds to a recursive residual less than .676sd. Although no comments are made in Hawkins

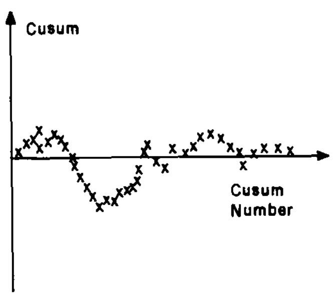

Figure 3 a

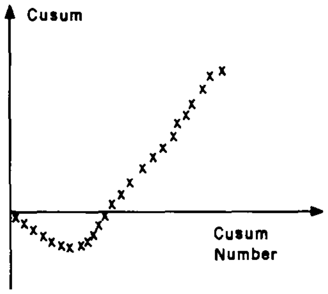

Figure 3 b

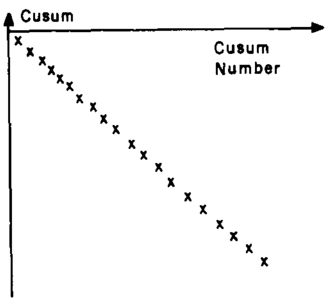

Figure 3 c

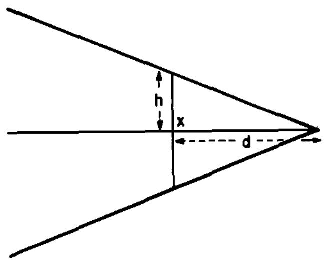

Figure 3 d

(1981) as to the effect of a nonzero mean of X on the distribution of Y, it can be shown that the distribution remains close to its nominal $N(0, 1)$ for small changes in the mean. This plot may also show an upsurge or downsurge in the case of a change of regime.

In the case of outliers, this plot might show merely a straight-line segment, as in Figure 3c, or an upsurge or downsurge. These problems should, however, be diagnosed using the first cusum plot, and not the second.

Thus a more or less random cusum plot, together with a cusum of square roots plot such as is shown in Figure 3b, is an indication of possible change of scale. This case is almost indistinguishable from the case of heteroscedasticity. (In the case of a change of scale, the plot will show straight-line segments, while in the case of heteroscedasticity the plot will show parabolic segments.)

3.4 Automation and Testing of the Recursive Residual Plots

The pattern on this plot (and on the cusum plot discussed above) can be tested for significance by means of a cusum mask. This mask has the form shown in Figure 3d. Hawkins (1981) recommended that a mask with $h = 6$ and $d = 24$ be used. For further references on cusum masks see Lucas (1976) or van Dobben de Bruyn (1968). For computational purposes it may be preferable to use this mathematically equivalent scheme: Define $s^+(0) = s^-(0) = 0$, $k = h/d$,

$$s^+(t) = \max(0, s^+(t-1) + w_j/sd - k),$$

$$s^-(t) = \min(0, s^-(t-1) + w_j/sd + k).$$

Trigger a signal if $s^+ > h$ or $s^- < -h$. This formulation leads to an easy automation of the cusum procedures as part of the standard regression software. The cusums may then be tested automatically and printed out for the user's attention *only* if significant, a possibility that avoids burdening the user with masses of uninteresting printout.

In consequence of this, it becomes possible to investigate all $2(p + 1)$ possibilities of ordering the data by each predictor and by the predicted value, and using both forward and backward deletion, with the computation and testing being invisible to the user except where a model misfit is indicated.

A similar remark applies to the probability plot, which need not be output unless the automated test indicates a departure from model.

4. EFFECTS OF VARIOUS MODEL DEFECTS ON THE THREE PROPOSED PLOTS

As we have seen, various model defects lead to specific patterns on the three plots. These are summarized in Table 1. A question mark indicates that there are several possible patterns.

In the case of omitted variables, the normal probability plot may either not pass through the origin or be curved, depending on which variable has been omitted. If the omitted variable is a squared (or higher-order) term in one of the variables in use, the cusum plot will show one (or more) straight-line segments, ending in a random sequence. The cusum of square roots will show a rise, followed by a dip (or several rises and dips). The plots for omitted variables may be similar to those for nonincluded outliers, but the cause of the problem can readily be ascertained if the recursive residuals are calculated twice, first using as base the first p observations, and then using the last p observations (assuming that they do not overlap). If both normal probability plots do not pass through the origin, then the problem is not outliers but an omitted variable. It is thus very important to eliminate outliers before checking the other model assumptions. (Note, however, that outlier testing assumes that the model under consideration is the correct one.) Harvey and Collier (1977) point out that the effectiveness of the recursive residuals to detect a misspecification of one of the variables is increased if the data are ordered on that variable. They also note that the effectiveness is increased if the base for the recursive residuals consists of the smallest $p/2$ and the largest $p/2$ data points (with respect to the variable under discussion).

Table 1. Summary of Model Deficiencies and the Typical Resultant Appearance of the Three Proposed Plots

	Normal Probability Plot	Cusum 1	Cusum 2
Nonnormality	1b–1d	2a	3a
Outliers ("In")	1e	2g	3b
("Out")	1f	2h	?
Change of Regime	2d	2c	3b
Heteroscedasticity	1b	2i	3b

Since the model defect may be related to one of the variables, some thought should be given to the ordering of the data set. The maximum benefit will be derived from the recursive residual plots if the data are in some natural order. If there are several possible natural orderings for the data, it may be beneficial to examine the recursive residual plots obtained for each ordering. This can be done very easily using the automatic testing procedures described in Section 3.4.

Another question of interest is of course whether the error terms are independent, or whether they are serially correlated. This can be checked by calculating the Durbin-Watson statistic, using the recursive residuals. Because of the distributional properties of the recursive residuals, the tail area of the Durbin-Watson statistic can be obtained exactly. (The standard distribution of the Neumann ratio as given by Hart 1942 is applicable.) Phillips and Harvey (1974) also recommend a split base for increasing the power of this test. However, Harvey and Phillips (1974) note that the effectiveness of the recursive residuals is maximized, when testing for heteroscedasticity, by taking the base as the central p observations.

In the more challenging case of multiple departures from the model, these plots also form a useful aid in detecting possible problems.

5. PRACTICAL EXAMPLE

We turn now to a practical example. These data concern the transportation of sand along a pipe, using various sizes of sand particles, two gradients for the pipe, various flow rates of the water carrying the sand, and four different heights of obstructions (placed in the pipe to simulate roughness). (Data obtained courtesy of the National Building Research Institute of the CSIR.) There were 88 data points, and the four predictors described above (plus an intercept term), in addition to all first order products, were used, so that $p = 11$. In an initial fit, the plot of residuals versus predicted values suggested that a square root transformation on the dependent variable (rate of transportation of the sand) was needed to obtain homoscedasticity. The data (before transformation) are listed in Table 2.

The regression model obtained was highly significant, with an F value of 437.56. The squared multiple correlation was .983. The regression equation appeared to be satisfactory. The plot of residuals versus predicted values seemed satisfactory, except for an indication of a possible outlier. This plot is shown in Figure 4a. Plots of the residuals versus the predictor variables sand class, flowrate, and height of obstruction (Figs. 4b–4d) did not show anything further. The time sequence plot (see Fig. 4e) also seemed acceptable, except for the outlier. A cusum plot of the residuals (Fig. 4f) did show some sort of model misfit, in that the test was significant, but it did not indicate what the problem was.

The recursive residuals for this regression were calculated using the initial observations as base, so that the first recursive residual calculated was that for observation 88. Owing to the design of the data set, several observations were encountered for which recursive residuals could not be calculated, since their deletion would have resulted in the design matrix becoming singular. These observations were retained in the base. (The base thus consisted of observations 1, 2, 4, 5, 15, 16, 18, 53, 54, 56, and 63.) The recursive residuals are listed in Table 3. The normal probability plot and the two cusums appear in Figures 4g–4i.

The normal probability plot showed an approximately straight line, passing approximately through the origin (as can be seen in Fig. 4g). The slight curvature towards the horizontal in the lower left-hand corner of the plot may indicate a slightly heavy tail, but the curvature is so slight as to be no cause for concern. The 5% critical value of the normal probability plot correlation coefficient is given by Filliben as .983 for $n = 75$ and .984 for $n = 80$, so that this correlation coefficient (.993, based on 77 observations) is not significant at the 5%

Table 2. Raw Data for Example Described in Section 5

	Gradient													
	.0067										.0040			
Data Point	2	3	4	5	Data Point	2	3	4	5	Data Point	2	3	4	5
1	600	1.64	.0	7.3	37	1.537	5.02	35.0	13.0	53	.600	3.38	.0	8.8
2	600	1.64	15.0	6.2	38	1.537	6.87	.0	32.0	54	.600	3.38	15.0	6.9
3	600	1.64	25.0	4.4	39	1.537	6.87	15.0	26.4	55	.600	3.38	25.0	5.8
4	600	3.56	.0	18.2	40	1.537	6.87	25.0	24.6	56	.600	5.07	.0	15.6
5	600	3.56	15.0	15.0	41	1.537	6.87	35.0	20.5	57	.600	5.07	15.0	12.4
6	600	3.56	25.0	12.2	42	2.360	1.64	.0	4.7	58	.600	5.07	25.0	10.4
7	600	3.56	35.0	10.2	43	2.360	1.64	15.0	3.0	59	.600	6.64	.0	25.1
8	600	5.02	.0	30.0	44	2.360	3.56	.0	11.7	60	.600	6.64	15.0	20.8
9	600	5.02	15.0	25.2	45	2.360	3.56	15.0	9.4	61	.600	6.64	25.0	17.8
10	600	5.02	25.0	19.7	46	2.360	5.02	.0	17.6	62	.600	6.64	35.0	16.0
11	600	5.02	35.0	17.2	47	2.360	5.02	15.0	14.8	63	1.180	3.38	.0	8.5
12	600	6.87	.0	53.3	48	2.360	5.02	25.0	13.4	64	1.180	3.38	15.0	7.6
13	600	6.87	15.0	40.0	49	2.360	6.87	.0	25.7	65	1.180	3.38	25.0	5.0
14	600	6.87	25.0	31.3	50	2.360	6.87	15.0	23.8	66	1.180	5.07	.0	15.1
15	1.180	1.64	.0	6.9	51	2.360	6.87	25.0	21.4	67	1.180	5.07	15.0	12.1
16	1.180	1.64	15.0	6.1	52	2.360	6.87	35.0	18.5	68	1.180	5.07	25.0	9.9
17	1.180	1.64	25.0	4.8						69	1.180	6.64	.0	22.0
18	1.180	3.56	.0	15.2						70	1.180	6.64	15.0	19.3
19	1.180	3.56	15.0	13.2						71	1.180	6.64	25.0	15.6
20	1.180	3.56	25.0	11.3						72	1.180	6.64	35.0	14.2
21	1.180	3.56	35.0	8.9						73	1.537	3.38	.0	8.7
22	1.180	5.02	.0	22.5						74	1.537	3.38	15.0	5.7
23	1.180	5.02	15.0	19.4						75	1.537	5.07	.0	13.7
24	1.180	5.02	25.0	17.0						76	1.537	5.07	15.0	10.9
25	1.180	5.02	35.0	14.5						77	1.537	6.64	.0	22.3
26	1.180	6.87	.0	35.0						78	1.537	6.64	15.0	16.7
27	1.180	6.87	15.0	31.2						79	1.537	6.64	25.0	14.6
28	1.180	6.87	25.0	27.4						80	1.537	6.64	35.0	10.0
29	1.180	6.87	35.0	23.9						81	2.360	3.38	.0	5.2
30	1.537	1.64	.0	6.7						82	2.360	3.38	15.0	4.4
31	1.537	1.64	15.0	5.1						83	2.360	5.07	.0	10.5
32	1.537	3.56	.0	14.3						84	2.360	5.07	15.0	8.5
33	1.537	3.56	15.0	11.5						85	2.360	6.64	.0	17.3
34	1.537	5.02	.0	21.2						86	2.360	6.64	15.0	11.8
35	1.537	5.02	15.0	16.4						87	2.360	6.64	25.0	9.1
36	1.537	5.02	25.0	14.9						88	2.360	6.64	35.0	8.3

NOTE: Variables are: 2=sand class; 3=flow rate; 4=height of obstruction; 5=rate of transportation. Other variables were: 6=gradient•sand class; 7=gradient•flow rate; 8=gradient•height; 9=sand class•flow rate; 10=sand class•height; 11=flow rate•height.

level. Thus the curvature may be ignored.

Let us now consider the first cusum plot. The leftmost point on the plot is the standardized recursive residual based on $n - 1$ observations (i.e., the recursive residual for observation 88). The next point is the sum of the standardized recursive residuals for observations 87 and 88. The first 52 observations of the data set were observations for the gradient .0067, while the other 36 observations were for the gradient of .004. The split between these two sets occurs on this plot between 32 and 33 (on the horizontal axis). The plot shows that the regression line for the two gradients is not the same: The first part of the plot has an underlying negative trend, while the second half has a positive trend followed by a random segment. Careful inspection of this plot and a listing of recursive residuals (shown in Table 3) revealed that all except one of the recursive residuals associated with the lowest flowrate were negative. Twelve of the 15 recursive residuals for observations for the highest flowrate were positive. It thus appears that the flowrates have not been modeled correctly. Since the flowrates differ for the two gradients, and since model misfit is apparent for both flowrates and gradient, it appears that the gradient-flowrate interaction term has also not been modeled correctly. The numerical cusum test triggered a signal on the 45th observation.

The second cusum plot shows an upward trend followed by a downward trend over the first 32 sums, indicating that the standard deviations for the points at gradient .004 vary. The second part of the plot shows that the recursive residuals for gradient .0067 are also variable. The variability appears to be related to the variable sand class. The numerical cusum test triggered a signal on the 40th observation.

No outliers are evident from this normal probability plot. Use of an outlier detection routine (based on the stepwise procedure described in Hawkins 1980, p. 65) showed, however, that the observation that gave a large (ordinary) residual (observation 12) was an outlier. The plots after deletion of this outlier are shown in Figures 4j–4n. The plot of residuals versus predicted values (Fig. 4j) and the time sequence plot (plot of residual versus order of entry, Fig. 4k) appear to be satisfactory. The normal probability plot again shows an approximately straight line passing approximately through the origin, with a slight upward curve in the third quadrant. The two cusum plots (Figs. 4m and 4n) are again similar to the two previous plots (Figs. 4h and 4i), with signals being triggered on the 40th and 17th observations, respectively.

These plots show that the removal of the outlier did not improve the fit of the equation. Thus, although the equation obtained is highly significant, with a high R^2, low residual sum of squares, and highly significant F value, and the plots of the (ordinary) residuals seem satisfactory, the analysis of the recursive residual plots shows that the equation does not fit the data. In the absence of the diagnostic information given by the cusum plot, one might try to remedy this problem by stepwise selection from higher-order polynomial terms

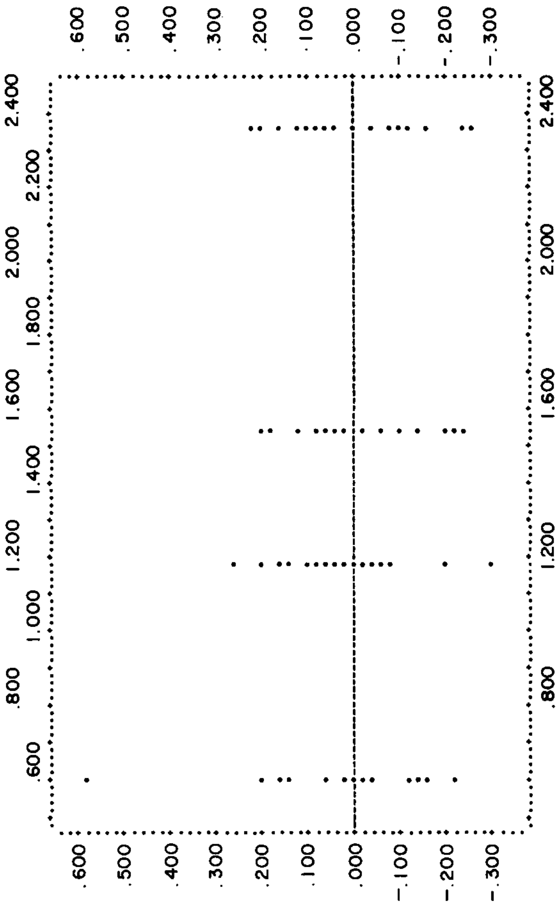

Figure 4b. Plot of residuals (vertical axis) versus sand size (horizontal axis).

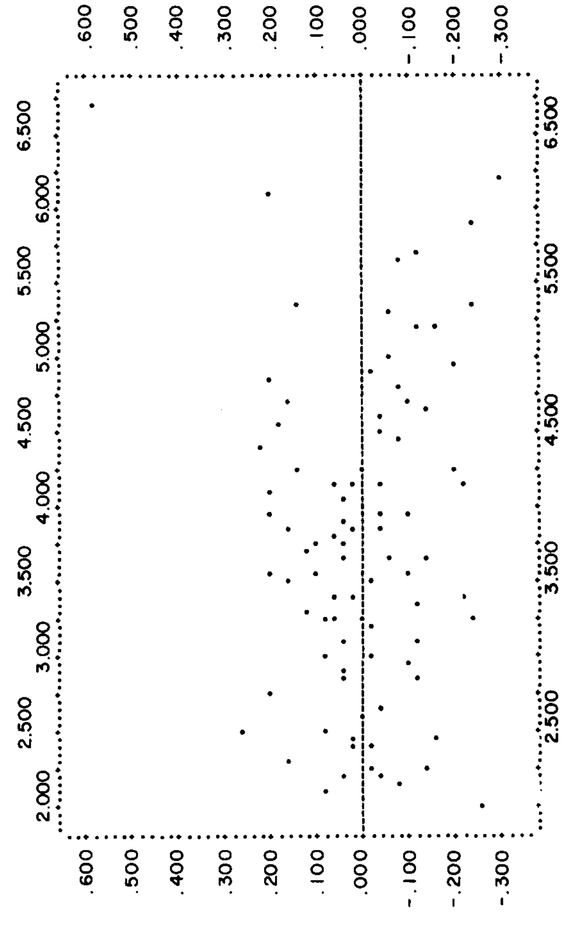

Figure 4a. Plot of residuals (vertical axis) versus predicted (horizontal axis).

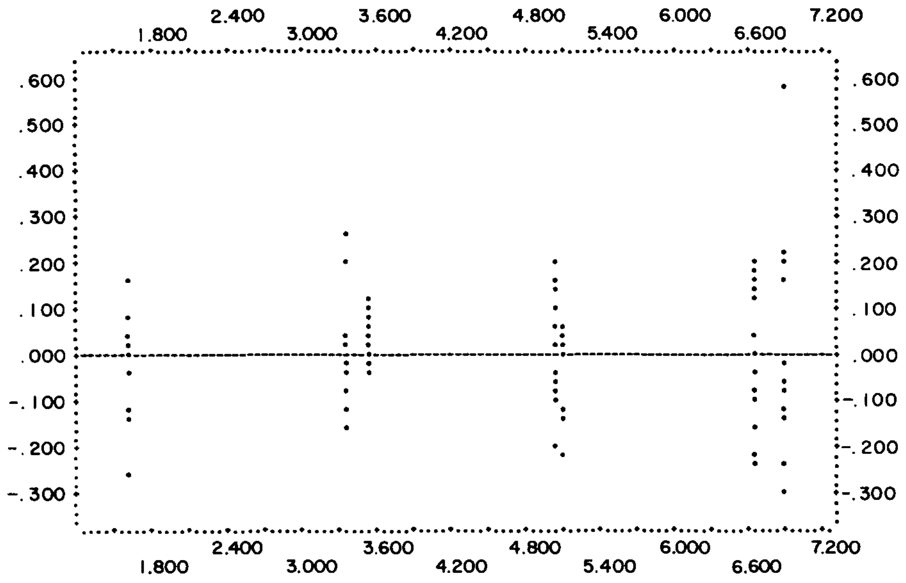

Figure 4c. Plot of residuals (vertical axis) versus flow rate (horizontal axis).

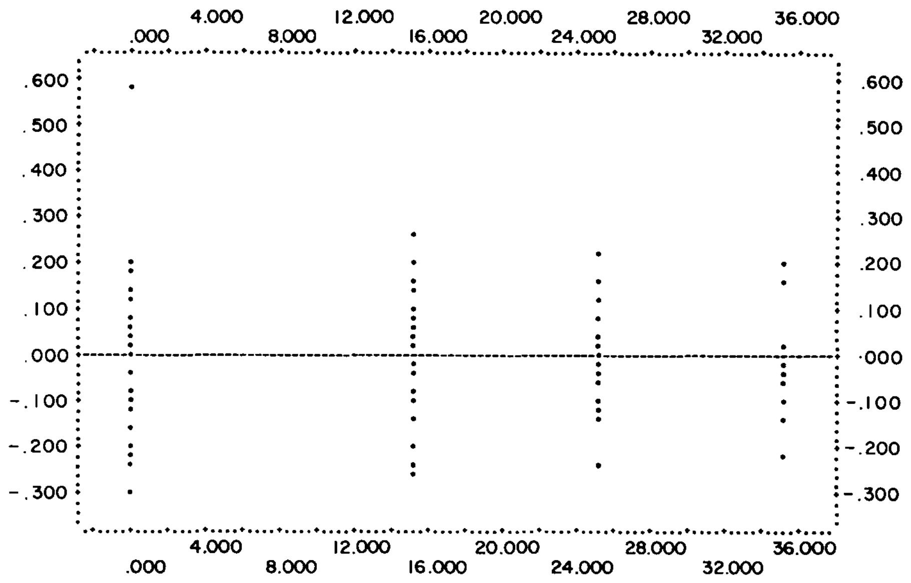

Figure 4d. Plot of residuals (vertical axis) versus height (horizontal axis).

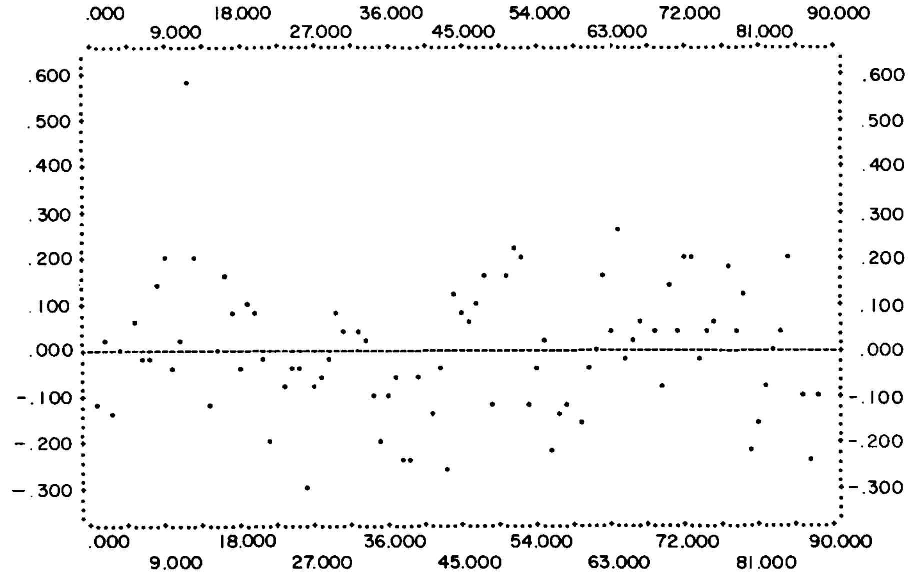

Figure 4e. Time sequence plot: Residuals (vertical axis) versus order of input (horizontal axis). Durbin-Watson test statistic is 1.322.

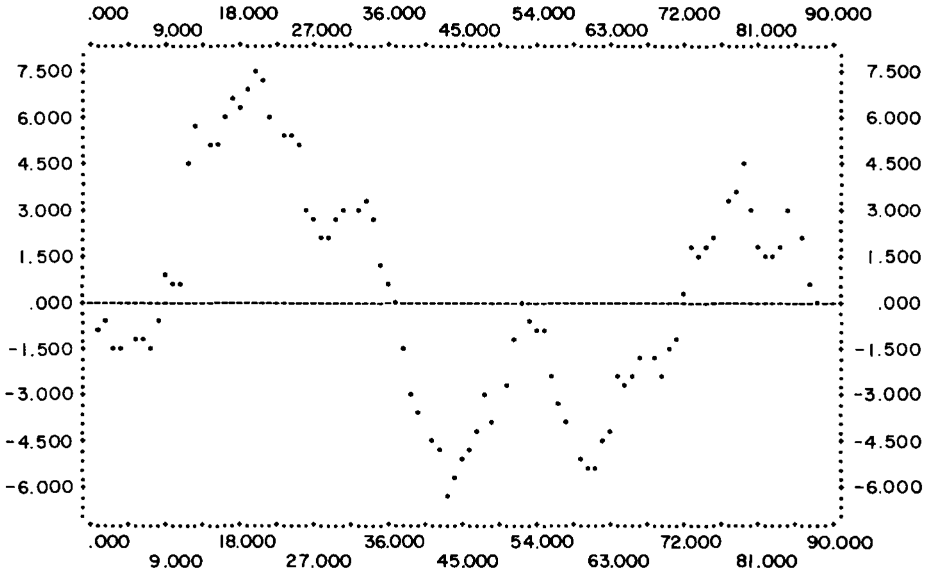

Figure 4f. Plot of cusums of residuals/sd (vertical axis) versus cusum number (horizontal axis).

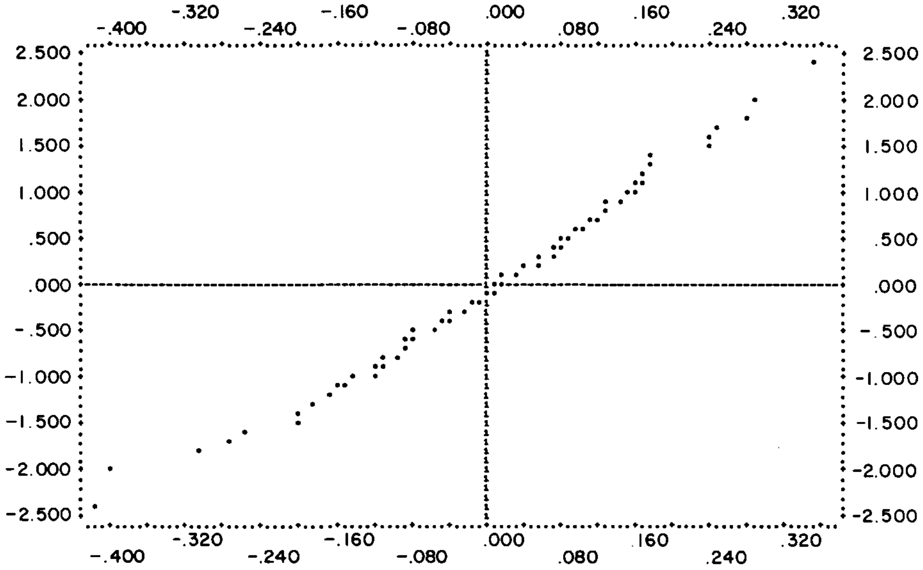

Figure 4g. Normal probability plot of recursive residuals: Standard normal ordinate (vertical axis) versus recursive residuals (horizontal axis). Recursive residuals calculated backwards. Normal probability plot correlation coefficient is .992555.

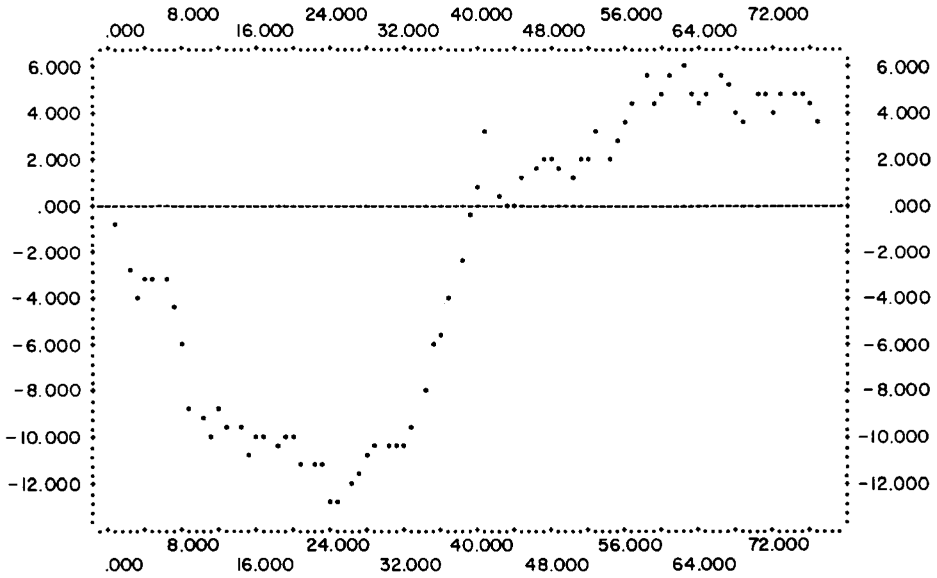

Figure 4h. Plot of cusums of recursive residuals/sd (vertical axis) versus cusum number (horizontal axis). Recursive residuals calculated backwards.

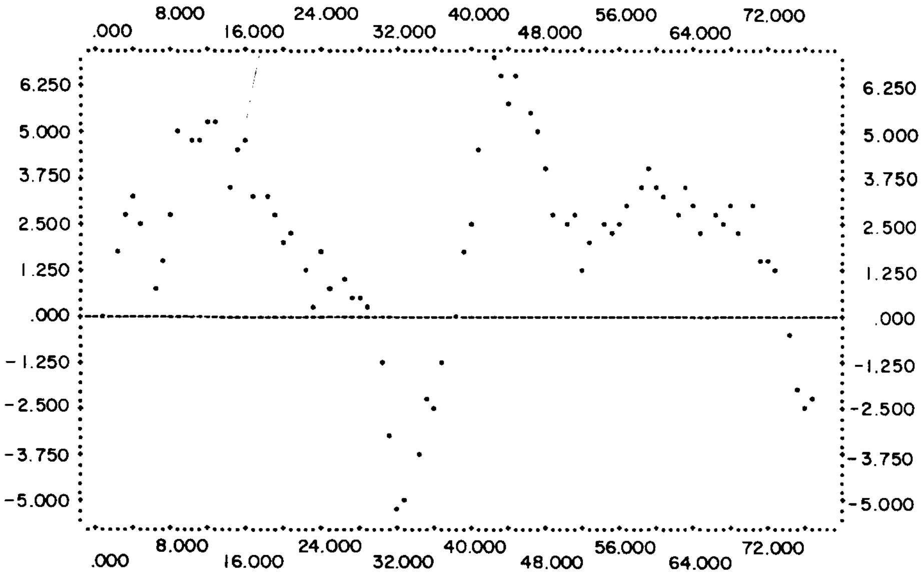

Figure 4i. Plot of cusums of ((rec. res/sd)$^{1/2}$ – .82218)/.34914 (vertical axis) versus cusum number (horizontal axis). Recursive residuals calculated backwards.

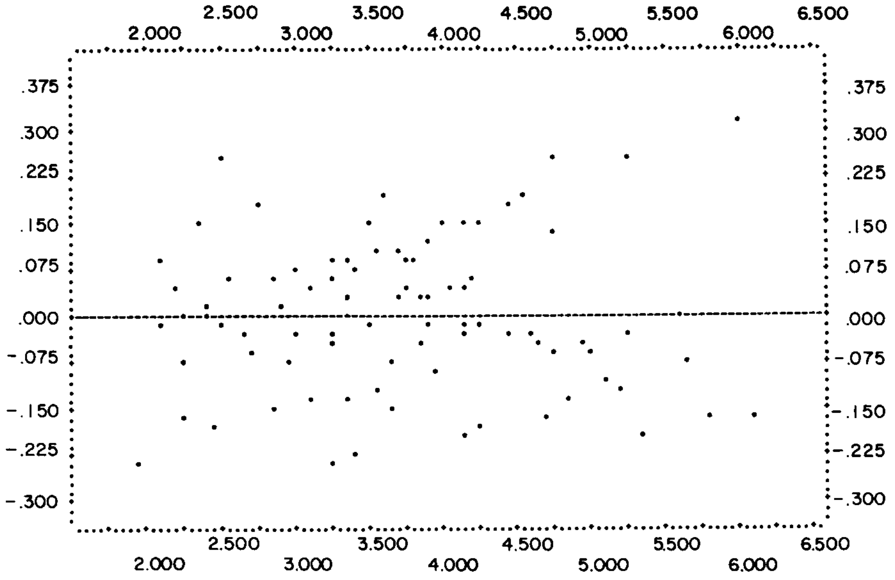

Figure 4j. Plot of residuals (vertical axis) versus predicted values (horizontal axis). One significant outlier has been deleted.

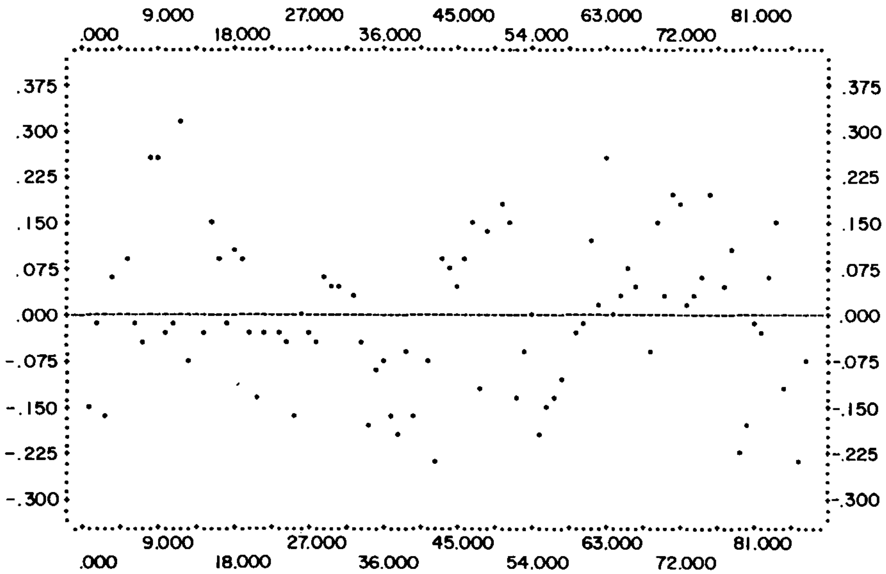

Figure 4k. Time sequence plot: Residuals (vertical axis) versus order of input (horizontal axis). One significant outlier has been deleted. Durbin-Watson test statistic is 1.379.

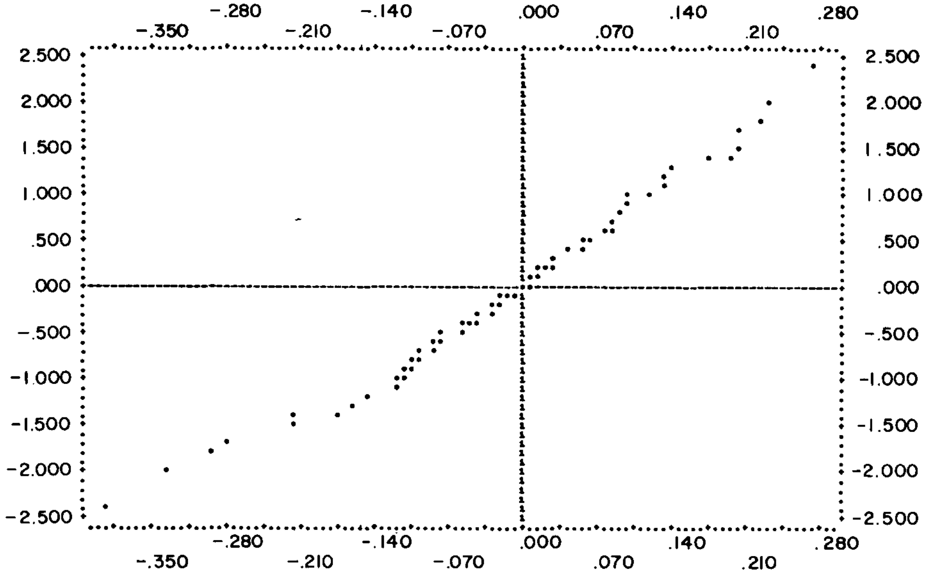

Figure 4l. Normal probability plot of recursive residuals: Standard normal ordinate (vertical axis) versus recursive residuals (horizontal axis). Recursive residuals calculated backwards. One significant outlier has been deleted. Normal probability plot correlation coefficient is .989627.

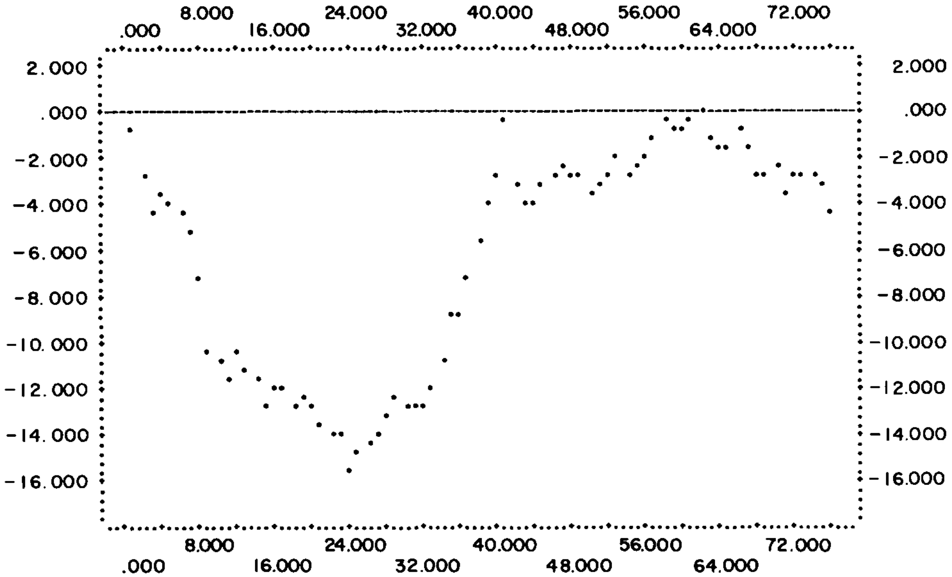

Figure 4m. Plot of cusums of recursive residuals/sd (vertical axis) versus cusum number (horizontal axis). Recursive residuals have been calculated backwards. One significant outlier has been deleted.

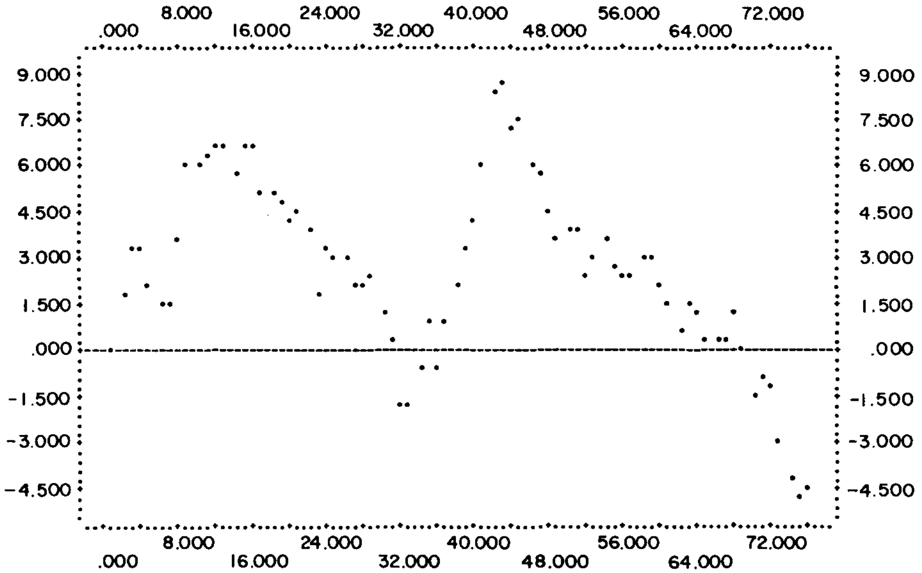

Figure 4n. Plot of cusums of ((rec. res./sd)$^{1/2}$ − .88218)/.34914 (vertical axis) versus cusum number (horizontal axis). One significant outlier has been deleted.

Table 3. Recursive Residuals (Calculated Backwards)

3	−.122218	47	.239797
6	−.057491	48	.236850
7	−.012806	49	.079838
8	.005653	50	.279080
9	.081653	51	.244244
10	−.113585	52	.125603
11	.014828	55	.003877
12	.165242	57	−.002122
13	−.044128	58	−.009334
14	−.161401	59	.091294
17	−.084793	60	.103812
19	.129665	61	.059837
20	.054977	62	.117169
21	−.054657	64	.030269
22	−.180447	65	−.272995
23	.059745	66	.031210
24	.090203	67	−.041778
25	.069877	68	−.123446
26	−.147688	69	−.039810
27	.151479	70	.075516
28	.142443	71	−.088766
29	.115407	72	.018940
30	.093407	73	.127335
31	−.152695	74	−.196343
32	.176315	75	−.008918
33	.013440	76	−.085904
34	.115276	77	.157378
35	−.086367	78	−.094931
36	−.022957	79	−.077953
37	−.024563	80	−.399645
38	.071299	81	−.258456
39	.037985	82	−.168631
40	.170738	83	.006582
41	.037449	84	−.039360
42	−.079537	85	.157269
43	−.412337	86	−.200767
44	.355871	87	−.305601
45	.174619	88	−.110219
46	.289084		

involving the predictors. This approach, however, does not improve the cusum, while the form of the cusum shows that when the data pertain to only one gradient, the model fit is satisfactory, but that when data for both gradients are used, the model undervalues for the two lower levels of sand class (.6 and 1.18) and overvalues for the two higher levels (1.537 and 2.36). The regression function must thus be modified in a way that corrects this deficiency. This suggests an attempt to model the effects of gradient (G) and sand class (S) by $\exp(aG - bS)$. This greatly improved the cusum, which still, however, remained significant. Further model refinement steps were therefore made, leading to the final model

$$\begin{aligned} R = {} & -10.4413 + 17.801 S^{.93} F^{.468} \exp(164.3971G \\ & - 2.5517S - .00665H + .0723F^{1.25} + .039S^2) \\ & + 10.889S - 3.705S^2 + .0044SQH - .0266FH \\ & + .1642S^2F + .114SF^2 - .06775S^2F^2, \end{aligned}$$

where F = flow rate, H = height, and R = rate of transportation. This model was received well by the user, who had previously settled on a 25-parameter model.

[*Received October 1981. Revised November 1983.*]

REFERENCES

ANDREWS, D.F., and PREGIBON, D. (1978), "Finding the Outliers That Matter," *Journal of the Royal Statistical Society*, Ser. B, 40, 85–93.

BARTLETT, M.S. (1951), "An Inverse Matrix Adjustment Arising in Discriminant Analysis," *Annals of Mathematical Statistics*, 22, 107–111.

BELSLEY, D.A., KUH, E., and WELSCH, R.E. (1980), *Regression Diagnostics*, New York: John Wiley.

BROWN, R.L., DURBIN, J., and EVANS, J.M. (1975), "Techniques for Testing the Constancy of Regression Relationships Over Time" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 37, 149–192.

COOK, R.D. (1977), "Detection of Influential Observation in Linear Regression," *Technometrics*, 19, 15–18.

——— (1979), "Influential Observations in Linear Regression," *Journal of the American Statistical Association*, 74, 169–174.

COOK, R.D., and WEISBERG, S. (1982), *Residuals and Influence in Regression*, London: Chapman and Hall.

FILLIBEN, J.J. (1975), "The Probability Plot Correlation Coefficient Test for Normality," *Technometrics*, 17, 111–117.

HART, B.I. (1942), "Tabulation of the Probabilities for the Ratio of the Mean Square Successive Differences to the Variance," *Annals of Mathematical Statistics*, 13, 207–214.

HARVEY, A.C., and COLLIER, P. (1977), "Testing for Functional Misspecification in Regression Analysis," *Journal of Econometrics*, 2, 103–119.

HARVEY, A.C., and PHILLIPS, G.D.A. (1974), "A Comparison of the Power of Some Tests for Heteroskedasticity in the General Linear Model," *Journal of Econometrics*, 2, 307–316.

HAWKINS, D.M. (1980), *Identification of Outliers*, London: Chapman and Hall.

——— (1981), "A CUSUM for a Scale Parameter," *Journal of Quality Technology*, 13, 228–231.

HOAGLIN, D.C., and WELSCH, R. (1978), "The Hat Matrix in Regression and ANOVA," *The American Statistician*, 32, 17–22.

LUCAS, J.M. (1976), "The Design and Use of V-Mask Control Schemes," *Journal of Quality Technology*, 8, 1–12.

PHILLIPS, G.D.A., and HARVEY, A.C. (1974), "A Simple Test for Serial Correlation in Regression Analysis," *Journal of the American Statistical Association*, 69, 935–939.

PLACKETT, R.L. (1950), "Some Theorems in Least Squares," *Biometrika*, 37, 149–157.

RIDDELL, W.C. (1980), "Estimating Switching Regressions: A Computational Note," *Journal of Statistical Computation and Simulation*, 10, 95–101.

THEIL, H. (1965), "The Analysis of Disturbances in Regression Analysis," *Journal of the American Statistical Association*, 60, 1067–1079.

VAN DOBBEN DE BRUYN, C.S. (1968), *Cumulative Sum Tests: Theory and Practice*, London: Griffin's Statistical Monographs and Courses.