Using Simulated Envelopes in the Evaluation of Normal Probability Plots of Regression Residuals

Virginia F. Flack
Division of Biostatistics
School of Public Health
University of California
Los Angeles, CA 90024-1772

Rafael A. Flores
Institute of Nutrition of
Central America and Panama
Guatemala City, Guatemala

We investigate properties of a diagnostic-envelope method for evaluating normal probability plots of regression residuals that was proposed by Atkinson (1981), implemented by BMDP (Hardwick 1987), and extended to logistic regression by Landwehr, Pregibon, and Shoemaker (1984). The envelope's stability properties and joint residual vector-inclusion levels, undocumented so far in the literature, are explored here with several examples. Alternative resistant techniques for creating envelopes are considered; interpretations that can be derived from these plots are discussed. A resistant version of the envelopes shows good stability and good sensitivity to outlying residuals; both full-normal and half-normal probability plots with this envelope method provide useful information to the analyst.

KEY WORDS: Diagnostics; Graphical method; Outlier; Residual correlations; Resistant rule.

1. INTRODUCTION

In usual least squares multiple-regression analysis, examination of the resulting residual vector when Y is regressed on X is a standard technique for detecting grossly deviant observations; often a normal probability plot of the residuals is used. The evaluation of normal probability plots is a subjective process; even experienced statisticians may have serious disagreements in the identification of data anomalies when using a probability plot as an analysis tool.

Normal plots can be used for detection of one or more bad data values and non-Gaussian error distributions (Daniel 1959). The extent to which these problems can be detected depends on the data set and the analyst's interpretive skills. Daniel (1959) and Daniel and Wood (1980) gave training plots of normal data for various sample sizes to demonstrate the "usual" amount of deviation from linearity in a Gaussian probability plot. Regression residuals have a correlation structure that depends on the spatial arrangement of the observed independent-variables vectors, however; this data-dependent correlation will influence the form of the deviations from linearity and, therefore, should be used to guide the statistician's corresponding probability-plot interpretation.

We have the regression model $Y = X\beta + \varepsilon$, with $Y(n \times 1)$, $X(n \times p)$, and $\varepsilon(n \times 1)$. The error terms ε_i $(i = 1, \ldots, n)$ are assumed to be iid according to $N(0, \sigma^2)$. The externally standardized ith residual is

$$r_i^* = (y_i - \hat{y}_i)/s_{(i)}\sqrt{1 - v_i},$$

where y_i is the ith observed y value, $\hat{y}_i$ is its ordinary least squares estimator, $s_{(i)}$ is the estimator of σ when the ith observation has been omitted in the estimation, and v_i is the ith diagonal element of the hat matrix, $X(X'X)^{-1}X'$.

The idea of using simulation as an aid to the interpretation of a particular residual plot is to give the statistician a basis for comparison of the observed plot with an *expected* plot, in which that comparison is derived from Gaussian-distributed residuals (or residuals from an appropriately chosen error distribution). The capacity for generating a large number of Gaussian-based plots gives the potential for many ways to compare these many plots to the one real-data plot. The choice of a measure that adequately summarizes the information available from the simulated residual plots should be made to highlight features that a trained statistician would identify as being unusual when viewing the normal probability plot.

In Section 2 of this article, we describe Atkinson's proposal for generating diagnostic normal probability-plot envelopes and also give a robust alternative method. In Section 3, we describe the simulation methods used for this study. Section 4 gives joint residual vector-inclusion rates, envelope-variability measures, and descriptive comparisons of the bounding methods for four data sets from the literature. A general discussion of the results and concluding remarks are given in Section 5.

2. SUMMARIZING SIMULATED RESIDUAL VECTORS

2.1 Atkinson's Envelopes

Atkinson (1981) introduced a simulation-based method for placing envelopes on the plot, thus giving the data analyst a yardstick for comparison when considering a plot. Atkinson used half-normal probability plots; we will use both full-normal and half-normal probability plots for this investigation. We call all of the reference envelopes discussed here *diagnostic envelopes* rather than using the *confidence envelopes* terminology that has been applied to this type of envelope in other settings.

Atkinson's prescription for establishing a diagnostic envelope for a given data set was to do so via simulation. First generate $M(n \times 1)$ vectors Z from the $N(0, I)$ distribution. Second, for each of these vectors, fit the model $Z = X\beta + \varepsilon$ to obtain M simulated vectors $r^*_{:}$. Third, order the elements of each simulated residual vector, and fourth, for each $i = 1, \ldots, n$, select $l_i = \min r^*_{:(i)}$ and $u_i = \max r^*_{:(i)}$. These upper and lower values for the ith-order statistic of the M simulated residual vectors form the upper and lower edges of the diagnostic envelopes, respectively.

Atkinson's suggestion of $M = 19$ gives envelope boundaries that are estimates of the 5th and 95th percentiles of the distribution of the ith-order statistic of the externally studentized residual vector, given X and iid $N(0, \sigma^2)$ errors. The preceding method of creating diagnostic envelopes is referred to here as the A method.

Diagnostic envelopes were described and used by Atkinson (1983, 1985) for various regression-analysis examples. The current release of BMDP2R has an option for plotting these diagnostic envelopes (Hardwick 1987). Dempster, Selwyn, Patel, and Roth (1984) applied this type of envelope, with $M = 119$ and $M = 8{,}000$, to the upper extremes of a half-normal plot to stabilize the bounds; they did not use extreme-order statistics to define thè envelope edges. They suggested that diagnostic bounds simulated from one data set can be used on other data sets having similar covariance structures, but a criterion to determine "similarity" is unclear. Landwehr, Pregibon, and Shoemaker (1984) extended the idea to logistic regression.

2.2 A Resistant Diagnostic Envelope

The A method of obtaining diagnostic envelopes is restricted because the envelope limits for the ith-ordered residual are estimates of a $100\,k/(M + 1)$ percentile of that residual's appropriate distribution, with k being an integer $\le M$. If percentiles further into the tail of the distribution are desired, M must be made larger, requiring a larger and perhaps prohibitive number of simulations to obtain the desired diagnostic envelopes. Moreover, the A procedure depends on extreme-order statistics among the simulated residuals, making this procedure susceptible to creating quite variable envelopes due to extreme values among the simulated data.

We also examine the performance of a diagnostic envelope extension of an outlier-resistant rule proposed by Hoaglin, Iglewicz, and Tukey (1986) for univariate data. From the set of M simulated ith-order statistics among the externally studentized residuals, we define the ith lower and upper diagnostic-envelope bounds as

$$l_i = F_i^L - k\{F_i^U - F_i^L\}$$

and

$$u_i = F_i^U + k\{F_i^U - F_i^L\},$$

where F_i^L and F_i^U are the upper and lower fourths of the M ith-order statistics and k is an appropriately chosen constant. For each i, these fourths are the f-order and $(M + 1 - f)$-order statistics, respectively, of the M simulated values of $\{r^*_{:(i)}\}$. The *ideal* f of Hoaglin and Iglewicz (1987) is used for estimating the fourths; it is the value $f = M/4 + 5/12$. If f is not an integer, linear interpolation between adjacent-order statistics is used to obtain the fourth. We label this form of diagnostic-envelope creation the R_k method.

The R_k bounds will be less sensitive to extreme values among the simulated residuals than the A bounds. In addition, varying k allows the estimation of virtually any percentile of the ith-order statistic distribution, although the relationship between k and a specific desired percentile may not be known. We consider the use of values of $k = 1.5$ and $(2.25\text{–}3.6/M)$ for diagnostic envelopes $R_{1.5}$ and R_*, respectively. Hoaglin et al. (1986) chose 1.5 in the univariate case as a good rule for flagging outliers in moderately sized samples. The second value for k has, in the univariate case, an approximate 5% probability that a Gaussian sample of M < 100 has one or more points falling beyond either of the associated upper and lower bounds (Hoaglin and Iglewicz 1987). A third value of k was chosen empirically for each data set to give R_A envelopes having approximately the same joint Gaussian inclusion fraction as the A method.

3. SIMULATION-ENVELOPE COMPARISONS

We examine variability and joint inclusion properties of each form of the diagnostic envelopes. For each method of constructing diagnostic envelopes. 1,000 envelopes were created. At each of the $2n$ points of the envelope, the pointwise variability was estimated by the sample standard deviation (SD) over the 1,000 replications. For each of the 1,000 envelopes, whether the upper or lower bounds at each x-axis index excluded the appropriate data residual was recorded. The *mean envelope* for each method is the pointwise average over the 1,000 replications of envelopes. A maximum pointwise SD of 1.0 gives a maximum pointwise standard error of .03 for the mean envelopes.

The mean envelopes are taken as *true envelopes* to estimate Gaussian joint inclusion levels. The fraction of 2,000 independent Gaussian-based simulated externally studentized residual vectors that fall entirely within the mean diagnostic envelope was tabulated. The maximum standard error of the inclusion proportions is .01.

4. APPLICATIONS TO DATA SETS

To study and compare the performance of Atkinson's diagnostic envelopes with the three R_k methods, we use several examples. We consider the data given by Atkinson (1985) for a 2^4 experiment with a center point added, variations of these data, and Brownlee's (1965) stack-loss data. Selected full-normal and half-normal residual plots of these data are given later in this article.

Three forms of the 2^4 experiment with a center point are analyzed. The first is the actual data as simulated with standard Gaussian errors. Atkinson's (1985) half-normal residual-plot envelopes show no abnormalities. The second form of the data has one miscoded response at the center point (point 17); the decimal is shifted to the right by one digit, giving an extremely large response. Atkinson (1985) showed that the half-normal residual-plot envelopes undoubtedly exclude the resulting residual, indicating clearly a data problem. The third form of the data has all responses correctly coded, but the center point is miscoded in all four variables by again shifting the decimal one digit to the right. For these data, the miscoded point 17 gives the smallest signed residual; it is large and negative and is the largest absolute residual. The half-normal residual plot shows one residual that is some distance outside the envelopes but not by as extreme a distance as the other plot.

Brownlee's (1965) data has been examined in detail by Andrews (1974), Cook (1979), Daniel and Wood (1980), and Atkinson (1981). The data consist of the response variable and three explanatory variables. We use the model having a linear term for each of the three explanatory variables. Daniel and Wood concluded from their ordinary least squares and time-ordering analyses that observations 1, 3, 4, and 21 are outliers. Andrews noted in his analysis that the normal probability plot of the residuals shows that "observation 21 has an abnormally large residual" (p. 529); the sign of this residual is negative. Atkinson stated that the A method gives a diagnostic envelope that entirely contains the actual residual vector. He found point 21 slightly outside a Cook's-distance envelope plot, indicating a possibly bad data point.

4.1 Joint Residual Inclusion Rates

In this section, we analyze the joint residual-vector inclusion rates of the actual residuals and of the Gaussian-simulated residuals. The *grand means* of the 1,000 simulated envelopes were used to determine residual-vector inclusion rates for the 2,000 Gaussian-based n vectors. The fractions of these vectors that fell entirely within each of the true envelopes are given in Table 1. This joint fraction is below 50% for the A-method and R_A envelopes. The full A-method bounds exclude the simulated ith residuals infrequently and at a uniform rate (3%) for each i. The full R_A upper-bound exclusions of the simulated residuals are higher at the upper indexes; their lower-bound exclusions of the simulated residuals are higher at the lower indexes. In the R_A half-plots, simulated residuals fall above the bounds more frequently at both the upper and lower extremes of i. Table 1 also shows the values of k used to obtain the R_A bounds. The $R_{1.5}$ method has a higher joint Gaussian inclusion level. The R_* method has a joint inclusion level near 95% as expected; these envelopes are considerably wider than the other envelopes. For each i, the ith simulated residuals fall outside the bounds at a rate of less than 1% for full-normal and half-normal plots using methods $R_{1.5}$ and R_*.

Table 2 lists the fraction of the 1,000 envelopes that included the entire actual-residuals vector in the full-normal and half-normal probability plots for each data set. For the correct version of the 2^4 data, most of the full-normal-probability-plots envelopes include the entire residual vector. Slightly less often,

Table 1. Joint Inclusion Fractions of 2,000 Simulated Gaussian Residual Vectors, Using the Grand Mean Envelope of the 1,000 Replications of Envelopes for Normal Probability Plots

	Full-normal plots				Half-normal plots			
Data set	Atkinson	$R_{A(k)}$	$R_{1.5}$	R_*	Atkinson	$R_{A(k)}$	$R_{1.5}$	R_*
2^4 design								
(x, y) correct	.455	$.488_{(1.0)}$	.819	.941	.488	$.498_{(1.0)}$	.789	.927
y incorrect	.455	$.488_{(1.0)}$	.819	.941	.488	$.498_{(1.0)}$	.789	.927
x incorrect	.457	$.448_{(1.0)}$	.783	.921	.490	$.437_{(.9)}$	.776	.914
Brownlee data								
linear fit	.402	$.416_{(1.0)}$	.785	.934	.424	$.420_{(.95)}$	.788	.930

NOTE: The maximum standard error of each estimate is .011.

the half-normal plots contain the vector; for all envelope methods, this is due to the smallest three absolute residuals falling above the upper envelope edge. The R_A-method half-normal plots have a very low inclusion rate, because the first and second smallest absolute residuals fell above the upper envelope edge 40% and 53% of the time, respectively.

For both types of plot with the two types of coding errors for the 2^4 design, the residual vector is almost never entirely within the simulated envelope. For the miscoded x data, the lower edge of the full-normal bounds (all methods) excludes the smallest residual almost uniformly; the rest of the residuals remain inside the bounds except that the A-method edges are above the largest residual 46% of the time. For the miscoded y data, the upper edge of the full-normal and the half-normal envelopes all exclude the largest residual. Several of the ordered residuals immediately preceding the largest are abnormally low and are below the lower edges a large percentage of the time. The lower edges of the full-normal plots often exclude the smallest four residuals also.

For the Brownlee data, the A method and R_A have low joint inclusion rates; the R_* method gives a fairly high joint inclusion, with the $R_{1.5}$ method falling somewhere in between. The full A-method envelopes are above the smallest residual (21) 29% of the time, and the second smallest residual is above the bound 82% of the time. In contrast, residual 21 is excluded from below by 67% of the R_A bounds, and the second smallest residual is excluded above on 48% of the time. For the half-normal envelopes, the largest absolute residual is frequently above only the R_A bounds (34%). Both the A method and the R_A bounds are above the third to sixth largest absolute residuals often (27% to 47%), with the R_A exclusion rates for these points being higher.

The average number of the actual residuals excluded by the different envelope-generation methods decreases as the joint vector inclusion rate increases. There is no consistent pattern distinguishing the number of residuals excluded by the A method compared to the R_A method.

4.2 Envelope Variability and Shape

Further characterization of the variability of the bounds is important, since each envelope is the result of only $M = 19$ replications of the Gaussian residual vector-generation process. The maxima of the upper and lower n pointwise standard deviations over the 1,000 replications of each type of envelope are displayed in Table 3. These are identical for the correct data form of the 2^4 design and that with the incorrect response, since the design matrix is identical for the two cases. The R_A bounds are from approximately 50% to 66% less variable than the standard A-method bounds, except for the lower bounds on the half-normal plots. The R_A half-normal lower edges are less variable than the associated upper edges. The $R_{1.5}$ and R_* bounds are slightly more variable

Table 2. Joint Inclusion Fractions of Residual Vectors, Based on 1,000 Replications of Envelopes for Normal Probability Plots

	Full-normal plots				Half-normal plots			
Data set	Atkinson	R_A	$R_{1.5}$	R_*	Atkinson	R_A	$R_{1.5}$	R_*
2^4 design								
(x, y) correct	.695	.649	.936	.992	.478	.243	.651	.838
y incorrect	.000	.000	.000	.000	.000	.000	.000	.000
x incorrect	.029	.000	.004	.058	.052	.000	.032	.159
Brownlee data								
linear fit	.097	.125	.520	.820	.375	.217	.672	.863

NOTE: The maximum standard error of each estimate is .016.

Table 3. Maximum Pointwise Standard Deviation of Envelope Boundary, Using 1,000 Replications of the Envelopes for Normal Probability Plots

	Full-normal plots				*Half-normal plots*			
Data set	*Atkinson*	R_A	$R_{1.5}$	R_*	*Atkinson*	R_A	$R_{1.5}$	R_*
				Upper-edge variability				
2^4 design								
(x, y) correct	.72	.38	.47	.58	.72	.38	.47	.58
y incorrect	.72	.38	.47	.58	.72	.38	.47	.58
x incorrect	.71	.41	.51	.63	.74	.41	.55	.67
Brownlee data								
linear fit	.58	.37	.46	.57	.52	.33	.42	.52
				Lower-edge variability				
2^4 design								
(x, y) correct	.70	.34	.42	.52	.12	.27	.36	.47
y incorrect	.70	.34	.42	.52	.12	.27	.36	.47
x incorrect	.72	.35	.43	.52	.12	.25	.38	.50
Brownlee data								
linear fit	.47	.38	.47	.58	.13	.23	.32	.41

NOTE: Asymptotic standard error for each estimate of σ is σ $2000^{1/2}$; a maximum of $\sigma = 1$ makes the maximum standard error of each tabled value .02.

than the R_A bounds, but these methods are estimating more distant quantiles of each distribution and, therefore, higher variability is not unusual. For all of the full-normal envelopes, the upper edge bounds are most variable at the highest index; the edges at both extremes are more variable than the inner-index bounds. The lower edges are most variable at the lowest index. For the half-normal envelopes, both the upper and lower edges are most variable at the highest index. There is a sharp drop in edge-point variability as the index decreases away from its maximum.

All four types of the full-normal grand mean envelopes for the Brownlee (1965) data are plotted with the observed data residuals in Figures 1 and 2. The A-method and R_A-method envelopes differ only in the extremes; the R_A extreme index bounds are closer to 0 than the A bounds. The $R_{1.5}$ and R_* bounds are slightly wider than the A-method envelopes, but the difference is not large. The envelopes in these figures have pointwise standard errors smaller than .02 due to averaging over 1,000 envelope replications.

The Brownlee half-normal plot (Fig. 3) does not show abnormalities; the full-normal plot has one residual that is slightly outside the A-method bounds, however. Of more interest in the full-normal plot is the fact that the lowest point (21) is almost outside the lower envelope bound, followed by the next point that exceeds the upper envelope bound. The patterning in the full-normal plot gives evidence for

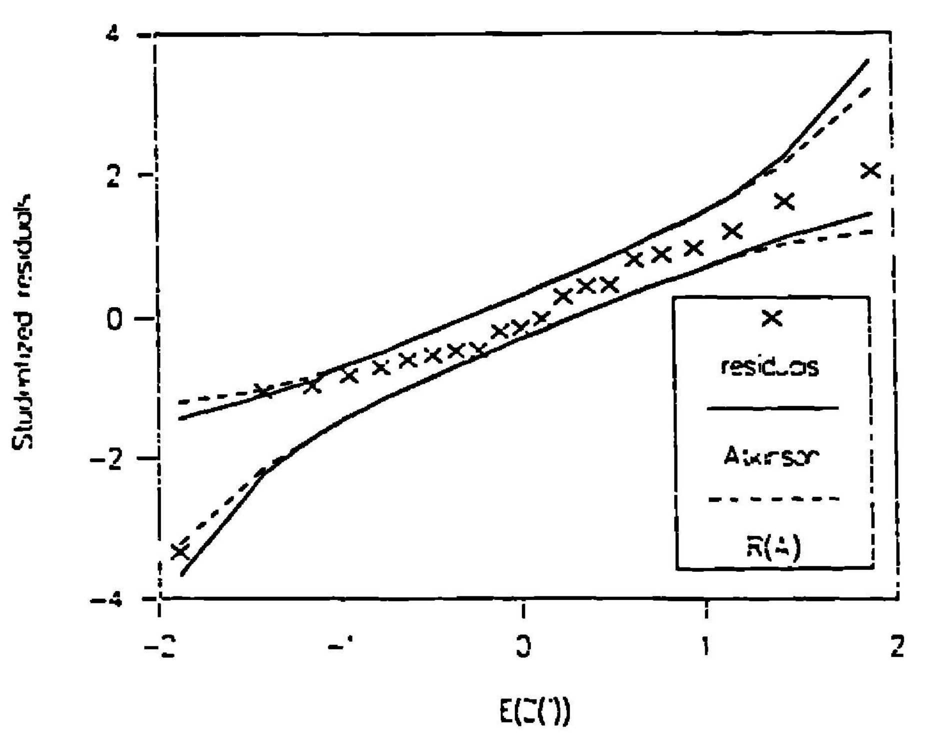

Figure 1. Full-Normal Probability Plot of Brownlee Residuals, With A-Method (M = 19) and R_A Diagnostic Bounds. The bounds are means over 1,000 envelope replications.

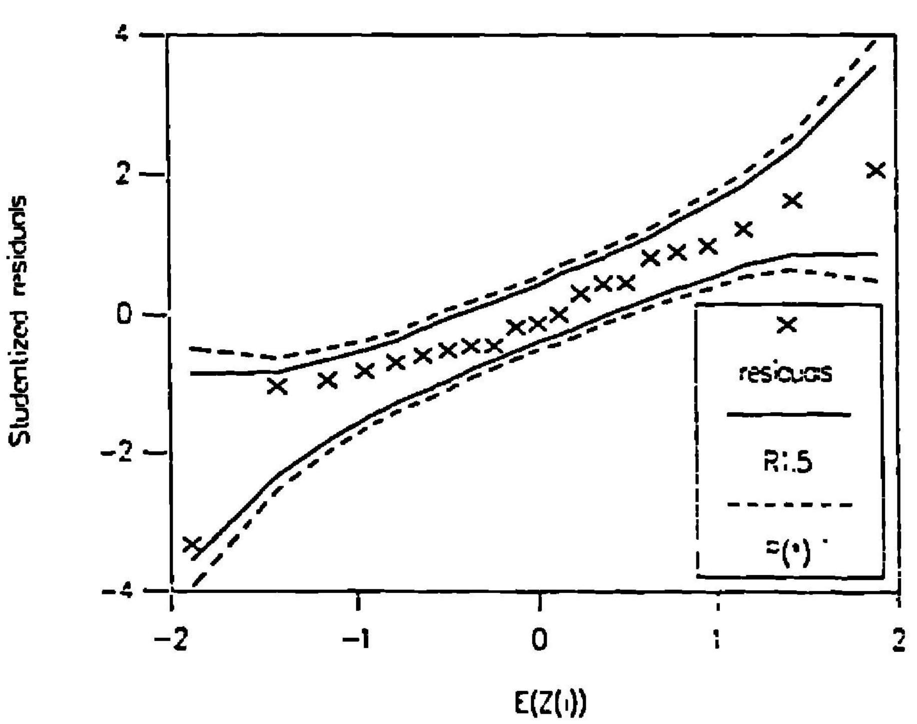

Figure 2. Full-Normal Probability Plot of Brownlee Residuals, With $R_{1.5}$ and R_* Diagnostic Bounds. The bounds are means over 1,000 envelope replications.

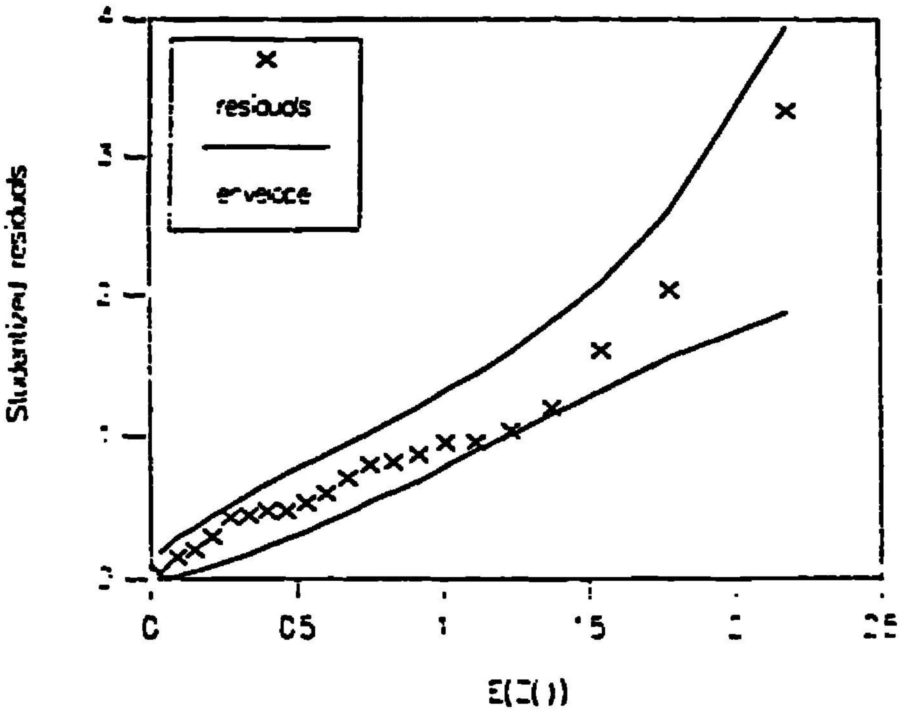

Figure 3. Half-Normal Probability Plot of Brownlee Residuals, With A-Method (M = 19) Diagnostic Bounds. The bounds are means over 1,000 envelope replications.

wariness about point 21 that is not available in the half-normal plot.

For the correct 2^4 design data, as expected, no strong abnormalities are evident. The miscoded response is immediately evident from both the full-normal and half-normal plots of that data (not shown). The miscoded x data shows an extreme residual in both the full-normal and half-normal probability plots; since the envelope boundary is known to be stable here, the outlying residual in either plot gives evidence for some concern. Figure 4 gives the full-normal A-method and $R_{1.5}$ plots for the miscoded x data; the errant residual value is excluded by both bounds.

5. DISCUSSION

While the diagnostic envelopes for normal probability plots of multiple-regression residuals are informative in a broad sense, their interpretation is still subjective, and patterns in the plot still should be considered as possible problem indicators. We reinforce others' cautions against using the diagnostic envelopes as pure acceptance/rejection indicators for testing the normality of errors assumption. The envelope edges can highlight patterns, such as one extreme point, that may fall outside the bounds and/or force nearby points outside their diagnostic limits. Evidence of plot curvature associated with skewness can be more clear with a comparison shape superimposed over the ordered residuals.

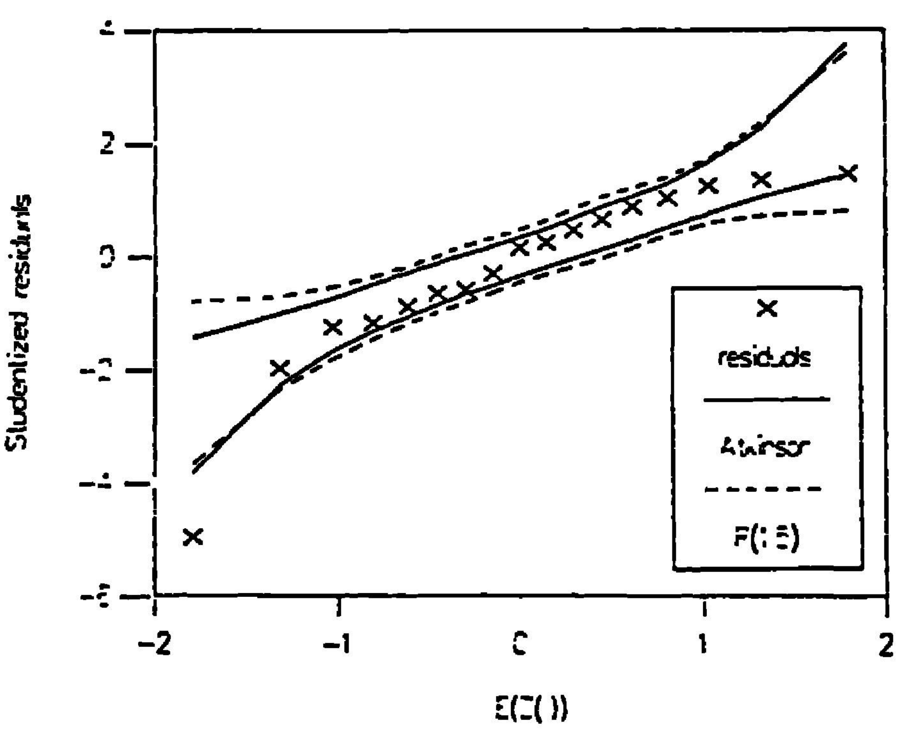

Figure 4. Full-Normal Probability Plot of Residuals of the 2^4 Design With Wrong x at the Center Point, With A-Method (M = 19) and $R_{1.5}$ Diagnostic Bounds. The bounds are means over 1,000 envelope replications.

The resistant alternatives to the envelope-creation method proposed by Atkinson (1981) show promise. The resistant methods give more stable bound estimates for a given number of simulations and have a more flexible range of levels of joint Gaussian residual-vector inclusion than the A method. The $R_{1.5}$ method, however, seems to be more useful than R_*, whose bounds may be too wide to be sensitive to data irregularities. Replication of the simulation process, followed by averaging the resulting envelopes, will stabilize the bounds even further; the computing required to replicate the envelopes is minimal relative to the current state of statistical computing. Use of a larger value of M is also possible; this is more computer-intensive, requiring storage and sorting of the M individual vectors. The conditional percentile being estimated is a function of M for the A method, so the characteristics of the A-method bounds may change with different M.

The joint inclusion fractions and the envelope variability observed for these data sets are, of course, not directly generalizable to other data sets. As comparative measures of the relative inclusion rates and variability of the different methods, they contribute valuable information.

In comparing these different envelope-generation methods, both full-normal and half-normal probability plots have been used. Atkinson (1981) stated that the half-normal plot can highlight extreme values more effectively than full-normal plots. When compared to the simulated-envelope boundaries for the data sets here that have outlying residuals, the full-normal plots exclude residuals more frequently than the half-normal plots. In one data example here, the half-normal plot masks irregularities that are evident in the full-normal plot—an argument for suggesting the plotting of both types for data analyses rather than one or the other.

ACKNOWLEDGMENTS

This work was supported in part by Grant BRSG S07 RR 5442 awarded by the Biomedical Research Support Grant Program, Division of Research Resources, National Institutes of Health. We also thank the referees and editors for suggestions resulting in improved presentation of this material.

[*Received March 1988. Revised August 1988.*]

REFERENCES

Andrews. D. F. (1974). "A Robust Method for Multiple Linear Regression." *Technometrics*. 16. 523–531.

Atkinson. A. C. (1981). "Two Graphical Displays for Outlying and Influential Observations in Regression." *Biometrika*. 68. 13–20.

——— (1983). "Diagnostic Regression Analysis and Shifted Power Transformations." *Technometrics*. 25. 23–33.

——— (1985). *Plots, Transformations, and Regression*. Oxford. U.K.: Clarendon Press.

Brownlee. K. A. (1965). *Statistical Theory and Methodology in Science and Engineering*. New York: John Wiley.

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

Daniel. C. (1959). "Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments." *Technometrics*. 1. 311–341.

Daniel. C.. and Wood. F. S. (1980). *Fitting Equations to Data* (2nd ed.). New York: John Wiley.

Dempster. A. P., Selwyn. M. R., Patel. C. M.. and Roth. A. J. (1984). "Statistical and Computational Aspects of Mixed Model Analysis." *Applied Statistics*. 33. 203–214.

Hardwick. J. (1987). "2*R* Diagnostics." *BMDP Communications*. 19. 12–14.

Hoaglin. D. C.. and Iglewicz. B. (1987). "Fine-Tuning Some Resistant Rules for Outlier Labeling." *Journal of the American Statistical Association*. 82. 1147–1149.

Hoaglin. D. C.. Iglewicz. B.. and Tukey. J. W. (1986). "Performance of Some Resistant Rules for Outlier Labeling." *Journal of the American Statistical Association*. 81. 991–999.

Landwehr. J. M.. Pregibon. D.. and Shoemaker. A. C. (1984). "Graphical Methods for Assessing Logistic Regression Models." *Journal of the American Statistical Association*. 79. 61–71.