Graphical Methods for Assessing Logistic Regression Models

JAMES M. LANDWEHR, DARYL PREGIBON, and ANNE C. SHOEMAKER*

In ordinary linear regression, graphical diagnostic displays can be very useful for detecting and examining anomalous features in the fit of a model to data. For logistic regression models, the discreteness of binary data makes it difficult to interpret such displays. Modifications and extensions of linear model displays lead to three methods for diagnostic checking of logistic regression models. Local mean deviance plots are useful for detecting overall lack of fit. Empirical probability plots help point out isolated departures from the fitted model. Partial residual plots, when smoothed to show underlying structure, help identify specific causes of lack of fit. These methods are illustrated through the analyses of simulated and real data.

KEY WORDS: Binary data; Goodness of fit; Residual analysis; Near neighbors; Probability plot; Partial residual.

1. INTRODUCTION

Logistic regression models are useful in problems where the dependent variable takes on only a few discrete values. Major fields of application include econometrics, biostatistics, and educational testing. We consider the special case in which the response is dichotomous (binary). An example considered in detail in Section 6 concerns patient survival for a specified period after surgery for breast cancer. The response variable can be coded as 1 if the patient survives (with probability p) and as 0 otherwise (with probability $1 - p$). The logistic regression model specifies that $\text{logit}(p) = \log p/(1 - p)$ is some linear combination of explanatory variables.

A now classical theoretical treatment of binary data is that of Cox (1970). In contrast, this paper focuses on more informal and exploratory techniques associated with fitting logistic models to binary data. The examination and plotting of residuals to detect inadequacies in a fitted linear regression model is a useful and highly recommended practice (e.g., Daniel and Wood 1971; Weisberg 1980). Analogous displays of residuals and related quantities for logistic regression would be most useful. In a similar vein, work on influential observations in linear regression (see Belsley, Kuh, and Welsch 1980) has been generalized to logistic regression by Pregibon (1981).

This paper proposes and discusses three graphical methods that can be used to help assess and possibly improve the fit of logistic regression models. These methods are modifications and generalizations of linear model displays, developed with the aim of alleviating the effect that the discrete response variable has on the associated plots.

The formulation and some basic features of the logistic regression model are considered in Section 2. Section 3 presents local mean deviance plots, a graphical method for assessing the overall adequacy of the fit. Section 4 concerns probability plots, which can be useful for detecting outliers and, sometimes, more general model departures. Section 5 describes the use of partial residual plots for investigating possible systematic departures. Each method is illustrated with examples using simulated data. In Section 6 all three methods are applied to the breast cancer data. Section 7 contains further discussion of the usefulness and application of the suggested methods.

2. BACKGROUND

Consider data of the form $\{y_i, x_{i1}, x_{i2}, \ldots, x_{im}; i = 1, \ldots, N\}$. The response y_i is assumed to be binomial $(1, p_i)$, and $X_1, \ldots, X_m$ are explanatory variables. The basic model is that

$$\text{logit}(\mathbf{p}_i) = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{im}\beta_m = \mathbf{x}_i^T\boldsymbol{\beta}, \quad i = 1, \ldots, N.$$

In an obvious matrix notation, the model can be written as $\text{logit}(p) = \mathbf{X}\boldsymbol{\beta}$. The log-likelihood function is

$$l\{\boldsymbol{\beta}; \mathbf{y}\} = \sum_{i=1}^{N} \{y_i \log p_i + (1 - y_i)\log(1 - p_i)\} = \sum_{i=1}^{N} \{y_i \mathbf{x}_i^T\boldsymbol{\beta} - \log(1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta}))\}. \quad (2.1)$$

The maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ maximizes (2.1) and satisfies $\mathbf{X}^T\mathbf{r} = 0$, where $\mathbf{r} = \mathbf{y} - \hat{\mathbf{p}}$ and $\hat{p}_i = \exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})/(1 + \exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}))$. This demonstrates one connection with

* James M. Landwehr and Daryl Pregibon are Members of Technical Staff, Statistics and Data Analysis Research Dept., Bell Laboratories, Murray Hill, NJ, 07974. Anne C. Shoemaker is Member of Technical Staff, Quality Assurance Center, Bell Laboratories, Holmdel, NJ, 07733. The authors wish to thank R. Gnanadesikan, J.R. Kettenring, C.L. Mallows, J.W. Tukey, and the associate editor for helpful comments. This paper was presented as the Theory and Methods Invited Paper at the ASA Annual Meeting, August 1983, Toronto, Canada.

© Journal of the American Statistical Association
March 1984, Volume 79, Number 385
Invited Papers Section

standard least squares estimation in linear regression. The equations here are nonlinear in $\hat{\boldsymbol{\beta}}$, however, and iterative methods are required to solve them. Second derivatives can be computed, yielding the information matrix $\mathbf{I} = \mathbf{X}^T\mathbf{V}\mathbf{X}$, where $\mathbf{V} = \text{diag}\{p_i(1 - p_i)\} = \text{cov}(\mathbf{y})$. The Newton-Raphson method expresses $\hat{\boldsymbol{\beta}}$ at the $(t + 1)$st iteration as

$$\hat{\boldsymbol{\beta}}(t + 1) = \hat{\boldsymbol{\beta}}(t) + (\mathbf{X}^T\mathbf{V}(t)\mathbf{X})^{-1}\mathbf{X}^T\mathbf{r}(t), \quad t = 1, 2, \ldots, \tag{2.2}$$

where the arguments of $\mathbf{V}$ and $\mathbf{r}$ refer to the values of these quantities evaluated at $\hat{\boldsymbol{\beta}}(t)$.

To assess the disagreement between the observed response $\mathbf{y}$ and the fitted values $\hat{\mathbf{p}}$, a goodness-of-fit statistic is required. A standard measure is the chi-squared statistic

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1 - \hat{p}_i)}.$$

This measure is unstable for fitted values near zero or one. Following Nelder and Wedderburn (1972), we prefer a measure that is more analogous to the residual sum of squares in linear regression. This measure, called the deviance, is given by

$$D(\hat{\mathbf{p}}; \mathbf{y}) = \sum_{i=1}^{N} d(\hat{p}_i; y_i) = -2 \sum_{i=1}^{N} \{y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)\}.$$

Since $D(\hat{\mathbf{p}}; \mathbf{y})$ is negative two times the maximized log-likelihood function, it will decrease as more parameters are fitted; this is not necessarily the case for χ^2. Associated with the deviance is the number of degrees of freedom upon which it is based. For an m parameter model, there are $N - m$ degrees of freedom.

As a starting point for the following sections, we assume a fitted model $\text{logit}(\hat{\mathbf{p}}) = \mathbf{X}\hat{\boldsymbol{\beta}}$. This could be obtained from any of several widely available programs, for example, Haberman (1979) or Baker and Nelder (1978). The standard statistical output from this fit consists of the estimates $\hat{\boldsymbol{\beta}}$, their estimated covariance matrix $\text{cov}(\hat{\boldsymbol{\beta}}) = \mathbf{I}^{-1}$, the fitted values $\hat{\mathbf{p}}$, and the deviance $D(\hat{\mathbf{p}}; \mathbf{y})$.

3. LOCAL MEAN DEVIANCE PLOTS

The deviance measures the global disagreement between the observed data and the fitted values. Currently, no adequate distribution theory for $D(\hat{\mathbf{p}}; \mathbf{y})$ exists. Even so, reduction of an entire sample to a single number is not an entirely adequate method of assessing goodness of fit. Our suggested procedure is based on a partition of $D(\hat{\mathbf{p}}; \mathbf{y})$ into a pure-error component and a lack-of-fit component. If the model accounts for the systematic variation in the data, then the latter component will be small. Otherwise it will be inflated because of systematic behavior that the model has failed to account for.

If there are exact replicates (identical rows $\mathbf{x}_i^T$) in the data, the pure-error component of the deviance is easily obtained. If there are no exact replicates, as is generally the case in observational studies, we propose a method based on near-neighbors and an approximate decomposition of the deviance. The method is a generalization of one proposed by Daniel and Wood (1971, Ch. 7) for assessing goodness of fit in linear regression, though several differences in detail are evident. Specifically, we propose the following procedure:

1. Partition the N observations into K nonoverlapping groups (clusters) with N_k observations in each. There is no limit to the size of the individual groups, though to retain local fits that are nearly model independent, it is best to use small groups of homogeneous points. Groups of size $N_k = 1$, can be omitted from further consideration as they contribute no information concerning local variation.

2. Form the $N \times K$ matrix Z with (ik)th element $Z_{ik} = 1$ or 0 according as the ith observation is or is not in the kth group.

3. Compute the local estimate $\hat{\mathbf{p}}^L$ by fitting $\text{logit}(\mathbf{p}^L) = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{X}\boldsymbol{\beta}$. The fit of this equation is close to that given by $\text{logit}(\hat{\mathbf{p}}^L) = \text{logit}(\bar{\mathbf{y}}) + (\mathbf{X} - \bar{\mathbf{X}})\hat{\boldsymbol{\beta}}$, where the average vectors are local averages determined by $\mathbf{Z}$. This form reduces to $\text{logit}(\hat{p}_k{}^L) = \text{logit}(\bar{y}_k)$ when the kth group consists of exact replicates, since $\mathbf{x}_{kj} = \bar{\mathbf{x}}_k$ for all j. When the kth group consists of near replicates the second term makes local corrections to the cluster mean.

4. Use the model fitted in (3) to compute the local deviance contribution of each observation, $d(\hat{p}_{kj}{}^L; y_{kj})$, and sum the deviances within each group giving $D_k{}^L = \sum d(\hat{p}_{kj}{}^L; y_{kj})$.

5. Reorder the K groups so that $S_1 \le S_2 \le \cdots \le S_k$, where S_i is a measure of group inhomogeneity. For example, if a hierarchical clustering method is used, S_i is the height (sometimes called level or distance) in the tree at which the group is formed. Thus, points in group 1 are more tightly clustered than those in group 2, which are more tightly clustered than those in group 3, and so on.

6. Compute running estimates of approximate pure-error

$$\bar{D}^L(t) = \sum_{k=1}^{t} D_k{}^L \Big/ \sum_{k=1}^{t} (N_k - 1).$$

The values $\bar{D}^L(t)$, $t = 1, \ldots, K$ represent the local mean deviance calculated from the tightest t groups.

7. Plot $Y = \bar{D}^L(t)$ against its degrees of freedom $X = \sum_{k=1}^{t} (N_k - 1)$, for $t = 1, \ldots, k$. Superimpose on this plot the line $Y =$ global mean deviance, $\bar{D} = D(\hat{\mathbf{p}}; \mathbf{y})/(N - m)$, and observe its position relative to the points plotted above.

We refer to the plot produced by these steps as the local mean deviance plot. Lack of fit is exhibited when the line in (7), which shows the general level of variability of the data about the current fitted model, is consistently

above the points plotted in (7), which suggest the level of local variability.

To correctly assess the fit of a model using this display, one must be aware of the tradeoff between insufficient degrees of freedom and group inhomogeneity. On the one hand, we would like to calculate the local mean deviance using only those groups that refer to exact replicate design points. This typically leads, however, to an estimate with too few degrees of freedom. On the other hand, as we start accumulating more and more groups, it is not clear that our estimate is reflecting only local variability. The reason we prefer presenting the results in graphical form, rather than as a statistic, is that by looking at the plot one can get more information about how this trade-off occurs for the data.

Example 1. This example involves two explanatory variables X_1 and X_2 each with 50 values generated from the $U(0, 1)$ distribution. Binary responses were constructed according to the model $\text{logit}(P) = 4X_1 + 4X_2 - 12X_1X_2$. The data are plotted in Figure 1 (ignore the loops at present).

First consider fitting the model including the intercept, X_1 and X_2 terms, but not the X_1X_2 interaction. The deviance for this fit is 60.92 on 47 df. To see if this model is adequate, we compute the local mean deviance plot. The rows of the (X_1, X_2) matrix were clustered using a hierarchical clustering procedure based on Euclidean distance, and the resulting tree was cut at the level giving 19 clusters. Points clustering together are indicated by the loops in Figure 1. The resulting local mean deviance plot is shown in Figure 2 (ignore the two horizontal lines at present). There are 17 points since two clusters are singletons. Note that the first four points from the left all have Y-value 0, indicating that each of the four tightest clusters has observations that are either all 0's or all 1's. Moving to the right, the local mean deviance values then fluctuate but settle down around .8 as the degrees of freedom increase. The height (level) S at which each cluster is formed is shown by the scale at the top of the plot; it increases nonlinearly as degrees of freedom increase. A sharp increase in S for large degrees of freedom would suggest that substantially larger groups were being included at this point, and these might not be measuring local deviance. Fortunately, there is no such increase here.

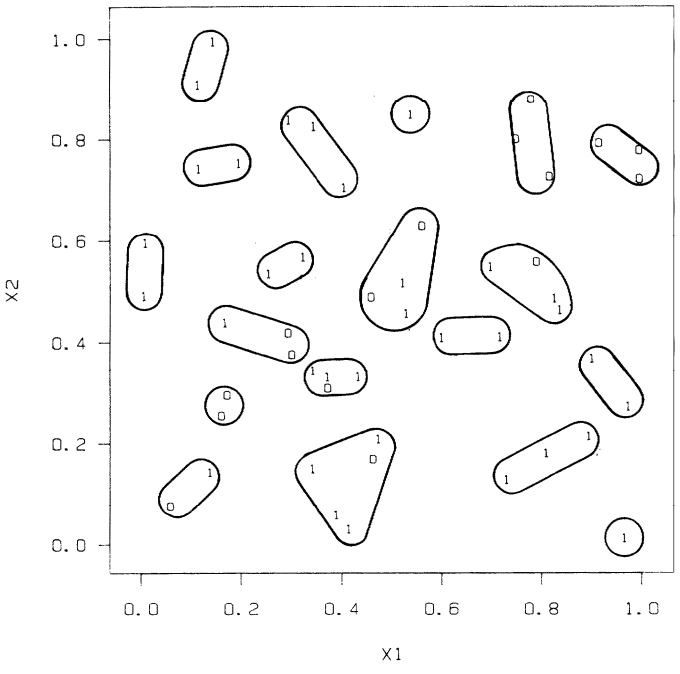

Figure 1. SCATTER PLOT OF DATA AND RESPONSE FOR EXAMPLE 1. The loops indicate the points that are clustered together.

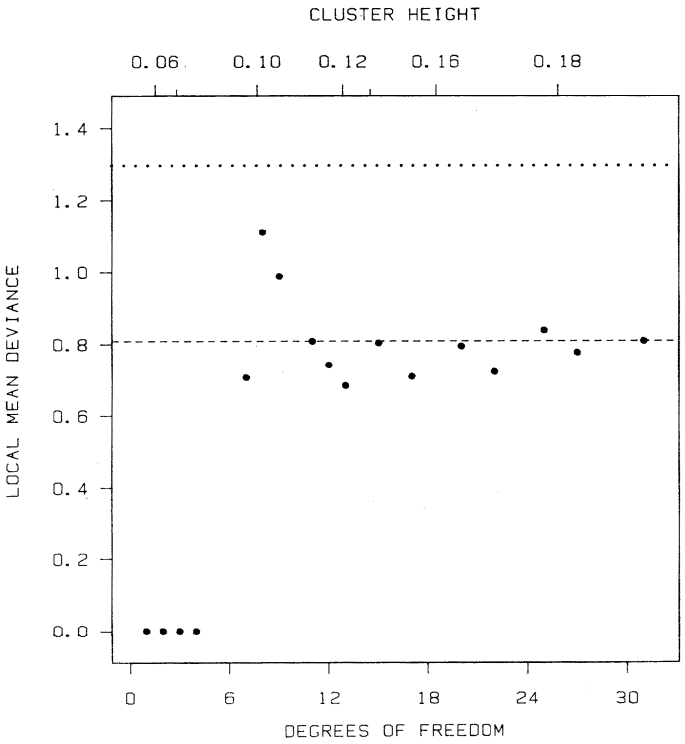

Figure 2. LOCAL MEAN DEVIANCE PLOT FOR EXAMPLE 1. The dotted line is the mean deviance from the model fitting X_1 and X_2. It is above the points giving the local mean deviance estimates, indicating lack of fit for this model. The dashed line is the mean deviance from the model which also includes X_1*X_2. It passes near the points with large degrees of freedom, so no lack of fit is indicated.

The global mean deviance $\bar{D}$ from the model with intercept, and X_1, and X_2 is 60.92/47 = 1.3, shown on Figure 2 by the dotted line. For moderate to large degrees of freedom, where local variability is estimated reliably, the line is consistently above the points. We conclude that this model is not adequate for these data. Fitting the model with intercept, and X_1, X_2, and X_1X_2, gives mean deviance 37.17/46 = .81, shown on Figure 2 by the dashed line. Contrary to the model omitting the interaction term, this line passes through the points, indicating a satisfactory fit.

The possible need for an interaction term here can be conjectured from Figure 1 by noting the predominance

of 1's in the NW and SE corners, with 0's in the NE and SW corners. This shows that even a basic display such as Figure 1 can be useful, although with more explanatory variables or a more complicated model we would not expect as much.

4. EMPIRICAL PROBABILITY PLOTS

In linear regression problems, normal probability plots of residuals are a useful tool for detecting outliers and examining distributional assumptions about the errors (e.g., Daniel and Wood 1971, Ch. 3; Weisberg 1980). This section develops a probability plot for logistic regression.

In linear regression, the basic idea is to plot ordered values of appropriately standardized residuals on the y axis against ordered quantiles of the standard normal distribution on the x axis. When the assumed model correctly fits the data this plot approximates a straight line. A nonlinear configuration of points can result from a nonnormal error distribution or the presence of outliers. For logistic regression, the analogous plot has an appropriate residual quantity on the y axis and quantiles from its reference distribution on the x axis. A natural residual to use is the deviance contribution $d_i = d(\hat{p}_i; y_i)$, standardized by its approximate standard error. Intuitive arguments suggest using a $\chi^2(1)$ distribution as the reference, but this was not supported in our empirical studies.

In order to obtain a probability plot with linear null configuration, we propose a simulation procedure for obtaining the reference distribution. The residual quantity for plotting on the y axis is taken to be simply $(y_i - \hat{p}_i)$, $i = 1, \ldots, N$. This is a monotonic function of the signed deviance contribution, and it is directly interpretable in terms of the observations and fitted model. The simulation procedure is based on $\hat{\mathbf{p}}$ and estimates the distribution that these residuals would have if the fitted model were correct. It involves the following steps:

1. From the fitted model $\text{logit}(\hat{\mathbf{p}}) = \mathbf{X}\hat{\boldsymbol{\beta}}$ obtain residuals $r_i = y_i - \hat{p}_i$, $i = 1, \ldots, N$.
2. Order the r_i, giving $r_{(i)}$.
3. Simulate data $\mathbf{y}^*$ from the fitted model, i.e., $y_i^* \sim \text{binomial}(1, \hat{p}_i)$, $i = 1, \ldots, N$.
4. Fit the model $\text{logit}(\mathbf{p}^*) = \mathbf{X}\boldsymbol{\beta}^*$ to these data.
5. Compute the residuals $r_i^* = y_i^* - \hat{p}_i^*$ and order them, giving $r_{(i)}^*$.
6. Repeat steps 3–5 M times independently.
7. Compute typical values (e.g., medians) of the ordered residuals over the M replications, say $T_{(i)} = \text{med}\{r_{(i)}^*\}$, $i = 1, \ldots, N$. (In fact, a slightly more complicated version of this calculation, as described at the end of this section, is used.)
8. Plot the ordered residuals from the original fit against the typical ordered residuals from the simulation, i.e., plot $(T_{(i)}, r_{(i)})$, $i = 1, \ldots, N$.
9. Obtain and plot upper and lower confidence bounds at each of the $T_{(i)}$ by taking upper and lower quantiles of the M values $\{r_{(i)}^*\}$ corresponding to the desired confidence coefficient. For example, with $M = 25$, taking the next-to-lowest and next-to-highest of the $\{r_{(i)}^*\}$ corresponds to an approximate confidence coefficient of 100(22/26) = 84.6% between these limits and 7.7% outside at each end, for each $i = 1, \ldots, N$.

We call this an empirical probability plot. When the model we are examining is indeed the one from which the data arose, this plot approximates a straight line through the origin with unit slope. It may seem that an enormous price, M iterative fits, has been incurred to obtain linearity when the model is correct. However, it would be appreciably more difficult to detect model inadequacies if the plot did not have this property. The cost of the M fits can be reduced by using $\hat{\boldsymbol{\beta}}$ as starting values and following only one or two iterations of the Newton-Raphson procedure. The following two examples show that the empirical probability plot has the ability to detect certain types of departures of the data from the hypothesized logit model.

Example 2. Fifty observations were generated from the model $\text{logit}(P) = -1 + 2X_3$, where $X_3 = -2.5(.1)2.5$, omitting 0. However, an observation that has true probability of being one equal to .002 was changed from zero to one. The model with intercept and X_3 was fitted, and Figure 3 shows the empirical probability plot with $M = 25$ and confidence coefficient = 84.6%. The fact that the model fits well is suggested by the points generally falling near the $Y = X$ line and within the confidence bands in both upper and lower portions. Although it may seem

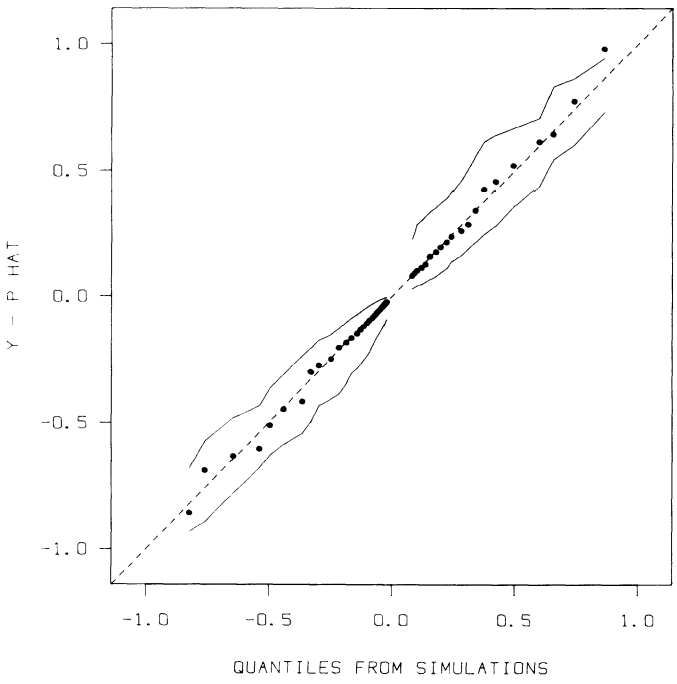

Figure 3. EMPIRICAL PROBABILITY PLOT FOR EXAMPLE 2. The solid lines are the confidence bands obtained from simulations of the fitted model. The dotted line is the Y = X line, shown for reference. Note that the top point is outside the confidence band, indicating a possible outlier. The rest of the points are inside the bands and do not indicate any lack of fit.

surprising to have a vertical and horizontal gap in the middle of the plot, this does not in fact indicate a problem. It is because the upper portion is $1 - \hat{p}$, and the largest $\hat{p}$ here is .92, while the bottom portion is $0 - \hat{p}$ and the smallest $\hat{p}$ is .02. The observation that was changed to be an outlier is the uppermost point, lying above the confidence band.

Even though this point is not much farther from the $Y = X$ line than several others, it clearly lies above the confidence band and is the most extreme observation. Our attention is directed towards it as a possible outlier, as in fact it is.

Example 3. This example considers a different situation, where the model being fitted is inappropriate in a certain interior region of the predicted value ($\hat{p}$) space. This could correspond to a missing explanatory variable that especially affects a certain subset of observations. One hundred observations were generated from the model $\text{logit}(P) = -1.5 + X_4$, where $X_4 = -2.5(.05)2.5$, omitting 0, except that 15 observations corresponding to true probabilities from .22 to .37 were all set equal to 1. We would have expected about 4.4 of these 15 observations to be equal to 1 by chance. The model including the intercept and X_4 was fitted, giving deviance equal to 105.9 on 98 degrees of freedom. The empirical probability plot is Figure 4. The most striking feature is the large gap in the lower section, corresponding to a deficit of $0 - \hat{p}$'s near $0 - \hat{p} = -.5$. That is, based on the simulation we would expect to have some observations with $y = 0$ corresponding to $\hat{p}$ from about .39 to .54 but there were none. The slight bulge and curvature in the upper part is due to the surplus of $y = 1$ observations, which results from the deficit of $y = 0$ observations in the specific region. These interpretations agree with the type of model misspecification built into this example. It is interesting that we have been able to detect a model deficiency caused by at most 15% of the 100 observations, and, furthermore, the misfitting is due to errors in the middle region of the fitted model rather than to extreme outliers.

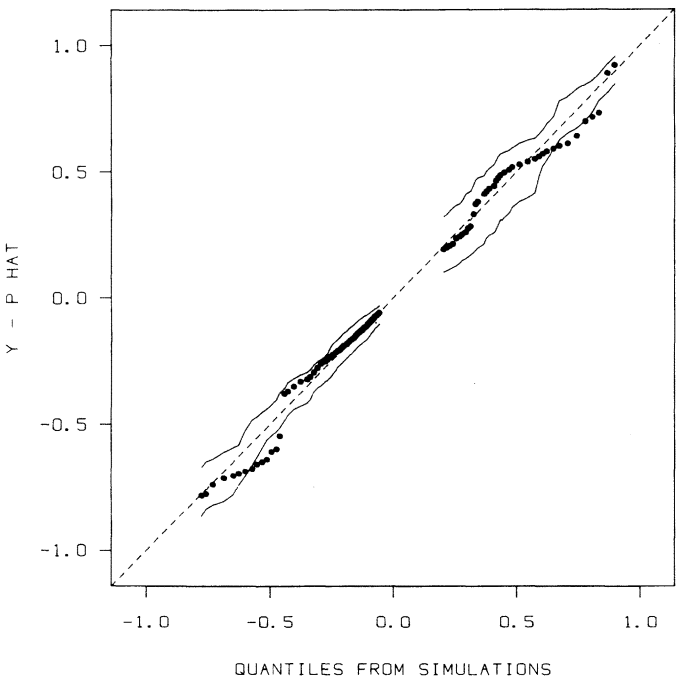

Figure 4. EMPIRICAL PROBABILITY PLOT FOR EXAMPLE 3. The solid and dotted lines are as in Fig. 3. Note the vertical gap in the lower section, where several points fall outside the confidence bands. This indicates a lack of fit in an interior region of the model. Specifically, there is a deficit of observations taking the value 0 in a region near $\hat{p} = .5$, compared with what would be expected based on simulations from the fitted model.

When one interprets an empirical probability plot, it is important to realize that all points and confidence bands must lie within $(-1, +1)$. This implies that the confidence bands will converge to the $Y = X$ line at the two ends of these plots, though we will not see this unless the fitted $\hat{p}$ get near 0 or 1. An outlier is constrained to lie within the plot so its distance from the $Y = X$ line might not be large, but it should lie outside the confidence band (as in Fig. 3) if it indeed corresponds to a significantly atypical observation. Other notions worth keeping in mind are that the confidence bands need not be symmetric about $Y = X$, nor of approximately uniform width. In addition, the bands will tend to be narrower in regions with many $\hat{p}$'s and wider where there are fewer $\hat{p}$'s.

The upper and lower confidence bands are crucial for detecting and interpreting lack of fit in empirical probability plots, both for the magnitude of extreme values and for the size of internal gaps. It is difficult to interpret the distance of particular points from the $Y = X$ line on an absolute basis, without the confidence bands. Note also that the confidence coefficients are local rather than global. The points in the plot are correlated, and we do not have a way to calculate a global coefficient. Even when the model is correct it is not surprising to have one or a few points near the edge of the confidence bands, so it can be difficult to distinguish slight lack of fit from adequate fit (as is also the case for other types of probability plots). Our view is that these plots should not be used as formal test statistics, but should rather be used to suggest one or several places where possibly the model is not fitting as well as it should.

In calculating the empirical probability plot we have slightly modified steps 7 and 9 in the recipe above. Rather than simply ordering the $(y^* - \hat{p}^*)$'s and proceeding as described earlier, we interpolate within the distribution $\{\hat{p}^*: y^* = 1\}$ to get the x-axis plotting positions for the upper part of the plot. Similar interpolations from $\{\hat{p}^*: y^* = 0\}$ are used to obtain plotting positions for $0 - \hat{p}$. This avoids having, for example, $1 - \hat{p}_i$ plotted against a negative value $0 - \hat{p}^*$ on the X axis, which would give a strange-looking point near the gap between the upper and lower portions that is not related to any deficiency in the fit. This would occur when there are unequal numbers of 1's within y and y*.

With this modification, the empirical probability plot is the same as two empirical quantile-quantile plots (Wilk and Gnanadesikan 1968) put on the same page. The top

section plots the $(1 - \hat{p})$ values against the $(1 - \hat{p}^*)$ values. This is the same as plotting the empirical quantiles from the set $\{-\hat{p}_i: y_i = 1\}$ against the empirical quantiles estimated from the sets $\{-\hat{p}_i^*: y_i^* = 1\}$. That is, the top part of the $(y - \hat{p})$ plot compares the distributions of $\hat{p}$'s for observation with $y = 1$ against the distribution of $\hat{p}^*$'s for observations with $y^* = 1$, and this second distribution is estimated from simulation. When the correct model is fitted to the data we should expect these two distributions to be about the same, so their empirical quantile-quantile plot should approximate a straight line through the origin with unit slope. Similarly, the bottom half of the plot can be thought of as comparing the distribution of $\hat{p}$'s for observations with $y = 0$ to the distribution of $\hat{p}^*$'s for observations with $y^* = 0$.

5. PARTIAL RESIDUAL PLOTS

After a specific linear regression model is fitted, it is worthwhile to examine relationships between the response and explanatory variables to see if improvements can be made. For example, nonlinear contributions might be required from variables currently included only linearly. Partial residual plots can help to detect the need for such improvements in linear regression, and this section adapts this technique to logistic regression. We first review the linear regression situation.

Consider data from the true model

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} + f(\mathbf{z}), \quad \text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}, \tag{5.1}$$

where $f(\mathbf{z})$ is some function of an explanatory variable Z. Often we might suspect that the model inadequacy is related to the variable Z without knowing what the function f is. We would like a plot that would suggest the form of f.

A naive approach would be to plot the residuals $\mathbf{y} - \hat{\mathbf{y}}$, against $\mathbf{z}$. To determine whether this is an appropriate plot for our purposes, the expected value of the ordinate of this plot is calculated under the true model (5.1). Using the notation $\mathbf{r}_{y|x}$ for $\mathbf{y} - P_x\mathbf{y}$ and P_x for orthogonal projection onto the columns of $\mathbf{X}$ (i.e., $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$), gives

$$\begin{aligned} E(\mathbf{r}_{y|x}) &= E(\mathbf{y} - P_x\mathbf{y}) \\ &= \mathbf{X}\boldsymbol{\beta} + f(\mathbf{z}) - P_x(\mathbf{X}\boldsymbol{\beta} + f(\mathbf{z})) \\ &= (I - P_x)f(\mathbf{z}). \end{aligned} \tag{5.2}$$

If this expectation were $f(\mathbf{z})$, then the relationship between the expected value of the ordinate and the abscissa $(\mathbf{z})$ would show the form of f, and looking at the actual plot might suggest a reasonable choice for the function f. Equation (5.2) shows, however, that this will not in general be the case, so this plot is not adequate for our purpose.

If f is linear, so that $f(\mathbf{z}) = \mathbf{z}\gamma$ in (5.1), Larsen and McCleary (1972) and Wood (1973) have argued that it is informative to plot the partial residual $\mathbf{r}_{\text{par}} = \mathbf{r}_{y|x,z} + \mathbf{z}\tilde{\gamma}$ against $\mathbf{z}$, where $\tilde{\gamma}$ represents the Gauss-Markov estimate of γ from the model fitting both $\mathbf{X}$ and $\mathbf{z}$. Then $E(\mathbf{r}_{\text{par}}) = \mathbf{z}\gamma$, assuming (5.1) with f linear. The expected value of the ordinate is linear in the abscissa $\mathbf{z}$ with slope γ, which is the desired property.

When $f(\mathbf{z})$ of model (5.1) is nonlinear, partial residual plots can be used to determine the form of f and to improve the model. In the nonlinear case $E(\mathbf{r}_{\text{par}})$ is generally not exactly $f(\mathbf{z})$, so accurate determination of the structure of f can be difficult; one cannot expect to do it perfectly even in the best of circumstances with no error (or equivalently, large sample size). A possible f may be worth pursuing even if it is only imperfectly suggested by the plot.

A partial residual plot in logistic regression can be found by exploiting the relationship between logistic regression and weighted linear regression. The iteration formula for $\hat{\boldsymbol{\beta}}$ at the $(t + 1)$st iteration, Equation (2.2), can be rewritten as

$$\hat{\boldsymbol{\beta}}(t + 1) = (\mathbf{X}^T\mathbf{V}(t)\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}(t)\mathbf{y}^*(t), \tag{5.3}$$

where

$$\mathbf{y}^*(t) = \mathbf{X}\hat{\boldsymbol{\beta}}(t) + \mathbf{V}^{-1}(t)\mathbf{r}(t), \tag{5.4}$$

and $\mathbf{r}(t) = \mathbf{y} - \hat{\mathbf{p}}(t)$ (Nelder and Wedderburn 1972). The analogy between expressions (5.3), (5.4), and weighted linear regression estimation suggests that, in terms of the logistic fit, $\mathbf{y}^*$, $\mathbf{X}\hat{\boldsymbol{\beta}}$, and $\mathbf{V}^{-1}\mathbf{r}$ can be thought of, respectively, as observation, fitted value, and residual. (Of course, logistic "observations" cannot be formed directly since they would involve log(0) and log(∞).) This analogy is therefore useful for binary data in order to construct partial residuals.

Consider the augmented model $\text{logit}(\mathbf{p}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}\gamma$. A maximum likelihood fit of this model leads to $\tilde{\boldsymbol{\beta}}$, $\tilde{\gamma}$, $\tilde{p}$, and $\mathbf{V}$, and hence to the "logistic observations"

$$\mathbf{y}^* = \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{z}\tilde{\gamma} + \mathbf{V}^{-1}(\mathbf{y} - \tilde{\mathbf{p}}).$$

Application of the linear regression material gives the logistic partial residual

$$\begin{aligned} \mathbf{r}_{\text{par}} &= \mathbf{V}^{-1}\mathbf{r} + \mathbf{z}\tilde{\gamma} \\ &= (\mathbf{y} - \tilde{\mathbf{p}})/(\tilde{\mathbf{p}}(1 - \tilde{\mathbf{p}})) + \mathbf{z}\tilde{\gamma}. \end{aligned}$$

Note that $\mathbf{r}_{\text{par}}$ involves standardizing $\mathbf{r}$ by $\tilde{\mathbf{p}}(1 - \tilde{\mathbf{p}})$ rather than by $\sqrt{\tilde{\mathbf{p}}(1 - \tilde{\mathbf{p}})}$, which might be more natural in other contexts. The partial residual plot is obtained by plotting $\mathbf{r}_{\text{par}}$ against $\mathbf{z}$.

Figure 5 is the partial residual plot for an example that is described further later (ignore the solid line for now). Note that the points fall into two separate clouds. This is because one cloud corresponds to observations with $y_i = 1$, the other cloud to $y_i = 0$. The expected value of the partial residual may approximate the linear or nonlinear structure $f(\mathbf{z})$, but this is obscured by the binary nature of $\mathbf{y}$.

An estimate of the expected ordinate that is less discrete can be found by smoothing the plot. Example 4 uses Cleveland's (1979) nonrobust locally weighted regression (i.e., zero iterations). In our experience smoothing is cru-

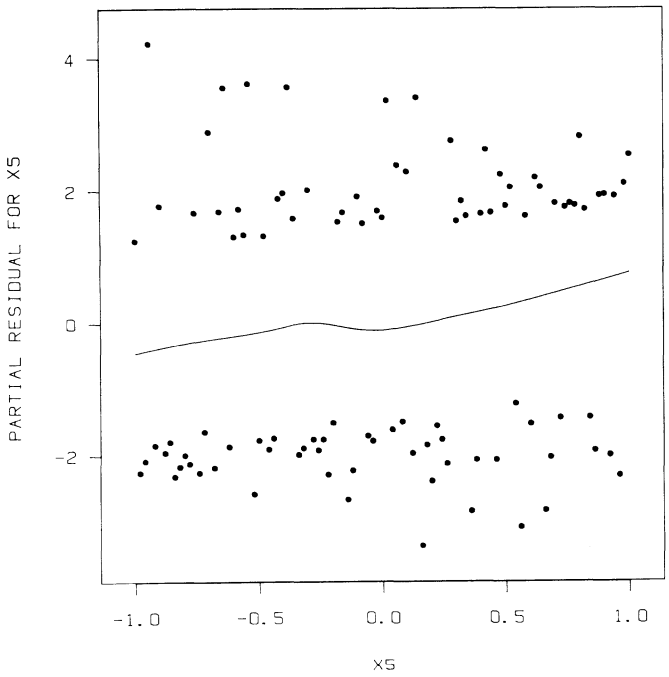

Figure 5. PARTIAL RESIDUAL PLOT FOR EXAMPLE 4: VARIABLE X_5. The smooth for these points is the solid line. Here it is reasonably straight and does not suggest any nonlinear transformation for X_5.

cial for these plots, as will be shown by the following examples and in Section 6.

Example 4. One hundred one observations were generated according to the model logit(P) = $-1 + X_5 + X_6 + 2X_6^2$, where $X_5 = -1(.02)1$ and the elements of X_6 were chosen iid $U(-1, 1)$. Presuming that we do not know

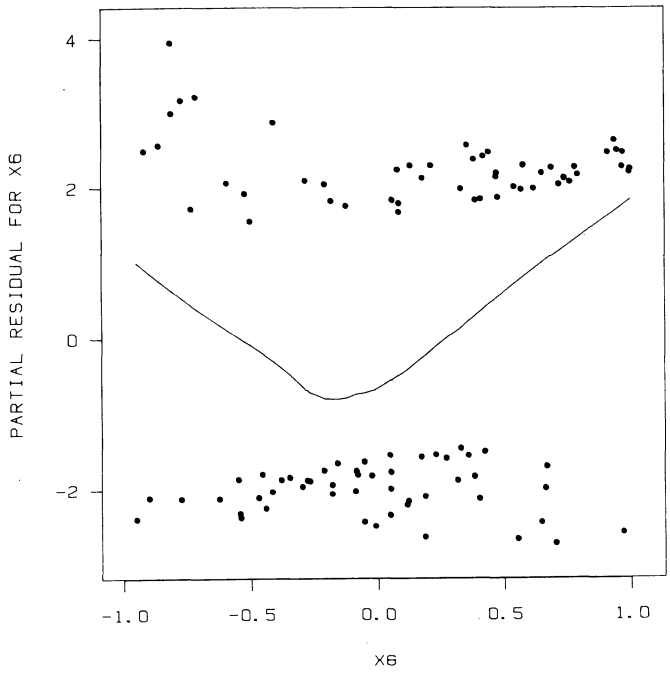

Figure 6. PARTIAL RESIDUAL PLOT FOR EXAMPLE 4: VARIABLE X_6. Here the smooth is definitely nonlinear and suggests a transformation such as $(X_6)^2$ or $|X_6|$.

the appropriate dependence on X_5 and X_6, we fit the model linear in X_5 and X_6 (D = 131.95 on 98 df), and calculate the partial residual plots for X_5 and X_6, Figures 5 and 6, respectively. Note that only one fit is required to construct both plots. Figure 5 suggests a basically linear dependence on X_5. The smooth in Figure 6, however, suggests that the dependence on X_6 is nonlinear and that an additional term such as X_6^2 or $|X_6|$ should be included in the model.

6. A REAL EXAMPLE

A study conducted between 1958 and 1970 at the University of Chicago's Billings Hospital concerned the survival of patients who had undergone surgery for breast cancer (Haberman 1976). There are 306 observations on four variables:

y_i = 1 if patient i survived 5 years or longer
0 otherwise;

x_{i1} = age of patient i at time of operation;

x_{i2} = year of operation for patient i (minus 1900);

x_{i3} = number of positive axillary nodes detected in patient i.

Initially, we entertain the model

$$\text{logit}(P) = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3,$$

where P represents the probability of at least five-year postoperative survival.

The local mean deviance plot, Figure 7, provides a way to examine the overall fit of the initial model. The horizontal dotted line is the mean deviance of the initial model, which is 1.087 from D = 328.25 on 302 df. (The horizontal dashed line refers to a model discussed later.) The peak near the left is based on only a few clusters and degrees of freedom, so it is of no special significance. As degrees of freedom increase the local mean deviance values decrease until they stabilize around 1.0. Moreover, the steady increase in group level from .65 to .85, followed by a more rapid increase, suggests using local mean deviance estimates corresponding to degrees of freedom from about 150 to 190. This suggests that the initial model does not fit the data as well as one would like.

To improve the model we next examine partial residual plots for each of the explanatory variables. Figures 8 and 9 display the partial residual plots for X_1 and X_3. For X_1 the smooth is mildly nonlinear and resembles a cubic function. For X_3 two smooths are used, one nonrobust and one robust. This is done to protect our interpretations of the plot against the effects of extreme points on the right of the plot. Both smooths definitely appear nonlinear. The robust smooth (dashed line) suggests that the upward swing of the nonrobust smooth is due entirely to the five observations with $X_3 > 27$, which are substantially larger than the rest. Taken together, this suggests as a first step trying a reciprocal transformation for X_3.

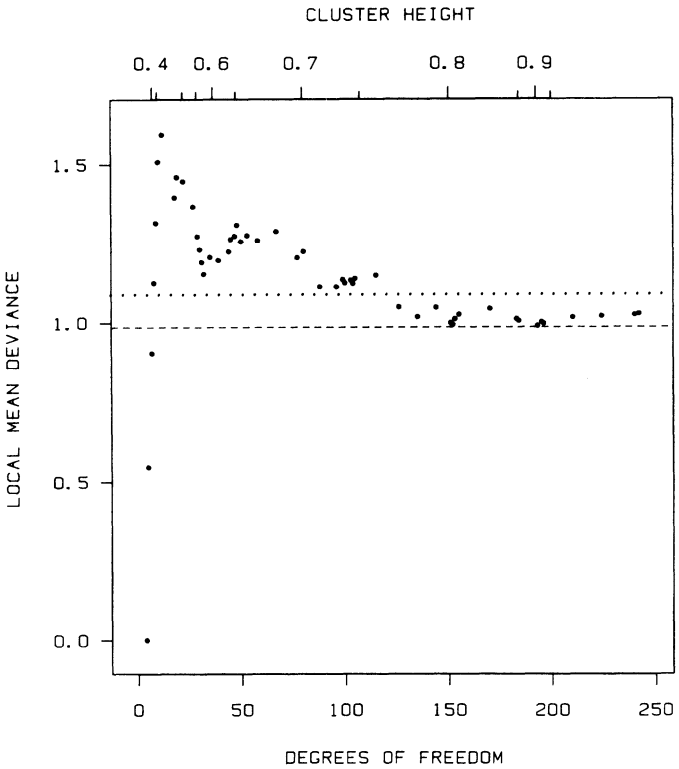

Figure 7. LOCAL MEAN DEVIANCE PLOT FOR BREAST CANCER EXAMPLE. The dotted line at Y = 1.087 is the mean deviance from the initial three-variable model. This line is above the points for moderately large degrees of freedom, indicating some lack of fit for this model. The dashed line at Y = .984 is the mean deviance from the six-variable model omitting one outlying observation. No lack of fit is indicated as this line passes near the points for moderately large degrees of freedom.

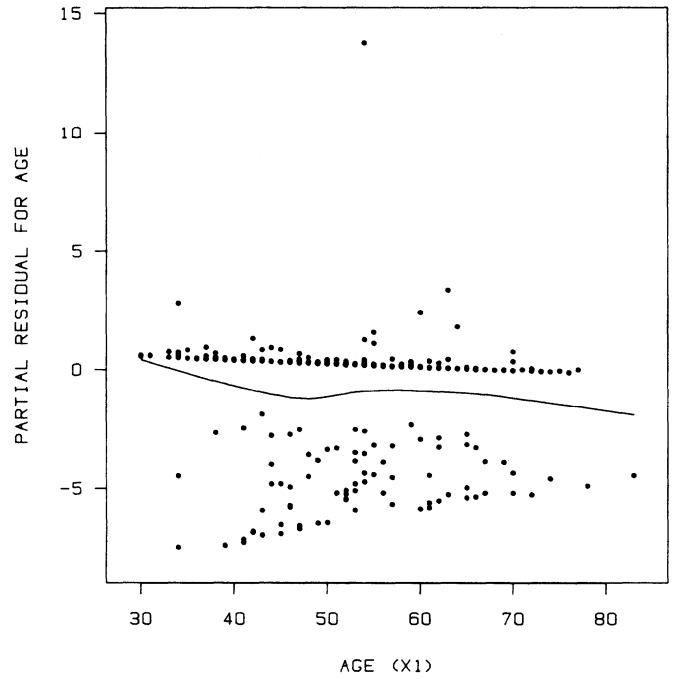

Figure 8. PARTIAL RESIDUAL PLOT FOR BREAST CANCER EXAMPLE: AGE. The smooth for these points is nonlinear. A cubic function of age is suggested.

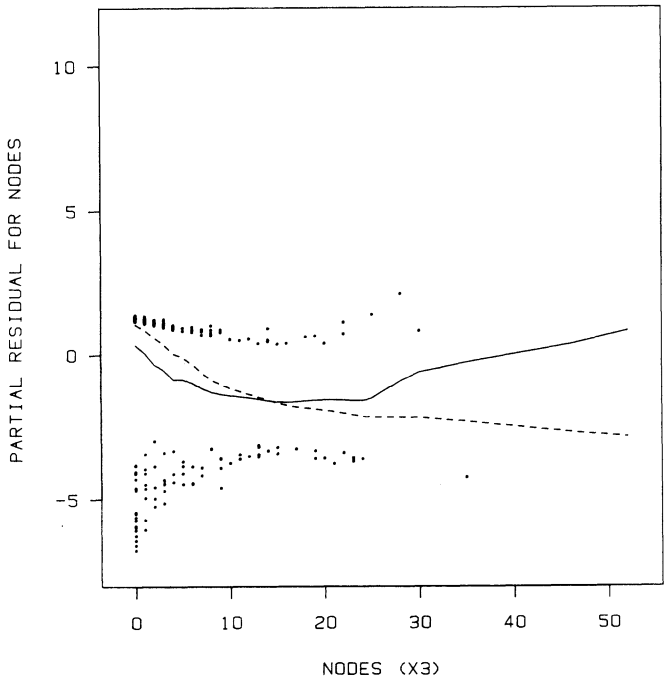

Figure 9. PARTIAL RESIDUAL PLOT FOR BREAST CANCER EXAMPLE: NODES. Two smooths are plotted for this variable, a nonrobust smooth (solid line) and a robust smooth (dashed line). The primary difference between these smooths is that the former is nonmonotone whereas the latter is monotone. This difference tends to discount a quadratic function in nodes. The overall shape of the robust smooth indicates that a reciprocal transform may be useful.

Fitting a model including X_1, X_2, and $1/(1 + X_3)$ (avoiding division by zero) gave a partial residual plot for $1/(1 + X_3)$ that was more linear than before but still noticeably nonlinear. It appeared that the reciprocal overcorrected the nonlinearity shown in Figure 9 and that a weaker transformation was needed. Figure 10 shows the resulting partial residual plot for $\log[1/(1 + X_3)] = -\log(1 + X_3)$. The smooth appears reasonably linear. Continuing in this way to examine models using these variables and their interactions led us to the model

$$\text{logit}(P) = \beta_0 + X_1\beta_1 + X_1^2\beta_2 + X_1^3\beta_3 + X_2\beta_4 + X_1X_2\beta_5 + \log(1 + X_3)\beta_6,$$

giving mean deviance of 1.011 (D = 302.33 on 299 df).

The empirical probability plot from this model is displayed in Figure 11, using 25 simulations and 84.6% confidence bands. The lowermost point is outside the band and suggests that there may be an outlier. This observation corresponds to a 34-year-old patient with no positive axillary nodes, but who died within five years of her operation. Her fitted probability of survival is .986. In contrast, of nine other patients of about the same age who also had no positive axillary nodes, none died.

It is of interest to refit the six variable model omitting this observation. This gives mean deviance of .984(D = 293.09 on 298 df). The estimated coefficients and standard

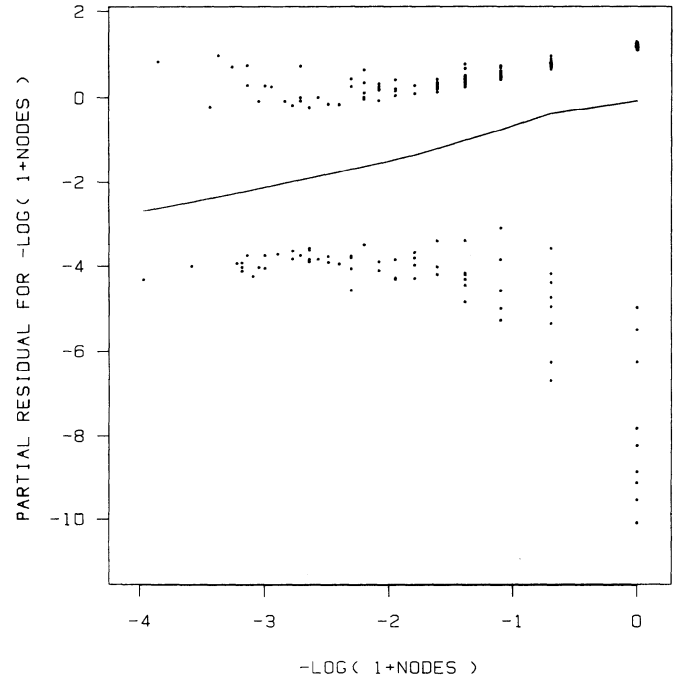

Figure 10. PARTIAL RESIDUAL PLOT FOR BREAST CANCER EXAMPLE: − LOG(1+NODES). The smooth is reasonably linear, indicating that the logarithmic transform was successful in removing the nonlinearity observed in Fig. 9.

errors for this model are

$$\text{logit}(P) = \underset{(.26)}{1.684} + \underset{(.028)}{.03315}\,Z_1 + \underset{(.0017)}{.003728}\,Z_1^2 - \underset{(.00011)}{.0002945}\,Z_1^3 - \underset{(.045)}{.01129}\,Z_2 + \underset{(.0049)}{.01374}\,Z_1Z_2 - \underset{(.13)}{.7916}\log(1 + X_3),$$

where $Z_1 = X_1 - 52$, $Z_2 = X_2 - 63$. The dashed horizontal line in Figure 7 corresponds to mean deviance from this model. This line lies very near the lowest points on the plot, indicating a global fit about as good as that from the local model. Without the aid of graphical methods, Haberman (1976) suggested a model including X_1, X_2, X_3, X_3^2, and X_1X_2, which gives mean deviance $1.046(D = 313.86$ on 300 df).

For our six-variable model used in the previous paragraph, Figure 12 gives the fitted probability of survival plotted against age (X_1) for the case of 0 nodes ($X_3 = 0$)—solid line—and 20 nodes ($X_3 = 20$)—dashed line. These calculations correspond to the mean year of operation in these data, which is 1963. The effect of 20 nodes is large, and the nonlinearity and nonmonotonicity in age is also apparent.

The observed effect of X_1 (age) in the probability of survival is described in our model by a cubic function. Based on this formulation, our analysis predicts decreasing survival probability from age 30 to 49, then increasing

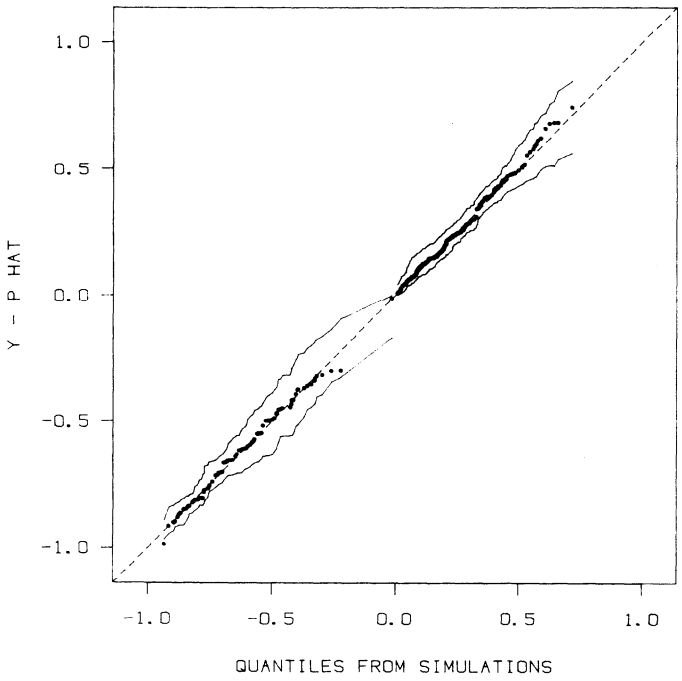

Figure 11. EMPIRICAL PROBABILITY PLOT FOR BREAST CANCER EXAMPLE. All points lie within the confidence bands (based on 25 simulations) except the most extreme point in the lower left corner.

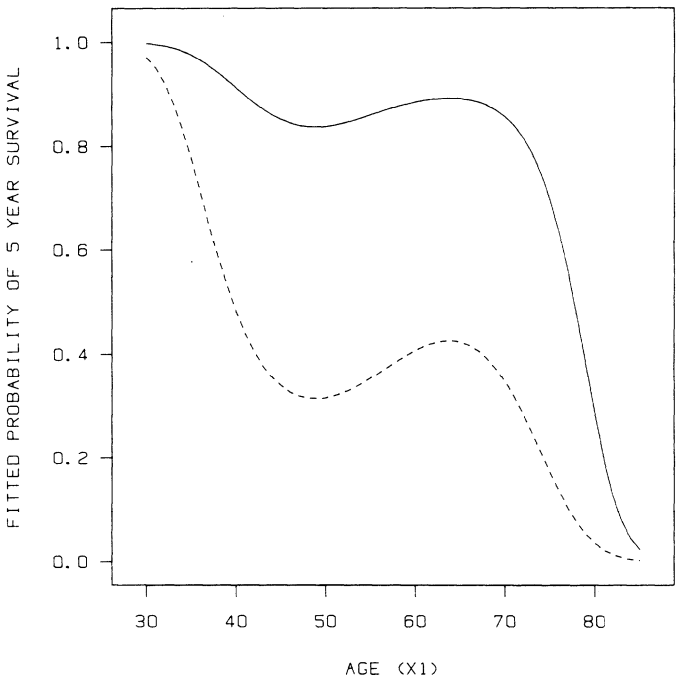

Figure 12. FUNCTION PLOT OF FITTED SURVIVAL PROBABILITY. The solid line is the estimated 5-year survival probability as a function of age, for a woman with no observed nodes. The dashed line is the estimated 5-year survival probability as a function of age, for a woman with 20 observed nodes. Both curves were computed for an operation performed in 1963. The cubic dependence on age is highlighted.

survival probability until age 63, followed by decreasing survival probability, as seen in Figure 12. This nonlinear behavior is similar to age-specific *incidence* rate curves for breast cancer in Western countries. In the epidemiological literature this phenomenon is called Clemmensen's Hook (Farewell 1979). We are aware that postoperative survival and age-specific incidence are two entirely different aspects of breast cancer, but the similarities in the curves are too great to pass without comment.

7. DISCUSSION

Because of the extreme discreteness of binary data, any progress towards understanding the fit of a particular logistic regression model must be based on pooling or grouping of some sort. This is one way in which the proposed methods are related, and it gives a general context for considering them. Local mean deviance plots explicitly involve grouping using the rows of **X**. Smoothed partial residual plots explicitly group using the variable Z. Empirical probability plots can be thought of as a more sophisticated version of grouping on $\hat{p}$ and comparing the observed proportions of successes with the fitted proportions. We now make a few more specific comments on the use of each of these methods.

The local mean deviance plot uses neighboring points, so the notion of which points are "close" is crucial. This is related to several specific details: the metric used for defining distances between rows of **X**; the choice of clustering algorithm; and, for a hierarchical clustering algorithm, the height at which the resulting cluster tree is cut to give a set of groups. As in any application of clustering the best choices depend on the underlying structure, which is unknown, and no panacea is available. We recommend the following. To calculate distances, use only those columns of **X** corresponding to linear effects of separate variables; that is, if **X** contains X_1, $\log(X_1)$, X_2, X_1X_2, X_2^2, use only X_1 and X_2. Standardize these variables, use Euclidean distance, and cluster using the complete-linkage hierarchical clustering algorithm (also called maximum distance or furthest neighbor; see Everitt 1974). Although one might want to compute distance using the globally fitted model (Daniel and Wood 1971), we prefer to obtain groups using a procedure less related to the specific hypothesized model when estimating local variability. The complete-linkage clustering algorithm is known to have a tendency to give spherically shaped clusters, which seem appropriate here. Now consider the number of clusters to use, which is determined by the height of the hierarchical tree at which the groups are obtained. The trade-offs are that with more groups each group is smaller on average and thus better represents only local variability; on the other hand, with more groups there will be more singleton groups, which do not contribute anything to the assessment of local variability. Our recommendation is to cut the clustering tree so that about 5% of the points (i.e., rows of **X**) fall into singleton groups. The remaining 95% of the points fall in larger clusters and help estimate local variability, but having 5% as singletons ensures that there will still be a relatively large number of groups. The examples in Sections 3 and 6 were constructed using these guidelines.

The local mean deviance plot is especially useful for detecting a missing interaction term (e.g., X_1X_2), since this is the kind of misspecification that can give, in certain regions, large differences between $\hat{p}$'s from the global model and $\hat{p}$'s from the local model. For misspecifications involving a specific variable (e.g., $\log(X_1)$ instead of X_1), partial residual plots are generally more useful.

A traditional way for checking goodness of fit with logistic regression is to first partition the $\hat{p}$ values into intervals such as (.00, .10), (.10, .20), . . . , (.90, 1.00). Then the observed proportions of successes in the intervals are compared with the expected proportions, as estimated by the model, using a chi-squared statistic. The empirical probability plot can be viewed as a sophisticated, graphical version of this procedure, since both use the underlying idea of comparing an observed distribution with an expected distribution from a fitted model. In the empirical probability plot, however, no formal grouping on $\hat{p}$ is required, the expected distribution is obtained from a simulation rather than from asymptotic theory, and comparisons are made visually rather than with a test statistic. These simulations are in the same spirit as the bootstrap (Efron 1979), which produces a standard deviation for some parameter by simulating from the data. Our goal, however, is to find possible outliers in the data or misspecifications of the model. To do this we simulate from the fitted model and make comparisons with the data. A related question concerns the number of simulations required to obtain stable confidence bands; in our experience at least 25 simulations are required, and this number is computationally reasonable.

Empirical probability plots are also related to probability plots of residuals in linear regression. Both plots are useful for identifying outliers, but empirical probability plots can serve broader purposes in logistic regression. In linear regression a probability plot with nonlinear shape is generally taken to imply that the error distribution is not modeled correctly. One does not hope to learn anything about the adequacy of the $\mathbf{X}\hat{\boldsymbol{\beta}}$ term from a probability plot. In logistic regression the situation is different. Because of the discrete nature of the response with expectation p and variance $p(1 - p)$, adequate modeling of the entire error distribution is inherently bound up with adequate modeling of the expectation term $\mathbf{X}\boldsymbol{\beta}$. Thus, differences that appear in an empirical probability plot from logistic regression might be related to an inadequacy in $\mathbf{X}\hat{\boldsymbol{\beta}}$, for instance in Example 3.

In examining an empirical probability plot, a key feature is the presence of a vertical gap together with points on at least one side of the gap falling outside the confidence bands. Such situations occur in Figures 3 and 4. This lack of $y - \hat{p}$ values in a region where some would be expected can indicate a problem with y (outlier) or a

problem with $\hat{p}$ (model inadequacy). In a related but different situation, suppose that for 10 observations with moderate $\hat{p}$, say $.10 < \hat{p} < .15$, there are 5 observations with $y = 1$. Taken singly, none of these points would be considered an outlier. Taken as a group, however, we would expect only one or two of the observations to give $y = 1$. This would show up on the empirical probability plot as a vertical gap to the left of the fifth point, which should lie above the confidence band. Taken together, these five points can be interpreted as outliers from this model, though taken separately none is an outlier.

The usefulness of partial residual plots depends on the magnitude of the effect of Z on the response probability. Important effects are likely to be detected whereas marginal effects are not. In practice, we recommend including the hypothesized nonlinear term $f(Z)$ in a new model, refitting, and plotting the partial residual for $f(Z)$. This should be continued until the smoothed plot appears linear for each variable in the model.

A smooth is crucial for interpreting partial residual plots. We recommend routine plotting of both a nonrobust and a robust smooth, as in Figure 9. The extra computational cost is small. Often there is not much difference between the two smooths. We suspect, however, that a nonrobust smooth may be more powerful for detecting small but real misspecifications, as it is calculating something like a local mean rather than a local median. On the other hand, a nonrobust smooth can be more affected by outliers. If few possible outliers appear, it is useful to compare the robust and nonrobust smooths to see to what extent these points might have affected the smoothed structure in the plot. A further aid in the interpretation of partial residual plots would be the addition of confidence bands about the smooth. Two methods seem feasible, (a) linear model theory methods based on the local linear smooth, and (b) methods based on the simulations used in generating the empirical probability plots. To date, these enhancements have not been explored.

All the methods introduced in this paper can be easily modified to handle binomial rather than binary responses. For small values of np (say <5) the methods presented are likely to be extremely useful. For larger values of np the discreteness problem tends to be less pronounced, suggesting that linear model methods can be used. For general multinomial and ordered multinomial data, we expect that extensions of our methods will be useful in assessing fitted regression-type models.

In conclusion, note that it is feasible to implement the proposed techniques in most computing environments. The main software requirements are a program for fitting the logistic model and a flexible plotting package. Much of our initial experience came from printer plots, although the versions shown here were produced using a flatbed plotter. Two required support programs are clustering and smoothing algorithms.

[*Received August 1981. Revised July 1982.*]

REFERENCES

BAKER, R.J., and NELDER, J.A. (1978), "The GLIM System—Release 3," distributed by Numerical Algorithms Group: Oxford.

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

CLEVELAND, W.S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

COX, D.R. (1970), *The Analysis of Binary Data*, London: Methuen.

DANIEL, C., and WOOD, F.S. (1971), *Fitting Equations to Data*, New York: John Wiley.

EFRON, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1–26.

EVERITT, B. (1974), *Cluster Analysis*, New York: John Wiley.

FAREWELL, V.T. (1979), "Statistical Methods and Mathematical Models For Research in Breast Disease," *Commentaries on Research in Breast Disease*, 1, 193–232.

HABERMAN, S.J. (1976), "Generalized Residuals for Log-Linear Models," *Proceedings of the 9th International Biometrics Conference, Boston*, 104–122.

——— (1979), *Analysis of Qualitative Data* (Vol. 2), New York: Academic Press.

LARSEN, W.A., and McCLEARY, S.J. (1972), "The Use of Partial Residual Plots in Regression Analysis," *Technometrics*, 14, 781–790.

NELDER, J.A., and WEDDERBURN, R.W.M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society* (Ser. A), 135, 370–384.

PREGIBON, D. (1981), "Logistic Regression Diagnostics," *Annals of Statistics*, 9, 705–724.

WEISBERG, S. (1980), *Applied Linear Regression*, New York: John Wiley.

WILK, M.B., and GNANADESIKAN, R. (1968), "Probability Plotting Methods for the Analysis of Data," *Biometrika*, 55, 1–17.

WOOD, F.S. (1973), "The Use of Individual Effects and Residuals in Fitting Equations to Data," *Technometrics*, 15, 677–695.