Smoothed Residual Plots for Generalized Linear Models

Rollin Brant *

University of Minnesota
School of Statistics

Technical Report #450

April 1985

University of Minnesota
School of Statistics
Department of Applied Statistics
St. Paul, Minnesota 55108

Supported by University of Minnesota Graduate School
Faculty Summer Research Fellowship, 1984.

Smoothed Residual Plots for Generalized Linear Models

Abstract

Methods for examining the viability of assumptions underlying generalized linear models are considered. By appealing to the likelihood, a natural generalization of the raw residual plot for normal theory models is derived and is applied to investigating potential misspecification of the linear predictor. A smoothed version of the plot is also presented, as are asymptotic standard errors appropriate for assessing the significance of apparent deviations from the presumed model.

Key Words: Generalized linear models; Lack of fit; Diagnostic procedures; Graphical methods

1. INTRODUCTION

The aim in this paper is to consider the development of procedures appropriate for the examination of the assumptions relevant to the application of generalized models. Methods derived from techniques applicable to the normal linear model have provided a useful starting point in past investigations. However, the diversity encompassed by the generalized linear model (hereafter GLM) demands that care be taken in delineating the aims of general purpose diagnostic procedures; only with clearly stated goals is it possible to verify that proposed procedures are reliable in the broad context. The claim here is that only limited goals can be served by a truly general methodology, and proposals are put forward to that end.

Uninitiated readers will find in McCullagh and Nelder (1983) a thorough introduction to the GLM, the basic characteristics of which are outlined below. The GLM has found widespread application as a flexible foundation for modelling response-explanatory relationships in a general setting. The data is taken to consist of independent observations on a response, y, together with covariates or explanatory variables, described by a p-vector $\underset{\sim}{x}$. The responses are assumed to have distributions belonging to an exponential family, with log densities

$$\ell(y;\theta,\phi) = a(\phi)^{-1}\{y\theta - b(\theta)\} + c(y,a(\phi))$$

The covariates, $\underset{\sim}{x}$, enter above through θ, which is taken to be a function of the linear predictor, $\eta=\underset{\sim}{\beta}^T\underset{\sim}{x}$, where $\underset{\sim}{\beta}$ is a p-vector of unknown parameters. ϕ is a nuisance parameter referred to as the dispersion.

Inference is simplest when $\eta=\theta$, the natural parameter. To allow for greater scope, however, it is convenient to reparameterize in terms of

$\mu = E(y) = b'(\theta)$, and to specify the model by a link function, g, which satisfies $g(\eta) = \mu$. When $g^{-1} = b'$, so that $\eta=\theta$, the link is termed canonical. To avoid notational complexity in relating the three parameterizations, we shall adopt a natural, if potentially ambiguous, notation for the various functional relationships, writing $\mu=\mu(\theta)$ or $\mu=\mu(\eta)$, $\eta=\eta(\theta)$ or $\eta=\eta(\mu)$, etc. as convenience dictates. Primes (') will denote differentiation with respect to the indicated argument.

The paradigmatic GLM is the normal linear model (NLM) given by $y=\underset{\sim}{\beta}^T\underset{\sim}{x}+\varepsilon$, where ε is $N(0,\phi/w)$. Here, as in many GLM's, $a(\phi) = \phi/w$, where w is a weight which may vary with observations. In other cases ϕ may be superfluous, as in the logistic regression model for binary data, which specifies Bernoulli responses with $\mu=p=P(y=1)$ and the canonical link $\eta = \log(p/1-p)$.

We shall restrict attention here to the case that $a(\phi)$ is constant over all observations; subsequent results generalized in a straightforward but notationally burdensome fashion. Given this restriction, the log likelihood function based on a sample, $(\underset{\sim}{x}_i, y_i)$, $i=1,\ldots,n$ takes the form

$$\ell_n = a(\phi)^{-1} \sum_{i=1}^{n} \{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^{n} c(y_i, a(\phi)) ,$$

and the maximum likelihood equations for $\underset{\sim}{\beta}$ are given by

$$\sum_{i=1}^{n} \{y_i - \mu_i\}\theta'(\eta_i)x_{ij} = 0 \qquad (j=1,2,\ldots,p)$$

where the x_{ij}'s are the elements of the matrix

$$X = (\underset{\sim}{x}_1 | \underset{\sim}{x}_2 | \ldots | \underset{\sim}{x}_n)^T = (x_{ij}) \quad .$$

The equations above, together with the associated asymptotic distribution theory form the conventional basis for inference regarding $\underset{\sim}{\beta}$.

An iterative solution to the joint equations above can be achieved in terms of weighted least squares calculations by defining

$$\tilde{y} = \tilde{\eta} + \{y - \tilde{\mu}\}\eta'(\tilde{\mu}) \ ,$$

where $\tilde{\mu}$ and $\tilde{\eta}$ derive from a preliminary estimate of $\underset{\sim}{\beta}$. An updated estimate is obtained by regression $\tilde{y}$ on $\underset{\sim}{x}$ with weights proportional to $\{\eta'(\mu)^2 \ \mathrm{var}(y)\}^{-1}$, the fully iterated estimate being the maximum likelihood solution.

2. DIAGNOSTIC AIMS

Given the formal resemblance of the GLM to the NLM, and the fact that GLM calculations can be viewed as iterative NLM calculations, it is natural to attempt the extension of NLM motivated procedures to GLM's as a whole. It is imperative, however, to first consider in some detail the woes which can afflict GLM's in general.

The aims of diagnosis in the NLM are firstly, the identification of unusual observations, and secondly, the detection of systematic deficiencies in the model. The first aim subsumes both the identification of outliers, as well as influential observations, observations whose impact on some aspect of the overall fit is large. The second aim addresses the potential for global misspecification of the model with regard to

i) the marginal distributions for the response

ii) the dependence structure of the response

iii) the specification of covariates in η.

With regard to GLM's, in general, the addition of

iv) the functional form of g, the link,

seems to complete the list of potential aberrations of interest. It is important to consider the implications of these various issues as they depend on specific GLM's. For the sake of concreteness, we will pay particular attention to the logistic regression model (LRM), which in some sense opposes the NLM on the extremes of the GLM "spectrum".

2.1 Outliers and influential observations

By outlier one generally means an isolated observation whose behavior is not in accordance with the model. While any data collection is susceptible to the presence of deviant observations, in the construction of diagnostics one is restricted to consider those discrepancies detectable by formal means. In particular, any usable characterization of such behavior will depend largely on the assumed model. With regard to GLM's, the natural course is to consider the possibility of isolated cases of misspecification for either the predictor, η, or the dispersion ϕ, corresponding to the mean shift and variance shift formulations in the NLM. Perturbations in the covariate observations are accounted for in the first instance.

Natural though the above may be, it is not universally applicable, the LRM being a case in point. Taking for simplicity a model where η is given, it is clear that only when $p = \exp(\eta)/\{1 + \exp(\eta)\}$ is near 0 or 1 is it possible to observe identifiably discrepant responses, and that the misspecification of η will be otherwise undetectable. The situation is essentially unaltered in

extended cases, for the inclusion of covariates and associated parameters only obscures the issue, while the existence of replicates allows for diagnosis of essentially systematic deficiencies, only. This case contrasts sharply with the NLM, and calls into question the universal utility of any outlier identification scheme following directly from methods developed for the NLM.

The detection of potentially influential observations presents a wholly different situation, since influence can be viewed as a problem of numerical stability. That fact that for GLM's, in general, contributions to the likelihood of individual observations depends largely on the linear form $\eta=\underset{\sim}{\beta}^T x$ implies that many of the ideas regarding influence in the NLM translate naturally to the general context, modulo some added computational complexity. Pregibon (1980) provides a useful approach along these lines.

2.2 Distributional misspecifications

The serious consequences of departures from independence in the NLM is not paralleled by a well developed methodology for detecting dependency, though certain structures, such as ordered dependency can be investigated. Similar remarks would appear to GLM's. The misspecification of the marginal distributions is an issue that has received more extensive treatment. Again, however, consideration of the LRM calls into question the value of a general attack, since it is not clear what sorts of departure from Bernoulli variation can sensibly be addressed. On the other hand, in many GLM's the inevitably approximate nature of the specified form for $\ell(y;\theta,\phi)$ must be acknowledged, and in some instances modification of the familiar procedures as quantile plots for residuals may be helpful. The difficulty here resides in relating the behavior of residuals, however they are defined, to the assumed distributional form; in the NLM the residuals are relatively easy to relate to

the "true" errors while in other models such an identification may be less useful.

2.3 The link and linear predictor

The important unification achieved by the GLM is its ability to describe a diversity of relationships in essentially linear fashion, and to this end, the validity of the specified form for the link and linear predictor are vital. Owing to problems of identifiability, it is not entirely natural to make a sharp distinction between misspecification of g and $\eta=\underset{\sim}{\beta}^T\underset{\sim}{x}$, especially in the case of a single covariate. This inherent ambiguity permits a degree of latitude with regard to the characterization of potential misspecifications. Experience with the NLM indicates that the most fruitful approach is through respecification of the predictor, rather than the link. This is due largely to the significant advantages in computational, inferential and interpretive ease offered by the canonical identity link. Similarly in applying other GLM's, the canonical link is the sensible preliminary choice in nearly all instances, and faced with an apparent need for respecification, it will generally be more convenient to consider respecifying the form of the covariates in η than to sacrifice the simplicity of the canonical form of g.

Thus, the development of methods for assessing the validity of proposed forms of the linear predictor, $\eta=\underset{\sim}{\beta}^T\underset{\sim}{x}$, is practically important, especially in connection with the use of canonical links. The feasibility of a more or less general diagnostic approach is considered in the sequel.

3. ASSESSING THE LINEAR PREDICTOR

A natural framework for the investigation of misspecification of the predictor in the NLM carries over directly to the GLM: we consider the

possibility of improving the model by augmentation of η to $\eta^{\alpha} = \underset{\sim}{\beta}^T\underset{\sim}{x} + \alpha(z)$, where z is an additional covariate, possibly dependent on $\underset{\sim}{x}$, and $\alpha(z)$ is a function of unspecified, but ostensibly simple, form. This weak specification warrants a flexible approach, and graphical methods are especially appropriate. In the case of the NLM, methods based on residuals may be motivated by noting that $e = y - \hat{\underset{\sim}{\beta}}^T\underset{\sim}{x}$ can be regarded as an estimate of $\alpha(z)$. Of course $E(e) \neq \alpha(z)$ in most cases, unless $\underset{\sim}{\alpha} = (\alpha(z_1),\ldots,\alpha(z_n))^T$ is orthogonal to the columns of X, hence the need for modifications such as the added variable and partial residual plots. In addition, case deletion based diagnostics arise naturally when z is a case indicator, equalling 1 for case i and 0 otherwise (Cook and Weisberg, p. 20).

If a generally applicable definition for residuals can be constructed in accordance with the above characterization, extensions of NLM type residual plots are feasible. To this end we begin by considering a possibly misspecified GLM with η given and let $\mu^{\alpha} = \mu(\eta + \alpha(z))$ denote the true value of E(y). The natural starting point is the raw residual, $y-\mu$, which estimates $\mu^{\alpha}-\mu$. Since the aim is inference about $\alpha(z)$, the information in the raw residual must be translated to the scale of linearity defined by the link. The naive choice of transformation is to take $g(y) - g(\mu)$; this, however, is not always well defined, as in the LRM though when it exists, it does have the appeal of being maximum likelihood (in the absence of replicate z's). A Taylor series expansion of $g(\mu^{\alpha})$ around μ yields $\alpha \approx (\mu^{\alpha}-\mu)g'(\mu)$, motivating the use of the locally linearized residual, $r = (y-\mu)\{\mu'(\eta)\}^{-1}$, which is essentially the form applied by Landwehr, et al (1984) in defining residual plots for the LRM. The fact that in the LRM, r is distributed on $\{p^{-1},(1-p)^{-1}\}$ can give rise to practical difficulty, especially when one attempts visual interpretation of such plots as r vs. z (see Section 8.2).

Landwehr, et al attempt to mitigate this difficulty by a more or less ad hoc form of smoothing that fails to account for the sampling behavior of the r's, notably the inhomogeneous variance, $\mathrm{Var}(r) = \{p(1-p)\}^{-1}$.

A more formal approach is motivated by reference to the NLM when one notes that in the absence of z replicates, $e = y-\mu$ is the maximum likelihood estimate of $\alpha(z)$, at least in the fully specified case considered here. In the replicated case, the m.l.e. is

$$\hat{\alpha}(z) = \underset{z_i=z}{\mathrm{Ave}}\{e_i\} ,$$

which is arguably the dominant perception conveyed by the ordinary residual plot. Accordingly one can consider the log likelihood for $\alpha(z)$ in the GLM, the relevant factor of which is given by

$$d_n(\alpha) = \sum_{i=1}^{n} d(y_i, \eta_i^{\alpha})\delta_z(z_i) ,$$

where $d(y,\eta) = y\cdot\theta(\eta) - b\{\theta(\eta)\}$ and

$$\delta_z = \begin{cases} 1 & z_i = z \\ 0 & \text{otherwise} \end{cases}$$

This leads to solving

$$\begin{aligned} \overset{\circ}{d}_n(\alpha) &= \frac{\partial d_n(\alpha)}{\partial \alpha} \\ &= \sum_{i=1}^{n} \{y_i - \mu(\eta_i+\alpha)\}\theta'(\eta_i+\alpha)\delta_z(z_i) \\ &= 0 \end{aligned} \qquad (3.1)$$

for α giving $\hat{\alpha}(z)$, the maximum likelihood, an estimate which is non-parametric to the extent that no functional form is assumed for $\alpha(z)$.

Though likelihood based inference has its finite sample justifications (see Godambe, 1960) its main appeal stems from its optimal asymptotic properties. Since the sample size relevant above,

$$n.(z) = \sum_{i=1}^{n} \delta_z(z_i)$$

may be small, large sample behavior is largely irrelevant, and it will likely be desirable to modify (3.1) to obtain satisfactory small sample behavior. Nonetheless, the likelihood framework is valuable in that it is at once general, and yet closely tied to the particular GLM under consideration.

With respect to our goal of defining an all purpose residual, consideration of (3.1) reveals that no single set of residuals, however defined, serve to determine $\hat{\alpha}(z)$, unless g is linear, as in the NLM. $\hat{\alpha}(z)$ is in some sense though a nonlinearly averaged residual, with weights determined by the likelihood, and as such, appropriate for the particular GLM under consideration.

4. Formal properties of $\hat{\alpha}(z)$

Assuming that $n.(z) \to \infty$, $\hat{\alpha}(z)$ enjoys the usual optimal properties; in particular $\sqrt{n}\{\hat{\alpha}(z)-\alpha(z)\}$ is asymptotically normal with mean 0 and variance given as followed. Letting $I(\eta) = a(\phi)^{-1}\mu'(\eta)\theta'(\eta)$, the information in a single observation, we define the limiting average Fisher information relative to $\alpha(z)$ by

$$\bar{I}.(z) = \lim_{n\uparrow\infty} n^{-1} \sum_{i=1}^{n} I(\eta_i+\alpha)\delta_z(z_i) \ .$$

By standard arguments, the desired variance is $\{\bar{I}.(z)\}^{-1}$. In the canonical

case, this leads to the approximation

$$\mathrm{Var}(\hat{\alpha}) \approx V(z) = a(\phi)\Big\{\sum_{i=1}^{n} \mu'(\eta_i+\alpha)\delta_z(z_i)\Big\}^{-1} .$$

Of course these results are of questionable relevance, for when $n.(z)$ is small the bias in $\hat{\alpha}(z)$ may be sizeable, and $\hat{\alpha}(z)$ may fail to be well defined. For example, in the LRM if η is initially specified as a known function of z, $\eta(z)$, with the correct specification being $\eta^{\alpha}(z) = \eta(z) + \alpha(z)$, then

$$\hat{\alpha}(z) = \mathrm{logit}\{y.(z)/n.(z)\} - \eta(z) ,$$

where

$$y.(z) = \sum_{i=1}^{n} y_i\delta_z(z_i) .$$

Difficulties arise if $y.(z)$ is 0 or $n.(z)$, and in any case the bias of $\hat{\alpha}(z)$ can be large (Cox, 1970, p. 33).

In general then, some modification to (3.1) is required. A simple course applicable to the LRM above is to take $\hat{\alpha}(z) = \hat{\eta}(z) - \eta(z)$, where

$$\hat{\eta}(z) = \log\Big[\Big\{y.(z) + \tfrac{1}{2}\Big\}/\Big\{n.(z) - y.(z) + \tfrac{1}{2}\Big\}\Big]$$

is the form of the empirical logit which eliminates the $O(n^{-1})$ bias. $\hat{\eta}(z)$ can also be motivated on Bayesian grounds, as the posterior mode under a Gamma prior for p proportional to $p^{1/2}(1-p)^{1/2}$, and falls within the scope of Joanes

and Peers (1976), in which certain Bayesian estimates are given frequentist justifications. The calculations there are for estimates based on identical replicates, but they generalize easily to the present case. Following their example, we consider the asymptotic bias in $\hat{\alpha}(z)$ as determined by the equation

$$S_n(\alpha) = a(\hat{\phi})S(\alpha) + \overset{\circ}{d}_n(\alpha) = 0 \qquad (4.1)$$

where $S(\alpha) = \partial \ln \pi(\alpha)/\partial\alpha$ represents the contribution of a prior $\pi(\alpha)$, to be determined so that the $O(n^{-1})$ bias in $\hat{\alpha}$ disappears. Assuming that $\hat{\phi}$ is asymptotically unbiased for ϕ, the relevant expansion is

$$E(\hat{\alpha}-\alpha) = (n\kappa_2^*)^{-1}\{S(\alpha) + (\kappa_{11}^* + .5\kappa_{001})(\kappa_2^*)^{-1}\} + O(n^{-1}) \qquad (4.2)$$

where

$$\kappa_2^* = \{na(\phi)\}^{-1} \sum_{i=1}^{n} E\{\overset{\circ}{d}(y_i, \eta_i^\alpha)^2\}\, \delta_z(z_i)$$

$$= n^{-1} \sum_{i=1}^{n} \mu'(\eta_i^\alpha)\theta'(\eta_i^\alpha)\delta_z(z_i)$$

$$\kappa_{11}^* = \{na(\phi)\}^{-1} \sum_{i=1}^{n} E\{\overset{\circ}{d}(y_i, \eta_i^\alpha)\,\overset{\circ\circ}{d}(y_i, \eta_i^\alpha)\}\delta_z(z_i)$$

$$= -\tfrac{1}{2} n^{-1} \sum_{i=1}^{n} \mu'(\eta_i^\alpha)\theta''(\eta_i^\alpha)\delta_z(z_i)$$

and

$$\kappa_{001} = n^{-1} \sum_{i=1}^{n} E\{{}^{\circ\circ\circ}d(y_i, \eta_i^{\alpha})\}\delta_z(z_i)$$

$$= -\, n^{-1} \sum_{i=1}^{n} \{\mu''(\eta_i^{\alpha})\theta'(\eta_i^{\alpha}) + 2\mu'(\eta_i^{\alpha})\theta''(\eta_i^{\alpha})\}\delta_z(z_i)$$

where $^{\circ}$ denotes differentiation with respect to α.

By choosing

$$S(\alpha) = -(\kappa_{11}^{*} + .5\, \kappa_{001})(\kappa_2^{*})^{-1} \tag{4.3}$$

the $O(n^{-1})$ bias in $\hat{\alpha}(z)$ is eliminated. If a canonical link is assumed the correction reduces to using

$$S(\alpha) = -\, .5\, \kappa_{001}(\kappa_2^{*})^{-1}$$

$$= .5\, \frac{\sum_{i=1}^{n} w(\eta_i^{\alpha})u(\eta_i^{\alpha})\delta_z(z_i)}{\sum w(\eta_i^{\alpha})\delta_z(z_i)} \tag{4.4}$$

where $w(\eta) = \mu'(\eta)$ and $u(\eta) = \mu''(\eta)/\mu'(\eta)$.

In many familiar instances, for example, the Poisson, binomial and gamma models, $u(\eta)$ is a linear function of $\mu(\eta)$, say $u(\eta) = 2[c_0 - c_1\mu(\eta)]$, so that

$$S(\alpha) = c_0 - c_1\, \mu_w^{\alpha}(z)\{w_{\cdot}^{\alpha}(z)\}^{-1} \tag{4.5}$$

where

$$\mu_w^{\alpha}(z) = \sum_{i=1}^{n} \mu_i^{\alpha} w(\eta_i^{\alpha}) \delta_z(z_i) \qquad \text{and}$$

$$w_{\cdot}^{\alpha}(z) = \sum_{i=1}^{n} w(\eta_i^{\alpha}) \delta_z(z_i) \quad .$$

Thus $S(\alpha)$ is essentially a function of a weighted average of the μ_i^{α}'s.

While the motivation in using $S(\alpha)$ above is asymptotic, its use in finite samples is still warranted by the practical need to stabilize $\hat{\alpha}(z)$'s behavior. For instance, in its application in the LRM, its net effect is to shrink the estimate of η_i^{α} towards 0 and to ensure the existence of finite estimates. Thus in what follows we shall take (4.1) as a basis for forming $\hat{\alpha}(z)$, with $S(\alpha)$ of the form given in (4.3). Note that the contribution of $S(\alpha)$ is asymptotically negligible, so that the asymptotic variances given previously still apply.

5. COVARIATE ADJUSTMENTS

The foregoing calculations have been based on a fully specified, i.e., $\underline{\beta}$ given, model. The extension of the previous methods to the full simultaneous estimation of $\alpha(z)$ and $\underline{\beta}$ is formally simple, but is not pursued here. Instead we proceed by replacing η in its occurrences in $S(\alpha)$ and $\mathring{d}_n(\alpha)$ in (4.1) by $\hat{\eta} = \hat{\underline{\beta}}^T \underline{x}$, where $\hat{\underline{\beta}}$ is the maximum likelihood estimate from the initial model. This use of $\hat{\underline{\beta}}$ corresponds to the formation of the ordinary residual in the NLM. As in that case, the resultant $\hat{\alpha}(z)$ may be severely biased, unless a suitable degree of orthogonality holds between $\underline{\alpha}$ and X. The only entirely reliable way of avoiding such bias is the simultaneous approach mentioned above, which is practically feasible but involves considerable increased computational

complexity. Given the exploratory motivation here, we opt for a less cumbersome approach, understanding that our results are to be taken as useful indicators rather than as basis for formal inference regarding $\alpha(z)$.

The previously given results for the bias and variance of $\hat{\alpha}(z)$ must be reconsidered in this new context. The bias introduced by "fixing" $\hat{\beta}$ well be non-vanishing except in certain balanced designs, and in the null case, $\alpha(z)=0$. This latter case being the tentative assumption, we may still be interested in the $O(n^{-1})$ bias in $\hat{\alpha}(z)$. The estimation of $\hat{\underline{\beta}}$ will, in general, contribute an additional $O(n^{-1})$ term to the bias of $\hat{\alpha}(z)$ given in (4.2). However, since $\hat{\underline{\beta}}$ will be derived from the full sample, its bias contribution will be proportionately smaller, on the order of $n.(z)/n$, than those terms given in (4.2). Thus, as a practical matter, the use of (4.3) is still indicated as a means of reducing bias, in addition to its utility in computational terms. We do not seek to refine it further.

With regard to variance, it is sensible to restrict our considerations to the behavior of $\hat{\alpha}(z)$ under the assumption of the initial models validity, the aim being to provide standard errors relevant to model assessment. The variance given previously must be modified to account for the variance of $\hat{\underline{\beta}}$; the modification resembles that which applies to Var(e) in relationship to $\mathrm{Var}(\varepsilon)$ in the NLM. Assuming, $\eta = \underline{\beta}^T\underline{x}$ is correctly specified, the asymptotic covariance of $\hat{\alpha}(s)$ and $\hat{\alpha}(t)$ is given by

$$a(\phi)\Big\{\sum_{i=1}^{n} w_i\delta_s(z_i)\sum_{i=1}^{n} w_i\delta_t(z_i)\Big\}^{-1}\Big\{\sum_{i=1}^{n} w_i\delta_s(z_i)\delta_t(z_i) - \underline{x}_w(s)^T I_\beta^{-1}\,\underline{x}_w(t)\Big\} \tag{5.1}$$

where $w_i = \mu'(\eta_i)\theta'(\eta_i)$,

$$\underline{x}_w(z) = \sum_{i=1}^{n}\delta_z(z_i)w_i\underline{x}_i \ ,$$

$$\text{and } I_\beta = \sum_{i=1}^{n} w_i \underset{\sim}{x}_i \underset{\sim}{x}_i^T .$$

In addition, it is useful to note that the asymptotic covariance of $\hat{\alpha}(z)$ and $\hat{\underset{\sim}{\beta}}$ is 0, implying asymptotic independence. Thus, in assessing the validity of the initial model, the approximation

$$\begin{aligned} \text{Var}(\hat{\alpha}(z)) \approx \hat{V}_0(z) = a(\hat{\phi})\Big\{ \sum_{i=1}^{n} \hat{w}_i \delta_z(z_i)\Big\}^{-2} \\ \times \Big\{ \sum_{i=1}^{n} \hat{w}_i \delta_z(z_i)^2 \\ - \underset{\sim}{x}_{\hat{w}}(z)^T I_{\hat{\beta}}^{-1} \underset{\sim}{x}_{\hat{w}}(z)\Big\} \end{aligned} \qquad (4.2)$$

may be useful.

6. SMOOTHING

The extent of replication in a data set is crucial to the utility of the previously outlined approach, as it is to many diagnostic procedures. The replication required may be termed "marginal" replication, in that only the degree of replication in z is significant; true replicates need not be present. The occurrence of marginal replicates is common enough that the method is useful (see Section 8.2). If the degree of replication is insufficient to provide stable estimates, $\hat{\alpha}(z)$, or if it is merely desired to impose some measure of smoothness on $\hat{\alpha}(z)$, δ_z can be relaxed to take on non-zero values in a neighborhood of z. Many choices are available for this kernel or windowing function; in the one dimensional case, at least, most are more or less equivalent. A convenient choice, employed by Cleveland (1979) is

the tricube function

$$\delta_w(t) = \begin{cases} 1 - \left|\frac{z-t}{\omega}\right|^3 & |z-t| < \omega \\ 0 & \text{otherwise,} \end{cases}$$

where ω is a window width to be selected at our discretion, trading off between increased stability for large ω and smaller bias for large ω. The resultant approach is a natural extension of Cleveland's locally weighted (non-robust) regression to the nonlinear setting of the GLM. Rather than adapting the "nearest neighbor" distance, used by Cleveland, we fix ω, allowing the data to speak loudest where it has the most to say. The varying precision of $\hat{\alpha}(z)$, measured crudely by $n_\cdot(z)$, is easily accounted for in that the previously given covariance and variance expressions (5.1) and (5.2) are still valid.

The bias introduced by smoothing can to some extent be mitigated by elaborating $\alpha(z)$ to higher order terms, for example, by linear adjustment to $\alpha(z,z_i) = \alpha_0(z) + \alpha_1(z)(z_i - z)$. Since the bias in $\hat{\alpha}(z)$ in the non null $(\alpha(z) \neq 0)$ is likely to be non-negligible even without smoothing, the practical utility of this added complication is difficult to assess, and we retain the simpler form.

Extension of the bias adjustment already given is possible. The result for $\hat{\alpha}(z)$ solving (4.1) is, assuming known β's and $\alpha(z)$ constant in $z \pm \omega$, that $E\{\hat{\alpha}(z) - \alpha\} = \{n\kappa_{01}\}^{-1}\{S(\alpha) + (\kappa_{11}^{*}\kappa_{01} + .5\,\kappa_{001}\kappa_{2}^{*})\kappa_{01}^{-2}\}$ where

$$\kappa_2^* = n^{-1} \sum_{i=1}^{n} \mu'(\eta_i^\alpha)\theta'(\eta_i^\alpha)\delta_z(z_i)^2$$

$$\kappa_{01} = -n^{-1} \sum_{i=1}^{n} \mu'(\eta_i^\alpha)\theta'(\eta_i^\alpha)\delta_z(z_i)$$

$$\kappa_{11}^* = -n^{-1} \sum_{i=1}^{n} \mu'(\eta_i^\alpha)\theta''(\eta_i^\alpha)\delta_z(z_i)^2$$

and

$$\kappa_{001} = -n^{-1} \sum_{i=1}^{n} \{\mu''(\eta_i^\alpha)\theta'(\eta_i^\alpha) + 2\mu'(\eta_i^\alpha)\theta''(\eta_i^\alpha)\}\delta_z(z_i)$$

By choosing

$$S(\alpha) = -\kappa_{01}^{-2} \{\kappa_{11}^* \kappa_{01} + .5\, \kappa_{001} \kappa_2^*\}$$

we can hope to reduce the bias due to nonlinearity. In the canonical link case, the result is

$$S(\alpha) = k \times [c_0 - c_1 \mu_w^\alpha(z)\{w_\cdot^\alpha(z)\}^{-1}]$$

$$\text{where} \quad k = \{w_\cdot^\alpha(z)\}^{-1} \sum_{i=1}^{n} w(\eta_i^\alpha)\delta_z(z_i)^2 \,. \tag{6.1}$$

Though it is arguable that the bias introduced by the δ-weight smoothing is likely to outweigh any reduction in bias yielded by the use of (6.1), the adjustment is warranted by the practical need to stabilize $\hat{\alpha}(z)$ discussed earlier.

7. COMPUTATION AND APPROXIMATIONS

It's natural to consider the few cases where $\hat{\alpha}(z)$ can be explicitly determined, as well as to derive approximations for when it may not. For simplicity, we consider only the case of the canonical link in what follows. In general, (4.1) can be rewritten as

$$S(\alpha)\, a(\hat{\phi}) + y_{\cdot}(z) - \mu_{\cdot}^{\alpha}(z) = 0$$

$$\text{where } \mu_{\cdot}^{\alpha}(z) = \sum_{i=1}^{n} \mu_i^{\alpha}\, \delta_z(z_i)\ . \tag{7.1}$$

A form that is useful in preliminary data exploration stems from beginning with a null model, i.e. $\underset{\sim}{\beta} = 0$, in which case (7.1) leads to solving

$$\mu(\alpha) + a(\hat{\phi})S(\alpha)\, n_{\cdot}(z)^{-1} = \bar{y}_{\cdot}(z)$$

where $\bar{y}_{\cdot}(z) = n_{\cdot}(z)^{-1} y_{\cdot}(z)$.

In the case that $S(\alpha)$ is linear of the form (4.5), this yields

$$\hat{\alpha}(z) = \tilde{\eta}(z) = g\left\{\frac{y_{\cdot}(z) + c_0 a^{*}(\hat{\phi})}{n_{\cdot}(z) + c_1 a^{*}(\hat{\phi})}\right\}$$

where $a^{*}(\hat{\phi}) = ka(\hat{\phi})$,

which is a bias reduced form of the naive estimate, $g(\bar{y}_{\cdot}(z))$. $\hat{\alpha}(z)$ is also tractable if z corresponds to the full set of covariates and δ is the Kronecker δ, in which case we have that $\hat{\alpha}(x_i) = \tilde{\eta}(x_i) - \hat{\eta}(x_i)$, i.e. the obvious residual based on the unadjusted (for covariates) smooth.

In the remaining instances, which will comprise the bulk of applications, (7.1) must be solved iteratively. One step versions, starting at $\alpha(z) = 0$, are useful, A Newton-Rafson approach applies generally, though in the "linear" case for $S(\alpha)$, it is more convenient to use a Taylor series in $\mu(\eta_i^{\alpha})$, which yields

$$\hat{\alpha}_1(z) = \frac{y_{\cdot}(z) - \mu_{\cdot}(z) + a^{*}(\hat{\phi})(c_0 - c_1 \tilde{\mu}_w(z))}{w_{\cdot}(z) + a^{*}(\hat{\phi}) c_1 w_{\cdot}^{2}(z)}$$

$$\text{where } \mu_{\cdot}(z) = \sum_{i=1}^{n} \mu(\hat{\eta}_i)\delta_z(z_i)$$

$$\tilde{\mu}_w(z) = \sum_{i=1}^{n} \mu(\hat{\eta}_i) w(\hat{\eta}_i) \delta_z(z_i)$$

$$w_{\cdot}(z) = \sum_{i=1}^{n} w(\hat{\eta}_i)\delta_z(z_i)$$

$$\text{and } w_{\cdot}^2(z) = \sum_{i=1}^{n} w(\hat{\eta}_i)^2\delta_z(z_i) \ .$$

A particularly simple and revealing form results from neglecting $S(\alpha)$ in (7.1) in which case one gets

$$\hat{\alpha}_1(z) = w_{\cdot}(z)^{-1} \sum_{i=1}^{n} r_i w(\hat{\eta}_i)\delta_z(z_i) \ .$$

$$\text{where } r_i = \{y_i - \mu(\hat{\eta}_i)\} / w(\hat{\eta}_i)$$

is the locally linearized residual discussed in Section 3. This form is asymptotically equivalent to the fully iterated form if $\alpha(z) = 0$, and is simple to compute. Since $w(\eta) = \{\text{var}(r)\}^{-1}$, $\hat{\alpha}_1(z)$ is just an efficiently weighted average of the r_i's.

It is informative to compare $\hat{\alpha}_1(z)$ above and the corresponding smoothed partial residual considered by Landwehr, Pregibon and Shoemaker (1984). This is possible under the assumption that z is incorporated linearly in η. The partial residual of Landwehr, Pregibon and Shoemaker is $r_{par} = r + \hat{\beta}_z z$. If δ-weight smoothing (the simplest version of Cleveland's smooth) is applied to the above, the result is

$$\bar{r}_{par}(z) = \bar{r}_{\cdot}(z) + \hat{\beta}_z \bar{Z}_{\cdot}(z)$$

$$\text{where } \bar{r}_{\cdot}(z) = n_{\cdot}(z)^{-1} \sum_{i=1}^{n} r_i \delta_z(z_i)$$

$$\text{and } \bar{Z}_{\cdot}(z) = n_{\cdot}(z)^{-1} \sum_{i=1}^{n} z_i \delta_z(z_i) \ .$$

The relevant comparison to be made is between $\bar{r}_{\cdot}(z)$ and $\hat{\alpha}_{\cdot}(z)$, both of which estimate the residual (nonlinear) contribution of z to the true linear

predictor, the latter incorporating the efficient weights neglected in the former. The gains incurred by efficient weighting will depend on the structure of the covariate. When $p = 1$, weighting may have only a slight effect, whereas when $p > 1$, the weights may vary sufficiently that the gain is substantial.

Turning to consider the estimated variance of $\hat{\alpha}(z)$, which is also relevant to $\hat{\alpha}_1(z)$, we note that $\hat{V}_0(z)$ requires an additional pass through the data, in order to obtain $\hat{\underline{x}}(z)$. An obvious, and easily calculated upper bound which is generally close enough to $\hat{V}_0(z)$ to be useful in calibrating the $\hat{\alpha}$'s is $\tilde{V}_0(z) = a^*(\hat{\phi})\{w_\cdot(z)\}^{-1}$, which is essentially $\hat{V}_0(z)$ ignoring the adjusting for estimated $\underline{\beta}$.

Owing to their simple form, the approximations attain appeal not only for computational reasons, but due to the fact that the asymptotic distributional results may be more reliable, since the central limit theorem applies more or less directly to these linear forms.

8. EXAMPLES

8.1 The Linear Model

Application of the methods described to the case of the usual regression model is straightforward. The bias correction is superfluous, and

$$\hat{\alpha}(z) = \bar{e}_\cdot(z) = \{n_\cdot(z)\}^{-1} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)\delta_z(z_i) \quad ,$$

a simple non-robust smooth. A more robust approach along the lines of M-estimation (Huber, 1964) arises if one assumes a distribution of the least favorable type for the errors, say with log density ρ. This takes us outside the class of generalized linear models, but the ideas here generalize easily

to suggest using $\hat{\alpha}(z)$ solving

$$\sum_{i=1}^{n} \psi(e_i - \alpha)\delta_z(z_i) = 0$$

or alternatively, the one-step version,

$$\hat{\alpha}'(z) = \sum_{i=1}^{n} w_i e_i \delta_z(z_i) / \sum_{i=1}^{n} w_i \delta_z(z_i)$$

where $\psi'(u) = p'(u/\hat{\phi})$ and $w_i = \psi(e_i)/e_i$

8.2 Logistic Regression

As already seen, $\hat{\alpha}(z)$ corresponds to a smoothed, covariate adjusted version of the empirical logit. To illustrate its application, we begin with a simple example in which $\eta = \beta_0 + \beta_1 x$ was fit to 100 observations artificially generated with the x values uniform on -1 to 1 and the true $\eta = -1 + x + 2x^2$, which is a simpler version of the example 4 in Landwehr, et al. A smoothed partial residual plot along the lines of their approach is reproduced in Figure 1. As noted in Section 2, the configuration of raw residuals, and correspondingly, the unaugmented partial residual plot, is completely uninformative; any judgement made rests solely on the smooth, which in this example differs little from the efficiently weighted $\hat{\alpha}_1(z)$. However, without the aid of auxiliary cues, such as standard errors, smooths themselves are difficult to assess. Figure 2 reproduces Figures 5 and 8 (reversed and inverted) of Landwehr, et al. Though Figure 2b clearly exhibits non-linearity to some greater degree than 2a, it is debatable whether either speaks convincingly for the need to transform. The authors, on the other hand, see a clear need in 2b but not in 2a; their expert judgement is not at issue here, but whether less experienced investigators may be as reliably lead by such plots. The application of $\hat{\alpha}(z)$ gives a similar smooth in Figure 3, similarly

indicating the potential need for inclusion of a quadratic term, while the studentized version, $\{\hat{V}_0(z)\}^{-1/2}\hat{\alpha}(z)$, confirms the "significance" of the lack of fit.

The data given in Haberman (1976), and analyzed as well in Landwehr, et al, illustrates a situation with a high degree of hidden replication (see Figure 4), where it is possible to apply the Kronecker δ and avoid the problems of smoothing. The data describes the 5-year survivorship of 306 breast cancer surgery patients, with

$$y = \begin{cases} 1 & \text{patient survived 5 years from the date of surgery} \\ 0 & \text{otherwise} \end{cases}$$

x_1 = age of patient at time of surgery

x_2 = year of operation (minus 1900)

x_3 = number of positive auxiliary nodes detected in the patient

Noteworthy is the presence of a number of patients with large counts for x_3, which are likely to be influential in any fit incorporating x_3 linearly. Figure 5 gives $\hat{\alpha}$ for $z = x_3$ based on fitting

$$\eta = \beta_0 + \sum_{i=1}^{3} \beta_i x_i \quad .$$

The apparently poor fit at $x_3 = 0$ is clearly significant, since, owing to the large sample size, $n.(0) = 134$, the asymptotic standard error is relevant. Following Landwehr, et al, a transformation of x_3 to $x_3' = \log(1 + x_3)$ seems to alleviate the problem (Figure 6); the effect of the transformation is to shrink the influential large counts to the extent that they no longer severely degrade the fit at $x_3 = 0$.

8. CONCLUDING REMARKS

In attempting to devise procedures appropriate for exploratory analysis,

it is important to balance the necessities of flexibility and computational economy against the desirability of a clear inferential framework. Methods based on residuals, particularly in combination with simple smoothing techniques are appealing on the first count, and in addition, can be placed within a more formal parametric framework of the GLM. This formalization, desirable in its right, provides indications towards procedures which are more efficient and informative. A completely satisfactory formal development remains a more or less distant, and perhaps, unrealistic goal.

REFERENCES

Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatter plots, **Journal of the American Statistical Association, 74**, 829-836.

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

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

Godambe, V. P. (1960) An optimum property of maximum likelihood estimation, **Annals of Mathematical Statistics, 31**, 1208-1211.

Haberman, S. J. (1976) Generalized residuals for log-linear models, **Proceedings of the 9th International Biometrics Conference**, Boston, 104-122.

Huber, P. J. (1964) Robust estimation of a location parameter, **Annals of Mathematical Statistics, 35**, 73-101.

Joanes, D. N. and Peers, H. W. (1974) On the sampling properties of Bayesian point estimators, **Journal of the American Statistical Association, 69**, 560-564.

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.

McCullagh, P. and Nelder, J. A. (1983) **Generalized Linear Models**, Chapman and Hall: New York.

Pregibon, D. (1980) Logistic regression diagnostics, **Annals of Statistics, 9**, 705-724.

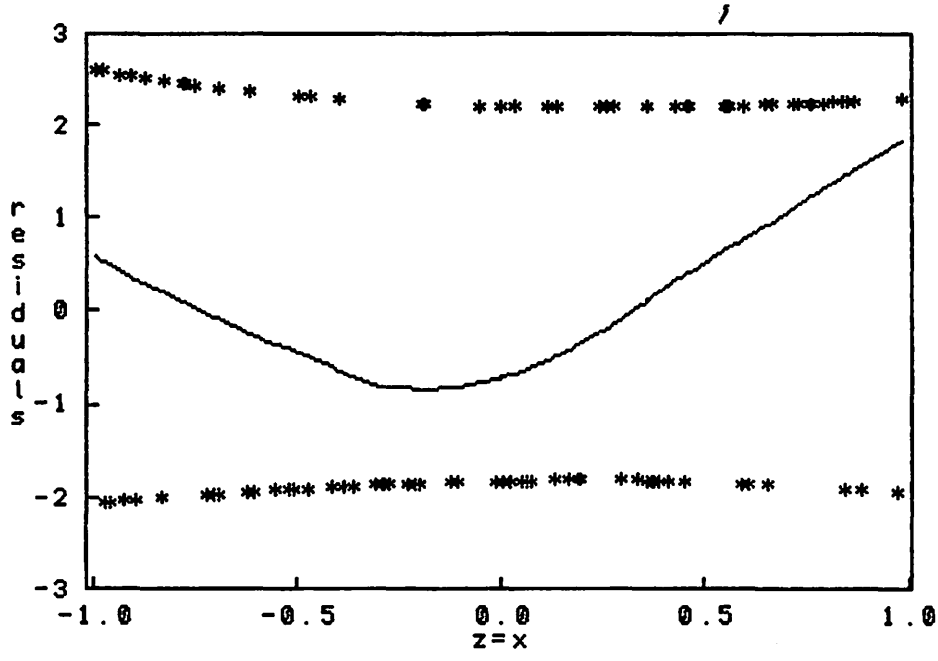

Figure 1. PARTIAL RESIDUAL PLOT. The residual $r_i = \{y_i - \mu(\hat{n}_i)\}/w(\hat{n}_i)$ has a two point distribution in the logistic regression model, which can lead to uninformative scatter plots.

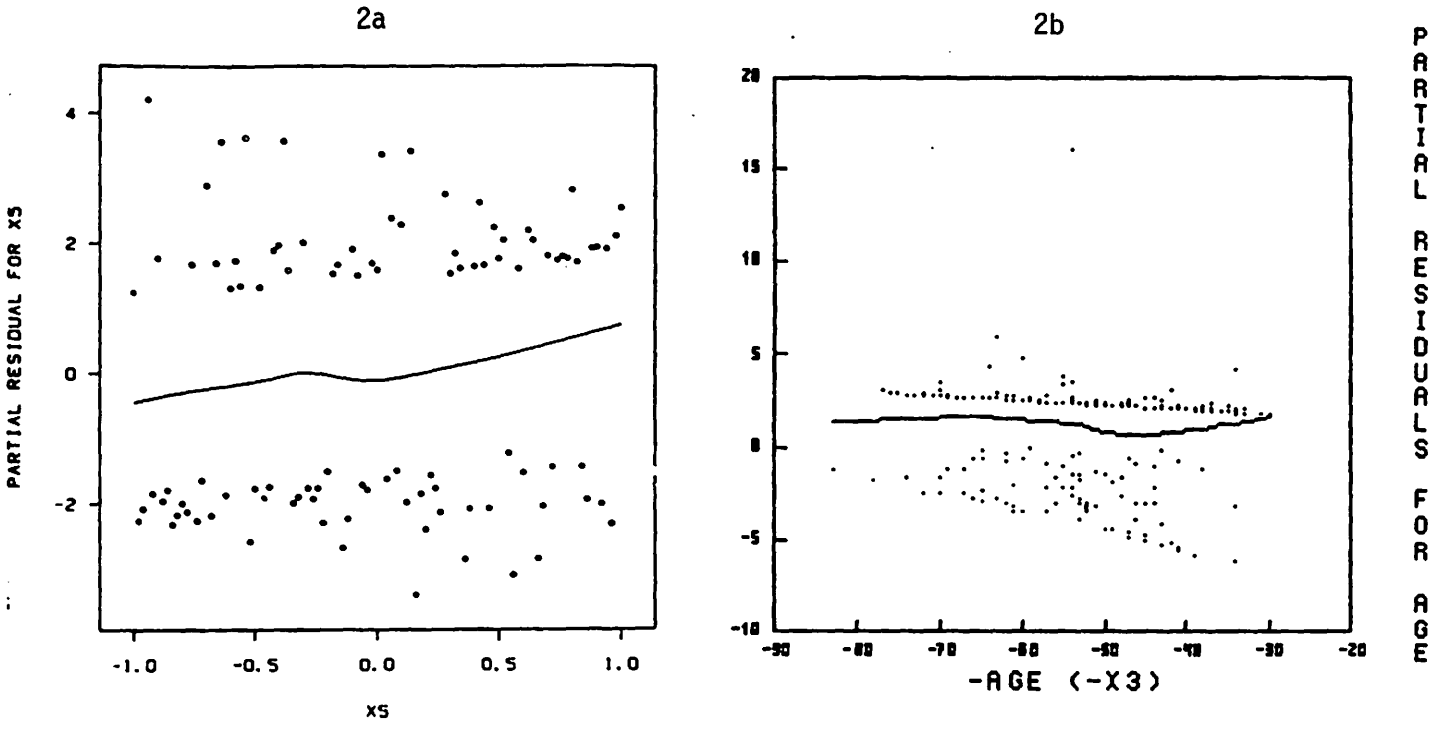

Figure 2. SMOOTHED PARTIAL RESIDUAL PLOTS. Figures 2a and 2b reproduce Figures 5 and 8 (reflected), in Landwehr, Pregibon and Shoemaker, 1984.

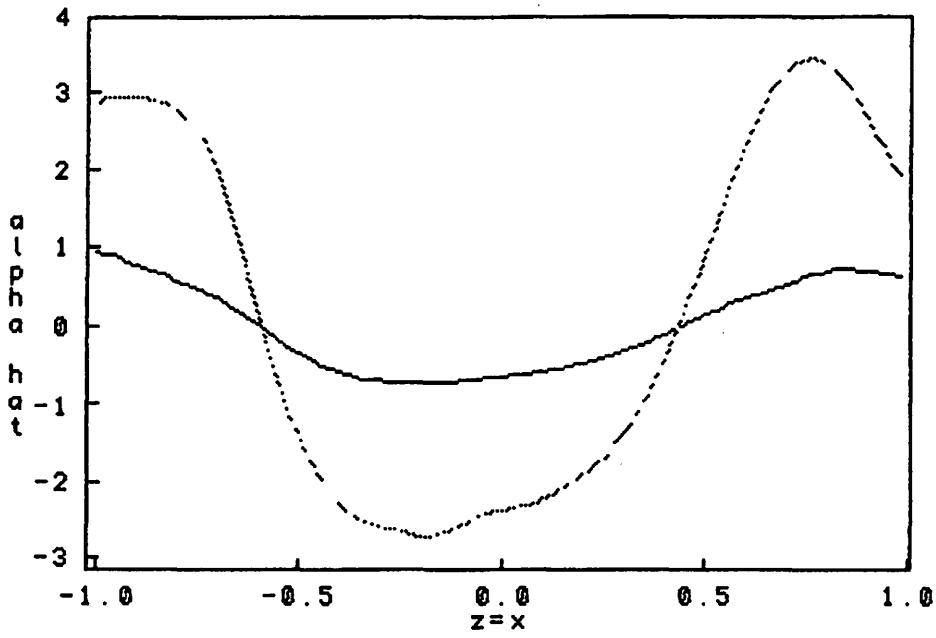

Figure 3. SMOOTHED RESIDUAL PLOT. Including a standardized version of $\hat{\alpha}(z)$ (dotted line) facilitates judgements as to the significance of the apparent lack of fit.

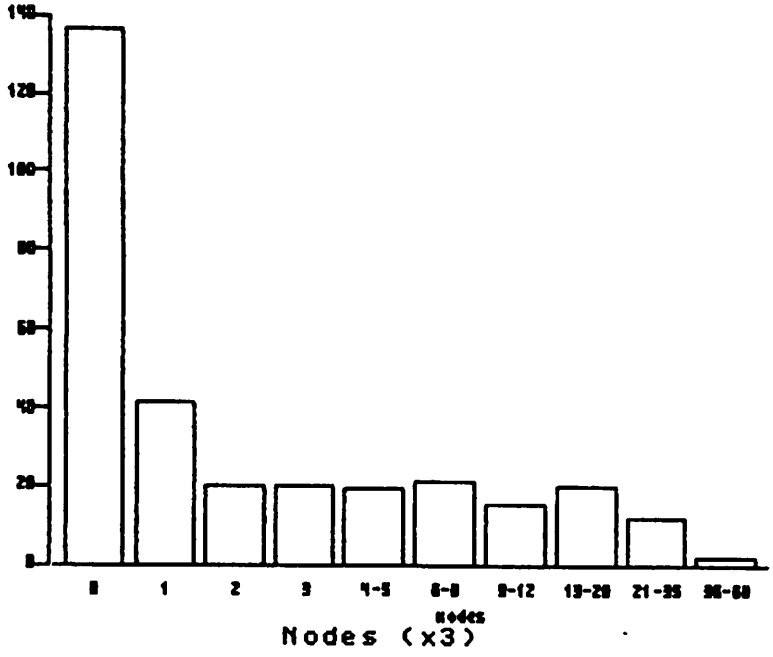

Figure 4. BAR CHART FOR X_3 IN CANCER EXAMPLE. The high degree of marginal replication facilitates diagnostic fits.

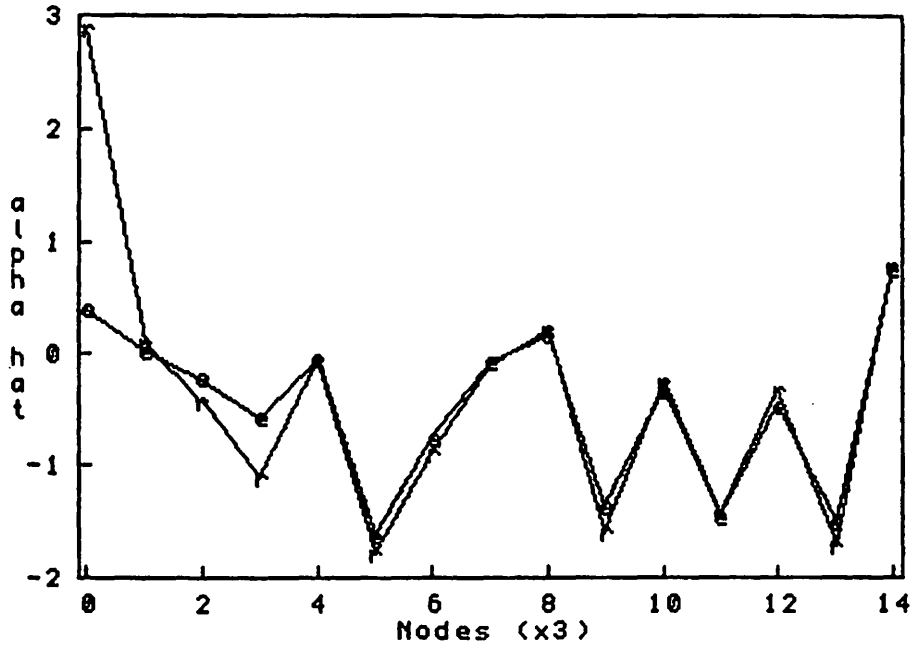

Figure 5. RESIDUALS FOR X_3 IN CANCER EXAMPLE. Both raw, $\hat{\alpha}(z)$, and standardized, $\hat{\alpha}(z)/\{\hat{V}(z)\}^{\frac{1}{2}}$ versions, given as e's and r's, respectively, are plotted.

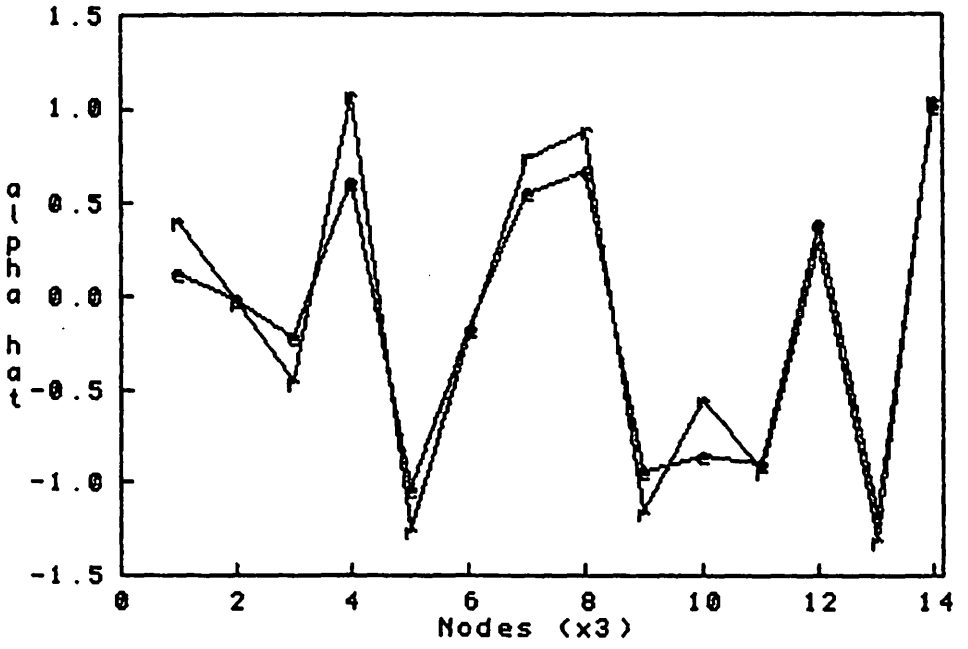

Figure 6: RESIDUALS FOR X3 IN AUGMENTED MODEL.

Smoothed residuals for Nodes (x3)			
z	residual	studentized residual	Std. error
0	0.4041	2.892627	0.1397
1	0.0354	0.1000283	0.3539
2	-0.2154	-0.418171	0.5151
3	-0.5553	-1.101786	0.504
4	-0.0412	-0.0650355	0.6335
5	-1.6334	-1.786308	0.9144
6	-0.7391	-0.884197	0.8359
7	-0.0728	-0.089391	0.8144
8	0.1635	0.2088923	0.7827
9	-1.3548	-1.596888	0.8484
10	-0.3198	-0.2607632	1.2264
11	-1.4573	-1.433786	1.0164
12	-0.468	-0.3268385	1.4319
13	-1.5133	-1.721811	0.8789
14	0.7623	0.776827	0.9813

Table 1. RESIDUALS FOR X_3 IN CANCER EXAMPLE.