Goodness of Link Tests for Generalized Linear Models

By DARYL PREGIBON

University of Toronto, Ontario, Canada

[Received July 1978. Final revision June 1979]

SUMMARY

Data analytic procedures are proposed to examine the adequacy of the hypothesized link used in fitting a generalized linear model. Through model expansion and linearization, tests and estimation techniques are provided. These procedures, along with the release of GLIM3, enable the user to examine routinely and objectively the fit of an hypothesized model. Examples are presented to illustrate the testing and fitting procedure.

Keywords: GENERALIZED LINEAR MODELS; FAMILY OF POWER TRANSFORMATIONS; MODEL FITTING AND TESTING; NON-ADDITIVITY

INTRODUCTION

The theory and use of generalized linear models was made explicit by Nelder and Wedderburn (1972). Since then, through the GLIM computing package, many have benefited from this unified and flexible approach to model fitting. In its simplest form, the generalized linear model is specified by:

(i) independent observations \( y_1, \ldots, y_n \) distributed according to an exponential family distribution,

(ii) a set of explanatory variables \( x \), available for each observation, describing the systematic linear component through \( Y = x^T \beta \), and

(iii) the link function \( g(\mu) = Y \), relating the mean of an observation to the systematic component.

From the user's point of view, the fitting procedure involves:

(a) choosing the relevant error distribution,
(b) determining which variables to include in the systematic component, and
(c) defining the link function \( g(\mu) \).

Within this framework, the generalized linear model can be fitted and its adequacy examined. In particular, residuals should be examined and plotted to determine if any systematic or unusual tendencies remain. This paper attempts to aid the user in this respect by outlining procedures which can be applied quite generally after the usual fitting and testing is completed. Emphasis is restricted to correct link function specification, assuming that both (a) and (b) are appropriate for the data under analysis.

As a by-product of the diagnostic tests, a modification of the link function will be indicated. In particular, this allows the link function itself to be estimated by the data at hand. The necessary computations are straightforward and relatively easy to implement on any system. For convenience, we have chosen to illustrate the methods using the new release of GLIM, namely GLIM3 (Baker and Nelder, 1978).
1. Basic Procedure: Link Expansion

In this section we consider the consequences of fitting a model with link function \( g_0(\mu) \) when in fact, the correct (but unknown) link function is \( g_*(\mu) \). This can be systematically studied if we can incorporate the hypothesized link into a family of link functions, of which \( g_*(\mu) \) is assumed to be a member. In most cases, this will involve the introduction of one or possibly several parameters (we will concentrate on two, say \( \alpha \) and \( \delta \)), so that we may write

Hypothesized link: \( g_0(\mu) = g(\mu; \alpha_0, \delta_0) \)

Correct link: \( g_*(\mu) = g(\mu; \alpha_*, \delta_*) \).

Now using a first-order Taylor series expansion about the hypothesized link, we have the approximate relationship

\[
g_*(\mu) \approx g_0(\mu) + (\alpha_* - \alpha_0) D\alpha(g_0) + (\delta_* - \delta_0) D\delta(g_0),
\]

(1)

where

\[
D\alpha(g_0) = \left( \frac{\partial}{\partial \alpha} g(\mu; \alpha, \delta) \right)_{|\alpha=\alpha_0, \delta=\delta_0},
\]

and similarly for \( D\delta(g_0) \). Note that we may now approximate the correct link function \( g_*(\mu) = x^T \beta \) by

\[
g_0(\mu) = x^T \beta + z^T \gamma,
\]

(2)

where \( z^T = (D\alpha(g_0), D\delta(g_0)) \) and \( \gamma^T = -(\alpha_* - \alpha_0, \delta_* - \delta_0) \). Hence, the problem of link modification has been reformulated in terms of the original hypothesized link, but now with additional factors in the systematic linear component describing the local differences between the link functions. The fact that the difference in link functions appears on the right side of the link defining equation should not be disturbing—indeed, this corresponds to the fact that the wrong link function is a systematic mis-specification of the model.

The practical advantage of the above formulation is now clear. Consider an initial fit of the model with link function given by \( g_0(\mu) = x^T \beta \). This fit yields estimates \( \hat{\beta} \) and fitted values \( g_0(\hat{\mu}) = x^T \hat{\beta} \), from which we can form \( \hat{z} = z_{g0=\hat{\beta}} \). To test the adequacy of the link then, one merely refits the model using \( \hat{z} \) as an additional set of explanatory variables. A significant reduction in the resulting deviance indicates a departure from the hypothesized link function. As a by-product, this procedure also yields an estimate of \( \gamma \) and hence of \( \delta_* \) and \( \alpha_* \), and therefore of the link itself. The difference between this method of estimation and testing and that of full maximum likelihood is given in a later section. We now illustrate the procedure by specific examples.

2. Application to Natural Linear Models

In this section, we apply the theory outlined above to the most commonly used generalized linear models. These are characterized by defining the link function as the natural exponential family parameter and, for which, sufficient statistics are available.

2.1. Normal Errors and the Identity Link: \( g_0(\mu) = \mu \)

A useful family of link functions incorporating the identity (and now available in GLIM3) is given by the family of power transformations

\[
g(\mu; \alpha, \delta) = ((\mu + \alpha)^\delta - 1)/\delta.
\]

(3)

Note that \( g_0(\mu) = g(\mu; 1, 1) = \mu \), so that the identity link is a member of the family. We now suppose that the true link function \( g_*(\mu) \) is given by \( g(\mu; \alpha_*, \delta_*) \) for some unknown values
\[ \alpha^* \text{ and } \delta^*. \] For this model, the appropriate series expansion is given by

\[
g^*_g(\mu) = g_0(\mu) + (\alpha^*_g - 1) \cdot \{1 + (\delta^*_g - 1) \cdot [(\mu + 1) \log(\mu + 1) - \mu]\}.
\]

A fit using the identity link yields estimates \( \hat{\beta} \) and fitted values \( g_0(\hat{\mu}) = \hat{\mu} \) leading to

\[
\hat{\varepsilon}^T = \{1, (\hat{\mu} + 1) \log(\hat{\mu} + 1) - \hat{\mu}\}.
\]

Further, the regression of \( r = y - \hat{\mu} \) on \( W\hat{\varepsilon} \) where \( W = I - X(X^T X)^{-1} X^T \), leads to

\[
\hat{\gamma}^T = -(\hat{\alpha}^*_g - 1, \hat{\delta}^*_g - 1) = (\hat{\varepsilon}^T W\hat{\varepsilon})^{-1} \hat{\varepsilon}^T W r,
\]

and the sum of squares \( \hat{\gamma}^T \hat{\varepsilon}^T W\hat{\varepsilon} \hat{\gamma} \) is distributed exactly as \( \sigma^2 \chi^2_1 \) under the null model that \( y \) is normally distributed with mean \( x^T \beta \) and constant variance \( \sigma^2 \). An F test, using the change in deviance, can be performed to judge the necessity of a link modification. This is essentially the same argument used by Andrews (1971), but taken from an entirely different point of view. In particular, one must not confuse link modifications with response variate transformations. In the latter, we hypothesize a linear model for the transformed response, say

\[
y^* = f(y; \alpha, \delta) = \{(y + \alpha)^\delta - 1\}/\delta
\]

such that on this scale of measurement, \( y^* \) is normally distributed with mean \( x^T \beta \). In the former (and the topic of this section), we suppose that \( y \) is normally distributed but with mean \( g^{-1}_g(x^T \beta) \). This is particularly relevant when the dependent variable is meaningful and not some transform of it.

To further contrast these different approaches, consider the survival time data reported by Box and Cox (1964). The data represent the survival times of animals in a \( 3 \times 4 \) (poison \( \times \) treatment) factorial experiment and a goal of the analysis is to achieve, if possible, a realistic model containing for simplicity, main effects only. For our purpose, it is useful to summarize the findings in terms of two plots: one displaying residuals vs fitted values, the other being the standard probability plot of residuals vs unit Gaussian quantiles. Figs 1 and 2 are the plots corresponding to the fit of the model

\[
data = \text{grand mean} + \text{row effect} + \text{column effect} + \text{Gaussian noise}.
\]

The inhomogeneity of variance is striking and the probability plot suggests that the disturbances (errors) are somewhat heavier tailed than Gaussian. Using the method outlined above, we may fit the model

\[
data = \text{grand mean} + \text{row effect} + \text{column effect} + \text{misfit} + \text{Gaussian noise},
\]

noting that the "misfit" term corresponds to the single parameter \( \delta \), since \( \alpha \) cannot be estimated when a grand mean is already included in the fit. The corresponding figures are not shown since there is no real improvement over the previous fit. It can be argued that since the initial fit (5) was such a poor one, the approximate method may not be very effective in detecting and correcting the observed problems. To illustrate that this is not the case, we fit the model

\[
data = g^{-1}(\text{grand mean} + \text{row effect} + \text{column effect}) + \text{Gaussian noise},
\]

with \( g \) given by (3). No improvement over the previous fits was evidenced. This indicates that link modifications (restricting oneself to the specified family \( g \)) will not alleviate the model mis-specifications.

The alternative procedure is to consider fitting the model

\[
f(data) = \text{grand mean} + \text{row effect} + \text{column effect} + \text{Gaussian noise}
\]

with \( f \) given by (4). Figs 3 and 4 demonstrate the effectiveness of this approach, as the aforementioned anomalies are not present in these plots.
FIG. 1. Residuals vs fitted values for the model $y = x^T \beta + N(0, \sigma^2)$.

FIG. 2. Gaussian probability plot of residuals from the model $y = x^T \beta + N(0, \sigma^2)$. 
ADEQUACY OF GLIM LINKS

Fig. 3. Residuals vs fitted values for the model $y^{-1} = x^T \beta + N(0, \sigma^2)$.

Fig. 4. Gaussian probability plot of residuals from the model $y^{-1} = x^T \beta + N(0, \sigma^2)$. 
The above example is one where a single transformation (of the data) has corrected both, the variance function and the scale for additivity. In general, this may not be the case, so that both, a variate transformation and a link modification, are required. Fisher’s tuberculin-test data (see Nelder and Wedderburn, 1972) provide such an example.

\[ g_\delta(\mu) = \log\{\mu/(n-\mu)\} \]

2.2. Binomial Errors and the Logit Link: 
As in the previous example, one could apply the power family to the quantity \( \mu/(n-\mu) \) and define a family of links accordingly. However, this is not very appealing from the viewpoint of bioassay, where tolerance distributions are usually taken as the mechanisms generating the observed responses. In particular, the logit link assumes that the underlying tolerance distribution is logistic—a symmetric distribution with somewhat heavier tails than the normal (probit) distribution. A versatile family of such tolerance distributions was given by Prentice (1976), but unfortunately, the necessary computations are unduly complicated. A more useful family is given by defining the tolerance distribution of \( d = (x-\mu)/\sigma \) by

\[ f(d) = \{P^\alpha-\delta -1 + (1-P)^{\alpha+\delta-1}\}^{-1}, \]

where \( P = \text{Pr}\{\text{reaction at dosage } x\} \) is defined as the solution to the implicit equation

\[ \frac{P^\alpha-\delta -1}{\alpha-\delta} - \frac{(1-P)^{\alpha+\delta-1}}{\alpha+\delta} - d = 0. \]

Note that symmetry results when \( \delta = 0 \), and that heaviness of tails is parameterized by \( \alpha \). In particular, as \( \alpha \) decreases, tail weight increases.

Accordingly, we consider the family of link functions defined by

\[ g(\mu; \alpha, \delta) = \frac{(\mu/n)^{\alpha-\delta} - 1}{\alpha-\delta} - \frac{(1-\mu/n)^{\alpha+\delta-1}}{\alpha+\delta}, \]

where we note that the logit link is given by

\[ g_0(\mu) = \lim_{\alpha,\delta \to 0} g(\mu; \alpha, \delta). \]

Once again, a fit of the model leads to estimates \( \hat{\beta} \) and fitted values \( \hat{g}_\delta(\mu) \), yielding

\[ \hat{\beta}_T = [\frac{1}{2}[\log^2(\hat{\mu}/n) - \log^2(1-\hat{\mu}/n)], -\frac{1}{2}[\log^2(\hat{\mu}/n) + \log^2(1-\hat{\mu}/n)]]. \]

From the fit of the extended model (2), the change in deviance from fitting these two additional variables can be computed and used as a guide to determine if the link function \( g_\delta(\mu) \) should be altered, to perhaps \( g_\gamma(\mu) \). As an example, consider the data (Table 1) of Bliss as reported

**Table 1**

<table>
<thead>
<tr>
<th>Log dosage</th>
<th>Number exposed</th>
<th>Number killed</th>
<th>Logit fit</th>
<th>Skew fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6907</td>
<td>59</td>
<td>6</td>
<td>3.45</td>
<td>7.03</td>
</tr>
<tr>
<td>1.7242</td>
<td>60</td>
<td>13</td>
<td>9.83</td>
<td>10.43</td>
</tr>
<tr>
<td>1.7552</td>
<td>62</td>
<td>18</td>
<td>22.48</td>
<td>19.07</td>
</tr>
<tr>
<td>1.7842</td>
<td>56</td>
<td>28</td>
<td>33.89</td>
<td>30.04</td>
</tr>
<tr>
<td>1.8113</td>
<td>63</td>
<td>52</td>
<td>50.11</td>
<td>49.16</td>
</tr>
<tr>
<td>1.8369</td>
<td>59</td>
<td>53</td>
<td>53.28</td>
<td>54.69</td>
</tr>
<tr>
<td>1.8610</td>
<td>62</td>
<td>61</td>
<td>59.22</td>
<td>60.83</td>
</tr>
<tr>
<td>1.8839</td>
<td>60</td>
<td>60</td>
<td>58.74</td>
<td>59.75</td>
</tr>
</tbody>
</table>
Fig. 5. Proportion killed vs log dosage. Observed (P) and fitted (L) for the link $g_0(\mu) = x^T \beta$.

Fig. 6. Proportion killed vs log dosage. Observed (P) and fitted (S) for the link $g_0(\mu) = x^T \beta + z_2 \gamma_2$. 
22

APPLIED STATISTICS

in Prentice (1976). A plot of the data and the logit fit are given in Fig. 5. A noticeable systematic departure is evident. Correspondingly, a fit of the extended systematic component indicates that the link function is not adequate (the observed $\chi^2$ value on two degrees of freedom is 8.2). In particular, fitting $\alpha$ and $\delta$ separately indicates that the reduction in deviance is due almost exclusively to $\delta \neq \delta_0$, demonstrating the skewness in the observed response curve. This finding coincides with that of Prentice, whose model gives a very similar fit to that of $g_0(\mu) = x^T\beta + z^T \gamma_2$. This "extended" fit is plotted in Fig. 6 and listed (as Skew Fit) in Table 1—the improvement is striking.

2.3. Other Error/Link Combinations

The previous two examples cover a broad range of applications but certainly not all. To accommodate the fitting of log-linear models, components of variance and a type of non-symmetric regression analysis (see Folks and Chhikara, 1978), the following error/link combinations are commonly used:

<table>
<thead>
<tr>
<th>Error</th>
<th>Link</th>
<th>$g_0(\mu)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>log:</td>
<td>$g_0(\mu) = \log\mu$</td>
</tr>
<tr>
<td>Gamma</td>
<td>inverse:</td>
<td>$g_0(\mu) = \mu^{-1}$</td>
</tr>
<tr>
<td>Inverse Gaussian</td>
<td>inverse square:</td>
<td>$g_0(\mu) = \mu^{-\delta}$</td>
</tr>
</tbody>
</table>

As in the normal/identity combination, the power family (3) provides a natural link generalization for these cases. However, since application of the method requires nothing novel for these cases, no further discussion will be given.

3. RELATIONSHIP AND COMPARISON TO MAXIMUM LIKELIHOOD ESTIMATION

The estimates and tests which result from applying the proposed theory will, in general, not coincide with those computed by full maximum likelihood estimation. In the former method, we estimate $(\beta, \alpha, \delta)$ from a fit of the model with link $g(\mu; \alpha_0, \delta_0) = x^T\beta + z^T(\alpha_0 - \alpha, \delta_0 - \delta)^T$, $\alpha_0, \delta_0$ fixed,

while in the latter, from a fit of the model with link $g(\mu; \alpha, \delta) = x^T\beta$.

Both of these fits are based on maximum likelihood estimation and since $g(\mu; \alpha, \delta) \equiv g(\mu; \alpha_0, \delta_0) + z^T(\alpha - \alpha_0, \delta - \delta_0)^T$,

we may regard the present method as a reasonable first approximation to the full set of likelihood equations. In particular, if $\hat{\beta}_0$ denotes the maximum likelihood estimate from a fit of the model $g(\mu; \alpha_0, \delta_0) = x^T\beta$, the first iteration of the method of efficient scores (starting from the initial guess $(\hat{\beta}_0, \alpha_0, \delta_0)$), yields results identical to the approximate method.

Provided that the initial guess $(\beta_0, \alpha_0, \delta_0)$ is good, further iteration on the estimates will produce the correct maximum likelihood estimate. This requires the sequence $g(\mu; \alpha_0, \delta_0) + z^T(\alpha - \alpha_0, \delta - \delta_0)^T$,

which can be conveniently incorporated in the usual iterative procedure for fitting generalized linear models provided the link function $g(\mu; \alpha, \delta)$ can be fit. This restriction poses no problems for GLIM users as the new release allows user-defined models (see Section 18 of the GLIM Manual, 1978).
ADEQUACY OF GLIM LINKS

In most applications, the benefits from iterating past the one-step approximate method will be small. Applied to the present problem, Bickel's (1975) results suggest that the one-step estimate is asymptotically equivalent to the fully iterated estimate.

Further, experience has shown that for both methods, the correlations between $\hat{\beta}$ and $\hat{\gamma}$ are usually quite high. This is primarily due to the fact that the data cannot provide the detailed information concerning the parameters that we require, and also the non-linear nature with which the parameters combine. These high correlations render individual comparisons with asymptotic standard errors misleading and necessitate a more elaborate procedure for constructing confidence intervals or sets (see Andrews, 1971). Typically, these procedures require considerable computation and are not suggested for routine use.

4. CONCLUDING COMMENTS

The techniques described in this paper can be interpreted as tests of model adequacy by removal of degrees of freedom from the usual fit. Perhaps the most common such degree of freedom for this purpose is due to Tukey (1949), which, in the general case, involves the additional variable $z = g_0(\mu)$. Although this single degree of freedom is often the most sensitive to model departures, it is not indicative of what corrective action should be taken for our application. The usual remedy is to transform the response variate $y$ to $f(y) = y^P$, and it is somewhat harder to interpret $(\log \mu)^P$ or $(\logit \mu)^P$ than the corrective action suggested by the proposed method.

Moreover, many other families can be chosen as link generators, each with their own corresponding degrees of freedom for link adequacy. One such family which is particularly useful, when $\mu$ is not necessarily positive, is given in Manly (1975) by

$$g(\mu; \alpha) = (e^{\mu \alpha} - 1)/\alpha.$$ 

In particular, the identity function is given by $g(\mu; 0) = \mu$. For the binomial case, a referee has suggested the one parameter link family

$$g(\mu; \alpha) = \log(\frac{([n/(n-\mu)]^\alpha - 1)}{\alpha}).$$

Although this family is sensitive to skewness only, it is important since it contains both, the logistic ($\alpha = 1$) and the complementary log–log ($\alpha \to 0$) transforms. Others, perhaps, will arise naturally from the particular application under study.

As stated in Section 1, the usefulness of the method depends upon the following assumptions:

(i) that the true link $g_\alpha(\mu)$ is in fact a member of the link family $g(\mu; \alpha, \delta)$ and

(ii) that $(\alpha_0, \delta_0)$ is sufficiently "close" to $(\alpha_*, \delta_*)$ such that the local linear expansion of $g_\alpha(\mu)$ about $g_0(\mu)$ is at least approximately adequate.

If (i) is not true, the tests can still point to improvements over $(\alpha_0, \delta_0)$ within the family, but not necessarily optimal. However, if (ii) does not hold, the method will most likely fail. For instance, if the hypothesized link leads to fitted values which are nearly constant, then $\hat{\beta}$ will be nearly constant and hence proportional to the already fitted grand mean. Again, this points to the fact that the method is likely to be most useful in determining if a reasonable fit can be improved, rather than the somewhat more optimistic goal of correcting a hopeless situation.

Finally, throughout this paper, the emphasis has been on correct link function identification. If the results of such an analysis indicate no anomalies in the current fit, one should not be overly confident in the results. More thorough analysis and re-analysis of the data and the fitted model is necessary, this being a logical first step.
ACKNOWLEDGEMENTS

The author would like to thank David Andrews, Mike Healy and Murray Aitken for comments on an earlier draft. The referees' criticisms served to make the final version a more comprehensive one.

This research was supported by the National Research Council of Canada while the author was visiting Imperial College, London.

REFERENCES


